

Thomas Niebler

Extracting and Learning Semantics from Social Web Data

Dissertation zur Erlangung des akademischen Grades eines Dok-
tors der Naturwissenschaften (Dr. rer. nat.)
in der Fakultät für Mathematik und Informatik der Universität
Würzburg
Würzburg, im Oktober 2018



For my family: my safe harbour and the source of my strength.

Acknowledgements

A long journey now comes to an end. I would have never arrived here, if not for the many, many wonderful people that supported me in writing this dissertation. Unfortunately, it is almost impossible to thank everyone in person: Still, I want to explicitly mention some of them.

First of all, I want to thank my supervisor Andreas Hotho for the extremely interesting topic of my thesis, the encouraging research environment as well as his guidance and belief in me throughout the years. Many thanks also go out to Robert Jäschke, who agreed to review this thesis. I also want to thank my former students Daniel Schlör, Luzian Hahn, and Tobias Koopmann, who contributed some parts to this thesis. At this point, I also want to thank my colleagues from far away, with whom I could collaborate on many parts of this thesis: Markus Strohmaier, Philipp Singer, and Stephan Doerfel.

I especially want to express my sincerest gratitude to the great band of people that I was allowed to call my colleagues: Albin Zehe, Alexander Dallmann, Christian Pölit, Daniel Schlör, Daniel Zoller, Lena Hettinger, Markus Ring, and Martin Becker from the DMIR Group, as well as the guys from the AI Chair, especially Georg Dietrich, Georg Fette, Jonathan Krebs, Marian Ifland, Markus Krug, and Maximilian Ertl. Your openness for interesting and fruitful discussions but also for a quick chat and last but not least, some casual table soccer sessions make me very happy to have you as my friends. Special thanks go out to my decade-old accomplice in many things, Martin Becker. You played an immeasurable role in my journey by always encouraging and supporting me, both in research and real life.

Speaking of decades, I also want to mention my friends that I was very lucky to meet and keep with me for over 12 years (and so many years to come!): Alexander Kleinschrodt, Andreas Freimann, Isabel Runge, Ralf Zeuka, Dogan Cinbir, Andreas Bauer, Thien Anh Le, Jonas Meier, and Sven Gehwald. Thank you for the many days (and nights) of big and small adventures!

Then there are those that, many, many years ago, laid the groundwork upon which I was able to build my life as it is now. Since my first thoughts, I knew that I could rely on my family in each and every aspect, most prominently Christine, Günther, Stefan, Rudolf, and Hannelore. You provided me a loving, safe, and stable environment in which I could explore my interests and to which I could retreat in bad times. Words cannot express how grateful I am for knowing you are there.

The last person that I want to thank, but can never thank enough, is my wife Nicola. Without your unconditional love, your unfaltering support, your constant encouragement, and finally your unshakeable belief in me, I would have never been able to complete this work. Thank you for being the most important part of my life. I love you.

*Es gibt nichts Schöneres, als geliebt zu werden, geliebt um seiner selbst willen
oder vielmehr: trotz seiner selbst.*

Victor Hugo

Abstract

Making machines understand natural language is a dream of mankind that existed since a very long time. Early attempts at programming machines to converse with humans in a supposedly intelligent way with humans relied on phrase lists and simple keyword matching. However, such approaches cannot provide semantically adequate answers, as they do not consider the specific meaning of the conversation. Thus, if we want to enable machines to actually understand language, we need to be able to access semantically relevant background knowledge. For this, it is possible to query so-called ontologies, which are large networks containing knowledge about real-world entities and their semantic relations. However, creating such ontologies is a tedious task, as often extensive expert knowledge is required. Thus, we need to find ways to automatically construct and update ontologies that fit human intuition of semantics and semantic relations. More specifically, we need to determine semantic entities and find relations between them. While this is usually done on large corpora of unstructured text, previous work has shown that we can at least facilitate the first issue of extracting entities by considering special data such as tagging data or human navigational paths. Here, we do not need to detect the actual semantic entities, as they are already provided because of the way those data are collected. Thus we can mainly focus on the problem of assessing the degree of semantic relatedness between tags or web pages. However, there exist several issues which need to be overcome, if we want to approximate human intuition of semantic relatedness. For this, it is necessary to represent words and concepts in a way that allows easy and highly precise semantic characterization. This also largely depends on the quality of data from which these representations are constructed.

In this thesis, we extract semantic information from both tagging data created by users of social tagging systems and human navigation data in different semantic-driven social web systems. Our main goal is to construct high quality and robust vector representations of words which can be used to measure the relatedness of semantic concepts. First, we show that navigation in the social media systems Wikipedia and BibSonomy is driven by a semantic component. After this, we discuss and extend methods to model the semantic information in tagging data as low-dimensional vectors. Furthermore, we show that tagging pragmatics influences different facets of tagging semantics. We then investigate the usefulness of human navigational paths in several different settings on Wikipedia and BibSonomy for measuring semantic relatedness. Finally, we propose a metric-learning based algorithm to adapt pre-trained word embeddings to datasets

containing human judgment of semantic relatedness.

This work contributes to the field of studying semantic relatedness between words by proposing methods to extract semantic relatedness from web navigation, learn high-quality and low-dimensional word representations from tagging data, and to learn semantic relatedness from any kind of vector representation by exploiting human feedback. Applications first and foremost lie in ontology learning for the Semantic Web, but also semantic search or query expansion.

Zusammenfassung

Einer der großen Träume der Menschheit ist es, Maschinen dazu zu bringen, natürliche Sprache zu verstehen. Frühe Versuche, Computer dahingehend zu programmieren, dass sie mit Menschen vermeintlich intelligente Konversationen führen können, basierten hauptsächlich auf Phrasensammlungen und einfachen Stichwortabgleichen. Solche Ansätze sind allerdings nicht in der Lage, inhaltlich adäquate Antworten zu liefern, da der tatsächliche Inhalt der Konversation nicht erfasst werden kann. Folgerichtig ist es notwendig, dass Maschinen auf semantisch relevantes Hintergrundwissen zugreifen können, um diesen Inhalt zu verstehen. Solches Wissen ist beispielsweise in Ontologien vorhanden. Ontologien sind große Datenbanken von vernetztem Wissen über Objekte und Gegenstände der echten Welt sowie über deren semantische Beziehungen. Das Erstellen solcher Ontologien ist eine sehr kostspielige und aufwändige Aufgabe, da oft tiefgreifendes Expertenwissen benötigt wird. Wir müssen also Wege finden, um Ontologien automatisch zu erstellen und aktuell zu halten, und zwar in einer Art und Weise, dass dies auch menschlichem Empfinden von Semantik und semantischer Ähnlichkeit entspricht. Genauer gesagt ist es notwendig, semantische Entitäten und deren Beziehungen zu bestimmen. Während solches Wissen üblicherweise aus Textkorpora extrahiert wird, ist es möglich, zumindest das erste Problem - semantische Entitäten zu bestimmen - durch Benutzung spezieller Datensätze zu umgehen, wie zum Beispiel Tagging- oder Navigationsdaten. In diesen Arten von Datensätzen ist es nicht notwendig, Entitäten zu extrahieren, da sie bereits aufgrund inhärenter Eigenschaften bei der Datenakquise vorhanden sind. Wir können uns also hauptsächlich auf die Bestimmung von semantischen Relationen und deren Intensität fokussieren. Trotzdem müssen hier noch einige Hindernisse überwunden werden. Beispielsweise ist es notwendig, Repräsentationen für semantische Entitäten zu finden, so dass es möglich ist, sie einfach und semantisch hochpräzise zu charakterisieren. Dies hängt allerdings auch erheblich von der Qualität der Daten ab, aus denen diese Repräsentationen konstruiert werden.

In der vorliegenden Arbeit extrahieren wir semantische Informationen sowohl aus Taggingdaten, von Benutzern sozialer Taggingssysteme erzeugt, als auch aus Navigationsdaten von Benutzern semantikgetriebener Social Media-Systeme. Das Hauptziel dieser Arbeit ist es, hochqualitative und robuste Vektordarstellungen von Worten zu konstruieren, die dann dazu benutzt werden können, die semantische Ähnlichkeit von Konzepten zu bestimmen. Als erstes zeigen wir, dass Navigation in Social Media-Systemen unter anderem durch eine semantische Komponente getrieben wird. Danach

diskutieren und erweitern wir Methoden, um die semantische Information in Taggingdaten als niedrigdimensionale sogenannte “Embeddings” darzustellen. Darüberhinaus demonstrieren wir, dass die Taggingpragmatik verschiedene Facetten der Taggingsemantik beeinflusst. Anschließend untersuchen wir, inwieweit wir menschliche Navigationsspfade zur Bestimmung semantischer Ähnlichkeit benutzen können. Hierzu betrachten wir mehrere Datensätze, die Navigationsdaten in verschiedenen Rahmenbedingungen beinhalten. Als letztes stellen wir einen neuartigen Algorithmus vor, um bereits trainierte Word Embeddings im Nachhinein an menschliche Intuition von Semantik anzupassen.

Diese Arbeit steuert wertvolle Beiträge zum Gebiet der Bestimmung von semantischer Ähnlichkeit bei: Es werden Methoden vorgestellt werden, um hochqualitative semantische Information aus Web-Navigation und Taggingdaten zu extrahieren, diese mittels niedrigdimensionaler Vektordarstellungen zu modellieren und selbige schließlich besser an menschliches Empfinden von semantischer Ähnlichkeit anzupassen, indem aus genau diesem Empfinden gelernt wird. Anwendungen liegen in erster Linie darin, Ontologien für das Semantic Web zu lernen, allerdings auch in allen Bereichen, die Vektordarstellungen von semantischen Entitäten benutzen.

Contents

1	Introduction	1
	1.1 Motivation	1
	1.2 Research Topics and Contributions	3
	1.2.1 Detecting Semantic Influence on Social Media Navigation	4
	1.2.2 Capturing Semantics in Social Tagging Data	5
	1.2.3 Extracting Semantic Relatedness from Social Media Navigation	6
	1.2.4 Relative Relatedness Learning: Learning Semantic Relatedness from Human Feedback	7
	1.3 Thesis Outline	8
2	Related Work	11
	2.1 Extracting Semantics from Structured and Semi-Structured Data	12
	2.1.1 Leveraging WordNet's Taxonomy to Compute Semantic Relatedness	12
	2.1.2 Capturing Semantic Information from Folksonomies	12
	2.1.3 Capturing Semantic Information from Wikipedia	14
	2.1.4 Evaluating Extracted Semantic Information	16
	2.2 Learning Semantics with and without Background Information	17
	2.2.1 Learning Word Embeddings	17
	2.2.2 Learning to Adapt Word Embeddings to Background Knowledge	20
	2.3 Mutual Influence of Semantics and User Behavior	21
	2.3.1 Modelling (Semantic) Influence on Human Navigation Behavior	22
	2.3.2 Folksonomy Usage Patterns and their Influence on Semantics	23
	2.3.3 Influence of Navigation Patterns in Wikipedia on Semantic Information	24
	2.4 Outlook: Applications of Semantic Information Models	26
	2.5 Summary	29

Foundations	31
3.1 Implicit and Explicit Forms of Semantic Knowledge in the Web	31
3.1.1 Social Tagging Systems and Folksonomies	32
3.1.1.1 Folksonomies	33
3.1.1.2 Social Tagging System User Interfaces	35
3.1.2 Wikipedia	37
3.1.3 The WordNet Taxonomy	39
3.1.4 Ontologies and Knowledge Graphs	40
3.2 Computational Methods to Model Distributional Semantic Knowledge	41
3.2.1 The Vector Space Model: Expressing Words as Vectors	42
3.2.1.1 Formal Definition of the Vector Space Model: Document-Level Semantics	43
3.2.1.2 Variants of the Vector Space Model for Word-Level Semantics	44
3.2.2 Choosing the Right Context	48
3.2.2.1 Context in Folksonomies	49
3.2.2.2 Context in Graphs	50
3.2.3 Measuring Semantic Relatedness and Similarity	50
3.2.3.1 Distinguishing Semantic Relatedness and Similarity	51
3.2.3.2 The Cosine Similarity Measure for VSMs	52
3.2.3.3 WordNet-Based Semantic Similarity Measures	53
3.2.4 Evaluating Semantic Relatedness Measures	54
3.2.4.1 Grounding on Semantic Relatedness Datasets	54
3.2.4.2 Grounding on Word Thesauri	57
3.2.5 Post-Processing and Applying Word Vector Representations for Semantic Information Gain	58
3.2.5.1 Improving Vector Representations using Lexicons: Retrofitting	58
3.2.5.2 Discovering Word Senses Using Vector Representations	59
3.3 Characterizing User Behavior in the Social Web	62
3.3.1 User Behavior Types in Folksonomies	62
3.3.1.1 Categorizers and Describers	62
3.3.1.2 Generalists and Specialists	63
3.3.2 Navigation Behavior in Social Media Systems	65
3.3.2.1 Navigation in Wikipedia	65
3.3.2.2 Navigation Behavior in Social Tagging Systems	67
3.3.3 Analyzing User Navigation using Navigational Hypotheses	67
3.3.3.1 Formulating and Comparing Navigation Hypotheses with Hyptrails	67
3.3.3.2 Standard Hypotheses	68
3.4 Summary and Relations to this Work	70

Datasets	71
4.1 Text-based Datasets	71
4.1.1 Delicious tagging data	72
4.1.2 BibSonomy tagging data	73
4.1.3 CiteULike tagging data	74

4.1.4	Wikipedia article texts	74
4.1.5	Wikipedia disambiguation pages	75
4.2	Link-based Datasets	76
4.2.1	Navigation Games on Wikipedia: WikiGame and Wikispeedia	76
4.2.1.1	WikiGame	77
4.2.1.2	Wikispeedia	78
4.2.2	Unconstrained Navigation on Wikipedia: ClickStream and WikiClickIU	79
4.2.2.1	ClickStream	79
4.2.2.2	WikiClickIU	79
4.2.3	Wikipedia's Static Link Network: WikiLink	80
4.2.4	BibSonomy Website Usage	81
4.2.4.1	BibSonomy Request Logs	82
4.2.4.2	Human Navigational Paths on BibSonomy	83
4.2.5	The BibSonomy Link Graph	83
4.3	Pretrained Word Embeddings	83
4.3.1	WikiGloVe	84
4.3.2	WikiNav	85
4.3.3	ConceptNet Numberbatch	85
4.4	Semantic Relatedness Datasets	85
4.4.1	RG65	86
4.4.2	MC30	87
4.4.3	WS-353	87
4.4.4	MTurk	88
4.4.5	MEN	88
4.4.6	SimLex-999	89

5

Detecting Semantic Influence on Social Media Navigation	91
5.1 Introduction	91
5.2 Wikipedia	92
5.2.1 Hypotheses	92
5.2.1.1 Standard Hypotheses	92
5.2.1.2 High and Low Degree Hypotheses	93
5.2.2 Results and Discussion	94
5.2.2.1 Game Navigation: WikiGame and Wikispeedia	94
5.2.2.2 Unconstrained Navigation	98
5.2.3 Conclusion	101
5.3 BibSonomy	101
5.3.1 Hypotheses	102
5.3.1.1 Standard Hypotheses	102
5.3.1.2 Social Tagging System Specific Hypotheses	103
5.3.1.3 Combined Hypotheses	104

5.3.2	Results and Discussion	106
5.3.2.1	Overall Request Log Dataset	106
5.3.2.2	Outside Navigation	108
5.3.2.3	User Gender	109
5.3.2.4	Usage Continuity	110
5.3.2.5	Tagger Classes	110
5.3.3	Conclusion	111
5.4	Summary	112

6

Capturing Semantics in Social Tagging Data	115
6.1 Introduction	115
6.2 A Critical Examination of WordNet-based Evaluation of Semantic Similarity	116
6.2.1 Experimental Setup	117
6.2.2 Results	118
6.2.2.1 Re-Evaluating the Jiang-Conrath Measure on Human Intuition	118
6.2.2.2 Vocabulary Coverage and Extensibility of Semantic Relatedness Datasets	119
6.2.3 Discussion	122
6.2.3.1 Comparison of Evaluation Approaches	122
6.2.3.2 Shortcomings of Evaluation on Human Intuition Datasets	123
6.2.4 Conclusion	124
6.3 Re-Evaluating Measures of Semantic Tag Similarity on Human Intuition	124
6.3.1 Preliminaries	125
6.3.2 Distributional Measures of Tag Similarity	125
6.3.3 Impact of User Pragmatics on Tagging Semantics	128
6.3.4 Summary	134
6.4 Influence of Tagging Pragmatics on Tag Sense Discovery in Folksonomies	134
6.4.1 Experimental Setup	136
6.4.2 Results and Discussion	138
6.4.3 Conclusion	141
6.5 Learning Tag Embeddings	142
6.5.1 Applicability of Embedding Algorithms on Tagging Data	143
6.5.1.1 Word2Vec	143
6.5.1.2 GloVe	143
6.5.1.3 LINE	144
6.5.1.4 Common Parameters	144
6.5.2 Experimental Setup	145
6.5.3 Results	146
6.5.4 Discussion	149
6.5.5 Conclusion	151
6.6 Summary	151

Extracting Semantic Relatedness from Social Media Navigation	153
7.1 Introduction	153
7.2 Wikipedia	155
7.2.1 Methods	156
7.2.1.1 Count-based Navigational Semantic Relatedness	156
7.2.1.2 Binary Navigational Semantic Relatedness	157
7.2.1.3 Evaluation on Human Intuition	158
7.2.2 Game Navigation	159
7.2.2.1 Contribution of Game Navigation to Semantic Relatedness	159
7.2.2.2 Benefit of Human Game Navigation over the Link Network	162
7.2.2.3 Path Selection to Improve Fit to Human Intuition	165
7.2.2.4 Conclusion	173
7.2.3 Unconstrained Navigation	175
7.2.3.1 Contribution of Unconstrained Navigation to Semantic Relatedness	176
7.2.3.2 Benefit of Human Unconstrained Navigation over the Link Network	179
7.2.3.3 Conclusion	181
7.2.4 Random Walks on the Wikipedia Graph	181
7.2.4.1 Contribution of Random Walks to Semantic Relatedness	183
7.2.4.2 Benefit of Biased Random Walks over Uniform Random Walks	185
7.2.4.3 Conclusion	192
7.2.5 Summary and Discussion	192
7.2.5.1 Summary	193
7.2.5.2 Discussion of the Results	193
7.3 BibSonomy	194
7.3.1 Preliminaries	195
7.3.1.1 Content-based Navigational Semantic Relatedness	195
7.3.1.2 Datasets and Evaluation	197
7.3.2 Unconstrained Navigation	197
7.3.2.1 Page Co-Occurrence Counting Semantics of BibSonomy Navigation	197
7.3.2.2 Leveraging the Tag Cloud of a BibSonomy Page	198
7.3.2.3 Removing Potentially Noisy Transitions	201
7.3.2.4 Discussion	204
7.3.2.5 Conclusion	205
7.3.3 Random Walks on the BibSonomy Graph	206
7.3.3.1 Experimental Setup	206
7.3.3.2 Results and Discussion	208
7.3.3.3 Conclusion	209
7.4 Summary	209

Relative Relatedness Learning: Learning Semantic Relatedness from Human Feedback	213
8.1 Introduction	213

8.2 The Relative Relatedness Learning (RRL) Algorithm	214
8.2.1 Motivation	215
8.2.2 Optimization Objective	216
8.2.3 Optimization	217
8.3 Experiments and Results	217
8.3.1 Preliminaries	217
8.3.2 Word Relatedness	218
8.3.2.1 Integrating Different Levels of Background Knowledge	219
8.3.2.2 Comparison and Combination with Retrofitting	220
8.3.2.3 Robustness of Pretrained Embeddings	222
8.3.2.4 Transporting User Intentions	223
8.3.3 Question Answering	224
8.4 Discussion	225
8.5 Conclusion	227

9 Conclusion and Future Perspectives	229
9.1 Summary	229
9.2 Open Problems and Future Perspectives	231

List of Notations	233
1 General Notation	233
2 Folksonomy Notation	233
3 Semantics Notation	233
4 Graph and Navigation Notations	234

Bibliography	235
---------------------	------------

List of Figures

1.1	General Thesis outline	9
3.1	Screenshot of BibSonomy	33
3.2	Tri-partite folksonomy hypergraph	34
3.3	Screenshot of the Wikipedia page about “Coffee”	38
3.4	An example for word-word co-occurrence and term-document matrices, constructed from a document corpus $\mathcal{D} = \{d_1, d_2\}$	44
3.5	Illustration of Word2Vec’s skipgram neural network architecture	46
3.6	Illustration of node neighborhoods in a graph	48
3.7	Illustration of the cosine similarity measure for vectors	52
3.8	Anscombe quartet	56
3.9	Distribution plot of human judgment in WS-353 and Bib100	57
3.10	Illustration of the tag sense discovery process	60
4.1	Distribution of Wikipedia senses per tag.	76
4.2	Available links between resource pages in the BibSonomy link graph.	84
5.1	Illustrations of navigation hypotheses on Wikipedia.	95
5.2	Marginal Likelihoods of navigation hypotheses on WikiGame navigation.	96
5.3	Marginal Likelihoods of navigation hypotheses on Wikispeedia navigation.	97
5.4	Hypothesis-based analysis of unconstrained navigation on the Click- Stream data.	99
5.5	Hypothesis-based analysis of unconstrained navigation on the WikiClickIU data.	100
5.6	Illustrations of the navigation hypotheses defined in Section 5.3.1.	105
5.7	Evidence chart for the navigational hypotheses on the complete request log dataset of BibSonomy.	107
5.8	Evidence curves for BibSonomy navigation outside of the user’s resources.	109
5.9	Evidence charts of BibSonomy navigation hypotheses split by gender.	113
5.10	Evidence charts for short-term BibSonomy users.	114
5.11	Evidence chart for different user types according to tagging behavior.	114
6.1	Impact of tagging pragmatics on tagging semantics in Delicious.	130
6.2	Impact of tagging pragmatics on tagging semantics in BibSonomy.	131

List of Figures

6.3	Impact of tagging pragmatics on tagging semantics, evaluated on the semantic similarity evaluation dataset SimLex-999.	133
6.4	Illustration of different word senses for <i>apple</i>	135
6.5	Evaluation of discovered tag senses on Wikipedia.	136
6.6	Tag Sense Discovery results for the Delicious dataset.	140
6.7	Tag Sense Discovery results for the BibSonomy dataset.	141
6.8	Evaluation results for embeddings generated by Word2Vec on MEN.	147
6.9	Evaluation results for embeddings generated by GloVe on the MEN dataset.	148
6.10	Evaluation results for embeddings generated by LINE on the MEN dataset.	149
6.11	Evaluation results for embeddings generated by the best parameter settings across the different tagging datasets.	150
7.1	A sketch of the sliding window approach and the resulting co-occurrence matrix.	157
7.2	A sketch of the binarization approach and the resulting co-occurrence matrix.	158
7.3	Illustration of the distribution of path lengths in the WikiGame and Wikispeedia.	167
7.4	Semantic relatedness calculated on different path selections on the WikiGame.	170
7.5	Semantic “fingerprints” for different concepts on WikiGame data.	171
7.6	Effect of successful / unsuccessful paths in the WikiGame.	174
7.7	Transition count distributions for the ClickStream, WikiClickIU, and WikiGame datasets	178
7.8	Spearman correlations on WS-353 for networks pruned to the top k -percent nodes according to Pagerank.	186
7.9	Evaluation results for random walks on the WikiGame graph.	189
7.10	Evaluation results for random walks on the ClickStream graph.	191
7.11	Illustration of the Multi-Tag-Co-Occurrence method for a window size of 2.	196
7.12	Results for the page co-occurrence counting method on BibSonomy navigation.	199
7.13	Results for the tagset weighting experiment on BibSonomy.	202
7.14	Results for the minimum transition count experiment.	204
7.15	Semantic relatedness results from paths when excluding all /user/USER pages.	205
8.1	A rough sketch of how RRL works.	214
8.2	Results on different levels of amounts of MEN training data.	220
8.3	Results on different amounts of WS-353 training data using the pretrained embeddings.	221
8.4	Results for training a measure on only 353 equidistant pairs from MEN.	226

List of Tables

3.1	The considered page types in the BibSonomy system.	36
3.2	Comparison of visited page types in game and unconstrained navigation on Wikipedia.	66
4.1	List of all datasets used in this thesis.	72
4.2	Statistics about the Delicious dataset.	73
4.3	Statistics about the BibSonomy dataset.	74
4.4	Statistics about the CiteULike dataset.	75
4.5	Properties of the Wikipedia text datasets.	75
4.6	Characteristics of the WikiGame navigation dataset.	77
4.7	Characteristics of the Wikispeedia navigation dataset.	78
4.8	Characteristics of the ClickStream navigation dataset.	79
4.9	Characteristics of the WikiClickIU navigation dataset.	80
4.10	Size comparison of all Wikipedia link datasets.	80
4.11	Characteristics of the BibSonomy navigation dataset.	81
4.12	BibSonomy request log and link network statistics.	82
4.13	Base statistics for the human intuition similarity datasets.	86
5.1	Transition statistics of different BibSonomy request log subsets.	106
5.2	Selected request log statistics.	108
6.1	Evaluation of WordNet semantic similarity measures on different datasets with human intuition of semantic relatedness.	119
6.2	Overlaps of the top tags from Delicious and BibSonomy with WordNet and WS-353.	120
6.3	Spearman correlation scores on WS-353 in other literature.	120
6.4	Comparison of different BibSonomy vocabulary subsets with the WS-353 and the Bib100 vocabularies.	121
6.5	Evaluation of the Jiang-Conrath and taxonomic path distance measures on Bib100.	122
6.6	Results for the WordNet-based evaluation method used by Cattuto et al. on the Delicious and BibSonomy datasets.	126

List of Tables

6.7	Results for evaluation on the human similarity intuition datasets for the Delicious and BibSonomy folksonomy dataset.	127
6.8	Examples of detected senses produced by hierarchical clustering with the optimal distance criterion threshold for the tag swing based on different partitions of included users of the Delicious dataset.	139
6.9	Spearman correlation values for the co-occurrence baseline.	145
6.10	Initial parameter values for each algorithm.	146
6.11	Best parameter values for each algorithm for the MEN dataset on Delicious data.	146
7.1	Semantic relatedness calculated on human navigational paths in Wikipedia games.	160
7.2	Semantic relatedness accuracy calculated in similar fashion as for Table 7.1.	161
7.3	Comparison of semantic relatedness calculations using a window size of $w = 3$ evaluated against WS-353 and MEN on all WikiGame paths with several baseline corpora.	164
7.4	Semantics in successful and unsuccessful game paths.	165
7.5	Performance comparison of our Wikipedia navigational datasets on common evaluation pairs.	180
7.6	Sampling results on baseline datasets based on WikiLink and ClickStream.	180
7.7	Spearman correlations achieved with co-occurrence counting and binarization by varying the number of walks started on each node in the network.	184
7.8	Overview of all link networks that we use to generate random walks.	185
7.9	Basic statistics for all random walk datasets generated from Wikipedia link networks.	185
7.10	Overlap of the Wikipedia page subsets according to their importance with respect to PageRank with the WikiGame and ClickStream datasets.	187
7.11	Evaluation results when characterizing a BibSonomy page by its single most frequently occurring tag.	200
7.12	Size comparison of the Wikipedia navigation datasets used in Section 7.2.3, together with their evaluation scores on WS-353.	203
7.13	Statistics about the random paths on BibSonomy generated from the navigation hypotheses presented in Section 5.3.	208
7.14	Evaluation of all random walks on BibSonomy for different window sizes.	211
8.1	Examples of word pairs with human assigned relatedness scores on a scale ranging from 0 to 1.	215
8.2	Baselines for word embedding fine tuning experiments.	222
8.3	Results for user intention transport experiments.	224
8.4	Evaluation results for the question answering task using DrQA.	225

Chapter 1

Introduction

User: “Siri, call me an ambulance!”

Siri: “From now on, I’ll call you ‘An Ambulance’. OK?”

1.1 Motivation

In 1966, Joseph Weizenbaum developed the computer program *ELIZA* [244], the first chatbot ever. Given any utterances from its users, it made use of a thesaurus to find direct and suitable answers from a large collection of possible phrases. It was specifically designed to represent a mock Rogerian psychotherapist [197], showing maximal contempt with a user while simultaneously allowing to show no knowledge of the real world, by simply paraphrasing the input of the user. However, this experiment only showed that by choosing very general responses the effect of assuming that a person is responding can be easily achieved. In fact, Eliza is simply looking up hypernyms or generalizations of encountered words in the user’s utterances. However, Eliza does not really *understand* what a user is really saying and thus cannot be seen as an intelligent system.

Nowadays, speech understanding systems, such as Amazon’s Alexa, Apple’s Siri or Google Now, are present on almost every smartphone and on their way into many households. With greatly increased processing power and sophisticated algorithms, computers are able to respond much better and much more directly to user’s needs and intentions¹. For example, modern systems like Alexa can inform a user about the current weather, calendar events and even perform simple actions like opening up apps with given parameters. This impression of “intelligent” machines is enabled by great strides in research since ELIZA on how to process and interpret natural language.

The Semantic Web, Ontologies and the Knowledge Acquisition Bottleneck

One key aspect responsible for this development is the evolution of the World Wide Web from static web pages to semantically supported dynamic user interfaces. In 2001, Berners-Lee, Hendler, and Lassila stated their idea of the *Semantic Web* [31].

¹especially to prevent situations like in the quote above.

1 Introduction

The ultimate goal of the Semantic Web is to make information machine-readable and -interpretable to enable *computers to understand humans*. Attributing semantic meaning to words and entities, while simultaneously connecting this information, is inherent to human thinking and verbal interaction: Mitchell et al. even could predict the activity in specific sections of the human brain which are associated to different meanings of nouns [162]. In the Semantic Web, this knowledge about the *semantics* of words is made explicit by extending the plain hyperlink structure between web pages prevalent in the World Wide Web to represent *semantic relations* between *semantic entities*.

The key structures encoding such knowledge are called *ontologies*. Ontologies are large databases of entities and contain explicitly named semantic relations between those entities [88, 224]. Concretely, ontologies contain common, real-world knowledge about one or many domains. The contents of ontologies are normally hand-crafted and need extensive expert knowledge to be curated and kept up-to-date. For example, WordNet, a large linguistic database of the English language, is curated by expert linguists and provides highly precise information about different meanings of words and their linguistic connections [72]. However, due to the manual curation process, WordNet adapts very slowly to recent changes in language. For example, the current version 3.1 of WordNet does not contain information about the meaning of *Python* as a widely-used programming language.² Especially in expert domains, such as medicine or science, the acquisition of exact and precise knowledge and its subsequent transformation into machine-readable form is a tedious and expensive task: Constructing and editing such specialized ontologies or knowledge graphs largely remains a task for domain experts, and is still unattractive for less knowledgeable and less engaged users. This problem is also often called the *knowledge acquisition bottleneck*.

Overcoming the Knowledge Acquisition Bottleneck using Computational Methods

Today, there exist several big ontologies that contain vast amounts of knowledge in the general domain, which used different ways to overcome the knowledge acquisition bottleneck: While Wikidata³ employs a knowledge acquisition strategy similar to Wikipedia, i.e., letting anyone contribute to its knowledge, the YAGO⁴ and DBPedia⁵ ontologies rely solely on computational methods to extend, enhance or construct their knowledge automatically [12, 225, 235]. The automated process to *construct, enhance and curate* ontologies can be described as “*Ontology Learning*” [145] and has been in the spotlight of the Semantic Web research community since a long time [54, 99, 145].

While all three ontologies mentioned above are largely constructed automatically, they cover only *general domain knowledge*, as that is available in abundance. However, in order to construct *expert domain ontologies* in an automated fashion, one often has to resort to unstructured sources. The extraction of structured knowledge from unstructured

²<http://wordnetweb.princeton.edu/perl/webwn?s=Python>

³<https://www.wikidata.org/>

⁴<https://github.com/yago-naga/yago3>

⁵<https://wiki.dbpedia.org/>

sources is a non-trivial problem, since it requires to address several inherent problems of human language that are not present in semi-structured sources. For example, the intended meaning of a word largely depends on its context, i.e., its surrounding words. In order to extract such knowledge from unstructured text it is possible to leverage *statistical natural language processing* (NLP) methods such as constructing representations of the *semantic meaning* of words and entities in natural language text [156, 184, 201, 205].

Social Media Data as a Source for Semantic Information

A rich source of data from which semantic information can be extracted on a large scale is represented by user contributed content in the Web 2.0. After a period of static web content, the Web 2.0 enabled users to not only *consume* provided content, but also *produce* and distribute it, without the need for access to expensive hardware or specialized knowledge. Thus, the Web 2.0 is also called the *Social Web*, and its prominent platforms (like Facebook⁶, Twitter⁷ and Wikipedia⁸) are often described as *Social Media*.

Especially the *social encyclopedia* Wikipedia symbolizes the collaborative spirit of the Social Web like no other social system. Because Wikipedia allows anyone to freely edit its content, Wikipedia grew to the largest encyclopedia available to date with a content precision that rivals that of Encyclopaedia Britannica [84]. Because of its size, scientific precision, and clarity of writing, Wikipedia's article collection quickly became one of the most studied textual corpora in today's NLP community. Another key technology of the Social Web is represented by *social tagging systems*, where users can publish any kind of content, but additionally annotate this content with so-called *tags*, i.e., short free-form textual descriptions that allow users to categorize content for ease of later retrieval. The collection of users, their resources and annotations is called a *folksonomy*. Similar to Wikipedia, the collaborative nature of social tagging systems and the emergence of stable semantic structures in their content attracted the interest of many behavioral and NLP researchers, but also from the Semantic Web community. Mika even went as far as describing folksonomies as "light-weight ontologies", because the emerging content structures showed inherent semantic hierarchical relations [155]. Well-known examples of social tagging systems are Delicious⁹, CiteULike¹⁰, Flickr¹¹, and BibSonomy¹².

1.2 Research Topics and Contributions

In this thesis, we present work on several research topics concerning semantic information in social media data and the mutual influence of user behavior and semantic

⁶<https://www.facebook.com>

⁷<https://twitter.com/>

⁸<https://www.wikipedia.org/>

⁹<https://del.icio.us/>

¹⁰<http://www.citeulike.org/>

¹¹<https://www.flickr.com/>

¹²<https://www.bibsonomy.org/>

1 Introduction

information. In the following, we will elaborate each of these topics separately by first describing the general setting and open problems, explicitly stating our approach to overcome these problems and finally summarize our findings. Each subsection refers to the corresponding chapter with the same name later in this work.

1.2.1 Detecting Semantic Influence on Social Media Navigation

Problem Setting. There is plenty of evidence that the navigational behavior of users is driven by semantic factors. For example, Chi et al. hypothesized that users navigating social media follow a so-called “information scent”, i.e., they follow information that might bring them nearer to their actual navigational goal [51]. West and Leskovec found indications for this behavior in navigation data obtained from the navigation game Wikispeedia [247]. It is however unclear if this also generalizes to other game navigation datasets, e.g., data from the WikiGame. Even more, we don’t know how users in *unconstrained settings* navigate Wikipedia, i.e., when lacking a predefined goal. While Doerfel et al. investigated the *usage* of the social tagging system BibSonomy in [64], they did not compare hypotheses about actual navigation behavior. Thus we do not know which factors drive navigation in social tagging systems.

Approach. In order to determine semantic influence on human navigation, we first have to understand human navigation processes and what drives them on a basic level. We propose several hypotheses about human navigation behavior in social media systems and compare them using Bayesian Statistics. For this, we analyze navigation on a wide range of human click trail datasets gathered from the social media systems Wikipedia and BibSonomy. While on Wikipedia, we distinguish *game* and *unconstrained navigation*, we investigate several *subsets of BibSonomy navigation data defined by user characteristics*. Because both systems possess an inherently semantic nature, it is a logical assumption that navigation in those systems is influenced by their semantic content. Thus, we explicitly *search for a signal of the influence of semantic information on navigation*. Furthermore, previous work [122, 176] found that groups of users with a certain tagging behavior also influence the semantic information in tagging data. It is thus only natural to expect that the *navigation behavior of those user groups differs somehow*.

Findings and Contributions. Generally, we find that *navigation in social media systems is influenced by semantic information*. However, for different datasets, we see varied impact. While navigation on Wikispeedia and the WikiClickIU dataset can be explained well by semantic navigation, WikiGame and ClickStream are rather dominated by degree-based navigation. In BibSonomy, navigation behavior can mainly be explained by *semantic navigation on a user’s own pages*. Furthermore, we can find *evidence for different navigation behavior of user groups defined by their tagging pragmatics*, e.g., categorizers navigate differently than describers. These results motivate the research later in Chapter 7.

1.2.2 Capturing Semantics in Social Tagging Data

Problem Setting. It has been repeatedly shown that *social tagging data* exhibit emergent structures that make it possible to extract semantic information [28, 29, 48, 86]. In [48], Cattuto et al. provided means to model semantic information as sparse, high-dimensional co-occurrence vectors and how to semantically grounding those models on semantic similarity derived from the WordNet taxonomy, concretely the Jiang-Conrath similarity [112]. There exist several issues both with the model and the evaluation approach. First of all, the vector representation of tags often suffer from their high dimensionality and sparsity, especially on small tagging datasets. Lately, several methods have been proposed to *learn* low-dimensional, dense word representations from unstructured text [156, 184, 227]. It is yet unclear if those methods can be used to also generate embeddings of tags and how this impacts the semantic information of tagging data. Furthermore, although Budanitsky and Hirst showed that the Jiang-Conrath measure exhibits a very high correlation with two small semantic relatedness datasets [44], it is unclear if the Jiang-Conrath similarity scores generalize well across other semantic relatedness datasets. If this would not be the case, this would affect the validity of other works that base their results on semantic similarity scores obtained from WordNet.

Approach. We discuss the WordNet-based evaluation approach of [48] and compare it with a direct evaluation on human intuition. Next to *concretizing drawbacks of the WordNet-based evaluation*, we propose to *additionally evaluate tagging semantics directly on human intuition*. We then re-examine two works that base their results on similarity scores from the WordNet-based Jiang-Conrath similarity measure. By re-evaluating their results on human intuition, we want to see if their *observations still hold or if they have to be reconsidered*. Concretely, we want to see if the *qualitative ranking of context choices for tag vector representations* as presented by Cattuto et al. [48] *can also be found when evaluating on human intuition*. Additionally, we repeat the experiments by Körner et al. [122] who measured the influence of tagging pragmatics on tagging semantics. We extend this experiment by also measuring the influence of pragmatics on a tag sense discovery process. Finally, we attempt to overcome the high dimensionality and sparsity issues of the model by Cattuto et al. For this, we explore the applicability of three well-known word embedding algorithms on tagging data, namely Word2Vec [156], LINE [227] and GloVe [184].

Findings and Contributions. First of all, we find that *the Jiang-Conrath measure does not generalize well across other semantic relatedness datasets*. The subsequent re-examination of the works by Cattuto et al. [48] and Körner et al. [122] shows that *while some of their observations are still valid in the evaluation setting on human intuition, we can see differences in some details*. For example, we can see that some types of user pragmatics are more useful to measure semantic similarity, while others generate semantic information that is beneficial for measuring semantic relatedness. We also find that *tagging pragmatics influence the tag sense discovery process*, where we can observe that a small portion of descriptors generates the most useful semantic information for this

1 Introduction

application. Finally, we find that generally, *embedding tags in low-dimensional vector spaces can improve the semantic quality of the tag representations*, regardless of the chosen embedding algorithm. Especially the GloVe algorithm captures folksonomy-specific semantics very well and can always outperform the sparse baseline representation from [48].

1.2.3 Extracting Semantic Relatedness from Social Media Navigation

Problem Setting. In Chapter 5, we will show that human navigation in information networks is partially driven by semantic incentives. Additionally, West, Pineau, and Precup already exploited human navigation on a small subset of Wikipedia to calculate semantic relatedness between Wikipedia concepts [249]. This indicates that human navigational paths in social media systems can be exploited as a valuable source of semantic information. Still, it is unclear if the semantic information is contained in the actual navigation behavior or in the underlying link network that only allows certain paths. Also, the results from [249] have been gathered in a game setting, which imposes an extrinsic bias on navigation behavior. We do not know if that bias influences the semantic information of those paths. Finally, in contrast to text and tagging data, information about human navigation is not readily available. However, navigation could be created artificially by performing random walks on the hyperlink graph. Again, it is unclear if the creation process of those artificial paths introduces unwanted bias that has a detrimental effect on the semantic information of those paths.

Approach. We want to determine the usefulness of human navigational paths in social media systems as a viable source of semantic information. For this, we propose several methods to model that information as co-occurrence based vectors. Again, we evaluate our model on game and unconstrained navigation in Wikipedia and as well as on unconstrained navigation in BibSonomy. As before, we investigate game and unconstrained navigation on Wikipedia separately. To overcome limitations in design and size of our navigation datasets, we also investigate if parameterized random walks on the social media link graphs can approximate the semantic information contained in human navigation.

Findings and Contributions. Our results show that *human navigation in Wikipedia can serve as a high quality source of semantic information*. Using the proposed co-occurrence count model, we achieve a high correlation of 0.709 with the human intuition contained in the semantic relatedness dataset WS-353 (cf. Section 4.4.3). When restricting ourselves to only a small amount of paths with a low average indegree, we can even achieve a correlation score of 0.760. Furthermore, we find that the extrinsic bias on navigation that is imposed by a game setting does not significantly influence the semantic information in navigation behavior, i.e., that *unconstrained navigation on Wikipedia is an equally viable source for semantic information as game navigation*. Here, we can achieve a correlation score of 0.640 on the semantic relatedness dataset MEN (cf. Section 4.4.5). As a final result for navigation on Wikipedia, we can even *simulate navigation data with*

*near-human semantic information by performing biased random walks on the Wikipedia link graph. Using human navigation on BibSonomy, we can however only extract a smaller amount of semantic information with a low correlation of scarcely above 0.4 with the human intuition in WS-353 and Bib100. Although we also experiment with simulated navigation on BibSonomy and find that *simulated semantic navigation generates more useful semantic information* in the resulting paths than *human navigation*, we cannot achieve correlation scores above 0.5 on Bib100 (cf. Section 6.2).*

1.2.4 Relative Relatedness Learning: Learning Semantic Relatedness from Human Feedback

Problem Setting. As seen in the previous chapters, methods to extract semantic information from semi-structured and unstructured data produce considerably good representations of human intuition of semantics. Still, there are certain drawbacks still to overcome. Although the captured information is relatively precise, fine-grained information in special cases or exclusive to rare topical domains is rarely expressed in natural language text and thus difficult to include in vector representations. Here it would help to adapt the learned word embeddings to human intuition by warping the surrounding vector space by exploiting background knowledge about semantic relatedness. While there also exist approaches to adapt word embeddings to semantic lexicons, i.e., information about synonymous and antonymous words, they do not capture the *degree of semantic relatedness* between words. However, this information is crucial for a deep understanding of human intuition of semantic relatedness. Intuitively, humans are more inclined to compare the strength of a relation to another reference value than providing absolute judgments of that strength.

Approach. Starting with the intuition that the *relative relatedness of word pairs* is more important for determining the semantic quality of word embeddings, we gradually develop an algorithm to adapt a pre-trained set of word embeddings to such relative constraints. Because we learn the relative relatedness of word pairs, we call our algorithm *Relative Relatedness Learning (RRL)*. RRL is based on a linear metric learning approach. In order to show the usefulness of RRL, we evaluate it on several pre-trained word embedding datasets and train it on several semantic relatedness datasets. Furthermore, we combine it with an algorithm that incorporates lexical knowledge into the embeddings to discover potential synergy effects. After purposefully training wrong information to determine if we can manipulate word embeddings also in a negative direction, we attempt to transfer the knowledge contained in a semantic relatedness dataset on another to see if we can generalize the obtained knowledge. Finally, we evaluate RRL in a question answering task.

Findings and Contributions. The main contribution is an *algorithm to adapt pre-trained word embeddings to human intuition of semantic relatedness*, the *Relative Relatedness Learning algorithm (RRL)*. We find that we can exploit relative comparisons of word relatedness scores (as provided by WS-353 or MEN, cf. Section 4.4) to significantly

1 Introduction

improve the fit of word vectors to human intuition of semantics and *even outperform the state-of-the-art correlation with the MEN dataset*. We also show that a *combination of RRL and another approach called Retrofitting is even more useful to improve the semantic information in word embeddings*. Here, we can achieve an unprecedented correlation score of 0.904 on our MEN test data. Unfortunately, we are unable to transfer the knowledge in one relatedness dataset to others, which we attribute to the different kinds of represented semantic information. However, we can also show that when using specific word embedding datasets, we can even improve the quality of a question answering task.

1.3 Thesis Outline

We will now describe the structure of this thesis to give the reader an overview of each chapter. A sketch of the interdependence of all chapters is given in Figure 1.1.

Chapter 1 motivates this work, describes the research topics and contributions for each topic and finally gives an overview of this work. Chapter 2 provides an extensive overview of the related work to the central topics of this thesis, namely extracting and learning semantics from social media data under consideration of user influence. We will survey approaches which extract semantic relatedness from several sources, such as folksonomies or Wikipedia, subsequent extensions which fine-tune distributional models, and then present works shedding light on the mutual influence of user behavior and semantics. In Chapter 3, we will cover the necessary theoretical and technical background to understand the methods discussed and introduced in this work. Here, we first introduce the data structures used in this thesis, the methods on which we build our research and methods in this thesis, and finally methods to measure and compare user behavior. Chapter 4 then introduces and detailly describes the used datasets. We distinguish between text-based and link-based datasets, pretrained word embedding datasets and ground truth datasets.

Chapters 5 to 8 constitute the main contribution of this work. Their contents have already elaborated in Section 1.2, so we will only briefly describe each chapter. Chapter 5 focuses on the existence and strength of a semantic component in human navigation both on Wikipedia and BibSonomy. Chapter 6 then extends upon established methods to extract and evaluate semantic information from social tagging data (as proposed in [26] and [48]), before Chapter 7 broadly introduces human navigation in social media as a viable source of semantic information. Finally, Chapter 8 addresses the problem of imprecise semantics by proposing an algorithm to learn semantic relatedness from human feedback.

Chapter 9 concludes this work with a recapitulation of all chapters and a discussion about potential applications where our research could be put to use. We give pointers to future research directions about influencing factors on extractable semantics, potential new approaches to extract high-quality semantics, and challenges in supervised learning of semantic relations using background knowledge.

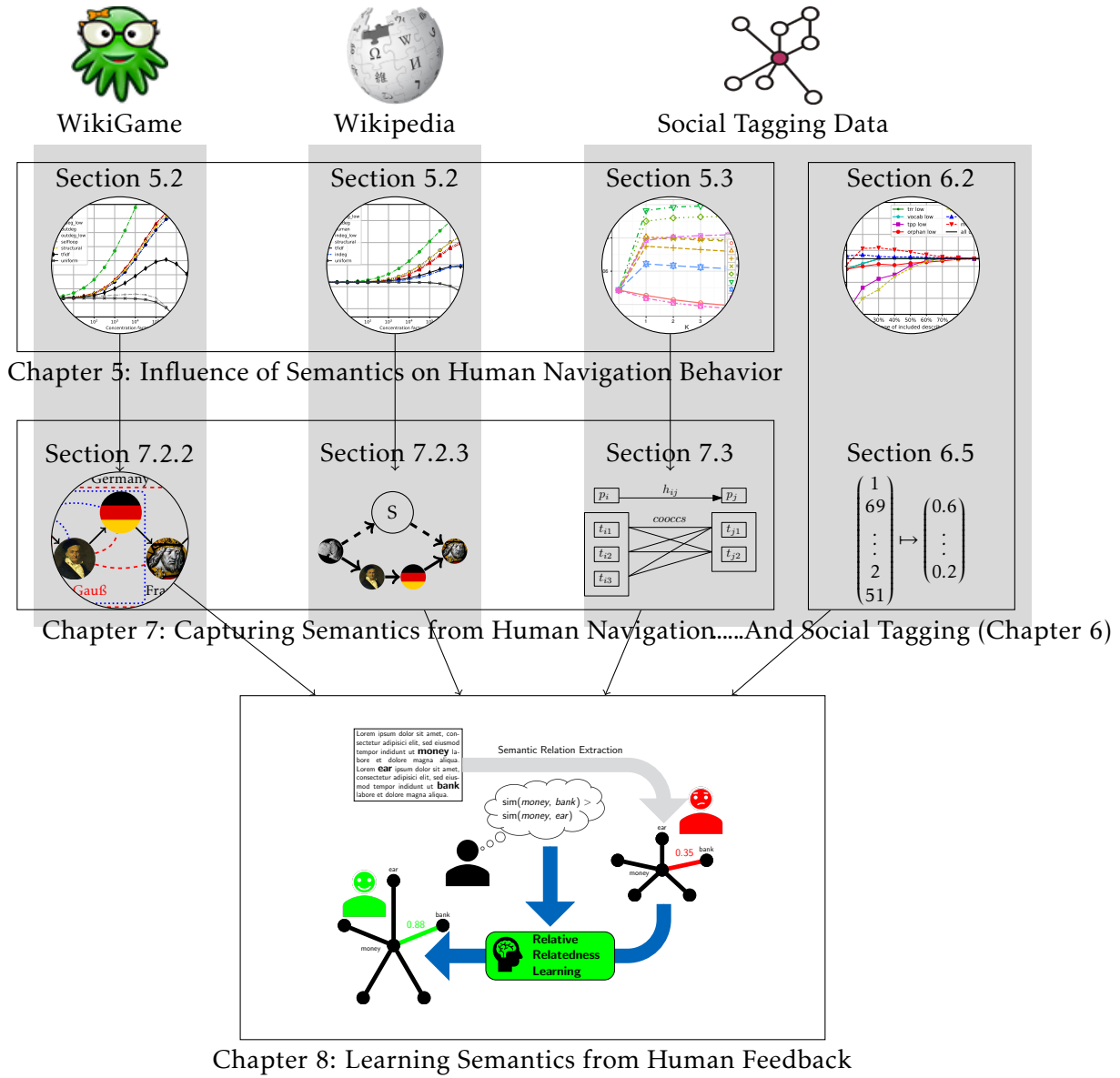


Figure 1.1: General thesis outline. In this thesis, we focus on methods to handle semantics in navigation and social tagging data. We contribute an analysis of navigation in social media, namely in Wikipedia and BibSonomy, with regard to a semantic component in navigation. After having shown the existence of this component, we propose and evaluate several methods to make the semantics in navigation visible. Additionally, we discuss how to evaluate semantics in social tagging data and propose a new way to represent them as vectors. Finally, we learn from human intuition to improve the semantic quality of word embeddings.

Chapter 2

Related Work

Although the basic idea of *computing semantics* and the application of computing technology on linguistics dates back much further to the beginning of the Cold War period [109], it only gained traction with the advent of more and more powerful computers with adequate processing capabilities to quickly iterate over large portions of text. Especially after Berners-Lee, Hendler, and Lassila presented their idea of a *Semantic Web* in 2001 [31], research has not ceased to present an unsurveyable amount of convincing results on how to extract and formalize semantic information from a wide range of sources, such as natural language text as well as many other forms of user contributed content.

Section 2.1 references different computational approaches to model the semantic information contained in various kinds of data. Because the main part of this thesis focuses on WordNet, Wikipedia and folksonomies as sources for semantic information, we only consider works that are based on data from these systems. After this, Section 2.2 presents the current state-of-the-art of methods which learn and adapt so-called word embeddings, that is, dense vector representations of words and their contexts. These methods are relevant for Sections 6.5 and 8, where we explore semantics of tag embeddings and propose a way to improve word embeddings to incorporate semantic background information. Section 2.3 discusses methods to measure the degree and the effects of the mutual influence of user behavior and semantics in social media. Again, we put a special emphasis on folksonomies and Wikipedia. In the final Section 2.4, we give a brief outlook on potential applications of computational semantics in several scenarios. Although this thesis mainly contributes methods to model semantic information, we feel that it is important for the reader to see how these models can actually be applied to a wide range of more practical uses, such as word sense disambiguation, sentiment analysis or query expansion problems. Section 2.5 concludes this chapter by recapitulating the presented topics and highlights the most important works for this thesis.

2.1 Extracting Semantics from Structured and Semi-Structured Data

This section will give an overview of related work that extracts semantic information from language and linguistic sources. Although there are many methods that deal with modelling semantic information in natural language, we will restrict ourselves on works utilizing data from WordNet, social tagging systems and Wikipedia, as these systems are heavily used in the main part of this thesis. We especially provide an extensive view on Wikipedia’s semantics, since a large portion of this work covers the extraction of semantic relatedness information from Wikipedia (e.g., Section 7.2). A detailed definition of the structure of all mentioned systems will be given later in Section 3.1.

2.1.1 Leveraging WordNet’s Taxonomy to Compute Semantic Relatedness

WordNet is a large lexical database of verbs, nouns and adverbs [72, 159]. Due to the highly precise semantic and linguistic information provided by WordNet, several researchers proposed semantic similarity measures based on WordNet. Some well-known and frequently used semantic similarity measures are Hirst and St-Onge’s [105] path-based measure, the information content based distance by Resnik [195] and lastly the Jiang-Conrath distance [112], which combines the information content nodes and path distance features. Budanitsky and Hirst analyzed and compared all of them by evaluating them on human intuition [44]. Their results show that the Jiang-Conrath measure outperforms its competitors in measuring the semantic similarity of WordNet concepts and additionally exhibits a very high correlation with human intuition. Consequently, we will utilize the Jiang-Conrath measure (or at least the similarity scores that it produces) in Section 6.2 as one alternative in a comparison of semantic similarity gold-standards in order to evaluate the semantic similarity of tags. A similarity measure that exploits WordNet’s graph structure more intelligently than [105] was proposed by Hughes and Ramage [108]. They treat the WordNet graph as a Markov chain and use a generalized PageRank variant to calculate a stationary distribution per word. Then they interpret a modified version of the Kullback-Leibler divergence between the distributions of two concepts as their semantic similarity score. Ramage, Rafferty, and Manning extend this approach on short text similarity [191]. Obviously, all those methods rely on the vocabulary and structure of WordNet. They are thus unable to deal with out-of-vocabulary (OoV) words. In order to deal with such OoV words, Agirre et al. combine a WordNet-based approach reminiscent to that of [108] with two distributional approaches, namely bag-of-words and window-based word co-occurrence counting on a corpus of 1.6 Terawords [2]. On a monolingual word similarity task, they achieve a relatively high correlation value with human intuition of 0.66.

2.1.2 Capturing Semantic Information from Folksonomies

Folksonomies are structures emerging from social tagging systems and have played an important role in the development of the Web2.0. However, they are also closely

connected with the Semantic Web. In this thesis, we work with both static folksonomy data as well as human navigation data on the link network of social annotation systems. This link network is also based on the folksonomy structures.

Connections between folksonomies and the Semantic Web. Wu, Zhang, and Yu aimed to enhance the Semantic Web vision of semantically annotated resources by utilizing the emergent semantics of social tagging systems [252]. Instead of *first* defining an ontology to *then* annotate web resources, they argue to exploit the emergent hierarchical and semantic structures of *already existing* annotations in social tagging systems to construct taxonomies. For this, they model the tagging process as a probabilistic generative model which is learned using an EM approach. Mika went even further and argued that folksonomies already exhibit hierarchical semantic structures on their own, thus being lightweight ontologies themselves [155]. Following this, Specia and Motta explored ways to directly integrate folksonomies and their inherent semantic information into ontologies [215]. Finally, Angeletou mapped tags to concepts in existing ontologies [5].

Capturing semantic information from tagging data. Instead of directly integrating existing ontologies and folksonomies, there also exist several methods to explicitly capture and model these emergent semantics. For example, Cattuto et al. proposed several ways to measure the semantic similarity of tags by modelling them as context vectors [48]. This paper is a central aspect to the way we model tag semantics in this thesis. We will explain the different context variants in Section 3.2.2.1, discuss their evaluation approach in Section 6.2, and finally extend upon their methodology in Section 6.5. These context vector representations played a central role in [148], where Markines et al. compared several vector similarity measures with regard to the measured semantic similarity. They found that although mutual information of tags provided the best results, the slightly worse cosine similarity of the corresponding tag vectors scales better and is thus recommended. Benz et al. also used the tag semantics model from [48] to evaluate a range of term abstractness measures and concluded that centrality as well as entropy based measures are good indicators for measuring the semantic generality level of tags [29]. They then used the best-performing generality measure as well as the tag-tag-context vector representations from [48] to induce a taxonomy from tags in [28]. This tag taxonomy construction algorithm that has been proposed by Benz et al. in [28] is actually an extension of a simpler version of the algorithm as presented in [103] by Heymann and Garcia-Molina. The added steps include both synonym combination as well as tag sense disambiguation. While synonyms are detected using the tag similarity measures presented in [48] and [148], a clustering approach is used to determine different tag senses that is reminiscent of that in [267]. In [26], Benz performed some experiments with that algorithm and found that it produces reasonable results, when compared to WordNet. We use that algorithm later on to determine the influence of user behavior on tag sense discovery in Section 6.4. In fact, we build on most of these works mentioned above to measure the semantic relatedness between tags, both in the static folksonomy (cf. Chapter 6) as well as in navigation on the link graph

2 Related Work

of the social tagging system (cf. Section 7.3).

2.1.3 Capturing Semantic Information from Wikipedia

Wikipedia is a large, publicly editable online encyclopedia with a wide content coverage in about five million articles. It often serves as a basis for a huge multitude of research articles, for example in computational semantics. As we will often use Wikipedia as a source for semantic information in this work, especially in Section 7.2, we will cover related work for approaches to extract semantic information from each of the following categories: text-based approaches, approaches based on Wikipedia’s category taxonomy, link network based approaches, usage based approaches, and hybrids thereof.

Text-based approaches. Wikipedia’s article content is one of the most used natural language text datasets in research [18, 81, 135, 154], not only because of its quantity of topics, but also because of the high article quality [84, 153] and semantic stability over time [219]. One of the most well-known works about semantic relatedness on Wikipedia was presented by Gabrilovich and Markovitch [81]. They propose the *Explicit Semantic Analysis* measure, which calculates semantic relatedness between Wikipedia concepts. For this, they use the plain Wikipedia article context to construct TF-IDF vectors. Concretely, they construct a term-document matrix, where the TF-IDF weighted occurrences in documents are seen as features. Hassan and Mihalcea [97] propose a similar notion by exploiting the textual context and the label of article links. A word is thus not expressed by its direct context, but rather only by the links in its immediate context. These links are then weighted using Pointwise Mutual Information. Both methods are based on the idea to facilitate Wikipedia articles as features, thus basically representing term-document approaches. Furthermore, the Wikipedia article corpus is an often used resource to learn word embeddings [18, 135, 184]. We will use a word embedding dataset trained on the Wikipedia corpus of 2014 in Chapter 8.

Category taxonomy based approaches. The potential of Wikipedia’s *category taxonomy* for calculating semantic relatedness was shown by Strube and Ponzetto [223]. They apply several measures originally proposed for the WordNet taxonomy on the Wikipedia category taxonomy and compare the results with several baseline approaches from WordNet. By using the category structure of Wikipedia, they outperform Google count based methods and a WordNet baseline. However, they obtain the best result using Wikipedia, Google and WordNet in combination. In fact, this has been amongst the first works to characterize the semantic relatedness of Wikipedia articles, without explicitly considering the context. Radhakrishnan and Varma exploit Wikipedia categories as semantic tags for articles, which they use to extract semantically related word pairs [188]. While we also relate Wikipedia articles with each other in Section 7.2, we use Wikipedia’s link network instead of the category taxonomy.

Exploiting the static link network. While the article texts contain an abundance of knowledge, another strength of Wikipedia is also the connection of knowledge by hyperlinks. These hyperlinks connect articles which are in some way related to each other. Since most articles represent semantic concepts, this link network also represents semantic knowledge, which can be extracted. Witten and Milne make use of the Wikipedia hyperlink structure in order to calculate semantic relatedness [251]. They propose the Wikipedia Link-based Measure (WLM), which is similar to TF-IDF and the Google distance measure [53], and evaluate a combination of both on article-link sets obtained from Wikipedia. Their results show that WLM outperforms both WikiRelate and ESA. Similarly, Agirre, Barrena, and Soroa [3] study Wikipedia links to determine semantic relatedness and disambiguation through random walks on the full Wikipedia link dataset instead of the subset of directed, non-reciprocal links. In this work, we only use the Wikipedia link structure as a baseline to show that *human navigation on these links* yields improved semantic information. Other applications that exploit Wikipedia’s link graph have been for example proposed by Guo et al., who also exploit the Wikipedia link graph, however for disambiguating and linking entities to knowledge base entries [92]. Zesch and Gurevych [261] investigate the temporal influence of change in Wikipedia’s link network on its semantic content. Similar to [238], they could see that with several semantic similarity measures, the content yielded stable semantics.

Hybrid approaches. Finally, there exist approaches which combine both the article contents and the link graph information. In [258], Yazdani and Popescu-Belis augment the Wikipedia link network with links based on semantic relatedness and then perform random walks on that network. Finally, they approximate the resulting visiting probability distribution for each node (and sets of nodes for text fragment similarity) in an embedding space. Their results indicate a significant improvement in correlation with human intuition compared to LSA, when they weigh the constructed semantic links higher than Wikipedia’s hyperlink structure in their relatedness measure. Zheng et al. make use of content and structure of Wikipedia, considering inlinks, outlinks, categories, Dice-, Google- and TF-IDF-based measures to compute semantic relatedness. They combine all measures using supervised machine-learning to calculate individual weights for all approaches [266]. Using the optimal weights, they achieve a correlation of up to 0.72 with WS-353. Unfortunately, the authors don’t provide information from where they obtained their training data. In [259], Yeh et al. apply the method from [108] on Wikipedia, i.e., calculating a Personalized PageRank on the Wikipedia graph. This resulted in small improvements upon other relatedness measures like ESA [81]. Lastly, Dallmann et al. [55] investigated random walks on the Wikipedia link graph in order to extract semantic relatedness information. They found that with a simple uniformly primed random walk, the extracted semantics produced already competitive evaluation scores. This work will be presented in Section 7.2.4. Additionally, we perform further experiments with non-uniformly primed random walks, amongst these a random walker that navigates towards articles, whose content is similar to the current article. This way, we try to artificially reproduce the semantic information content in human navigation.

2.1.4 Evaluating Extracted Semantic Information

Naturally, there exists a plethora of approaches to evaluate semantic information models if they fit to human intuition. In [44], Budanitsky and Hirst identified three types of evaluation scenarios. First, a theoretical evaluation of mathematical properties of a measure, as for example proposed in [137]. Second, comparison with human judgment of Word Similarity. Third, the use of an extrinsic task, i.e., in an external framework that somehow makes use of models of semantic information. In the following, we will present works about the latter two approaches, but we additionally also describe two other approaches that are found in literature, namely analogy testing, which has only recently been proposed, and multiple choice testing.

Word Similarity. Probably the most widespread evaluation approach is to compare similarity scores of word pairs, as produced by a semantic representation model, to human judgment of the semantic similarity of those pairs [18, 44, 81, 97, 210]. The first datasets containing such scores have been proposed in 1965 [199] and 1991 [160]. After that, it took some years to develop the bigger and today very famous WS-353 (short for WordSimilarity-353) dataset [75]. After this, several other datasets have appeared [43, 104, 143, 171]. The actual evaluation score is determined by computing the correlation between human and model scores. We will use this method heavily in this work and will describe it in detail in Section 3.2.4, as well as the used human intuition datasets in Section 4.4. Furthermore, we introduce a new semantic relatedness dataset with human judgment to explicitly measure the semantic information of domain-specific vocabulary from BibSonomy. In fact, there also exists another evaluation approach that is based on measuring word similarity, albeit in an indirect way. Cattuto et al. proposed to interpret the scores of the Jiang-Conrath measure on WordNet as a close resemblance of human intuition [48]. However, they did not compute the correlation of their scores with Jiang-Conrath scores, but rather the *average Jiang-Conrath distance of tags to their closest concept as determined by the model*. In Section 6.2, we compare this evaluation approach with evaluation on human intuition. In [148], Markines et al. then again computed correlation between model scores and Jiang-Conrath scores, i.e., they combined both approaches.

Analogy Testing. One standard way of evaluating semantic relatedness measures or word embeddings is to measure how well analogies are found in vector space [135, 156]. For example, given a 4-tuple of words (a, b, c, d) , the goal is to match d as closely as possible with the vector $b - a + c$. This task was inspired by Turney [229]. The most prominent example and motivation for many word embedding papers was given in [156] by the illustrative equation $king - man + woman \approx queen$. Technically, this simply tests if the embedding algorithm can also encode implicit relations into the word embeddings. Idea-wise, this then leads to embedding explicitly named relations, as done in Knowledge Graph embedding methods, such as TransR, TransE or TransH, to just name a few.

Multiple Choice Tests. Multiple Choice Tests are a standard question type, for example in IQ tests. A variant of this is to let the word embeddings take a TOEFL-test [77], which is basically a similar scenario as the analogy tests, except that now out of word groups, the odd-one-out is to be found instead of the correct word. Levy, Bullinaria, and McCormick recently proposed a new vocabulary test to evaluate semantic vectors on, as they identified some practical problems in the use of multiple choice tests as general evaluation measures [133].

Extrinsic Evaluations. Finally, it is also possible to indirectly evaluate semantic word embeddings by applying them in an extrinsic task. For example, Wieting et al. used their word embeddings in a sentiment analysis task by training a convolutional neural network with them [250]. In turn, this neural network had to classify sentences regarding their sentiment [118]. In a different setting, Both, Thoma, and Rettinger [39] applied word embeddings to a knowledge base completion task. Lastly, Chen et al. proposed a framework for question answering by reading Wikipedia articles [49]. Their framework allows to take a set of word embeddings as input, so this can also be used as an extrinsic evaluation task. We will present a broader overview of several applications of semantic information models in Section 2.4.

2.2 Learning Semantics with and without Background Information

Opposed to the simple extraction of semantic relatedness information from text or knowledge bases, as done by the approaches presented in the previous section, there has been a lot of work done lately towards *learning* such representations, in both unsupervised and supervised settings. We present approaches to learn and refine word embeddings using external knowledge, thus covering both unsupervised and supervised learning approaches. Here we also talk about approaches that include external knowledge directly into the embedding process. Because co-occurrence information between words can also be depicted as a graph, we additionally presented graph embedding approaches. In Section 6.5, we first train embeddings from tagging data to investigate if these representations can capture the semantic information of tagging data better than high-dimensional co-occurrence count vectors. Finally, we propose an algorithm to adapt word embeddings to external knowledge of semantic relatedness in Chapter 8 that is based on a linear metric learning approach.

2.2.1 Learning Word Embeddings

While count-based methods as presented in Section 2.1 are able to capture semantic information from language corpora pretty well, they are nothing else than more or less sophisticated heuristics. The following word embedding approaches on the other hand are *optimized* for a given task, e.g., to predict the context of a word or to maximize the information content of the vector model. All of the models in this section can be seen

2 Related Work

as unsupervised approaches, i.e., they do not exploit external knowledge to learn word embeddings.

Encoding semantic information in word embeddings. Ever since Mikolov et al. published Word2Vec in 2013 [156], the success of so-called *word embedding* models could not be halted anymore. In his seminal work, Mikolov showed that constructing low-dimensional word models could be done without costly factorizing huge, sparse matrices as they appear in co-occurrence models [46, 230] and also without training a complete language model [25] with a lot of overhead and a very long training time. For this, he used two shallow neural network with only one layer to predict words from their contexts or vice versa. Since then, Word2Vec has established itself as the de facto baseline for each and every NLP paper in existence that only touches word embeddings. This is hardly surprising, as Baroni, Dinu, and Kruszewski showed in [18] that word *prediction* models generally outperform co-occurrence counting vector representations on a number of tasks including determining word relatedness, synonym detection, concept categorization and analogy detection. In their work, they argue that *“this new way to train [distributional semantic models] is attractive because it replaces the essentially heuristic stacking of vector transforms in earlier models with a single, well-defined supervised learning step”* [18].

Count-based vs predictive embedding models. On the other hand, Pennington, Socher, and Manning proposed a simple model called GloVe [184], which makes use of global first-order co-occurrence statistics to construct word vectors. In their experiments, they again could outperform Word2Vec, thus directly contradicting Baroni’s results. In [135], Levy and Goldberg then performed an extensive comparison of count-based and prediction-based word representation models and explore the parameter space for PPMI matrices, SVD, Word2Vec’s Skip-Gram method and the GloVe algorithm and compare the resulting vector representations regarding to their performance on word relatedness and analogy tasks. They find that in contrast to the findings of [18] and [184], predictive embedding methods such as Word2Vec do not necessarily outperform count-based methods, but both are rather on par with each other, if the training parameters are chosen correctly. With these results, they finally gave recommendations how to get the most out of self-trained word embedding models [134].

Encoding semantic information using graph embeddings. Another potential way of embedding words is by constructing a co-occurrence graph and then applying a graph embedding algorithm, such as DeepWalk [185], LINE [227], node2vec [87], or HARP [50]. The DeepWalk [185] algorithm by Perozzi, Al-Rfou’, and Skiena also aims at embedding nodes in low-dimensional vector spaces. However, instead of sophisticated learning algorithms, DeepWalk conducts uniformly random walks on the graph and then trains a Word2Vec model on these walks. This way, they generate context for nodes and receive embeddings whose similarities are relative to node distances in the graph. Building on that, [87] by Grover and Leskovec modifies DeepWalk in such a way that the

random walks are not completely random anymore, but follow a depth- or breadth-first search algorithm. Tang et al. proposed the LINE algorithm [227], which addresses the problem of embedding very large information networks with up to millions of nodes into low-dimensional network spaces. LINE also preserves both the local and global properties of the graph, i.e., first-order node proximities (edges) as well as second-order proximities, which are represented by the shared neighborhoods. They perform several experiments on large real-world datasets and outperform DeepWalk as well as Mikolov's SkipGram, while also using less training time. Finally, the HARP algorithm proposed in [50] by Chen et al. builds upon DeepWalk, LINE, and Node2Vec by not only embedding the neighborhood of a node, but also the hierarchical structure of the graph. Using this approach, they are able to achieve an improvement of up to 14% of the Macro F1 score compared to the original implementations.

Combining embeddings with taxonomies and ontologies. Lastly, there have been attempts to not only induce semantic relations between words, but even combine these relations into complete taxonomies. Although While Fu et al. proposed a vector space transformation approach to determine hierarchical relations from word embeddings [78], Ristoski et al. built on that approach to induce large-scale taxonomies. Nickel and Kiela went a step further and developed an algorithm to directly embed such hierarchical relations onto a Poincare disk space [168]. However, this discards information about the semantic similarity relations between words. Lately, some researchers extended the idea of graph embeddings to knowledge graphs, i.e., ontologies, with different and explicitly named relations. For this, several algorithms have been proposed to embed the relations between two entities in the entity space as well. Well-known examples are TransE [37], TransG [256], Flexible Translation [73] and TransH [241]. We will not explain them here, since this is out of scope for this work. While embedding ontologies relies on the existence of ontologies, a possible option to actually find potential new relations between entities in an embedding space to enrich an ontology embedding is proposed in [80]. In this work, Fulda et al. extract "*feasible applications*" for entities, i.e., how these entities are used in the real world, from plain Wikipedia article text using a reinforcement learning setting and pre-trained word embeddings.

Out-of-vocabulary problems in word embeddings. A major issue of word embedding algorithms is that they learn the embeddings in an offline manner, which limits the available vocabulary. Adding new words often makes retraining the model necessary. To solve this, Luo et al. proposed an approach to learn word embeddings in an online manner [142]. Bahdanau et al. [14] use the description text of potential dictionary definitions to handle OOV words. Another option would be to train *sub-word embeddings*, i.e., embeddings of word parts instead of the whole words, as proposed by Pinter, Guthrie, and Eisenstein [186]. This way, at least new words which are constructable from those word parts could be added without retraining the whole model.

2.2.2 Learning to Adapt Word Embeddings to Background Knowledge

While the task to correctly determine the semantic relatedness of words or texts has been around for a long time, there are still few approaches which actually learn semantic relatedness in supervised settings. Bridging the gap between unsupervised relatedness learning approaches and human intuition by injecting background knowledge can be accomplished with post-processing methods or with directly injecting this knowledge in the embedding process.

Incorporating synonymy and antonymy in word embeddings. Faruqui et al. and Halawi et al. presented methods to fit pre-trained embedding vectors to their neighborhood defined by synonymy relations [69, 93]. Additionally, Mrkšić et al. proposes to additionally inject antonymy constraints [163]. The aim of all three works is to maximize the similarity of synonymous words, and in the case of [163], simultaneously minimizing the similarity of antonyms. Hereby, synonymy constraints acted as attractors in the semantic vector space, while the antonymy constraints acted as repellants. These constraints are often taken from so-called “lexicons”, i.e., lists of synonymous words. For example, synonym information can be taken from WordNet or the Paraphrase Database [82]. However, there rarely exist curated lists of antonyms. Wieting et al. [250] presented paraphrase models using Recursive Neural Networks to fine-tune existing word embeddings with knowledge extracted from paraphrases, i.e., synonymous segments of text despite differences in structure and wording. This is a similar approach to using lexicons, except that the paraphrases that Wieting uses are actual texts, as opposed to singular words as in the approaches of Halawi, Mrksic and Faruqui. Yu and Dredze directly include similarity constraints from the PPDB database into the word embedding process [260]. However, they also experiment with pretrained word embeddings, on which they then additionally train their model. In their experiments, their proposed model initialized with pretrained word embeddings outperforms all other models in a synonym prediction task.

Incorporating the strength of semantic relations into word embeddings. One fundamental issue of the abovementioned approaches is that they cannot distinguish the actual *degree of similarity or dissimilarity*. For example, *great* and *good* are semantically similar, as are *good* and *amazing*. However, the *intensity* of similarity, i.e., the strength of the relation, is greater between *great* and *good*, as for example the sentiment distance between both words is smaller than between *good* and *amazing*.

The following approaches are able to differentiate better between similarity and dissimilarity, as they also include the “intensity” or degree of similarity. Kim, Marneffe, and Fosler-Lussier propose to learn a vector space transformation, where on the one hand similar words are closer in space, but they also include “intensity information” in their learning process [117]: In a cluster of closely related words, they order them by intensity and base their metric on that. The relative intensity of a word is found by counting the number of n-grams “*x but not y*” in google searches for two words *x* and *y*, e.g., *good but not great*. This means that *y* is more intense than *x*. Niebler et al.

proposed an approach to learn semantic *relatedness* from human judgment of word pair relatedness [170]. However, they do not use the actual judgment scores, but, in a similar fashion as the Spearman rank correlation coefficient, exploit the relative order of word pairs, determined by the judged scores. Their results show significant improvements in rank correlation upon those of the vectors they learned from. However it is not clear if this improvement would also notably benefit a human-centered application, i.e., if humans would even note the quality increase of the applied word embeddings. We will introduce this algorithm in detail in Chapter 8.

Incorporating external semantic knowledge in the embedding process. There also exist methods which incorporate background knowledge directly into the embedding process. In [177], Ono, Miwa, and Sasaki proposed a word-embedding based approach to automatically detect antonyms in a supervised setting. For this, they include antonym and synonym constraints into the embedding process and then use these embeddings to predict word similarity. For example, Bian, Gao, and Liu leverage morphological knowledge, i.e., information about word prefixes or syllables, in the embedding process to overcome vocabulary incompleteness and word ambiguities [33]. While this approach improved results greatly on the word analogy task compared to Word2Vec, it performed similarly well on word relatedness and sentence completion tasks. In [167], Nguyen, Walde, and Vu integrate antonym and synonym relations directly into a Skip-Gram model.¹ They do this by additionally strengthening the most salient features for determining word similarity.

Relation of word embedding post-processing methods and metric learning. Finally, it is interesting to point out that, while most authors do not mention it, these approaches are very similar and sometimes even inspired by well-known metric learning algorithms. The counterfitting approach by Mrkšić et al. in [163] uses a very similar objective to the Large-Margin Nearest Neighbor Metric Learning algorithm proposed by Weinberger and Saul [243], as in both works similar vectors are attracted and dissimilar vectors repelled from each other. Also, in Chapter 8, we explicitly state that the Relative Relatedness Learning algorithm is largely inspired by the Least-Squares Metric Learning algorithm by Liu et al. [139].

2.3 Mutual Influence of Semantics and User Behavior

Language is an intrinsic part of how humans express their thoughts about the world. It is driven by expressing the conceptualizations that humans make when trying to understand their surroundings. While the semantics of human-generated web content are obviously influenced by human behavior, it is not as trivial to characterize semantics as an influence factor on behavior. Addressing this question, Pirolli and Card [187] proposed their *Information Foraging Theory*, which describes how humans search for information. It is based on the assumption that humans follow an *information scent* to find

¹<https://github.com/nguyenkh/AntSynDistinction>

the most lucrative gain in information and are able to adapt to external circumstances to maximize that gain [51]. Since we will investigate user behavior and its influence on semantic information in many parts of this thesis, we will give an overview of works regarding the mutual influence of semantics and user navigation.

2.3.1 Modelling (Semantic) Influence on Human Navigation Behavior

As already mentioned in the introduction, Chi et al. argued in [51] that users follow an *information scent* to reach their goals in navigational settings. Thus, Berendt [30] used site semantics to support and improve navigation on business websites. Juvina and Oostendorp show that navigational models should also consider semantic and structural knowledge in order to find suitable models of human navigation in the web [114].

Modelling navigation behavior using Markov chains. A promising approach to model human navigation is provided by Markov models, which have been previously used for clickstream data by Bollen et al. [36]. However, Singer et al. noted that although Markov chains of higher order are too complex and inefficient to be actually useful, the memoryless Markov model of order 1 does account well for navigation on the Web [208]. Singer et al. then presented the HypTrails method which is used to compare different hypotheses about user navigation [207]. The HypTrails approach has also been successfully applied by Becker et al. on navigational paths on the MicroWorkers crowdsourcing platform [19] as well as on real-life movements derived from temporal and geospatial information from FlickrR photos [22]. Furthermore, it has also been implemented in a distributed fashion [21]. Finally, Becker et al. extended this model to mixed hypotheses of potentially longer orders [20]. Both the HypTrails and MixedTrails approaches make it possible to measure the semantic influence on navigation. In fact, in [20] and [207], the authors show that game navigation on Wikipedia, last.fm and Yelp are all driven by semantic navigation, at least to some extent. In Chapter 5, we utilize HypTrails to find and quantify a potential semantic component that drives human navigation behavior in information networks, namely Wikipedia and BibSonomy.

Structural features of human navigation. Lamprecht et al. [128] identify features in the structure of Wikipedia articles which influence user navigation on Wikipedia. In general they find that users tend to click links which are located near the top of an article page, which are at the same time more general than links in the lower parts of articles. Dimitrov et al. finally investigate navigational behavior on unconstrained user navigation on Wikipedia [60]. There, they find that humans prefer navigation towards less connected links in the network but still navigate semantically. Similar to [128], they also find that the position of links in an article influences navigation, indicating a “visual top-left preference” of links. Finally, in [61], Dimitrov et al. extend the analyses from [60] with a hypothesis-based investigation of general user navigation behavior on Wikipedia. They find that users navigate more towards the periphery of the Wikipedia link network, i.e., towards less connected pages as well as towards semantically similar articles, but mainly if they are visible to the user. In Section 5.2, we build upon some

of these works, where we analyze and compare the navigation behavior of users in Wikipedia in a game setting and an unconstrained setting. This in part overlaps with some of the cited works above, however we put a special emphasis on the strength of semantic navigation in Wikipedia.

2.3.2 Folksonomy Usage Patterns and their Influence on Semantics

The semantic knowledge contained in the tag distributions of folksonomies is as diverse and subjective as the users that create this knowledge. This knowledge emerges and stabilizes through constant usage, after which we can extract semantic information, but also analyze user behavior and motivation and its influence on this semantic information.

Stabilization of semantic information through regular usage and contribution. In [149], Mathes hypothesized that tag distributions follow a power law distribution. Consequently, Golder and Huberman showed in [85] that after a certain time span, regularities in user activity, tag frequencies and relative frequency proportions could be observed. Wagner et al. [238] stated that social interaction in tagging systems allowed the inherent semantics to stabilize faster than in tagging systems without social interaction. This is largely attributed to the fact that social interaction options allow users to copy other users' posts, thus easily contributing to the stabilization process. In turn, this stabilization motivated further investigations of tagging systems, especially about the effective extraction of semantically stable content [28, 29, 48] and motivation of tag usage [4, 94, 122, 221].

Analyzing usage of social tagging systems. Heckner, Heilemann, and Wolff conducted a user survey on their motivation of using social tagging systems, that is, if users store resources for either their own retrieval or social sharing purposes [100]. That survey showed that on Delicious, retrieval of own resources was a more important purpose of using social tagging systems. Fu et al. [79] argued that in social tagging systems, the fact that users can see the posts of other users leads to imitation effects: In their experiments, they show that if users can see the tags of like-minded users, they tend to assign the same or at least semantically similar tags. While there exists a large amount of literature on tagging systems, to the best of our knowledge, there exists only a small amount of work utilizing and analyzing log data from a tagging system. Millen and Feinberg [157, 158] investigated user logs of the social tagging system Dogear, which is internally used at IBM. They could find strong evidence for social navigation, i.e., users are looking at posts from other people instead of mainly their own. Damianos et al. also proposed a social tagging system for a corporate intranet and to utilize log data to characterize users according to their behavior [56]. In [64], a thorough study of user behavior in BibSonomy and its potential influences was presented by Doerfel et al. The study is based on four assumptions about the usage of social tagging systems, e.g., the assumption that users are *social* and provide content to share it with other users instead of storing it for self-retrieval. Furthermore, Niebler et al. investigated several hypotheses about navigation in a social tagging system [173]. While (similar to the

2 Related Work

results of [100]) users indeed mostly navigated on their own pages, they did so by using semantic aspects, thus uncovering a semantic component in human navigation on social tagging systems. We will extend the findings of [64] and [173] in Section 5.3, mainly focusing on the explanation of navigation behavior with hypotheses.

Identifying and classifying tagging motivations and their impact on tagging semantics. To characterize *usage pragmatics* in social tagging systems, i.e., the *tagging motivations* of users, Strohmaier, Körner, and Kern introduced two user prototypes [222]: First, the *categorizers*, which assign selected tags from a carefully designed and controlled vocabulary, meant to categorize the posted resources for better retrieval afterwards. Second, the *describers* with completely opposite behavioral traits, such as rather describing resources with a free and large vocabulary, but without any structure. In consequence, Körner et al. explored the impact of categorizers and describers on the quality of semantic representations extracted from tagging data [122]. Their results show that describers, due to their rich vocabulary and expressiveness generate more useful data for semantic extraction. In [271], Zubiaga, Körner, and Strohmaier explored the influence of tagging pragmatics on emergent social classification, finding that categorizers produce more useful tags than describers for this task. While most of the works rely on the assumption that many users contribute to the contents of social tagging systems, Lorince et al. argue that most of the content is actually produced by “supertaggers”, i.e., extremely active users [141]. Their results show that for example in Delicious, roughly 5% of all users account for more than 50% of social annotations. They could also show that supertaggers are rather counted as describers. Niebler et al. defined two additional classes of users, based on the semantic generality of their tags [176], namely *generalists* and *specialists*. We will cover this work in detail in Section 6.4, where we will show that tagging pragmatics exerts influence on the performance of tagging semantics, in this case on the discovery of different tag senses. We found that small portions of the most extreme describers and generalists produce the best tagging data for word sense disambiguation. Moreover, we also use these two new tagger classes to show that they are in some cases even more useful for extracting semantic relatedness than describers and categorizers in Section 6.3.3, thus extending the work of Körner et al. [122]. As a last type of users that have large impact on the quality of social tagging systems in several aspects, if not detained, research has also investigated the behavior of *anti-social taggers*, i.e., spammers. Krause et al. proposed several algorithms to detect spammers in social bookmarking systems [125], considering several features, such as profile features, location based features, activity based features, and semantic features.

2.3.3 Influence of Navigation Patterns in Wikipedia on Semantic Information

Millions of users use the collaborative encyclopædia Wikipedia on a daily basis to satisfy their information needs. Naturally, most users do not contribute information, but only consume the provided contents. Understanding the users’ needs to consume that information, especially their exploration behavior in Wikipedia’s hyperlink network to

access connected information, has been extensively researched throughout many works.

Analysis of human navigation behavior on Wikipedia. West, Pineau, and Precup created a game called Wikispeedia to supervise user navigational behaviour [249]. They used the data collected from this game to examine the taken paths when provided with a navigation task. It was shown that most navigation followed a “zoning-out-homing-in pattern”, i.e., users first navigate to general pages, before narrowing down on the target page. They also proposed a new semantic relatedness measure for pairs encountered in paths, but it only can measure relatedness between concepts if they appeared on a path together. In [247], the authors examine a bigger dataset from Wikispeedia than in [249] and develop a method to predict the next target in a path. Their findings in human navigation analysis support their previous results, which they successfully apply to target prediction, what in turn could be used to improve website navigation design. In [203], Scaria et al. predicted user abandonment of navigation games in Wikispeedia after a few clicks. In their analyses, they also took the semantic distance between the current page and the goal concept into account and showed that unfinished paths often do not come anywhere near the target page in a game, when considering the semantic distance between the current page and the target page, which in turn leads user to give up navigation. West and Leskovec proceeded to design heuristic-driven random walkers [246]. Their goal was to optimize finding the shortest possible paths for Wikipedia navigation games. Although humans were already relatively good at this task, random walkers primed with sophisticated heuristics were often able to outperform humans. Using unconstrained Wikipedia navigation data, West, Paranjape, and Leskovec recommended links that should be inserted in Wikipedia [248] to improve user navigation and information gathering. It exploits human wayfinding strategies to circumvent missing hyperlinks to give recommendations for potential links to be inserted. However, by inserting hyperlinks, they change the data needed for link-based semantic relatedness approaches (see Section 2.1.3). Lastly, Singer et al. analyzed webserver logs of Wikipedia together with results from a user questionnaire to find out why people actually use Wikipedia [209]. They found that there exists a large spectrum of usage reasons with the most picked questionnaire response being that people heard about a topic in the media.

Extracting semantic information from human navigation on Wikipedia. Improving upon the semantic relatedness measure as proposed in [249], Singer et al. built an approach to exploit the sequential nature of pages in navigational paths to calculate semantic relatedness [210]. Here, they considered a large dataset of human navigational paths from the WikiGame². Additionally, they show that human navigation contains a significantly increased amount of semantic information, compared to the plain link network or to artificially created paths, thus making the suitability of human navigation for semantic relatedness extraction visible. We will cover [210] in Section 7.2.2, where we will additionally compare the original results from the WikiGame to those

²<http://www.thewikigame.com>

we obtain by applying our methodology on Wikispeedia data. We find that while game navigation data in general contains very useful semantic information, a small portion of low-indegree paths contains even more precise semantics. Niebler et al. [175] built on the results from [210] and explored unconstrained navigation using the ClickStream data from the Wikimedia foundation [255]. We will discuss this approach and its results in more detail in Section 7.2.3. We first compare game and unconstrained navigation by evaluating the semantic quality of both navigation variants. After showing that unconstrained navigation also differs from artificial navigation, we concluded that unconstrained navigation is similarly well fit to provide semantic information. However, we cannot apply the exact same methodology to extract the semantic information as we did in [210]

Simulating human navigation with random walks on the Wikipedia link graph. However, the investigated dataset of unconstrained Wikipedia navigation in [175] only contains transition information instead of full-fledged navigational paths or even navigation sessions by specific users as in the WikiGame. A potential idea is to conduct random walks on the Wikipedia link structure to artificially augment navigational data. In [246], West and Leskovec compared the ability of random walks to those of humans to find shortest paths. While humans already are quite good at this task, by drawing on huge amounts of background knowledge, West and Leskovec could show that algorithms with sophisticated heuristics are able to surpass humans on this task. In [55], Dallmann et al. then exploit random walks to extract semantic relatedness. While on the one hand random walks generated by uniformly random navigation yield sufficiently good semantics, they actually do not encode human navigation. However, when priming the random walker with weights obtained by human navigation, the resulting pseudo-human random walks achieved competitive results in a semantic relatedness task. A related approach to [55] was presented by Zhao, Liu, and Sun [265]. They apply DeepWalk [185] on the Chinese Wikipedia and evaluate their findings on a smaller Chinese variant of WS-353, a standard evaluation dataset for semantic relatedness of words. However, there are several key differences to the approach by Dallmann et al.: For example, Zhao, Liu, and Sun consider a network not only consisting of Wikipedia pages, but also of Wikipedia categories and even words from the articles [265]. In Section 7.2.4, we will present the experiments described in [55] and extend them by priming the random walker with hypotheses used in Chapter 5 to simulate pseudo-human navigation, in order to investigate the influence on the extractable semantics.

2.4 Outlook: Applications of Semantic Information Models

Semantic representations of words and sentences are widely needed and used in a variety of applications and studies, e.g., correcting word spelling errors [45], text segmentation using lexical cohesion [124, 146], image [211] or document [218] retrieval, cognitive science [226], fact checking [52], and many more. It is thus important to also emphasize potential *applications* for models of semantic information, since they actually are the

reason why we want to make computers understand language. Here, we largely focus on word sense disambiguation, since this is the main topic in Section 6.4, where we additionally measure the influence of tagging behavior on tag sense discovery, as mentioned above. Nonetheless, we also give pointers to other interesting fields of applications, such as query expansion, tag recommendation or sentiment analysis.

Word Sense Disambiguation and Discovery. One core problem of computational semantics is to resolve word ambiguity, that is, distinguishing different word senses. In statistical natural language processing, there exist supervised, dictionary-based and unsupervised word sense disambiguation approaches [147]. In the following, we only focus on dictionary-based and unsupervised algorithms.

Dictionary-based clustering algorithms rely on sense definitions defined in dictionaries or thesauri. One of the first dictionary-based algorithms for word sense disambiguation is the Lesk algorithm [131], which was adapted by Banerjee and Pedersen so it can be applied to WordNet [16]. It is based on word overlaps between the sense glosses³ of the disambiguated word and the sense glosses in the dictionary. Using tagging data, Angeletou, Sabou, and Motta [6] first identify a set of candidate senses for a given tag within WordNet, interpret co-occurring tags as context and then use a measure of semantic relatedness to choose the most appropriate sense. In a similar manner, Garcia-Silva et al. use cosine similarity between tag co-occurrence vectors and a bag-of-words representation of Wikipedia pages to identify the most suitable sense definition within DBPedia⁴ [83]. Lee et al. also computes a relevance score between tags and Wikipedia articles for the same purpose [130]. Liu, Yu, and Meng [140] present a new approach to evaluate senses of words in search queries by facilitating WordNet. Using a three step method the authors first use WordNet to gather information about the hyponyms, synonyms etc. of the terms found in the system. In case the correct senses of all terms cannot be determined a frequency based approach and a web search is conducted to try to find the remaining missing senses of a term.

Unsupervised approaches attempt to partition the context of a given term into clusters corresponding to its different senses. [11] analyzed several folksonomy-derived networks with regard their suitability to discover word senses by graph clustering algorithms. These networks are constructed by correspondingly omitting one of the three folksonomy building blocks of users, resources or tags. Zhang, Wu, and Yu [262] proposed an entropy-based metric to capture the level of ambiguity of a given tag. Si and Sun take into account a web-based measure of semantic relatedness as well as textual article content to disambiguate tags in weblogs by spectral clustering [206]. Au Yeung, Gibbins, and Shadbolt [10] use tags from Delicious to disambiguate search queries from a search engine. For a given tag, Au Yeung, Gibbins, and Shadbolt identify tags that are related in different contexts by applying a community detection algorithm. In [28], Benz et al. apply hierarchical agglomerative clustering of co-occurring tags to identify different senses and to ultimately construct a simple *is-a* taxonomy from tagging data. Finally,

³A *gloss* is a short description of a word sense.

⁴<http://www.dbpedia.org>

2 Related Work

Ratinov et al. [193] proposed an WSD algorithm which leverages the Wikipedia link network. Finally, Arora et al. used the linear algebraic structure of word embeddings to improve word sense disambiguation [8].

Tag Recommendation. Especially on tagging systems, recommending fitting tags for new posts is a very important task and crucial for offering users usage comfort, so they keep using the system. Also, by recommending fitting and already used tags, recommender systems help in stabilizing tagging semantics in social tagging systems. Here, Jäschke et al. described and compared several recommendation algorithms with regard to their applicability to tagging data [111]. Doerfel and Jäschke analyzed recommendation evaluation systems [62]. Additionally, Doerfel, Jäschke, and Stumme looked at the role of k -cores in recommender system benchmarks and extended the notion of cores to *set-cores* [63]. In a recent work, Zoller et al. exploited log data from the social tagging system BibSonomy to recommend recently used and semantically suitable tags [268].

Search and Query Expansion. Search plays an exceptional role in the web and in fact enabled lots of semantic research through fields like information retrieval [42] and query expansion [234]. One very important work on search in folksonomies is presented in [106], where Hotho et al. proposed the *FolkRank* algorithm, enabling efficient search and ranking in folksonomies. As long as search engines served as mere sophisticated information retrieval engines based on keyword matching, the enhancement of searching could be overcome by query expansion techniques. There are works based on mining semantics from search logs for query expansion [240], using semantically close terms to enrich queries [233], using knowledge bases to find suitable terms [90, 165], generating complete query substitutions [91, 113], or including social aspects into search [34, 232]. Naturally, there are also hybrid efforts to improve search by combining keyword based and semantic search [32, 35]

Multilinguality and Translation. Multilinguality and translation are still open problems, since there are no easy ways to project one language into another. Next to that, synonyms in one language are not necessarily synonyms in another. There has been a big amount of research addressing such issues. For a recent survey of multilingual word embeddings, see [200]. Koehn, Och, and Marcu first proposed a phrase-based translation model [121]. Hassan and Mihalcea [98] exploit Wikipedia's inter-language links to calculate semantic relatedness across languages. [269] introduce bilingual word embeddings, regardless of exact word co-occurrences in parallel training text. Faruqui and Dyer used correlations between different languages to improve performance of semantic relatedness measurements in the single languages and named the model BiCCA for *Bilingual Correlation Based Embeddings* [70]. In [231], Upadhyay et al. empirically compare several bilingual word embedding frameworks, namely BiSkip [144], BiCVM [101], BiCCA [70], and BiVCD [236]. On 4 tasks, the BiSkip model was largely able to outperform its competitors. In a very recent work, Zhang et al. managed to construct bilingual lexica in an unsupervised way by training adversarial neural networks. Lately,

DeepL managed to perform machine translation with extremely high quality⁵. However, the details of their translation are not published due to commercial reasons.

Sentiment Analysis. Sentiment is closely related to word semantics. As sentiment is a In [213], Socher et al. proposed a deep neural network model of semantic compositionality over a sentiment treebank. They use pretrained word embeddings to determine the sentiment of a sentence. This approach has been chosen to evaluate [69]. Kim [118] proposed a CNN architecture for text classification, which accepts pretrained word embeddings as input, thus providing a practical extension to any kind of word embeddings. In fact, it can then directly be used to classify the sentiment of texts, as e.g., done in [250].

2.5 Summary

In this chapter, we addressed many points related to computational semantics on social media data. We will now conclude this chapter by recapitulating the presented topics and highlight the most important works for this thesis.

In Section 2.1, we covered work on the *extraction* of semantic relatedness from WordNet, folksonomies and Wikipedia. Since we use the WordNet-based Jiang-Conrath similarity measure in Section 6.2 as a baseline for semantic relatedness, we provided an overview of works that compute semantic similarity on WordNet. Especially on Wikipedia, we presented many variants on how to extract semantics, e.g., by exploiting the link structure, the Wikipedia categories, the article texts or combinations of all sources. The results from these previous works provide a strong argument for the choice of Wikipedia as a research corpus. Additionally, they show that not only the article texts, but also the link structure connecting these articles shows great potential as a source for semantic information. We will use that result in Section 7.2, where we use the semantics contained in the Wikipedia link structure both as a baseline as well as a basis to perform parameterized random walks. Furthermore, because we discuss the evaluation and representation of tagging semantics in Chapter 6, we gave pointers to the most relevant work in this field.

Next, we focussed on ways to *learn semantic information* in Section 2.2. Concretely, we described algorithms to embed semantic information in low-dimensional vector spaces. In Section 6.5, we make use of some of those methods to learn tag embeddings. Additionally, we perform several experiments using pre-trained word embedding datasets in Chapter 8. We also presented several works on including external knowledge into such pre-trained embedding datasets, as well as in the embedding process itself. This is especially relevant for Chapter 8, where we propose an algorithm to incorporate the knowledge from semantic relatedness datasets (cf. Section 4.4) into pre-trained word embedding datasets.

Section 2.3 gave an overview of works regarding the *mutual influence* of semantics and user behavior in the web. We argued that although it is obvious that user behavior

⁵In fact, roughly 3 times better than Google's translation engine: <https://www.deepl.com/press.html>

2 *Related Work*

influences the semantics of web content, it is not as trivial to see if semantics influence user behavior. Again, we put an emphasis on folksonomies and Wikipedia. The mutual influence of semantics and behavior is a recurring theme in this work. First, we analyze and quantify a semantic influence on navigation on Wikipedia and folksonomies in Chapter 5. Then we exploit the existence of semantics in navigation to construct word representations both from BibSonomy and Wikipedia navigation in several settings in Chapter 7. Finally, in Section 6.4, we measure the influence of tagging pragmatics on tag sense discovery in social tagging systems.

In the last part of this section, we gave an outlook of different applications of semantic information models. Section 2.4 starts with a description of several word sense disambiguation algorithms, before we presented works on tag recommendation, multilinguality and translation, query expansion and sentiment analysis. In this thesis, we make use of a tag sense discovery algorithm to show the influence of tagging pragmatics on different facets of tagging semantics. Additionally, we use a question answering framework that is based on word embeddings to evaluate the quality of our Relative Relatedness Learning algorithm to adapt word embeddings to external knowledge in a realistic setting.

Chapter 3

Foundations

In this thesis, we focus on extracting semantic information from social media data. Social media data are available in various forms, mainly semi-structured or completely lacking any explicit structure. For example, tagging data from folksonomies can be seen as semi-structured data, since the folksonomy determines the assignment of tags to resources, but does not impose any other restrictions on the user. Methods from computational semantics aim to model the semantic information contained in such data. In contrast to lexical or formal semantics, computational semantics aim to represent semantic meaning and structure as mathematical structures, e.g., vectors or matrices. We can then compute semantics of more complex structures, such as navigational paths or even whole link graphs. This gives us the advantage to automatically analyze user-contributed textual content to better understand and support users in their goals in the social web.

This chapter explains and defines many basic concepts of computational semantics which we apply in the remainder of this thesis. First, we introduce several types of data structures from which we will extract semantic knowledge, such as folksonomies or the Wikipedia information network. After this, we go into detail about how we can model that semantic information and how we can compute semantic relatedness, i.e., the intensity of a latent semantic relation between words, given the introduced models of semantic knowledge. In the last part of this chapter, we will introduce different models of user behavior in tagging data. Also, we describe the method that we used to analyze human navigation in social media. At the end of this chapter, we summarize the presented content.

3.1 Implicit and Explicit Forms of Semantic Knowledge in the Web

Ever since the advent of the World Wide Web in 1990, millions of users added content for public consumption. With the evolution of interactive technologies, user engagement skyrocketed and let the Web 2.0 or Social Web emerge, which is characterized by very easy content contribution and social collaboration possibilities, even for layman users. Representative systems and websites for the Social Web already mentioned in the

introduction are Twitter¹, Facebook², or YouTube³ to just mention a few. These systems can be described by the abstract concept of *social media*. By actively participating in social media systems, users leave imprints of their subconscious, semantic knowledge in the contributed content. Thus, the WWW can be seen as a huge repository of human knowledge.

However, dependent on the nature of the used social media, this knowledge can be semi-structured, like in Wikipedia or social tagging systems, or unstructured, i.e., data with no explicit formal structure and which is available mostly in the form of free text, for example as posts on Twitter or Facebook. In the course of this thesis, we will be working on *making this implicit semantic knowledge explicit*. While there also exist highly-structured knowledge repositories of explicitly characterized human knowledge, such as DBPedia⁴, WordNet⁵ and ConceptNet⁶, these are hardly social, as they are either carefully hand-crafted by experts or created semi-automatically. Still, we can use these hand-crafted knowledge bases as a valid gold standard of semantic knowledge.

In this section, we will present different types of data structures which contain human knowledge in descending order of structuredness. First, we introduce two semi-structured types of datasets, on which we will perform the majority of our experiments in this thesis, namely Folksonomies and Wikipedia. After this we introduce WordNet, a large lexical database of English terms. At the end of this section, we introduce ontologies and knowledge graphs, which are, although hardly constructed by social means, the most structured knowledge representation entities currently in existence. Although we do not directly use them, they motivate the analyses and methodological contributions of this thesis.

3.1.1 Social Tagging Systems and Folksonomies

Social Tagging Systems have played an important role in the development of both the Semantic Web and the Web 2.0. In these systems, users collect resources and annotate them with freely chosen keywords, called tags. Examples are BibSonomy⁷, for collecting web links and scholarly publications, Delicious⁸ for bookmarks, Flickr⁹ for images, and last.fm¹⁰ for music. Figure 3.1 shows a screenshot of BibSonomy to illustrate an exemplary user interface of a social tagging system. Although Twitter, Instagram¹¹ and meanwhile even Facebook also support adding so-called “hashtags” to their posts, it is not compulsory to do so, as opposed to the systems mentioned above. However, Wagner

¹<https://www.twitter.com>

²<https://www.facebook.com>

³<https://www.youtube.com>

⁴<https://wiki.dbpedia.org/>

⁵<https://wordnet.princeton.edu/>

⁶<http://conceptnet.io/>

⁷<http://www.bibsonomy.org>

⁸<https://www.delicious.com>, ceased service in 2017

⁹<http://www.flickr.com>

¹⁰<http://last.fm>

¹¹<https://www.instagram.com>

3.1 Implicit and Explicit Forms of Semantic Knowledge in the Web

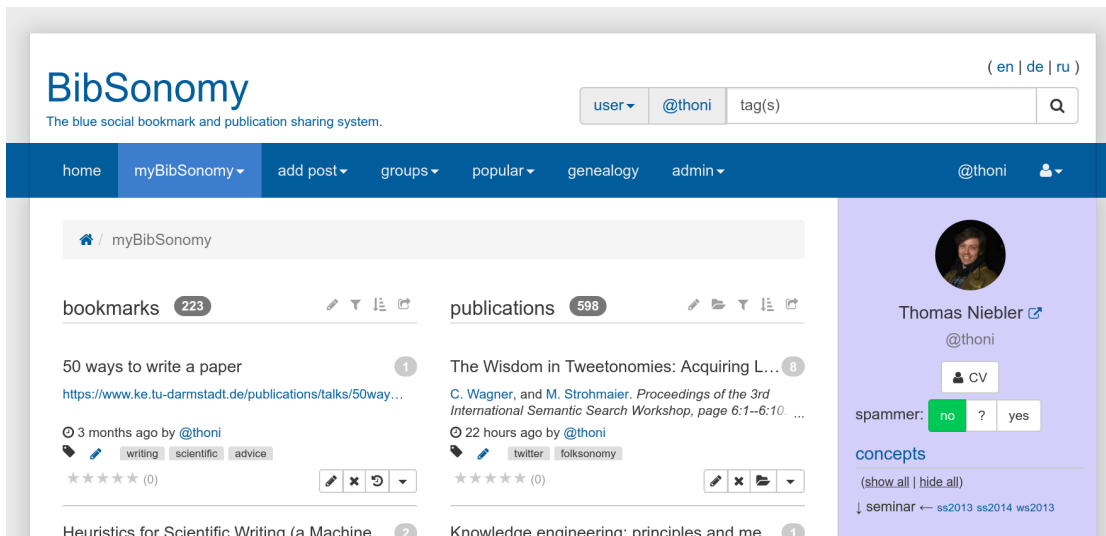


Figure 3.1: Illustrative screenshot of the social tagging system BibSonomy. The image shows the *user page* of the user *thoni* that contains all posts in the personomy of that user.

and Strohmaier argue that even those systems could be considered as social tagging systems [239].

In this thesis, we analyse two aspects of social tagging systems: First, we work on methods to model and evaluate the semantic information emerging from actively contributed content. For this, we leverage the structure emerging from social tagging systems, the so-called *Folksonomy*, in Chapter 6. Second, we investigate in how far passively contributed information, concretely human navigation data in social tagging systems, can be used to build models of semantic knowledge. We do so by analyzing log data about the browsing behavior of users on the *user interface* of social tagging systems in Section 5.3.

3.1.1.1 Folksonomies

The data structures emerging from social tagging systems are called *Folksonomies*. The term “folksonomy” was first coined in 2004 by Thomas van der Wal in a mailing list post¹². There he introduced the term as a portmanteau of “folk” and “taxonomies”, analogous to the description of folksonomies as a “*user-created bottom-up categorical structure with an emergent thesaurus*”. A thesaurus is a form of a controlled vocabulary that organizes words using semantic relations of synonymy. Together with the emerging hierarchical category structure, folksonomies can thus be interpreted as *taxonomies* generated by *folks*.

A first systematic analysis of these emergent semantic structures has been performed by Golder and Huberman [86]. One core finding was that the openness of these systems

¹²<http://vanderwal.net/folksonomy.html>

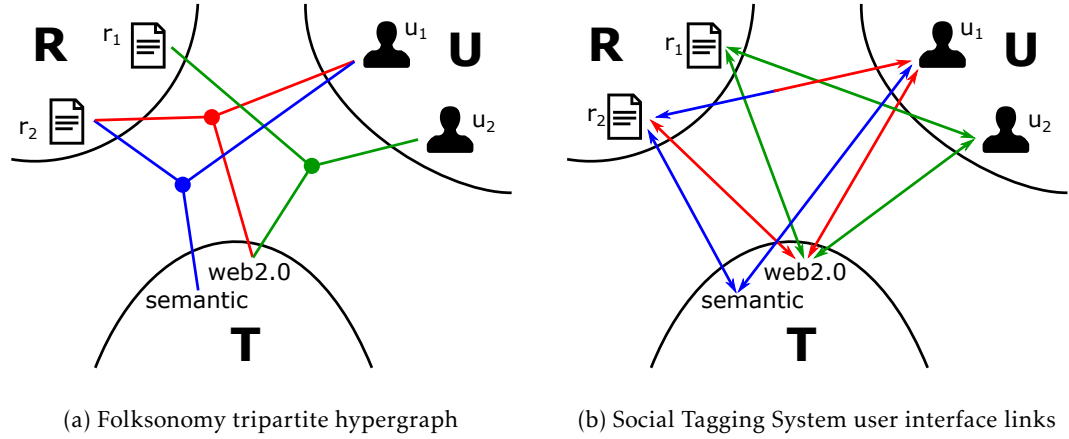


Figure 3.2: Representation of a folksonomy as a tri-partite hypergraph and its corresponding links in the user interface of a social tagging system, e.g., BibSonomy. The hyperedges are defined by the tag assignment set Y , which is a ternary relation between the entity sets tags (T), users (U) and resources (R).

did not give rise to a “tag chaos”, but led to the emergence of stable semantic patterns. Cattuto reported similar results and denoted the emerging patterns as “*semantic fingerprints*” of resources [47]. Finally, Wagner et al. presented several measures for semantic stability in social tagging streams, e.g., from Twitter, but also from Delicious [238]. Using these measures, they showed that tagging activity quickly led to the stabilization of the respective semantics, thus supporting the choice of tagging data as a stable source for semantic relatedness extraction. Following up, Mika even described folksonomies as light-weight ontologies [155], thus underlining the quality of the emergent semantic structures. According to Hotho et al., social tagging systems thus are able to overcome the knowledge acquisition bottleneck [106]. Consequently, there exist several works that propose methods to extract that emergent semantic information from folksonomies (cf. Section 2.1.2).

In this thesis, we follow the folksonomy definition by Hotho et al. given in [106]:

Definition 1 (Folksonomy). *A folksonomy is a tuple $\mathbb{F} := (U, T, R, Y)$, of finite sets U , T , R , and Y . The elements of U , T , and R are called users, tags, and resources, respectively. Y is a ternary relation between them, i.e., $Y \subseteq U \times T \times R$, and denotes the set of tag assignments.*

One corollary result from this definition is that folksonomies possess a *tri-partite structure*, i.e., they can be modeled using a hypergraph with three partitions. The nodes of this graph are represented by the tags T , users U and resources R , while the edges are defined by Y . Figure 3.2 shows an illustration of the folksonomy hypergraph.

We can now also define *Posts* and a *Personomy*, which we will especially need later for the definition of BibSonomy web page representations:

3.1 Implicit and Explicit Forms of Semantic Knowledge in the Web

Definition 2 (Post). A post $P_{ur} \subseteq Y$ in a folksonomy \mathbb{F} is a collection of tag assignments (u, t, r) for a given user u and a given resource r , i.e., a non-empty subset

$$P_{ur} := \{(u', t, r') \in Y \mid u' = u, r' = r\}. \quad (3.1)$$

The tags of a post are described by $T_{ur} = \{t \in T \mid (u, t, r) \in P_{ur}\}$.

Definition 3 (Personomy). A personomy \mathbb{P}_u is the set of all tag assignments by a certain user u , defined as

$$\mathbb{P}_u := \{(u', t, r) \in Y \mid u' = u\}. \quad (3.2)$$

The tags and resources used by user u are given by $T_u := \{t \in T \mid (u, t, r) \in \mathbb{P}_u\}$ and $R_u := \{r \in R \mid (u, t, r) \in \mathbb{P}_u\}$, respectively.

A trivial relation between the personomy \mathbb{P}_u of a user and her posts is given by

$$\mathbb{P}_u = \bigcup_{u'=u} P_{u'r} \quad (3.3)$$

Figure 3.1 shows the personomy of user thoni. As can be seen, the personomy consists of many singular posts. For example, the first publication post of user thoni “The Wisdom in Tweetonomies” has been annotated with the tags twitter and folksonomy. This could be formally expressed as

$$P := \{(\text{“thoni”}, \text{“twitter”}, \text{“The Wisdom in Tweetonomies”}), \\ (\text{“thoni”}, \text{“twitter”}, \text{“The Wisdom in Tweetonomies”})\}$$

These definitions will play a role in defining our models of *context* for tags in order to determine the semantic relatedness between tags in Section 3.2.2.1. We will further use them in Section 6.2, Section 6.5, and Section 6.4.

3.1.1.2 Social Tagging System User Interfaces

As social tagging systems and folksonomies share many similar traits, it is obvious that also the *user interface* of social tagging systems bears semantic features. Especially for the context of navigation, we will now define what we understand as a (content) page in a social tagging system and how we can characterize it. The following definitions are constructed on the example of the user interface of BibSonomy. However, with minor modifications, the definitions still hold for any other social tagging system user interface.

Definition 4 (Page in a Social Tagging System). A content page or simply page in the user interface of a social tagging system shows information about sets of posts. These sets of posts are determined by different filters, such as a user account, a set of tags, a resource key or a combination of those. Formally, a page can be expressed as a set of posts that share at least one coordinate:

$$p^{(u', t', r')} := \{(u, t, r) \in Y \mid u = u' \vee t = t' \vee r = r'\} \quad (3.4)$$

Table 3.1: The considered page types in the BibSonomy system. All page types are retrieval types, that is, they return entries in the folksonomy tag assignments. The formal description of each page type denotes the corresponding folksonomy subsets.

page type	category	description	folksonomy subset
/user/USER	<i>user</i>	contains all posts from USER	\mathbb{P}_{USER}
/user/USER/TAG	<i>tag</i>	contains posts from USER which have been tagged with TAG	$\{P \in \mathbb{P}_{\text{USER}} : (\text{USER}, \text{TAG}, r) \in P\}$
/tag/TAG	<i>tag</i>	contains posts which have been tagged with TAG	$\{P \in \mathbb{F} : (u, \text{TAG}, r) \in P\}$
/url/RES	<i>resource</i>	description page of a bookmark RES not specific to any user	$\bigcup_{u \in U} P_{u, \text{RES}}$
/bibtex/RES	<i>resource</i>	publication page of RES not specific to any user	$\bigcup_{u \in U} P_{u, \text{RES}}$
/bibtex/RES/USER	<i>resource</i>	describes a page of a publication RES specific to USER	$P_{\text{USER}, \text{RES}}$

Although most social tagging systems contain pages that do not carry any content, we focus only on pages that carry explicitly named folksonomic content in this thesis. Additionally, the landing page of BibSonomy or full-text search results are not considered as content pages, although they show folksonomic content, because the content shown is not a folksonomy subset determined by a folksonomy entity. For BibSonomy, Table 3.1 lists the considered page types as well as their corresponding folksonomy subsets.

In general, each page in a social tagging system can be semantically represented by its *tag cloud*:

Definition 5 (Tag Cloud of a Page in a Social Tagging System). *A tag cloud of a page in a social tagging system is the multiset¹³ of all tags viewable on that page, resulting from a multiset union of the tags of the posts in the page's posts set:*

$$\text{TagCloud}(p^{(u', t', r')}) := \text{multiset} \left\{ t \mid (u, t, r) \in p^{(u', t', r')} \right\} \quad (3.5)$$

Furthermore, a page in a social tagging system can have an *owner*, that is, the user who posted the contents on that page. For example, the page /user/thoni from Figure 3.1 shows all posts of the page's owner, the user thoni. We denote the set of pages that belong to a user u as V_u . However, there also exist pages with posts from different users which consequentially do not have an owner. Such a page is /tag/web: It shows all posts of *all* users that have been annotated with the tag *web*. This distinction as well as the

¹³A multiset is a set which can contain many instances of the same item, e.g., $\{\text{web}, \text{web}, \text{semantic}\}$.

3.1 Implicit and Explicit Forms of Semantic Knowledge in the Web

definition of pages in social tagging systems are important in Section 5.3.1, where we analyze and discuss user navigation in BibSonomy. Additionally, we make use of the tag cloud in Section 7.3 to extract semantic information from user navigation in BibSonomy.

We also can now define a *Social Tagging System Graph* as follows:

Definition 6 (Social Tagging System Graph). A social tagging system graph is a finite graph $G_{\mathbb{F}} = (P(u, t, r), E)$ where \mathbb{F} is the underlying folksonomy as given in Definition 1, $P(u, t, r)$ denote the folksonomy pages from Definition 4 and E are the links between those pages. The existence of links between two pages strongly depends on the folksonomy structure of the social tagging system.

Usually, the page of a resource r will contain actual hyperlinks leading to the page of the resource's owner u as well as to the pages of the tags t that are assigned to the resource. However, although we state in Definition 6 that links in the user interface originate on the folksonomy data of the social tagging system, the actual form of these links depends on the implementation of the social tagging system. For example, BibSonomy tag pages also link to related tags, although the folksonomy structure does not directly connect tags with each other (see Figure 3.2).

Analogous to pages that are owned by a user, there also exists a subgraph of a user's own pages:

$$G_{\mathbb{P}_u} = (V_u, E_u), \quad (3.6)$$

where V_u defines all pages which have the same owner u and E_u is the subset of E that connects these pages. \mathbb{P}_u denotes the personomy as described in Definition 3. We will use that subgraph as well as its complement to characterize human navigation behavior in social tagging systems in Section 5.3.

3.1.2 Wikipedia

The Wikipedia encyclopedia is one of the figurehead projects of the Social Web. By encouraging all its users to contribute and edit its content, it quickly grew to one of the most frequently visited webpages in the world, while maintaining an encyclopedic precision that rivals that of Encyclopaedia Britannica [84], as already mentioned in the introduction of this thesis. At the same time, it also is the biggest encyclopedia in the world with up-to-date content in about 5 million articles in the English Wikipedia alone. Wikipedia is also the most trusted internet resource in Germany, according to a survey by the GPRA in October 2017 with roughly 1000 German citizens¹⁴. Not only this, it is also one of the most visited resources for scientific researchers¹⁵, as 84,7% of a group of 1 354 consulted researchers used Wikipedia as a knowledge source in their daily work.

What makes Wikipedia such an interesting resource for researchers is the fact that anyone can contribute to its content and link that content to other pages with related

¹⁴<http://www.horizont.net/agenturen/nachrichten/GPRA-Vertrauensindex-Facebook-und-Yahoo-fallen-durch-Wikipedia-raeumt-ab-163201>

¹⁵<http://www.zbw.eu/de/ueber-uns/aktuelles/meldung/news/social-media-forschende-nutzen-am-haeufigsten-wikipedia/>

3 Foundations

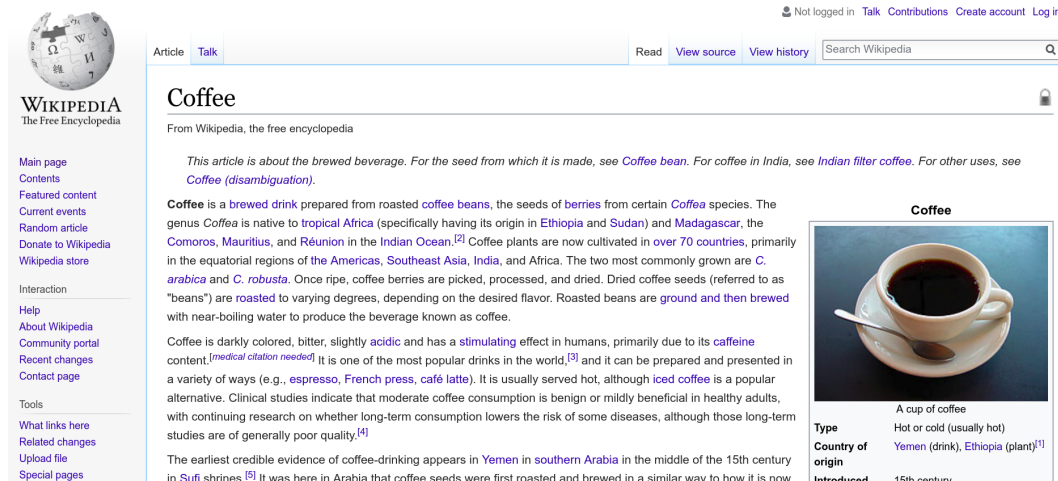


Figure 3.3: Screenshot of the Wikipedia page about “Coffee”. We can see a textual description of the general concept, links to the sense disambiguation page, and links to other related pages.

content. Figure 3.3 shows a screenshot of the Wikipedia page about “coffee”. At the same time, the Wikimedia foundation regularly provides dumps of the whole Wikipedia universe and a large amount of usage information¹⁶. This way, a huge and freely accessible knowledge pool is created, containing not only text data about a multitude of topics, but also explicit linkage of somehow semantically related concepts, as well as its creation and usage history, such as page views and navigation information.

Contrary to taxonomies or ontologies, Wikipedia is rather an information network, since its main goal is to enable users to discover, contribute and disseminate information instead of providing an exact semantic representation of concepts and their explicit relations. We first define a Wikipedia page as follows:

Definition 7 (Wikipedia page). *A Wikipedia page or Wikipedia article*

$$p_{title} = (title, content)$$

is a tuple of a string title, denoting the title of the page (name) as well as another string content, which contains a definition as well as a description of the concept given by the title. Each article in Wikipedia represents a single semantic concept.

This concept is then connected to other concepts via hyperlinks, thus making the information in Wikipedia better accessible for humans. Consequently, we define the Wikipedia information network as a graph in a similar manner as for folksonomies (Definition 6):

¹⁶<https://dumps.wikimedia.org/>

Definition 8 (Wikipedia graph). We define a Wikipedia \mathbb{W} graph G as a graph $G_{\mathbb{W}} = (V_{\mathbb{W}}, E_{\mathbb{W}})$ with vertices – i.e., pages or concepts – $V_{\mathbb{W}} = \{p_1, \dots, p_n\}$ and directed edges – i.e., links between those pages – $E_{\mathbb{W}} = \{(p_i, p_j) | p_i, p_j \in V_{\mathbb{W}}\}$.

The content of page also contains all the links which define the edges originating from this page. In fact, an edge (p_i, p_j) can only be contained in $E_{\mathbb{W}}$, if and only if the *content* of page p_i contains a hyperlink to page p_j . We can now define $inlinks(p_i)$ and $outlinks(p_i)$ for a given page p_i . The set of outlinks contains all links originating from p_i and is easily deduced as $outlinks(p_i) = \{(p_i, p_j) \in E_{\mathbb{W}} | p_j \in V_{\mathbb{W}}\}$. The set of inlinks contains all links pointing from different pages to page p_j and is defined analogously as $inlinks(p_j) = \{(p_i, p_j) \in E_{\mathbb{W}} | p_i \in V_{\mathbb{W}}\}$. The cardinality of inlinks and outlinks of a page p_i are called the *in-* and *outdegree* of p_i , respectively. The sum of in- and outdegree gives the *degree* of p_i .

3.1.3 The WordNet Taxonomy

WordNet is a well-known lexical database of English words, verbs, adverbs and adjectives [72]. It organizes words with a shared meaning in so-called “synsets”. These synsets thus represent *semantic concepts*. On the one hand, they can be used to resolve *synonymy*, i.e., that different words describe the same concept, but also *polysemy*, i.e., that a word can describe different concepts, depending on its context. Additionally, WordNet contains a brief description of each meaning of a synonym, the so-called *gloss*.

WordNet is not only a thesaurus, i.e., a lexicon of synonyms, but also a *semantic network*. This means that in WordNet, *semantic concepts* are linked if they are semantically related. WordNet covers *meronymy* (part-of relations) and *antonymy* (opposite meanings) relations. Additionally, all synsets are hierarchically organized in a directed, acyclic graph structure, where most of the edges represent semantic generality relations between synsets. This hierarchical *is-a* relation is called *hyperonymy* or *hyponymy*. For example, the synset dog has a hypernym canine and a hyponym pug. There have also been attempts to include relations about the semantic relatedness of concepts that are weighted according to the intensity of the relatedness [40]. Unfortunately, they finally proved to be unsuccessful due to the high amount of possible relations and the great uncertainty of human perception.

The greatest strength of WordNet is that it contains highly accurate linguistic knowledge about the English language, because it was manually created and curated by linguists. Additionally, WordNet provides easy-to-use programmatic access, for example on its website¹⁷. For these reasons, it has been widely used in research for many purposes, such as gold standard for the evaluation of knowledge base construction algorithms [28, 168], word sense disambiguation approaches [], and even as source of semantic similarity measures, as we will see in Section 3.2.

The fact that WordNet is curated manually is also one reason for its potentially greatest shortcoming, the slow ingestion of new knowledge into WordNet. For example, the

¹⁷<http://wordnetweb.princeton.edu/perl/webwn>

synset for “Python” does not cover the well-established meaning of the Python programming language. It can thus only serve as a source of general linguistic knowledge. However, this knowledge is still present in a very precise form, which leaves WordNet as one of the best sources of semantic information in existence. We will use WordNet as the basis for measures of semantic similarity in Section 7.3.

3.1.4 Ontologies and Knowledge Graphs

For sake of completeness, we will lastly introduce ontologies and knowledge graphs, although we do not directly use them in this thesis. As these structures provide the central motivation for the methods and analyses presented in this thesis, it is important to provide a description as well as a formal definition of each.

The most structured form to encode semantic knowledge today is given by *Ontologies*. Studer, Benjamins, Fensel, et al. define an ontology as follows [224]:

Definition 9 (Ontology). *An ontology is a formal, explicit specification of a shared conceptualisation.*

We will now explain the atomic parts of that definition briefly. A *conceptualisation* is an abstract description of an observed, real-world entity. By postulating that it is *shared*, we expect that conceptualisation to not be subjective, but objective, or at least in some way more broadly accepted. By describing that shared conceptualization, we obtain an *explicit specification* of it. Finally, we want that specification to be as strict and exact as possible, so we would formulate it in a *formal* way, i.e., mathematically. In summary, ontologies allow us to store and access human common-sense knowledge in machine-readable form. For a longer, in-depth discussion of the parts that constitute Definition 9, we refer the reader to [88].

While ontologies define the abstract structure of knowledge, *knowledge graphs* collect and organize actual *instances* of that knowledge. In [68], Ehrlinger and Wöß identify and compare several definitions and assumptions of what a knowledge graph is seen as in literature. They conclude that a commonly accepted definition could be as follows:

Definition 10 (Knowledge Graph (Ehrlinger and Wöß)). *A knowledge graph acquires and integrates information into an ontology and applies a reasoner to derive new knowledge.*

This means that a knowledge graph makes use of an ontology as an abstract *schema* of information, while it contains that explicit information itself.

Paulheim characterizes a knowledge graph in a slightly different way, [180]:

Definition 11 (Knowledge Graph (Paulheim)). *A knowledge graph*

1. *mainly describes real world entities and their interrelations, organized in a graph.*
2. *defines possible classes and relations of entities in a schema.*
3. *allows for potentially interrelating arbitrary entities with each other.*
4. *covers various topical domains.*

3.2 Computational Methods to Model Distributional Semantic Knowledge

In contrast to the definition by Ehrlinger and Wöß, Paulheim phrases his definition of an knowledge graph in such a way that an ontology is only a part of a knowledge *graph* instead of a separate entity. It has also to be noted that Paulheim does not consider WordNet as a knowledge graph, since WordNet only describes semantic relations between words, not things. This is largely based on his decision that words are not real-world entities, although he admits that this is a rather “philosophical discussion” [180]. However, besides this point, WordNet largely fulfils all criteria of a knowledge graph as postulated in Definition 11. For this reason, we will still consider WordNet as a knowledge graph, where needed.

Besides the well-known knowledge graphs Wikidata, DBPedia, YAGO, and WordNet, there also exists ConceptNet, which originated from the crowdsourcing project *Open Mind Common Sense* [217]. It collects knowledge from various sources, such as Wiktionary, WordNet, and a subset of DBPedia. In this thesis, we will use the ConceptNet Numberbatch word embeddings, which are constructed from the ConceptNet knowledge base [216]. We will introduce them in Section 4.3.

3.2 Computational Methods to Model Distributional Semantic Knowledge

Computational Semantics is a subfield of Computational Linguistics and focuses on the study of automatically computing the *meaning* of words or sentences and the subsequent processing of that information in several applications in Natural Language Processing and Understanding. Such applications are for example determining the semantic relatedness of words and texts [129, 156, 230, 266] Text Classification and Clustering [118, 264], Sentiment Analysis [138, 178], Translation [15, 253], and many others. The term *Computational Semantics* also encompasses many other research fields, such as *Lexical Semantics* and *Distributional Semantics*. *Lexical semantics* techniques are highly related to *Semantic Web techniques*, where entities are represented as nodes in ontologies and knowledge graphs (as presented in Section 3.1) and are connected by explicit semantic relations. For example, in the statement

```
dbp:apple rdf:type yago:fruit113134947,
```

dbp:apple denotes the DBPedia entity *apple*, rdf:type makes clear that the *type* of the apple is a *fruit*. Fruits are denoted by the Yago entity yago:fruit113134947. So, in human language, the statement above is the same as

An apple is a kind of fruit.

Such statements in ontologies and knowledge graphs are also called *facts*. Inserting new entities and facts is often done manually. As mentioned before, WordNet is completely hand-crafted by linguistic experts. As a consequence, constructing and updating WordNet (and other ontologies) requires considerable effort and expert knowledge and is thus very tedious and time-consuming. This especially becomes a problem if we want to create ontologies that cover specialized domain knowledge. The difficulties in acquiring

valid, timely, precise, and structured knowledge, for example in such expert domains, is generally known as the *knowledge acquisition bottleneck*.

In order to overcome the knowledge acquisition bottleneck for domain knowledge, an option is to *automatically* gather information about unknown entities and to deduce new facts from known ones. For this, we explore *distributional semantics* approaches to extract such information from natural language. Research about *Distributional Semantics* concerns itself with determining words with similar or related *meaning* by studying their *contexts* and their distribution within those contexts. The following two quotes offer a very intuitive description of the central goals that we want to achieve in distributional semantics. In 1954, Harris stated this in his “Distributional Hypothesis” [96]:

Words that occur in the same contexts tend to have similar meanings.

In [76], Firth coined one of the most well-known quotes in this context:

“You shall know a word by the company it keeps”.

In this thesis, we focus on computing distributional semantics from social media data. More specifically, we develop distributional word representations from tagging data, navigation data on Wikipedia and navigation data on BibSonomy. Our **goal** is to obtain *highly precise semantic relatedness measures* on top of these representations, which *reflect human intuition as closely as possible*.

In this section, we will first introduce the vector space model (VSM) to model distributional semantic information in linear vector spaces. This model provides the groundwork for all models of semantic information that we present in this thesis. As the quality of the vector space model depends on a sensible choice of *context*, we will present several options to design that context in folksonomies and in navigation. After this, we explain how we actually *measure semantic relatedness* using the just defined word vector representations. For this, we also discuss the notions of *semantic relatedness* and *semantic similarity*. To assess the *quality of the measured semantic relatedness scores*, we describe two variants of how to compare these scores with human intuition and calculate a corresponding *evaluation score*, on which we will base our results. At the end of this section, we will introduce two *applications of the vector space model*. While the first method aims at enriching and refining the vector representations constructed by a VSM model, the other method exploits the semantic knowledge in the vectors to discover and disambiguate multiple meanings of words.

3.2.1 The Vector Space Model: Expressing Words as Vectors

In order to “compute” with words, they need to be somehow modeled in a computable form. In Distributional Semantics, the meaning of a word is mostly computed from the distribution of its semantic or syntactic *context* in large collections of texts. One way to model words by their context in a computable form is to construct *vector representations*. This approach is called the *Vector Space Model* (VSM) and was first introduced by Salton, Wong, and Yang [201]. It allows us to apply linear algebra methods in order to perform

3.2 Computational Methods to Model Distributional Semantic Knowledge

mathematical operations with the semantic information of the entities, or documents in our case, which are represented by those vectors.

The first applications of the vector space model lie in the *SMART information retrieval system* to retrieve suitable candidate documents from a large document corpus [202]. Later adaptations of the VSM then allowed to represent words by the documents they have been used in, or by their textual context. Turney and Pantel provided an extensive survey of VSM applications in literature and identified three classes of VSMs: *term-document*, *word-context* and *pair-pattern* matrices, resulting in three classes of applications [230]. In the course of this thesis, we will however only focus on the first two classes.

In the following, we will first formally introduce and explain the vector space model for documents. After this, we describe variants of the VSM to construct vector representations of words.

3.2.1.1 Formal Definition of the Vector Space Model: Document-Level Semantics

Salton, Wong, and Yang used vectors to represent *text documents* d_i by the terms t_j that the document d_i contained. If a term t_j occurs k times in a document d_i , the corresponding cell d_{ij} in the document vector equals k :

$$d_i := (\dots, d_{ij}, \dots) = (\dots, k, \dots) \quad (3.7)$$

The length of that vector depends on the size of the *vocabulary* $\mathcal{V}_D := \{t_1, \dots, t_n\}$ of a given *corpus* $\mathcal{D} := \{d_1, \dots, d_m\}$ of documents. The vocabulary \mathcal{V}_D contains all terms that occur at least once in a document from \mathcal{D} . The idea behind this document representation is that the words contained in a document determine its topic. This can also be seen as some kind of document-level semantics: Two documents are semantically similar, if they have similar topics, i.e., their content overlaps. It has to be noted that these vectors pay no regard to the actual sequential structure of words inside the documents. This way of expressing documents by words is also called a *bag-of-words* approach.

With these document vectors, we can construct a matrix $TD \in \mathbb{N}_0^{m \times n}$ from a document corpus \mathcal{D} , where m is the size of \mathcal{D} , and n is the size of the set of all occurring terms, i.e., the *vocabulary* of \mathcal{D} . Thus, a whole document corpus $\mathcal{D} := \{d_i | 1 \leq i \leq m\}$ can be expressed as the *term-document matrix* $TD_{\mathcal{D}}$:

$$TD_{\mathcal{D}} := \begin{matrix} & t_1 & \cdots & t_j & \cdots & t_n \\ \begin{matrix} d_1 \\ \vdots \\ d_i \\ \vdots \\ d_m \end{matrix} & \left(\begin{array}{ccccc} & & & & \\ & & & & \\ & & & & \\ \cdots & \cdots & d_{ij} & \cdots & \cdots \\ & & & & \\ & & & & \end{array} \right) \end{matrix} \quad (3.8)$$

The rows of this matrix are the *document vectors* for each document d_i , as given in Equation (3.7). This matrix is usually very *sparse*, as documents rarely share a high

d_1 : “The **quick brown fox jumps**.”

d_2 : “A **brown fox is quick**.”

(a) Example corpus \mathcal{D}

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

(b) Term-document matrix $W_{\mathcal{D}}$

$$\begin{matrix} \text{quick} \\ \text{brown} \\ \text{fox} \\ \text{jumps} \end{matrix} \begin{pmatrix} 0 & 2 & 2 & 1 \\ 2 & 0 & 2 & 1 \\ 2 & 2 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

(c) Word-Word Co-occurrence matrix $X_{\mathcal{D}}$

Figure 3.4: An example for word-word co-occurrence and term-document matrices, constructed from a document corpus $\mathcal{D} = \{d_1, d_2\}$.

amount of vocabulary with each other. Figure 3.4b shows the resulting document-term matrix for the document collection \mathcal{D} from Figure 3.4a.

A severe issue of the standard term-document vector space model is that common words that occur in many documents show a high impact on document retrieval. This however contradicts their overall importance to characterize a fraction of documents. To counter this effect, the TF-IDF model adds a weighting factor to each term-document entry that also takes the impact of a word on a limited number of documents into account [13]. TF-IDF stands for *Term-Frequency with Inverted Document Frequency* and is formally defined in Equation (3.9)

$$d_{ij}^{tfidf} := d_{ij} \cdot \log \left(\frac{|\mathcal{D}|}{|\{d_k \in \mathcal{D} | d_{ik} > 0\}|} \right) \quad (3.9)$$

This rules out “common” words, which do not contribute much to the characterization of a fraction of documents, since they occur so often, but also emphasizes the *important* terms, which are most descriptive for the document at hand. TF-IDF is by far the most used weighting scheme in information retrieval. We apply TF-IDF in Chapter 5 to semantically characterize web pages, as well as in Section 7.3 to extract relevant terms of BibSonomy pages.

3.2.1.2 Variants of the Vector Space Model for Word-Level Semantics

The term-document matrix from the original vector space model was mostly intended for document retrieval based on a coarse “semantic” document representation. We are however interested in determining the semantics of the words inside such documents. In the following, we will present several variants of the vector space model that enable us to determine such word-level semantics.

Word-Word Co-Occurrence Counting In this thesis, we mainly construct *word-context* or *word-word co-occurrence count* matrices. In general, these matrices are constructed in the same way as term-document matrices, only that we now count the *co-occurrences*

3.2 Computational Methods to Model Distributional Semantic Knowledge

of words in a pre-defined *context*. This allows us to represent a word w_i as a vector v_i , where

$$v_i := (v_{i1}, \dots, v_{ij}, \dots, v_{in}). \quad (3.10)$$

Here, v_{ij} denotes the co-occurrence count of w_i with w_j in a predefined context. For a discussion of potential context choices we refer to Section 3.2.2. The resulting matrix W is trivially symmetric, as can be gathered from Equation (3.11):

$$W := \begin{matrix} & w_1 & \cdots & w_j & \cdots & w_n \\ \begin{matrix} w_1 \\ \vdots \\ w_i \\ \vdots \\ w_n \end{matrix} & \begin{pmatrix} \ddots & & & & \\ & \ddots & & & \\ & & v_{ij} & & \\ & & & \ddots & \end{pmatrix} \end{matrix} \quad (3.11)$$

Figure 3.4c shows how a constructed word-word co-occurrence matrix W of a document corpus \mathcal{D} would look like after stopwords removal. Because *jumps* only occurs in document d_1 , it also co-occurs only once with all other words.

Additionally, we use *Word-Word Co-Occurrence Binarization*. This is the most simple variant of word-word co-occurrence, as it only denotes *if* two words co-occurred instead of *counting* these co-occurrences. This straightens out the effect of words that occur very often and thus extremely dominate the co-occurrence distribution. At the same time, rarely occurring words show a higher impact on the vector representation. The formal definition is the same as of word-word co-occurrence, only that w_{ij} can exclusively assume the values 0 or 1. We will apply binarization in Chapter 7 to extract improved semantics from unconstrained navigation both on Wikipedia and BibSonomy.

Word Embeddings While the co-occurrence counting approaches presented in the preceding section are a very intuitive way to represent words or documents as vectors, they suffer from sparsity and very high dimensionality. This impacts machine learning algorithms, since a large number of weights has to be learned from only very few examples, since the vectors are extremely sparse. Also, sparsity itself poses a problem, since it originates from insufficient or missing information and could lead to inaccurate or even wrong information. Concretely, the *curse of dimensionality* describes the problem that with increasing number of dimensions, the distances between any pair of vectors tend to approximate the same value, i.e., in very high dimensions, vectors are distributed equidistantly. In NLP, this can for example occur if a text contains many synonymous words that have each been used only rarely. Each of these words however defines a new dimension, although they would formally describe the same feature.

A possible way to alleviate the problems of high-dimensional, sparse vectors as described above is to *embed* vectors in a *low-dimensional space*. While the idea of dimension reduction for co-occurrence matrices by factorizing them has been around for some time (see for example Latent Semantic Analysis [58] or Principal Component Analysis [182]),

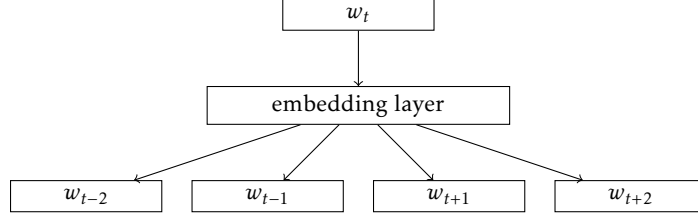


Figure 3.5: Illustration of Word2Vec’s skipgram neural network architecture. The words w_{t-2} , w_{t-1} , w_{t+1} , and w_{t+2} are used to predict the word w_t . The word embedding for w_t is finally represented by the weights of the neural network’s hidden layer.

it recently gained a lot of traction with the advent of neural embedding models. The idea here is not to costly factorize a huge word co-occurrence matrix, but instead directly learn the low-dimensional *word embeddings* that approximate the columns of the factorized matrix. In [25], Bengio et al. proposed a language model based on neural networks. Later on, Mikolov et al. proposed Word2Vec [156], a fast and easy-to-use embedding method to predict a word from its context. This training objective is based on the intuition that words with similar contexts are in turn also similar. There have been other embedding learning algorithms mimicking Word2Vec’s behaviour, such as GloVe [184] or LINE [227]. In this work, we make use of GloVe, LINE and Word2Vec. In the following, we briefly describe the intuition behind each of these, their training objectives and their relation to each other.

The most well-known word embedding algorithm today, *Word2Vec*, makes use of a shallow neural network, i.e., with only a single hidden layer, to produce word embeddings. It comes in two variants: predicting the context of a given word (SkipGram) or predicting a word from a given context (CBOW) [156]. The word representations are extracted from the hidden layer of the neural network. For an illustrative example, see Figure 3.5. The Skip-Gram optimization objective is to maximize the prediction probability $p(w_{t+j}|w_t)$ of a context word w_{t+j} , given a center word w_t .

$$\frac{1}{|\mathcal{V}|} \sum_{t=1}^{|\mathcal{V}|} \sum_{\substack{-c \leq j \leq c, \\ j \neq 0}} \log p(w_{t+j}|w_t) \quad (3.12)$$

Here, c is the context size and $p(w_{t+j}|w_t)$ is defined via the softmax function

$$p(w_{t+j}|w_t) = \frac{\exp(v_{t+j}'^T v_t)}{\sum_{i=1}^{|\mathcal{V}|} \exp(v_i'^T v_t)} \quad (3.13)$$

The vectors v_i are the actual learned embeddings of words w_i , which we will use later in this work, concretely in Section 6.5. Because Equation (3.13) depends on the size of the vocabulary in the denominator, computing its value is expensive. In order to decrease

3.2 Computational Methods to Model Distributional Semantic Knowledge

computation time, Mikolov et al. experimented with different ways to approximate the denominator term. Discussing those is however irrelevant to the scope of this thesis and we refer the reader to the original paper [156].

In contrast to Word2Vec, the *GloVe* algorithm by Pennington, Socher, and Manning exploits the *global* context of words instead of the *local* context, i.e., the make use of the vast amount of word repetition in the global corpus context [184]. Thus, the resulting word embeddings are similar when the mutual information of two words is high. GloVe takes a word-word co-occurrence matrix $W = (w_{ij})$ as input, as described in Equation (3.11). In the optimization objective, i.e., minimize

$$\sum_{i,j=1}^V f(w_{ij}) (v_i^T \tilde{v}_j + b_i + \tilde{b}_j - \log w_{ij})^2, \quad (3.14)$$

v_i and \tilde{v}_i are the resulting embedding vectors for a word and its context, respectively. b and \tilde{b} are bias vectors, again for a word and its context. Finally, the weight function f downweights rare words and thus reduces the impact of data noise on the resulting model. In the conducted experiments in [184], Pennington, Socher, and Manning argue that the GloVe model outperforms Mikolov's Word2Vec as well as a set of other embedding algorithms in analogy and word relatedness tasks. Still, they note that by choosing a set of right parameters, this advantage can be diminished, and that this needs a more thorough analysis. Later, Levy, Goldberg, and Dagan showed in [134] that the performance of both algorithms is indeed heavily influenced by the choice of hyperparameters.

Finally, *LINE* [227] is a *graph embedding* algorithm to obtain vector representations for nodes in a graph instead of words. However, the global context of a word can then be encoded in the neighborhood of a node in the co-occurrence graph. It is thus still possible to learn word embeddings with this method. LINE attempts to reflect the proximity of a node in the graph in the distance of the generated embeddings. The algorithm relies on two definitions of node proximity. The *first-order proximity* in a network is the *local* pairwise proximity of two nodes, i.e., the weight of an edge connecting these two nodes. For each pair of vertices linked by an edge (i, j) , the weight on that edge ω_{ij} indicates the first-order proximity between i and j . If there is no edge between i and j , their first-order proximity is 0. Thus, the first-order neighborhood of a node i contains all nodes j which are directly connected with an edge to i . The strength of this correlation is defined by the weight ω_{ij} of the edge. To achieve this objective, LINE minimizes the following objective function:

$$O_1 = - \sum_{(i,j) \in E} \omega_{ij} \log \left(\frac{1}{1 + \exp(-\langle v_i, v_j \rangle)} \right) \quad (3.15)$$

Here, v_i and v_j are the embeddings of nodes i and j in the graph. The *second-order proximity* of two nodes between a pair of vertices (i, j) in a network is the similarity between their *first-order neighborhoods*. Mathematically, let $p_i = (w_{i,1}, \dots, w_{i,|V|})$ denote

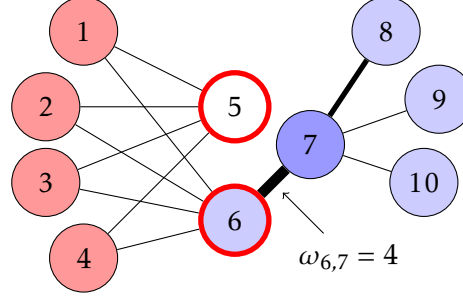


Figure 3.6: Illustration of node neighborhoods in a graph as defined in [227]. The light-blue filled nodes are considered first-order neighborhood for node 7, while the thickness of the edges denote the weight of the connection. For example, the edge between nodes 6 and 7 has weight $\omega_{6,7}$. The nodes filled with red define the most amount of the second-order neighborhood of nodes 5 and 6 (with a red border).

the first-order proximity of i with all the other vertices, then the second-order proximity between i and j is determined by the similarity between p_i and p_j . If no vertex is linked from/to both i and j , the second-order proximity between i and j is 0.

$$O_2 = - \sum_{(i,j) \in E} \omega_{ij} \log \left(\frac{\exp(\langle v_i, v_j \rangle)}{\sum_{k=1}^{|V|} \exp(\langle v_i, v_k \rangle)} \right) \quad (3.16)$$

Both types of node proximity are illustrated in Figure 3.6.

LINE embeddings are typically constructed by separately learning embeddings for the first-order and second-order proximity objectives and by subsequently concatenating them. This way, both proximity variants can influence the similarity of nodes, or in our case words.

We will discuss the applicability of each word embedding algorithm on tagging data in Section 6.5, before we evaluate the learned tag embeddings on human intuition. In Chapter 8, the low dimensionality of those embeddings enables us to design an algorithm that refines the word vectors to better reflect human intuition of semantic relatedness.

3.2.2 Choosing the Right Context

What is common to all models in the preceding section, is that they rely on word co-occurrences in a given *context*. Obviously, the quality of a model relies heavily on the amount of provided information by a) choosing the right context and b) the right size of the context. In this section, we will briefly present how the context is chosen in in folksonomies and in any kind of graph.

3.2.2.1 Context in Folksonomies

Because folksonomies exhibit a *tripartite* structure (see Section 3.1.1) of users, tags and resources, the choice of a suitable context for extraction of semantic knowledge has been extensively discussed [48, 262]. In this thesis, we will stick to the context definitions by Cattuto et al. [48], which we will re-evaluate in Section 7.3.1.2.

Given a folksonomy $\mathbb{F} = (U, T, R, Y)$, Cattuto et al. define three natural first-order co-occurrence context descriptions for tags based on the posts in a social tagging system, all of which can finally be used in the vector construction methods we introduced in Section 3.2.1.1. The context of a tag can be described by either its co-occurring *tags* in the same post, by the *users* who have used this tag, or by the *resources* that this tag has been assigned to. They employ a vector space model to represent words, i.e., tags, by their contexts, so the corresponding feature spaces are of the dimensions $|T|$, $|U|$, and $|R|$, respectively. Thus, each context description can also be represented as a matrix $C \in \mathbb{R}^{k \times |T|}$, where k denotes the dimension of the corresponding feature space. The column vectors v_j of these matrices represent the corresponding semantic contexts of each tag t_j , $j = 1, \dots, |T|$.

Tag Context The tag context matrix $C^{tag} \in \mathbb{R}^{|T| \times |T|}$ describes the context of a tag based on the posts in a folksonomy, i.e., the distinct annotations of resources by users.

$$C_{ij}^{tag} := \left| \left\{ (u, r) \in U \times R \mid (u, t_i, r), (u, t_j, r) \in P_{ur} \right\} \right| \quad (3.17)$$

The entry in a matrix cell C_{ij}^{tag} in the tag context matrix is hence the number of posts P_{ur} , where a user u assigned both tags t_i and t_j to the same resource r .

Resource Context If we consider the resource context of tags, we obtain a matrix $C^{res} \in \mathbb{R}^{|R| \times |T|}$, which is defined as follows:

$$C_{ij}^{res} := \left| \left\{ u \in U \mid (u, t_j, r_i) \in Y \right\} \right| \quad (3.18)$$

This measure counts how often a tag occurred with the same resource.

User Context In the user context matrix $C^{user} \in \mathbb{R}^{|U| \times |T|}$, the entries describe how often a tag has been used by the same user.

$$C_{ij}^{user} := \left| \left\{ r \in R \mid (u_i, t_j, r) \in Y \right\} \right| \quad (3.19)$$

Cattuto et al. found that resource context produced the best results for the construction of meaningful co-occurrence vectors, closely followed by tag context. Still, the tag context similarity yielded the most clear-cut results in many other experiments, so we will use this context type in the remainder of this work, if not specified otherwise. We use all those context definitions in Section 7.3.1.2.

3.2.2.2 Context in Graphs

The context of a node in graphs is not as easily defined as in text. Textual context, e.g., in a sentence, is one-dimensional: There are either words in front or behind the current position. In graphs, there is no concept of “before” or “behind”.

Node Neighborhood Instead, one can define a *Neighborhood* of a node. In [227], this neighborhood is characterized in two ways. The *first-order* neighborhood is defined as all nodes directly connected to the current node, while the *second-order* neighborhood includes all nodes that share a similar first-order neighborhood with the current node. Figure 3.6 shows an illustration of both neighborhood variants. However, as we already introduced that in Section 3.2.1.2, we will not explain either concept here.

Paths in a Graph In contrast to the notion of node neighborhoods in a graph, another option is to *generate* context for nodes by *walking* the graph. This way, a one-dimensional context is artificially generated (similar to a sentence in natural language), on which we can subsequently apply any methodology from Section 3.2.

Definition 12 (Navigational Paths in a Graph). *Given a graph $G = (V, E)$ with vertices V and directed edges $E = \{(v, w) | v, w \in V\}$, we define a path $\mathbf{p} \in \mathbb{P}$ from the set of all paths \mathbb{P} on G as a sequence of vertices $\mathbf{p} = (v_1, \dots, v_n)$ with $v_i \in V, 1 \leq i \leq n$ and $(v_i, v_{i+1}) \in E, 1 \leq i \leq n - 1$. The length of a path \mathbf{p} is defined as the length of the corresponding sequence of vertices.*

Paths in a Graph can originate from various sources, such as a collection of all potential node sequences of a given length [210], a set of randomly generated paths [55, 87, 185], or human navigation [175, 210, 249]. In Chapter 7, we will propose methods that exploit human navigation as well as artificially created random walks in social media systems to model semantic knowledge. Consequently, navigational paths play a central role in this thesis.

Additionally, we show that human navigation is a valuable source of semantic information, when compared to the static graph structure of social media pages. In this chapter, we will generally work on *link graphs*, i.e., graphs of web pages p that are connected through directed links (p_i, p_j) .

3.2.3 Measuring Semantic Relatedness and Similarity

For many applications in NLP and Ontology Learning, it is necessary to determine if two words are related to a certain extent. While determining the nature of that relation is also important for many tasks, we exclusively focus on measuring the *strength* of such a relation in this thesis.

Before we show how we can use the semantic models defined in the previous sections, we introduce the notion of a *similarity measure*. We define a similarity measure between two entities i and j , each represented by a vector v_i and v_j , as follows:

Definition 13 (Similarity Measure). *Given a coherent interval $S \subseteq \mathbb{R}$, a function*

$$\text{sim}(v_i, v_j) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow S \subseteq \mathbb{R} \quad (3.20)$$

is called a similarity measure between two feature vectors $v_i, v_j \in \mathbb{R}^n$, if it suffices the following conditions:

1. $\text{sim}(v_i, v_i) = \max\{\text{sim}(v_i, v_j) | v_j \in \mathbb{R}^n\}$.
2. $\text{sim}(v_i, v_j) = 0$, iff v_i and v_j are completely unrelated.
3. $\text{sim}(v_i, v_j) > \text{sim}(v_i, v_k) \Leftrightarrow v_j$ is more similar to v_i than v_k .

While S can span the whole of \mathbb{R} , it can be easily normalized to the interval $[0; 1]$. In a semantic interpretation a similarity value of 0 between two word vectors v_i and v_j means complete unrelatedness, while a value of 1 means semantic synonymy. However, scores near 1 indicate that both words might be semantically connected by some kind of semantic relation. We also interpret the similarity score as the *strength* or *intensity* of the relation.

3.2.3.1 Distinguishing Semantic Relatedness and Similarity

For the remainder of this work, we feel it is important to point out the difference between semantic *relatedness* and *similarity*. Although both concepts are based on the notion that entities are to some extent semantically related to each other, the notion of actual similarity is a much more strict one.

Generally speaking, *semantic relatedness* between concepts represents the degree of “closeness” between those concepts [205]. While closeness is to a great extent subjective, there exists at least a consensus margin (see e.g., the singular voting results for the WS-353 dataset, Figure 3.9). An example of two semantically related words are coffee and cup, because one pours coffee into a cup, so both words are in some way related.

There also exists the notion of *semantic similarity*. In [137], Lin et al. gave an information-theoretic definition for general similarity, based on rather abstract propositions of “commonality” and “descriptions” of two objects. The object similarity is then defined as the ratio of the commonality information and the description information. Similarity is much more narrow than relatedness, in a sense that while coffee and cup are related, they are in no way similar, because one can not usually roast or drink a cup. However, tea and coffee are similar in a way that you can use both to make a very pleasant drink. But even here, the degree of similarity is subjective to the judging person, as a tea drinker with a distaste of coffee might rule them both more different than someone who likes to drink both. However, even the person disliking coffee cannot rationally judge it total dissimilar to tea, so with similarity, there also exists a subjectivity margin.

Both semantic relatedness and semantic similarity have different applications. For example, semantic similarity is useful to resolve synonymy or to construct synsets in

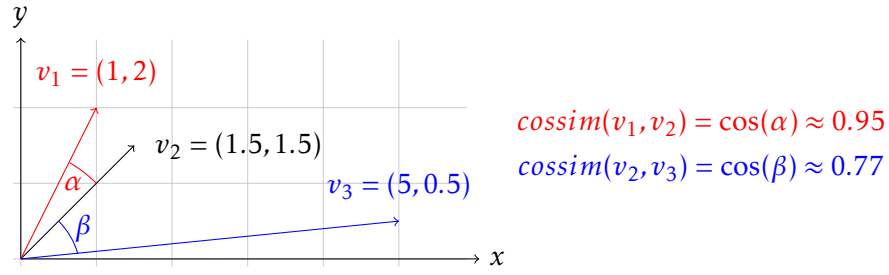


Figure 3.7: Illustration of the cosine similarity measure for vectors. The similarity between two vectors only depends on the angle between both vectors. Thus, similar vectors that point in the same direction have a high cosine similarity, regardless of their Euclidean distance.

ontologies, while semantic relatedness is important to construct semantic relations between concepts, including semantic similarity.

In this thesis, we will mostly focus on determining the semantic *relatedness* between terms. Wherever necessary, we will explicitly state if we aim to model semantic *similarity*.

3.2.3.2 The Cosine Similarity Measure for VSMs

The cosine measure is one of the most used relatedness measures in literature [13]. Intuitively, it measures the “closeness” of two points on a unit sphere in an Euclidean space. This means that the smaller the angle between two vector is, the higher the cosine similarity of two vectors. This is illustrated in Figure 3.7 Given two feature vectors v_i, v_j , the cosine measure is defined as follows:

$$\text{cosine}(v_i, v_j) = \frac{\langle v_i, v_j \rangle}{|v_i| \cdot |v_j|} = \left\langle \frac{v_i}{|v_i|}, \frac{v_j}{|v_j|} \right\rangle. \quad (3.21)$$

The cosine measure is directly related to the \mathcal{L}_2 norm and, written as a distance $d(x, y) := 1 - \text{cosine}(x, y)$, even equivalent to the Euclidean distance on normalized vectors. Since the calculation of the cosine measure is linear, it is very efficient to process and thus widely used. *In this thesis, we use the cosine measure as the main vector similarity measure and as a measure of semantic relatedness.*

It is possible to parameterize the cosine measure (in fact, we propose to do this in Chapter 8) by incorporating a positive semidefinite matrix (PSD) M as follows:¹⁸

$$\text{cosine}_M(v_i, v_j) := \frac{\langle v_i, v_j \rangle_M}{\|v_i\|_M \|v_j\|_M} = \frac{v_i^T M v_j}{\sqrt{v_i^T M v_i v_j^T M v_j}} \quad (3.22)$$

The PSD condition for M is necessary to enable cosine_M to satisfy the conditions of a scalar product. Since M is PSD, it is also symmetric and thus can be expressed as

¹⁸A quadratic matrix is positive semidefinite iff its eigenvalues are equal or greater than 0.

3.2 Computational Methods to Model Distributional Semantic Knowledge

a product of a matrix D with its own transpose: $M = D^T D$. If we insert that into Equation (3.22), we get

$$\text{cosine}_M(v_i, v_j) = \frac{v_i^T D^T D v_j}{\sqrt{v_i^T D^T D v_i v_j^T D^T D v_j}} = \frac{(Dv_i)^T Dv_j}{\|Dv_i\| \|Dv_j\|} = \frac{\langle Dv_i, Dv_j \rangle}{\|Dv_i\| \|Dv_j\|}, \quad (3.23)$$

i.e., we receive the original cosine measure, but of the transformed vectors Dv_i . Technically speaking, this even means that not necessarily the cosine measure is adapted, but the word vectors themselves.

3.2.3.3 WordNet-Based Semantic Similarity Measures

As already described in Section 3.1.3, WordNet encodes several explicit semantic relations between entities. Usually, those relations are not weighted. For applications, for example semantically supported search engines, a weighted notion of similarity or relatedness is however important to find more relevant results. Intuitively, WordNet is especially well suited to calculate semantic relatedness between its entities, with hypernymial and hyponymial links between synonym sets and instances. Because of this intuition, many semantic relatedness measures based on WordNet have been proposed. For an extensive overview and comparison of WordNet based similarity measures, the reader is referred to [44]. In the following, we will introduce two well-known measures, which we will discuss in more detail in Section 7.3.1.2.

The first and most simple WordNet based semantic similarity measure is the taxonomic shortest-path similarity measure by Hirst, St-Onge, et al. [105]. Jiang and Conrath proposed the Jiang-Conrath distance [112], which combines Resnik's information content similarity measure [195] and the taxonomic shortest-path similarity measure. In the following, we will specifically define those two prominent semantic similarity measures on WordNet.

Taxonomic Shortest-Path Similarity The *taxonomic shortest-path similarity* by Hirst, St-Onge, et al. depends on the length of the shortest path between two concepts x and y in the WordNet taxonomy [105]:

$$\text{sim}_{\text{path}}(x, y) := \max_{p \in \mathbb{P}(x, y)} \frac{1}{\text{len}(p)} \quad (3.24)$$

Here, $\mathbb{P}(x, y)$ denotes the set of all paths in the WordNet taxonomy between x and y . This measure is very intuitive, since the farther two entities are away in the taxonomy tree, the less related they are.

Jiang-Conrath Distance and Similarity In [112], Jiang and Conrath proposed a distance measure on the WordNet hierarchy, which depends both on the aforementioned path-based similarity as well as on an information content criterion inspired by Resnik's measure [195]. It is defined as

$$d_{\text{jcn}}(x, y) := IC(x) + IC(y) - 2 \cdot IC(L\text{Super}(x, y)), \quad (3.25)$$

where $IC(x)$ is the information content of x in WordNet and $LSuper(x, y)$ is the least common subsumer, i.e., the lowest node in the WordNet hierarchy which both subsumes x and y . In [44], Budanitsky and Hirst showed that the Jiang-Conrath distance yields the highest correlation with human intuition among all of the compared similarity measures. It is important to note that Equation (3.25) defines a *distance* measure. A corresponding *similarity* measure can be obtained by e.g., calculating

$$sim_{jcn}(x, y) := \frac{1 + \alpha}{d_{jcn}(x, y) + \alpha} \quad (3.26)$$

with $\alpha > 0$ to avoid division by zero.

3.2.4 Evaluating Semantic Relatedness Measures

The goal of creating a semantic relatedness measure is always to mimick human intuition of semantic relatedness. To this end, it is necessary to ground the measure on a gold standard of relatedness information. Additionally, we perform a statistical significance check to see if the produced scores are indeed valid. Over time, literature provided many options to do so. Schnabel et al. [204] listed a set of evaluation approaches and discussed each of them briefly. We will focus on two approaches, which “ground” a semantic relatedness measure on human intuition.

3.2.4.1 Grounding on Semantic Relatedness Datasets

In this work, we mainly focus on evaluation on semantic relatedness datasets, as done throughout many works in literature [43, 44, 75]. While this way of evaluating semantic relatedness has some flaws [71], it is one of the most direct options to evaluate the modeled semantic information on human intuition. *In fact, we will use this method throughout the whole thesis as the main performance indicator of our models.*

In order to correctly assess how well the semantic relatedness scores produced by a relatedness measure fit to human intuition of semantic relatedness, we can compare them to human judgment. As described in Section 4.4, there exist semantic relatedness datasets which contain lists of word pairs together with human-assigned relatedness scores, i.e., human judgment of semantic relatedness. The procedure of evaluating semantic relatedness measures on those scores is then as follows:

1. For each pair of words in the human intuition dataset where both words also possess a vector representation, calculate the cosine similarity of those vector representations.
2. Calculate the correlation for the word pair ranking induced by the human-assigned scores and the word pair ranking induced by the cosine scores.
3. The nearer the correlation score of both rankings is to 1, the better can the vector representations be used to approach human intuition of semantic relatedness.

3.2 Computational Methods to Model Distributional Semantic Knowledge

It is furthermore important to note the number of pairs on which the correlation score is calculated. A higher coverage of word pairs in an evaluation dataset also means a stronger claim of validity for the calculated scores (see below). Additionally, correlation scores from different word pair subsets are hardly comparable, since the actual gold standard that we evaluate against changes. The reason why one can not always evaluate against each single pair of concepts is that the concept might not even be present in the underlying corpus.

We now briefly discuss the choice of the method to calculate correlation scores. In literature, there largely exist two popular choices of how to determine the correlation score. The first is the Pearson correlation coefficient, which determines a potential *linear* correlation between two random variables X and Y [183]:

$$r := r_{X,Y} := \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} \quad (3.27)$$

Cov denotes the covariance of X and Y , while σ_X is the standard deviation of X , which equals the square root of its variance $\text{Var}(X)$. The Pearson coefficient takes into account the *actual values* of the random variables and compares them with each other. In contrast, the Spearman rank correlation coefficient, relies on the *ranking* of the values, i.e., the fact *that* one value is higher than another is more important, but not *how much* higher it is. The value ranking of a random variable X is denoted by rg_X and describes the list position of each value in a sorted list of values. It thus measures the *monotonic* correlation between two random variables and is defined as follows [214]:

$$\rho := \rho_{X,Y} := \frac{\text{Cov}(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}} \quad (3.28)$$

Concretely, this also constitutes a Pearson correlation, but between two *rankings* of scores instead of the scores themselves. The constituted difference and the need for a variation of the Pearson correlation coefficient can be seen in the Anscombe quartet. The Anscombe quartet consists of 4 different distributions of data points with the same Pearson correlation, but different Spearman correlations [7]. It is depicted in Figure 3.8.

Applied to the evaluation of semantic relatedness scores, the Spearman correlation intuitively captures what measuring semantic relatedness is about: We want to be able to validate an interpretable monotonic claim such as

money and *bank* are more closely related than *drink* and *ear*

than an uninterpretable judgment about an arbitrary absolute score, such as

alcohol and *chemistry* are 54.8% correlated.

We support this intuition by looking at the score standard deviation of the human judgement scores in both WS-353 and Bib100, which will be introduced in detail in Section 4.4 and Section 6.2, respectively. Figure 3.9 shows a joint scatter plot of mean human judgment of relatedness together with its standard deviation. Especially for

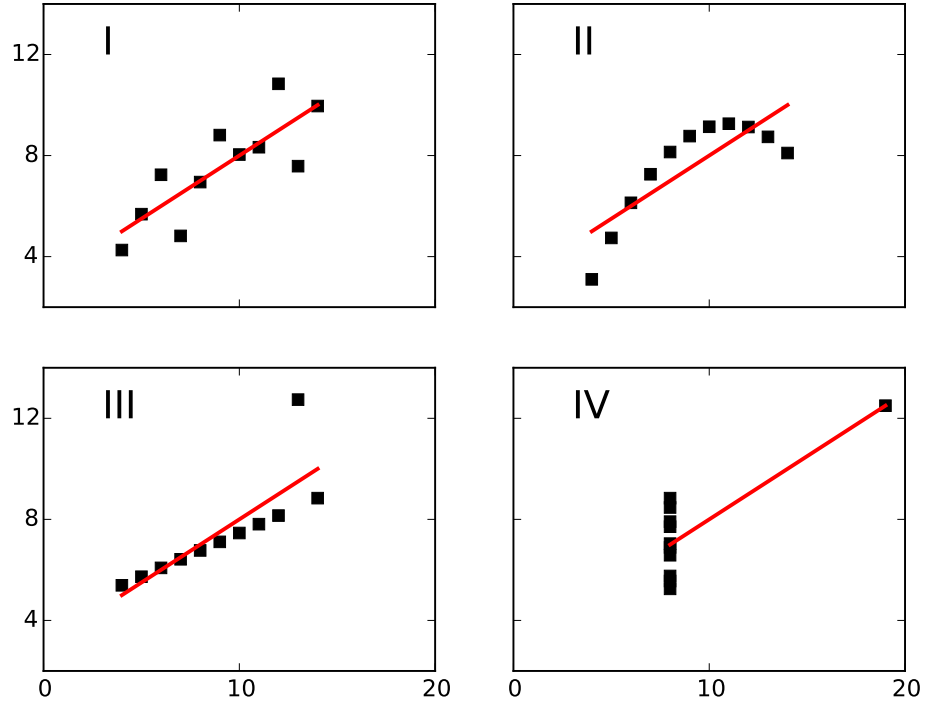


Figure 3.8: Anscombe quartet. There are 4 distributions depicted, each with the same Pearson correlation score, but different Spearman correlations. While the Pearson correlation measures a *linear* correlation, Spearman denotes the *monotonic* correlation between two random variables X and Y .

the lesser related pairs, human judgment sometimes shows differences of more than 4 score points and is only relatively low near completely unrelated and highly related word pairs. We also plotted the density of judgments to underline that observation. Thus, if humans are unsure how a medium-related pair of words should be judged via an absolute score and thus give scores spanning a great range, it is not necessary for a computer to return absolute values of relatedness, either.

Lastly, to determine the validity of a correlation score or an improvement upon another score, it is necessary to get a notion of their statistical significance. For this, we can either determine if the score is significantly different from zero, i.e., if the observed correlation score is actually present, or if two scores are significantly different. To calculate a p -value, we use Fisher's z -transformation for Spearman correlation scores [74]:

$$p(\rho_1, \rho_2, n) := 1 - \text{error} \left(0.5 \cdot (z(\rho_1) - z(\rho_2)) \cdot \sqrt{\frac{n-3}{1.06}} \right) \quad (3.29)$$

3.2 Computational Methods to Model Distributional Semantic Knowledge

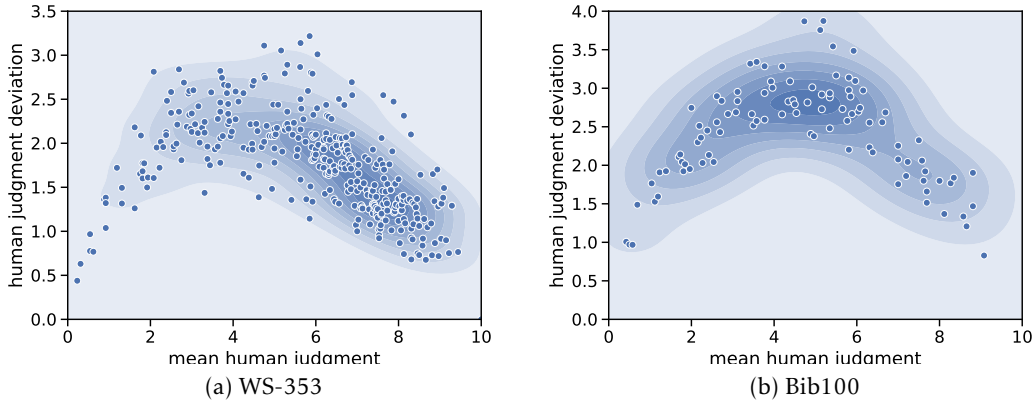


Figure 3.9: Joint plot of mean human judgment with standard deviation for both WS-353 and Bib100 together with judgment density. For both datasets, the standard deviation is lower towards the extremes of the judgment range, as can be seen from the underlying density plot. The datasets will be introduced in Section 4.4.

Here, z denotes the z-transformation $z(\rho) := \log \frac{1+\rho}{1-\rho}$ and $error$ is the Gaussian error function $error(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-\tau^2} d\tau$. Finally, n denotes the number of samples on which the two correlation coefficients ρ_1 and ρ_2 have been calculated. The constant 1.06 is determined empirically [74].

Naturally, the higher n , the more valid is the claim. For example, the Spearman correlation score of a ranking $X = [1, 2, 3]$ with itself is $\rho(X, X) = 1$. A simple swap of two items, e.g., $Y = [1, 3, 2]$, results in a Spearman correlation score of $\rho(X, Y) = 0.5$, which is a difference of 0.5. However, if conduct the same experiment with a list of 100 items and again switch only the last two items, we obtain a score difference of 0.05. Consequently, the more items are compared using the Spearman correlation coefficient, the more difficult it is to create big score differences. Thus, the significance of a high score difference is higher if the compared rankings consist of a large number of items.

3.2.4.2 Grounding on Word Thesauri

The semantic similarity measures proposed in Section 3.2.3.3, especially the Jiang-Conrath measure, exhibit a very high correlation with human judgment [44]. In [48] and [148], the Jiang-Conrath measure has been used as a proxy of an evaluation ground truth for semantic similarity measures. We will describe that evaluation approach here, because we will compare it to the evaluation on human judgment in Section 6.2.

Given a set of tags T and a context-based semantic similarity measure sim_{ctx} (cf. Section 3.1.1) on these tags, we first calculate for each tag $t \in T$ the most closely related tag $t' \in T \setminus \{t\}$ according to sim_{ctx} . For each of these pairs (t, t') , we then calculate

the mean of the WordNet-based semantic relatedness scores $sim_{wn}(t, t')$. Under the assumption that $sim(x, y)$ is bounded, this can be formally expressed as:

$$eval(sim_{ctx}, rel_{wn}) := \frac{1}{|T|} \sum_{t \in T} sim_{wn} \left(t, \arg \max_{t' \neq t} sim_{ctx}(t, t') \right) \quad (3.30)$$

Here, sim_{wn} is any similarity measure defined on WordNet. In this work, we use both the taxonomic shortest path similarity (Equation (3.24)) as well as the Jiang Conrath distance (Equation (3.25)), or more concretely, its similarity version, as defined in Equation (3.26). We discuss this evaluation approach in Section 6.2 and compare it with the evaluation on human judgment described in Section 3.2.4.1.

3.2.5 Post-Processing and Applying Word Vector Representations for Semantic Information Gain

Often enough, word representations capture human intuition rather well, although their abilities to reflect that intuition are always limited. While this on the one hand depends on the algorithm to construct word embeddings, this representation is also heavily influenced by the underlying corpus. Even corpora with standard vocabulary and general topics that cover most of the vocabulary contained in human evaluation datasets cannot exactly capture the actual relations and score rankings defined in the evaluation datasets. For example, a corpus containing scientific papers will hardly be able to correctly assess the semantic similarity of dog and cat. On the other hand, the exact meaning of the word paper depends on the context it is used in. In order to fix such problems, there have been some efforts to *adapt* or fine-tune existing word representations in a *post-processing step*, as well as methods that *apply* the vector representations to discover additional knowledge.

In the following, we will first introduce the Retrofitting algorithm by Faruqui et al., which is used to adapt vector representations to external semantic resources, in this case to semantic lexicons [69]. After this, we describe an approach to discover and disambiguate the different meanings of social tags, that is based on clustering the context of those tags. This tag disambiguation algorithm has been proposed by Benz [26].

3.2.5.1 Improving Vector Representations using Lexicons: Retrofitting

Faruqui et al. introduced a method called *Retrofitting*, where a set of existing word vectors \hat{Q} is adapted to a semantic lexicon, i.e., a dictionary of word-word relations. For example, the synset relations in WordNet can be interpreted as such a lexicon. In [69], they aim to infer new vectors Q from the old ones \hat{Q} by ensuring that the new vectors q_i stay close to their old representations \hat{q}_i , but also close to the vectors q_j of neighboring words according to the semantic lexicon. The minimization function is thus

$$\Psi(Q) = \sum_{i=1}^n \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j) \in E} \beta_{ij} \|q_i - q_j\|^2 \right] \quad (3.31)$$

By varying α_i and β_{ij} , it is possible to put different weight on each criterion. The set of edges E , which define the neighborhood of a word, can for example be taken from an ontology, such as WordNet. While the resulting word vectors improved representation of semantics, Faruqui et al. could show that their results even hold in other languages. Also, retrofitting only affects vectors of words that occur in the lexicon, while the remaining vector space remains untouched. We will use Retrofitting as a baseline in Chapter 8.

Usually, those methods use external knowledge to either move some word vectors nearer towards others or even change the complete vector space. Technically, this process is called *Metric Learning* and has been an active field of research since 2003 [257]. The goal is to learn a distance metric according to a given set of constraints. Bellet, Habrard, and Sebban provided an exhaustive survey of metric learning algorithms in [23].

3.2.5.2 Discovering Word Senses Using Vector Representations

An important problem in modelling semantic information as vectors is polysemy, i.e., when a single word may have multiple meanings. Polysemy clearly affects functions of social media such as information retrieval or browsing: Because of the different senses of a word may be semantically unrelated (e.g., *apple* is both a fruit and a computer manufacturing company), the user is presented with irrelevant content. While this problem is present basically within all systems dealing with natural language, we consider it especially in the context of social annotation systems, as the open vocabulary as well as the lack of structure (compared to, e.g., the syntax of a written text) makes this issue in social annotation systems unique and interesting.

NLP approaches in the field of *word sense discovery* like [66, 179] are typically applying clustering approaches to divide a suitable context of a given term into partitions which correspond to its senses. When transferring this idea to social annotation systems, the problem can be divided into the following subproblems, (i) *context identification*, i.e., how to construct a “suitable” context and (ii) *context disambiguation*, i.e., how to subdivide this context into senses.

In the following, we describe a two-step approach for context identification and disambiguation for tagging data, as described by Benz in his thesis [26]. We will use the same approach later in Section 6.4, however this time in order to measure the impact of user pragmatics (see Section 3.3.1) on the tag disambiguation process.

Sense Context Identification. In prior work, [9] performed extensive studies on the characteristics of different context definitions for the task of tag sense discovery. The authors examined tag- and user-based document networks, as well as tag co-occurrence and similarity networks. It was found that tag similarity networks provided “*the most clear-cut results among all the network types*”.

The next question is which tags to include in the context of a given tag t . The goal here is to choose a sample of context tags which are representative for t ’s main senses. Here, Benz follows the heuristic described by [192], who found that the *20 strongest first-order associations [...] are [...] a good mix of the two main senses for each word*. In the case of tagging systems, first-order associations correspond to tag-tag co-occurrences.

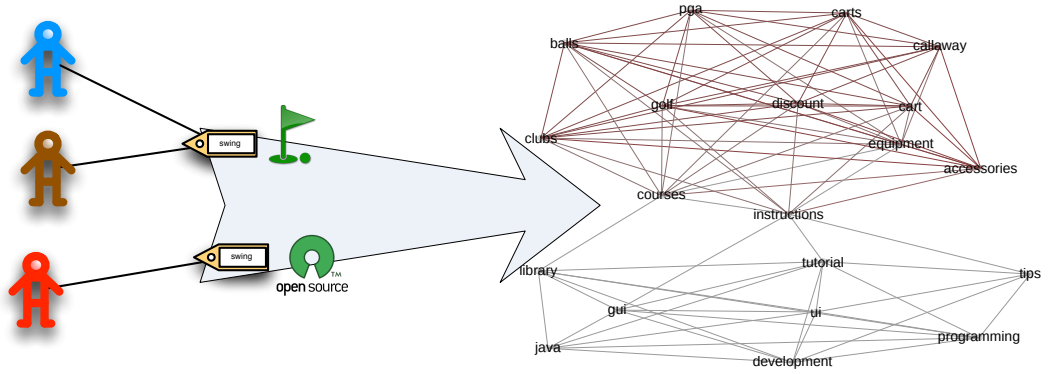


Figure 3.10: The process illustrated in this section and an exemplary sense context graph of the tag *swing*. While the upper tag cluster loosely corresponds to the *golf* sense of the word, the lower tag cluster corresponds to the *programming library* sense of the word. For readability reasons, only edges with weights > 0.18 are included in this illustration.

Although Benz does not necessarily target to discover *two* main senses, he follows the following steps to construct a context for a given tag t :

1. Let $t \in T$ be a tag whose senses are to be discovered.
2. Let $SC_t = (V_t, E_t)$ be an initially empty undirected graph, whose edges are weighted by a weighting function $w : V_t \rightarrow \mathbb{R}$. This graph is called the *sense context graph* for t .
3. The vertices V_t are constructed by adding those 20 tags $t_i \in T, t_i \neq t, i = 1, \dots, 20$ which co-occur most often together with t .
4. The edges are constructed by computing the pairwise tag context relatedness as described above among all $t \in V_t$; an edge is added between t_i and t_j if their similarity is greater than zero. The weights of the edges are given by the corresponding similarity value.

Figure 3.10 shows the process of inferring senses from tagging data and an exemplary sense context graph for the tag *swing*.

Sense Context Disambiguation. Given this graph representation of the context, the next problem is how to divide it into partitions which denote different meanings. Amongst others, clustering techniques have been used to this end, e.g., Clustering By Committee [179], Markov Clustering [66] or graph clustering [9]. In a different context, this problem is similar to community detection, for which modularity-based clustering techniques showed promising results [132]. These techniques were also applied to sense detection by [9].

3.2 Computational Methods to Model Distributional Semantic Knowledge

In order to better examine the influence of pragmatic factors, Benz adopted hierarchical agglomerative clustering as used in [28] as a representative of a standard sense discovery algorithm. Based on the similarities among the context tags which form the edges of the sense context graph SC_t , the hierarchical clustering procedure can be directly applied to form “sense clusters” in the following way:

1. Compute the distance matrix containing the distance between each pair of tags. Similarity values s are transformed into distances d according to $d = 1 - s$. Treat each tag as a cluster.
2. Find the least distant pair of clusters using the distance matrix. Merge these two clusters into one cluster. Update the distance matrix to reflect this merge operation.
3. If all objects are in one cluster, stop. Otherwise, go to step 2.

This procedure results in a so-called *dendrogram*, which graphically depicts the level of distance at which each merging step took place. Hierarchical agglomerative clustering algorithms can mainly be differentiated based the scheme by which the distance matrix in step 2 is updated. The core question here is how the distance between merged clusters is computed. Let $U = \{u_1, \dots, u_i\} \subseteq O$ be a newly created cluster by merging two existing clusters $S \subseteq O$ and $T \subseteq O$, and let $V = \{v_1, \dots, v_j\} \subseteq O$ be another cluster. Typical standard schemes are single link, complete link or average link clustering. In [26], Benz found that *Ward’s minimum variance method* showed the most promising results [242]. It updates the distance matrix according to:

$$dist(U, V) = \sqrt{\frac{|V| + |S|}{x} dist(V, S)^2 + \frac{|V| + |T|}{x} dist(V, T)^2 + \frac{|V|}{x} dist(S, T)^2} \quad (3.32)$$

(Here, $x = |V| + |S| + |T|$).

As stated above, the outcome of the clustering step is not a fixed set of clusters, but rather a dendrogram which captures the agglomerative merging steps. In order to derive clusters (which is desirable in this case), this dendrogram needs to be further parameterized. One method is to “cut” the latter into a set of flat sense clusters by using a distance threshold τ .

Once sense clusters are identified, the computed senses need to be labeled. In the literature, the most popular tags within the resulting clusters are sometimes used [9]. Benz chose a single label by selecting the tag t_i within the sense cluster which maximizes the tag context relatedness to the tag t which is to be disambiguated. More formally, let $S = \{t_1, \dots, t_i\}$ be a sense cluster of t , and let sim denote one of the three relatedness measures mentioned above. Then Benz chose the label t_S for S by $t_S = \operatorname{argmax}_{t' \in S} sim(t, t')$

While no tags were eliminated from the sense cluster, t_S can be taken as a rudimentary description of what the cluster is about. We will refer to the remaining tags as *sense tags* because they serve as additional descriptions.

3.3 Characterizing User Behavior in the Social Web

Analyzing the behavior of humans in the web is since long a highly interesting topic for a large part of the research community. Not only do researchers want to know *how* humans make use of the resources in the web [38, 67, 141], but also *why* they act like they do [4, 24, 221]. This knowledge can then be used to improve user interfaces or the kind and amount of provided information so that humans are enabled to exploit the whole potential of the World Wide Web. Additionally, there exists evidence that user behavior also influences the semantic information of social media data. In the example of social tagging systems, users who assign many different tags to their resources contribute more valuable semantic information than users who make use of an organized, yet limited vocabulary [122].

In this section, we first introduce several measures to categorize users of social tagging systems according to their tagging behavior. After that, we describe a method to compare hypotheses about user navigation. We apply that method in Chapter 5 on user navigation in social media systems. In that context, we will also show that the tagging behavior of users, as defined by the introduced measures, also impacts navigation behavior in social tagging systems.

3.3.1 User Behavior Types in Folksonomies

As mentioned in Section 2.3, users in social tagging systems can be characterized by how they assign tags to resources. For example, Strohmaier, Körner, and Kern defined the two classes of *categorizers* and *describers* [221], while Lorince et al. identified *supertaggers* and *non-supertaggers* [141]. In this work, we use categorizers and describers to measure an effect of user tagging pragmatics on navigation behavior (Section 5.3) as well as on word sense disambiguation (Section 6.4). In the following, we will describe some pragmatic user measures which help us characterize users either as categorizers or describers, as defined in [122].

3.3.1.1 Categorizers and Describers

The notion of *categorizers* and *describers* was initially presented by Strohmaier et al. in [221] and further elaborated in [123] by introducing and evaluating different measures for tagging motivation. We now introduce each measure, since we will also be using them in Section 6.3.3 and Section 6.4 to measure the input of users on tagging semantics.

Vocabulary Size The first measure is the *vocabulary size* of a user:

$$\text{vocab}(u) := |T_u| \quad (3.33)$$

T_u is the set of tags used by user u and has already been defined in Definition 3. Intuitively, categorizers are characterized by having a smaller, controlled vocabulary, while describers freely assign tags to resources, thus probably possessing a large vocabulary.

3.3 Characterizing User Behavior in the Social Web

Tag/Resource Ratio (trr) The *tag/resource ratio* measures how many unique tags on average were assigned to resources.

$$trr(u) = \frac{|T_u|}{|R_u|} \quad (3.34)$$

This follows the same intuition as the vocabulary size, but might provide a clearer picture of the verbosity of users than just the vocabulary size.

Average Tags per Post (tpp) Similar to the tag/resource ratio, the *average tags per post* measure now measures how many tags were assigned per post. Essentially, this should be the same as the tag resource ratio, except that we now also count multiple tag occurrences, e.g., the same tag on several posts.

$$tpp(u) = \frac{\sum_r |T_{ur}|}{|R_u|} \quad (3.35)$$

This addresses the verbosity nature of describers, who assign lots of tags to any resource.

Orphan Ratio Finally, the *orphan ratio* measures the number of abandoned tags.

$$orphan(u) = \frac{|T_u^o|}{|T_u|}, T_u^o = \{t : \|R(t)\| \leq n\}, n = \left\lceil \frac{|R(t_{max})|}{100} \right\rceil \quad (3.36)$$

Here, t_{max} denotes the tag that was used the most and defines a certain threshold for each users. Tags below this threshold are considered “orphaned” tags, as they are rarely used.

3.3.1.2 Generalists and Specialists

Another interesting distinction between different kinds of users and user behavior is based on the diversity of interests. For example in [220], Stirling proposed a general framework to measure diversity. The author suggests three fundamental properties to measure diversity heuristically: 1. *Variety* targets the number of categories, 2. *Balance* describes the distribution of items and 3. *Disparity* measures the degree of dissimilarity of the elements.

Identifying topical interests is an important task in order to recognize “valuable” users for different tasks. Kang and Lerman [115] for example introduce an automatic approach to distinguish between expert and novice users based on the knowledge structure the users express in social annotation systems. The authors define several features related to the properties introduced by Stirling [220] described above.

In our work, we aim to distinguish between *specialists* who exhibit a narrow topical focus when annotating resources and *generalists* who exhibit an interest in a wide variety of topics. Although there is preliminary research on this distinction, no valid measures for making this distinction automatically are available today. For this reason, we adopt a set of four metrics – motivated by the work of Stirling [220] and others – that capture

3 Foundations

some high-level intuitions about generalists and specialists in social annotation systems in general. We do not require these measures to represent valid measures for these distinctions, because for the anticipated experiments it is sufficient that they capture pragmatic factors. We leave the task of evaluating these measures in - for example - human subject studies to future work.

Mean degree centrality This measure calculates the *mean degree centrality* (based on the tag-tag co-occurrence graph) of all tags in a personomy and is determined by

$$mdc(u) = \frac{\sum_{t \in T_u} deg(t)}{|T_u|}. \quad (3.37)$$

The calculation is based on the degree of a tag measured on the tag-cooccurrence vector space of the folksonomy. The sum of the degrees of all tags is divided by the total number of distinct tags T_u of this user. The intuition behind this measure is that generalists would use more tags that co-occur with many other tags throughout the folksonomy. Hence, generalists would get a high degree centrality whereas specialist would keep this measure low. This metric covers the proposed *disparity* by Stirling [220].

Mean Q1 degree centrality This measure is similar to the mean degree centrality described before. The equation for this metric is

$$mqdc(u) = \frac{\sum_{t \in Q1_u} deg(t)}{|Q1_u|}. \quad (3.38)$$

The difference is that we just consider the most used tags of a user. To do this we calculate the first quartile $Q1_u$ of the tag usage vector (i.e., the number of occurrences of the tags) of a personomy. With this measure we want to remove the long tail of the tag usage vector of a personomy and just focus on the most used tags. Again generalists would acquire a high value whereas for specialists this measure should stay low.

Tag entropy The *tag entropy* characterizes the distribution of tags in a personomy and is defined by

$$ten(u) = - \sum_{i=1}^{|T_u|} p(t_i) \log_2(p(t_i)). \quad (3.39)$$

It can help us to understand user behavior based on tag occurrence distribution. Each tag occurrence count in a personomy is normalized by the total number of occurrences and stored in the probability vector p . A user can either use the tags of her personomy equally often or can focus on some few tags very often. In the first case the tag entropy would be high which would indicate that the user is more of a generalist whereas in the second case the value would be lower and the person would provide more of a specialist behavior.

Similarity score The *similarity score* calculates the average similarity of all tag pairs of a personomy. The formula for this final measure is

$$ssc(u) = \sum_{t_1 \in T_u, t_2 \in T_u, t_1 \neq t_2} \frac{sim(t_1, t_2)}{|T_u| * (|T_u| - 1)} \quad (3.40)$$

The similarity $sim(\cdot, \cdot)$ of the tag pairs is measured by the cosine similarity of the tags' vector representations. A high value would indicate that a person uses many closely related tags and this would display that she focuses just on a topical sub-field of the folksonomy leading to specialist behavior. In the other case the value would be low if a user uses very dissimilar tags and this would describe a typical generalist of such systems. This metric again covers the *disparity* of the tags within a personomy.

3.3.2 Navigation Behavior in Social Media Systems

The analysis of navigation behavior in the web is a wide field of research that has been around as long as the web itself. Next to the analysis of navigation in the entire web graph [150–152], researchers also aimed to understand navigation in closed systems like Wikipedia [245, 247] or social tagging systems [64, 65].

3.3.2.1 Navigation in Wikipedia

Although Wikipedia can be used as a simple source of knowledge to satisfy an imminent information need, its dense and meaningful link structure also invites to explore new things. Sometimes, this serendipitous browsing behavior can lead to fascinating new discoveries.¹⁹ Both the goal-oriented navigation and browsing navigation represent the duality of *navigation* or *searching*, according to Guha, McCool, and Miller [89].

Navigating Wikipedia for fun: The Wikipedia Games A very concrete variant of goal-oriented navigation manifested itself in the *Wikipedia game*, a navigation game based on Wikipedia's link structure, which we will describe in the following. Playing a navigation game on Wikipedia dates back to at least 2004^{20 21} and has been first conceived by students at Amherst College²². In its original form, the main objective was to navigate from a random source page to a random target page, which is picked first. The source page is then determined by following a series of random links starting from the target page. The actual game was then to find the way back using as least clicks as possible. In order to prevent cheating, a number of rules had to be followed. For example, it was only allowed to use links from the page's content. Still, those navigation games were played directly on the Wikipedia web page, so determining a winner or preventing fraud was done manually.

¹⁹This has even been acknowledged in a very prominent web comic: <https://xkcd.com/214/>

²⁰https://en.wikipedia.org/w/index.php?title=Wikipedia:Wiki_Game&offset=&limit=500&action=history

²¹Keep in mind that Wikipedia only exists since 2001

²²https://en.wikipedia.org/w/index.php?title=Wikipedia:Wiki_Game&oldid=7731042

Table 3.2: Comparison of visited page types in game and unconstrained navigation on Wikipedia. We can see that unconstrained navigation, as represented by WikiClickIU and ClickStream, occurs more often on pages about Persons and Movies. In contrast, game navigation from the WikiGame is less focused on pages about these concrete instances.

	WikiGame	WikiClickIU	ClickStream
Person	76	295	496
Movie	56	218	331
Other	858	476	172

Subsequently, two prominent Wikipedia navigation games have emerged that provided appropriate user interfaces to prevent cheating and to compete against players on the whole world. The only purpose of the *WikiGame*²³ was to create a fun game, easily accessible for everyone. In contrast, *Wikispeedia*²⁴ was explicitly created as a “game with a purpose”, to collect navigation data in order to infer semantic relations between Wikipedia articles [249]. The underlying network is from a static corpus designed for schools [249]. The page corpus of WikiGame is not restricted to just a few pages, but uses the whole available Wikipedia, but preprocesses each page before presentation, as mentioned above.

Unconstrained Navigation on Wikipedia In contrast to the goal-oriented navigation of the Wikipedia games, users often also browse randomly across Wikipedia’s link network. Singer et al. conducted a study about the motivations of people to use Wikipedia [209] and found that a large portion of users visited Wikipedia to improve their knowledge about a given topic by discovering related facts. Other users tended to navigate Wikipedia out of boredom, where navigation was less targeted, but rather serendipitous, i.e., about discovering entirely new things. Still, another group of users visited Wikipedia to directly satisfy an information need, i.e., they are looking for a specific concept and want to get to that information as fast as possible without randomly navigating along the Wikipedia graph. Obviously, these types of behavior fundamentally differ from behavior in a game setting. This can also be seen in the types of visited pages. For this, we compared the entities of the top 1000 visited pages in the WikiGame, the ClickStream and the WikiClickIU datasets, which we will still introduce in more detail later in Section 4.2. Table 3.2 shows the distribution of page categories across the three datasets. The majority of pages in WikiClickIU (~53%) can be assigned to persons and movies, as is the case in our baseline, ClickStream (~83%). Opposed to that, the WikiGame top pages mostly focus on other topics (~86%).

²³<http://thewikigame.com>

²⁴<http://www.wikispeedia.net>

3.3.2.2 Navigation Behavior in Social Tagging Systems

While social tagging systems have long been in the focus of research about the emergent, individual and collective processes that can be observed in such systems, there have been only rare attempts to understand the actual *usage and navigation processes*.

In [64], Doerfel et al. analyzed a large dataset of usage logs of BibSonomy, mainly with regard to several assumptions about the way that users collect and share information. However, they also gave a closer look to the retrieval behavior of people in social tagging system, i.e., how they retrieve stored information. While in BibSonomy, most information is shared publicly, Doerfel et al. noticed that the majority of navigation and thus retrieval happens on a user's own pages. While this first systematic analysis of folksonomy usage shed some light on the behavioral properties of navigation, it only focussed on local transitions. In order to obtain new insights about the *global* navigation behavior, i.e., how users actually navigate BibSonomy, we extend upon this work in Section 5.3.

3.3.3 Analyzing User Navigation using Navigational Hypotheses

Understanding human trails on the Web and the underlying generative processes is an open and complex challenge for the research community. Previous work that addresses these problems has been covered in Section 2.3.

A promising approach to generally model human navigation is provided by Markov models, which have been previously used for clickstream data in [36]. Singer et al. noted in [208] that while Markov chains of higher order are too complex and inefficient to actually be useful, the memoryless first-order Markov model does account well for human navigation on the Web. Singer et al. then presented the *HypTrails method* in [207]. Hyptrails is used to compare different hypotheses about user navigation and has been applied in both web [19] and real-life contexts [22].

In this section, we will first give a short introduction to Hyptrails. We explain how we formulate hypotheses about human navigation and then describe how Hyptrails is used to qualitatively compare two hypotheses. After that, we define several standard hypotheses from literature that serve as a basis for the analysis of user behavior in Wikipedia (Section 5.2) and BibSonomy (Section 5.3).

In Chapter 5, we will apply HypTrails on human navigation data from Wikipedia and BibSonomy in order to find a semantic component in navigation on these networks. We will use the standard hypotheses defined below, adaptations thereof to the special structure of the social tagging system BibSonomy as well as new hypotheses.

3.3.3.1 Formulating and Comparing Navigation Hypotheses with Hyptrails

We will now describe how we formulate hypotheses about how users navigate a specific system and how HypTrails is applied to compare the evidence of these hypotheses to the evidence of the actual navigation, using the Bayes factor. The higher the Bayes factor, the more probable is a given hypothesis to have generated the navigational data.

HypTrails utilizes first-order Markov chain models and Bayesian inference for expressing and comparing hypotheses about human navigation behavior. In our case, the states of the Markov Chain are pages in a social media system, while our hypotheses represent ideas about how users navigate these pages. Specifically, hypotheses are formulated as transition probabilities. That is, given a page $p_i \in G$ in a link graph G , we define the probability to choose any other page $p_j \in G$ as $P(p_j|p_i)$.

We use transition functions \bar{P} to represent hypotheses, which can easily be converted into the required probability distributions by normalizing the values for each source state p_i :

$$P(p_j|p_i) = \frac{1}{\sum_{p_k \in G} \bar{P}(p_k|p_i)} \bar{P}(p_j|p_i) \quad (3.41)$$

To obtain insights into the relative plausibility of a set of hypotheses $\mathcal{H} = \{\mathcal{H}_1, \dots, \mathcal{H}_n\}$ given data D , HypTrails resorts to the Bayes factor, which compares the marginal likelihoods

$$P(D|\mathcal{H}_i) = \int P(D|\theta) \cdot P(\theta|\mathcal{H}_i) d\theta, \quad (3.42)$$

also called *evidences*, of the different hypotheses \mathcal{H}_i . Generally, a hypothesis H_i is more “plausible” than another hypothesis H_j if its marginal likelihood $P(D|\mathcal{H}_i)$ is greater than the likelihood $P(D|\mathcal{H}_j)$. According to the reference table in [116], the difference of two hypothesis evidences is decisive, if this difference is greater than 10. Thus, all differences reported in Chapter 5 are decisive. Hypotheses are encoded into the marginal likelihood via the prior $P(\theta|\mathcal{H}_i)$ by eliciting the parameters of a Dirichlet distribution from the corresponding transition probabilities of a hypothesis \mathcal{H}_i and a concentration factor K .

The higher we set K , the more we believe in a given hypothesis. This means that the higher K , the stronger we believe in the actual transition probability distribution specified by the hypothesis. With lower values of K , the Dirichlet prior also assigns probability mass to other probability distributions similar to the original one, thus, we give the hypothesis some “tolerance”. To understand the difference of our hypotheses in detail, we compare hypotheses based on different values of K . We always express evidences on a log scale.

Note that, while HypTrails gives significant results regarding the ordering of the plausibility of different hypotheses, it is not an absolute measure. We can thus only give a qualified statement regarding the *qualitative*, not the *quantitative* difference in performance between the evidence of two hypotheses.

3.3.3.2 Standard Hypotheses

In the following, we define several standard hypotheses for navigation that have been proposed by Becker et al. [20] and Singer et al. [207]. They will serve as a basic characterization of navigation in social media systems in Chapter 5. While Dimitrov et al. also proposed some hypotheses for navigation on Wikipedia [61], we do not use them since a main result of that work is that users preferably click links that they can see.

3.3 Characterizing User Behavior in the Social Web

Self-Loop Hypothesis The assumption here is that users would never leave the page that they are currently on, because they keep refreshing the page. This hypothesis is the absolute baseline, since it does not use the underlying network at all.

$$\bar{P}_{selfloop}(p_i, p_j) = \begin{cases} 1, & \text{if } p_j = p_i \\ 0, & \text{otherwise} \end{cases} \quad (3.43)$$

Uniform Hypothesis This hypothesis expresses the belief that transitions from any page to every other page are equally likely. This enables potential “navigation” to otherwise not directly reachable pages, also called teleportation. The uniform hypothesis also serves as a baseline, since it does not incorporate any knowledge of the network.

$$\bar{P}_{uniform}(p_i, p_j) = 1 \quad (3.44)$$

Structural Hypothesis Similar to the uniform hypothesis, this hypothesis represents the belief that people navigate randomly, but this time they follow the provided link structure, instead of teleporting to any possible page.

$$\bar{P}_{structural}(p_i, p_j) = \begin{cases} 1, & \text{if } (p_i, p_j) \in E \\ 0, & \text{otherwise} \end{cases} \quad (3.45)$$

Semantic Hypothesis Finally, the semantic hypothesis is based on the most amount of knowledge about the system that users navigate on. Not only does it make use of the link structure, but it also takes the *semantic similarity* of pages into account. Each page p_i is then represented by a vector v_i . The semantic hypothesis thus postulates that users on p_i navigate towards pages p_j whose vector representations v_j are similar to v_i .

$$\bar{P}_{tfidf}(p_i, p_j) = \text{cossim}(v_i, v_j) \quad (3.46)$$

The construction process of the vector representation v_i of page p_i is determined specifically for each system.

Due to the usually very high number of states in navigation datasets, we additionally restrict the semantic hypothesis to transitions between linked pages. This means that users on p_i navigate only towards pages p_j , which are linked to p_i , but prefer pages whose vector v_j is more similar to v_i .

$$\bar{P}_{tfidf}(p_i, p_j) = \begin{cases} \text{cossim}(v_i, v_j), & \text{if } (p_i, p_j) \in E \\ 0, & \text{otherwise} \end{cases} \quad (3.47)$$

This way, we have to compute significantly fewer transition probabilities than if we assumed that semantic navigation was independent of the underlying link structure.

The semantic hypothesis will be in the focus of interest in Chapter 5, as it captures the semantic component in human navigation of social media systems.

3.4 Summary and Relations to this Work

The Web is a large fountain of knowledge. Especially with the advent of the Web 2.0 and its possibilities to easily contribute content, the interest rose in tapping into this knowledge. While some of this knowledge is already captured in highly structured databases, such as Knowledge Graphs or WordNet, the overwhelming majority is still implicitly encoded in either slightly or completely unstructured text.

We introduced several potential sources of semantic information with varying amounts of structure in Section 3.1, namely WordNet as a knowledge base, folksonomies such as BibSonomy or Delicious, and Wikipedia as an information network. In order to make this knowledge explicit and exploitable, we described methods to capture semantic relatedness information from text and tagging systems in Section 3.2. Here, we talked about context-based word representation methods which construct sparse, high-dimensional co-occurrence counting matrices as well as word embeddings methods, i.e., which embed words in low-dimensional vector spaces. After discussing the choice of a suitable semantic context to construct word representations, we introduced several similarity measures that we use in this work. Finally we talked about potential evaluation approaches to ensure the quality of the word representations introduced above. Here we emphasized statistical tests to validate evaluation performance improvements. Lastly, Section 3.3 introduced and discussed several ways to measure the influence of human behavior on web data. We concentrated on tagging behavior in folksonomies and on user navigation in the web. For each, we introduced existing approaches, e.g., measures to determine the type of a user determined by her tagging behavior in folksonomies and the HypTrails approach to compare hypotheses about navigation behavior [207].

After having reviewed the methods and approaches presented in this chapter, we are now able to define and present our advances in measuring semantic relatedness from tagging data and navigation in social media systems.

Chapter 4

Datasets

This chapter describes all datasets used in this work. Although we also use WordNet, we already introduced it in Section 3.1.3. While Chapters 5 and 7 mainly focus on the analysis of human navigation on BibSonomy or random walks on the Wikipedia link network, Chapter 8 applies the results from Chapter 7 to learn improved semantic word representations, but also uses folksonomy data and pretrained word embedding datasets. The semantic evaluation is always performed on semantic relatedness datasets as the most direct representation of human intuition.

Consequently, we partition the presentation of the used datasets into four categories: We describe *text*-based datasets, i.e., folksonomy and Wikipedia article data, in Section 4.1. Section 4.2 introduces all *link*-based datasets, such as WikiGame navigation data but also the static BibSonomy link graph. In Section 4.3, we give an overview of the *pretrained word embeddings* that we use in Chapter 8, while Section 4.4 finally describes the *semantic relatedness* datasets that we use throughout this work for evaluation, but also as a learning resource in Chapter 8. Table 4.1 lists all used datasets, their type and the chapter where they are applied.

4.1 Text-based Datasets

The following section describes all *text-based* datasets used in this work. A dataset is text-based, if it is somehow based on any kind of free-form word sequences, be it tagging data from folksonomies or natural language texts as in Wikipedia articles. In the case of folksonomy data, we cover three different datasets with tagging data from three big social tagging systems, namely from Delicious, BibSonomy, and CiteULike. Delicious is focused on information technology and design related fields, while BibSonomy and CiteULike rely on a rather research-oriented audience. For each folksonomy dataset, we provide a short overview of the dataset contents together with some basic statistics.

We also use three Wikipedia-based text datasets, the first two containing Wikipedia article texts, while the third contains a lexicon of word senses, as provided by the Wikipedia disambiguation pages. The Wikipedia article texts are used for the construction of the semantic hypotheses for the WikiGame, Wikispeedia, ClickStream, and WikiClickIU datasets in Section 5.2, while the Wikipedia disambiguation lexicon

Table 4.1: List of all datasets used in this thesis. We also give the dataset type and the chapters where each dataset is used.

Name	Type	Used in chapters
BibSonomy folksonomy	text	5, 6
Delicious folksonomy	text	6, 8
CiteULike folksonomy	text	6
Wikipedia senses	text	6
Wikipedia abstracts	text	5, 7
WikiGame	link	5, 7
Wikispeedia	link	5, 7
WikiLink	link	5, 7
WikiClickIU	link	5, 7
ClickStream	link	5, 7
BibSonomy request logs	link	5, 7
WikiGloVe	embeddings	8
WikiNav	embeddings	8
ConceptNet Numberbatch	embeddings	8
WS-353	relatedness	6, 7, 8
MEN	relatedness	6, 7, 8
Bib100	relatedness	6, 7, 8
MTurk	relatedness	6, 8
SimLex-999	relatedness	6, 8

provides the evaluation basis for the sense discovery process in Section 6.4.

4.1.1 Delicious tagging data

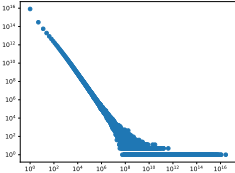
Delicious¹ is a social tagging system, where users can share their webpage bookmarks and freely annotate them with tags. The system has been started in late 2003, pioneering social bookmarking in the web and up until recently, remained one of the most widely used social tagging applications. However, after being repeatedly sold to different companies, the effort spent on maintaining Delicious slowly deteriorated and in the summer of 2017, it was finally shut down. Its userbase was mainly interested in design and computer science topics, as can be seen from the list of the ten mostly used tags. Furthermore, the used tags cover rather “normal” words instead of serving as functional elements of the corresponding system. Still, we analyze the tagging data from Delicious because it has been thoroughly studied in the literature and offers sufficiently meaningful and fine-grained emerged semantics and also covers slightly different and more mainstream topics than the other two tagging datasets that we analyzed. In this

¹<http://www.delicious.com>

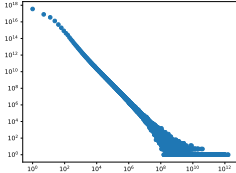
Table 4.2: Statistics about the Delicious dataset.

popular tags	Statistics				
design, blog, tools, webdesign, software, web, reference, programming, video, music	Complete dataset				
	$ Y $	$ U $	$ T $	$ R $	$ P $
	1 026 152 357	1 951 207	14 782 752	118 520 382	339 383 297
	Restricted dataset (top10k most popular tags)				
	$ Y $	$ U $	$ T $	$ R $	$ P $
	813 996 501	1 890 707	10 000	93 842 540	289 573 985

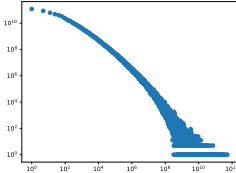
Entity frequency distributions



Tags



Users



Resources

thesis, we use a freely available dataset from 2011, which was provided by [270]. The dump is freely available for download at ². We list some basic statistics about the Delicious tagging data in Table 4.2.

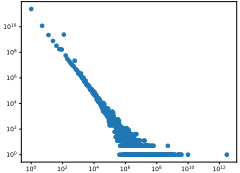
4.1.2 BibSonomy tagging data

The social tagging system BibSonomy provides users with the possibility to collect bookmarks (links to websites) or references to scientific publications and annotate them with tags.³ It was created in 2006 as a practical implementation of the folksonomy model and to explore new research ideas hands-on [27]. Since then, a great ecosystem for integration with Google Scholar, WordPress, Typo3 and other publishing and publication management tools has grown both in possibilities and interest. Because tagging systems have long been in the focus of spammers (cf. [125]), we filtered this dataset to only include data of manually classified non-spammers. The used BibSonomy dump covers all non-spam tagging assignments from 2006 till the end of 2015.⁴ Most notably, Table 4.3 shows that the BibSonomy dataset is the smallest of all the folksonomy datasets, which partially impacts results when using this dataset in experiments [26]. We will however see that BibSonomy folksonomy data still provide a valuable asset to calculate semantic relatedness.

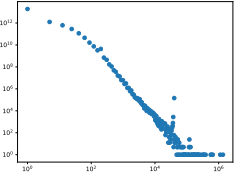
Table 4.3: Statistics about the BibSonomy dataset.

popular tags	Statistics				
imported, software, web, deutschland, programming, blog, tools, myown, research, internet	Complete dataset				
	$ U $	$ T $	$ R $	$ Y $	$ P $
	9 401	186 228	1 204 544	3 529 046	1 304 919
	Restricted dataset (top10k most popular tags)				
	$ U $	$ T $	$ R $	$ Y $	$ P $
	8 632	10 000	1 019 473	2 770 082	1 113 707

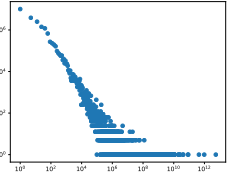
Entity frequency distributions



Tags



Users



Resources

4.1.3 CiteULike tagging data

We took a snapshot of the official CiteULike page from September 2016.⁵ Since CiteULike describes itself as a “free service for managing and discovering scholarly references”, it contains tags mostly centered around research topics. From Table 4.4 it can be seen that it is especially popular with biologists, because tags describing the worm “*Caenorhabditis elegans*” are frequently occurring, although also some functional tags like “no-tag” and “bibtex-import” appear often. While notably bigger in size than the BibSonomy folksonomy dataset, the specialized vocabulary of the CiteULike folksonomy data could possibly impact any kind of semantic relatedness calculations. We will see evidence for this in Section 6.5.

4.1.4 Wikipedia article texts

While we do not directly leverage the content part of Wikipedia in this work to calculate semantic relatedness, we still exploit it for e.g. creating hypotheses about navigation in Wikipedia or to create baselines for semantic relatedness extraction from navigation (see Chapters 5 and 7).

We employ two differently dated snapshots of Wikipedia articles: The first snapshot is from 2011 and corresponds to the version of Wikipedia where most of the games in our WikiGame dataset have been played on. The second snapshot dates to 2015 and corresponds to the links present in the ClickStream dataset, which will be described in

²<http://www.zubiaga.org/datasets/socialbm0311/>

³<http://www.bibsonomy.org>

⁴<http://www.kde.cs.uni-kassel.de/bibsonomy/dumps/>

⁵<http://www.citeulike.org/faq/data.adp>

Table 4.4: Statistics about the CiteULike dataset.

popular tags	Statistics				
no-tag, bibtex-import, qchem, nucl review, humans, elegans, celegans, c_elegans, nematode	Complete dataset				
	$ U $	$ T $	$ R $	$ Y $	$ P $
	151 417	976 799	5 206 525	22 712 888	6 710 449
	Restricted dataset (top10k most popular tags)				
	$ U $	$ T $	$ R $	$ Y $	$ P $
	141 398	10 000	4 548 320	15 988 259	5 780 554

Entity frequency distributions

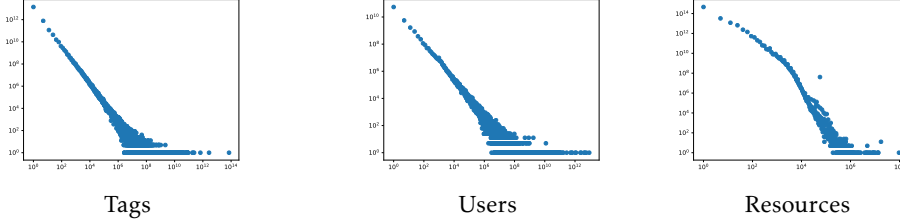


Table 4.5: Properties of the Wikipedia text datasets.

Dataset	Covered Pages/Concepts	Raw Vocabulary size	used in chapter
Wikipedia 2011	3 727 955	2 245 332	5, 7, 6
Wikipedia 2015	5 057 105	6 191 635	5, 7

the next section. Statistics about both datasets are given in Table 4.5.

4.1.5 Wikipedia disambiguation pages

Furthermore, we use the 2011 Wikipedia dataset to collect a large collection of possible word senses for tags in Delicious and BibSonomy by parsing the so-called *disambiguation pages* in Wikipedia. These pages can either be identified by their URL (containing the suffix *_(disambiguation)*), or via their membership to the Wikipedia category of disambiguation pages. Such pages list several possible senses or meanings of a given word, together with a brief sentence that characterizes that sense. A *sense* is one of possibly many different meanings of a word, depending on its context, e.g., *banks* could mean *the banks of a river* or *a number of financial institutes*. For a polysemous term, such a page typically contains an enumeration of its senses in form of a bulleted list, with each item containing a short description of that sense. If no disambiguation page was available, we used the first paragraph of a page as description, because it often describes the “standard meaning” of that term. The dataset at hand contains 9820 word sense lists, where however 6315 only have a single sense. Figure 4.1 shows the distribution of senses in Wikipedia for those 9820 words. Most words have between 10 and 20 different meanings.

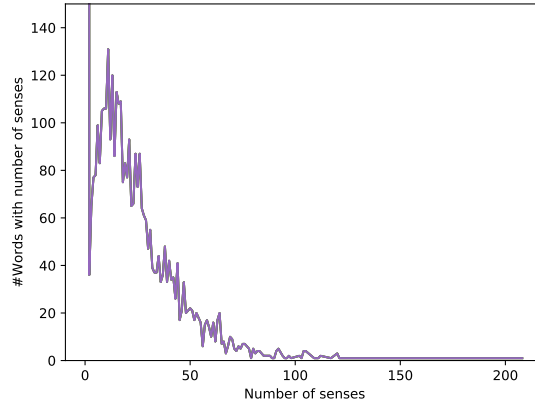


Figure 4.1: Distribution of Wikipedia senses per tag for 9820 tags from Delicious and BibSonomy. While 6315 tags only have a single sense, most words have between 10 and 20 different meanings.

4.2 Link-based Datasets

The following section covers all *link-based* datasets used in this work. We call a dataset link-based, if it does not directly contain any textual information and can directly be expressed as an unweighted or weighted graph. This includes both static link networks as well as transition graphs collected from human navigation on the web.

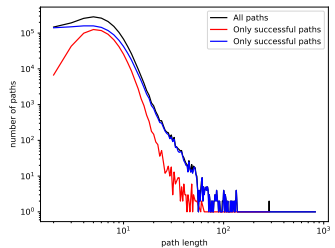
The investigation of these datasets in order to extract semantic relatedness from them is especially intriguing, since (i) we do not have to rely on natural language text anymore to extract semantic information between concepts and (ii) open up another dimension where semantic relatedness information is subconsciously applied by humans.

In this section, we describe several Wikipedia-based link datasets, including four datasets of actual human navigation on Wikipedia, both in a game and an unconstrained setting, and the raw link networks corresponding to the page datasets presented in the previous section. Furthermore, we introduce the BibSonomy folksonomy link network as well as a dataset with human navigation on BibSonomy spanning several years. The datasets introduced in this section are in the central spot of Chapters 5 and 7, where we will first analyse the datasets for a potential semantic component that drives human navigation and then attempt to extract that information using the vector space model introduced in Section 3.2.1.1.

4.2.1 Navigation Games on Wikipedia: WikiGame and Wikispeedia

The first available datasets containing information about navigation on the Wikipedia graph were collected in game settings. The game hoster can collect data in a controlled navigation scenario and does not have to be overly concerned with privacy issues, as users do not follow their own, private interests. Thus, access to such data is easier and less cumbersome. Furthermore, as people have fun playing such a game, they do so

Table 4.6: Characteristics of the WikiGame navigation dataset.

	Pages	Sessions	Users	Paths per user
	360,417	361,115	260,095	6.92
		Paths	Nodes	\emptyset path len.
All	1,799,015	10.758,242	5.98	
Succ.	653,081	4,116,879	6.30	
Unsucc.	1,145,934	6,641,363	5.80	

Most popular pages
 United_States, United_Kingdom, England, Europe, North_America,
 New_York_City, California, France, Japan, U.S._state

repeatedly and finally generate lots of data. In this work, we use two different datasets containing game navigation, namely data from the WikiGame and from Wikispeedia.

4.2.1.1 WikiGame

The WikiGame dataset is based on the online game “*The WikiGame*”⁶. This platform offers users a multiplayer game, where the goal is to navigate from one Wikipedia page (the start page) to another Wikipedia page (the target page) which is linked to the start page through Wikipedia underlying topological link network. The users can leverage most of Wikipedia’s directed link structure to reach their target node. However, the WikiGame removes certain links (such as category links or links to page lists). Other competition modes than randomly generated games are for example such as “5 clicks to Jesus”, where players have to reach the Jesus page in 5 clicks or less.

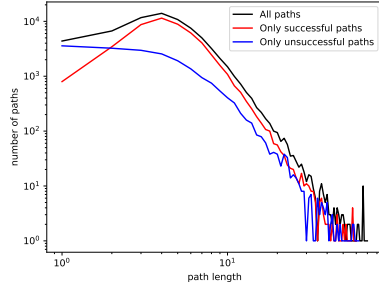
In some cases, the WikiGame data suggest that users follow links in their paths between articles that might not yet exist in Wikipedia’s topological link network. This can happen when, for example, users use the back button in their browser to navigate to a previous article, and the current article does not have a link back to the previous one. We cannot easily determine the page to which the user has backtracked, since the WikiGame data only log the visited page, not where the click originated. One explanation for such behavior could be that users originally end up at a concept they are not happy with and decide that going another route may be a better idea. This is a rich feature of this dataset as it enables us to establish relations between concepts that we normally would not see using Wikipedia’s link network.

Furthermore, we know which paths are *successful* – i.e., the user has reached the target concept – and which are *unsuccessful* – i.e., the user has failed to find a route to the target in the given timeframe. Table 4.6 shows some main characteristics of the WikiGame dataset. The adjusted dataset at hand consists of 1 799 015 navigation paths captured between 2009-02-17 and 2011-09-12. We can see differences in the path length

⁶<http://thewikigame.com/>

4 Datasets

Table 4.7: Characteristics of the Wikispeedia navigation dataset. Please note that the number of users in Wikispeedia and all dependent statistics are based on the number of unique ip address hashes, so the number of actual users might be lower than that.



Sessions	Pages	Users	Paths per user
42,631	4,182	20,868*	3.65*
	Paths	Nodes	\emptyset path len.
All	76,193	442,605	5.81
Succ.	51,318	4,169	6.36
Unsucc.	24,875	4,061	4.68

Most popular pages

United_States, Europe, United_Kingdom, England, Earth, Africa, World_War_II, North_America, Animal, Brain

distribution for successful, unsuccessful paths and all paths, but each distribution exhibits a peak at a length of around six (Table 4.6).

4.2.1.2 Wikispeedia

Wikispeedia is an online game similar in nature to the WikiGame. It has been created by Robert West as part of his research on human navigation in online networks [203, 246, 247], but more specifically to extract semantic relations between words from human navigation [249]. In this game, users are again provided with a start and a target page. Contrarily to the WikiGame, users do not compete with each other to e.g., be the fastest person to reach the goal page, but rather play the game individually. An interesting feature that this dataset possesses is that backclicks are explicitly recorded, in contrast to the WikiGame data, where only the visited pages are recorded. To assimilate the structures of both datasets, we remove the backclick information in the Wikispeedia data and only leave the page visit information. The collected navigation data can be freely downloaded⁷. Compared to the WikiGame dataset, the Wikispeedia dataset is very small in size and restricted to only a very small body of preprocessed articles, based on an excerpt of Wikipedia created for schools⁸. However, it still can serve as a fully-featured dataset containing human navigation in a game setting, and acts another basis for the claims made about constrained vs unconstrained navigation later on in Section 5.2. Again, Table 4.7 shows some main characteristics of the Wikispeedia dataset. The distribution of path lengths is depicted in Table 4.7.

⁷<http://snap.stanford.edu/data/wikispeedia.html>

⁸<http://schools-wikipedia.org/>

Table 4.8: Characteristics of the ClickStream navigation dataset. Please note that ClickStream only consists of transition data, so the path length distribution consists of only one data point.

Pages	2,193,539	Nodes	984,979,972
Paths	12,366,773	\emptyset path len.	2
Most popular pages			
Deaths_in_2015, Fifty_Shades_of_Grey_(film), 87th_Academy_Awards, Bird-man_(film), Islamic_State_of_Iraq_and_the_Levant, Dakota_Johnson, Fifty_Shades_Darker, Jamie_Dornan, Fifty_Shades_Freed, Jane_Wilde_Hawking			

4.2.2 Unconstrained Navigation on Wikipedia: ClickStream and WikiClickIU

While game navigation data are collected in a controlled setting, as mentioned above, real-world navigation does not impose any constraints on users. Consequently, the behavior in unconstrained settings is fundamentally different to that of users in game settings. For example on the public version of Wikipedia, we expect browsing users to rather explore new knowledge and to be interested in concrete information about books or actors rather than pages about common knowledge, such as *tree* or *apple*. Our results about unconstrained navigation are obtained from the analysis of two datasets: ClickStream, which contains navigation collected directly on the Wikipedia page, and WikiClickIU, collected at the University of Indiana.

4.2.2.1 ClickStream

To approximate large-scale, unconstrained human navigation behavior on Wikipedia originating from the whole web, we use a dataset extracted from the Wikipedia web-server logs in February 2015 [255]. This dataset contains an accumulation of transitions with their respective occurrence counts, i.e., how many users used a particular transition between two Wikipedia pages in the whole month. Transitions with less than 10 occurrences have been removed. We only used the transitions with both source and target pages inside the main namespace of Wikipedia. This dataset is by far the biggest dataset with human navigation data that we have access to and spans a large part of the actual Wikipedia link graph (cf. Table 4.10). However, it is also the most anonymized dataset, where we get no information about individual users. With over 984 979 972 requests across 12 366 773 links, originating from 1 383 301 pages to 2 046 154 target pages on 2 193 539 pages overall (as shown in Table 4.8), we receive a very clear representation of at least aggregated human navigation on Wikipedia.

4.2.2.2 WikiClickIU

The WikiClickIU dataset contains navigational information from real-world traffic on Wikipedia, originating from the University of Indiana. It is part of a larger dataset of

Table 4.9: Characteristics of the WikiClickIU navigation dataset.

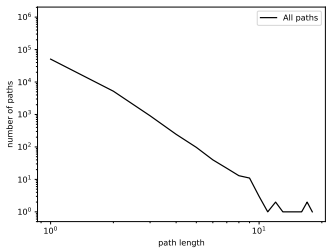
	Pages	810,626
	Paths	1,295,550
	Nodes	2,669,004
	\emptyset path len.	2.06
	Most popular pages Main_Page, Harold_Shipman, Phobia, Tycho_Brahe, QR_Code Archive, Computer_printer, Napster, Doreen_Miller, Baroness_Miller_of_Hendon, 1994_Scotland_RAF_Chinook_crash	

Table 4.10: Size comparison of all Wikipedia link datasets. We report the number unique pages, the number of uniquely visited links, as well as the unique source and target pages of these links.

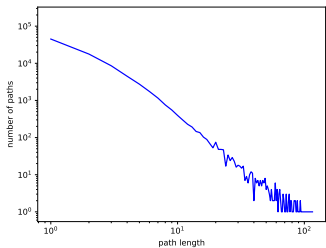
Dataset	Pages	Links	Sources	Targets
WikiLink 2011	3 594 896	208 276 022	3 587 386	3 577 026
WikiLink 2015	4 801 501	315 049 408	4 797 593	4 437 093
ClickStream	2 193 539	12 366 773	1 383 301	2 046 154
WikiClickIU	810 626	910 170	338 552	492 601
WikiGame	360 417	2 266 623	330 693	357 068
Wikispeedia	4 182	58 043	4 177	3 768

about 53.5 billion HTTP requests originating from Indiana University between 2008 and 2010, which was published by Meiss et al. to study structure and dynamics of Web traffic networks [151]. The data have been strongly anonymized, so there is no information about the IP where the request originated or the currently navigating user. However, referer information for each log entry is provided, so we can at least partially reconstruct the subset of the Wikipedia graph actually traversed by users and make out individual sessions. All requests originating from and targeting the Wikipedia domain were extracted, thus about 4 million requests with 1.3 million distinct target Wikipedia content pages were retained. We were furthermore able to reconstruct a set of 1 295 550 navigational paths with an average path length of 2.06, as shown in Table 4.9.

4.2.3 Wikipedia's Static Link Network: WikiLink

The WikiLink datasets represent the plain link network of Wikipedia in two snapshots, corresponding to the Wikipedia page datasets presented in the previous section. The datasets are restricted to only direct links (no redirects) between pages in the main namespace of Wikipedia. Table 4.10 displays basic statistics for both WikiLink datasets

Table 4.11: Characteristics of the BibSonomy navigation dataset.

	Pages	181,974
	Paths	263,373
	Nodes	742,844
	\emptyset path len.	2.82
	Most popular pages	
	/user/acka47,	/user/ls_leimeister, /user/powidl,
	/user/ls_leimeister/itegpub,	/user/hotho,
	/user/gottfriedv,	/user/avs, /user/thorade,
	/user/jaeschke, /user/trude	

in comparison to all other Wikipedia navigation datasets that we introduced until now.

We use two snapshots of the Wikipedia link network. The first WikiLink snapshot is from July 2011. The reason for this choice was that this was the dump closest to the timestamps in the WikiGame dataset that was publicly available⁹. We obtained the present page-to-page network provided by this dump and limited it to links between pages from the main namespace and also to links between the distinct pages available in our WikiGame dataset. The reason for this is that in Section 7.2.2.2, we compare the human navigational paths through the network with the corresponding topological network to see if human influence in the form of navigation impacts the quality of the semantic model upon the link network. Additionally, we use the 2011 WikiLink snapshot to analyze navigation in the WikiGame and WikiClickIU datasets (see Section 5.2). We also used a WikiLink dump from February 2015. This dataset served on the one hand again to provide the basis for a structural analysis of unconstrained navigation for the ClickStream navigation dataset in Section 5.2, but also as a basis to generate random walks across the whole of Wikipedia as well as some subsets of the link network in Section 7.2.4. Finally, to analyze navigation on Wikispeedia, we used the link graph provided by West and Leskovec, which was extracted from the “Wikipedia for schools” project, on which Wikispeedia is based (cf. Section 4.2.1.2).¹⁰

4.2.4 BibSonomy Website Usage

We were able to collect a large dataset of webserver request logs on BibSonomy that we use to shed light how people make use of the system. In this thesis, we analyze the motivations of user navigation behavior in BibSonomy in Section 5.3. Additionally, we attempt to extract semantic information from human navigational paths on BibSonomy in Section 7.3. In the following, we will first describe the *raw log files* of BibSonomy, before provide some information about the navigational paths.

⁹Wikipedia only makes a specific amount of recent dumps available for download

¹⁰<http://snap.stanford.edu/data/wikispeedia.html>

Table 4.12: BibSonomy request log and link network statistics. We report both the statistics for the BibSonomy request logs, collected from 2006 to 2012, and statistics about the static BibSonomy link network that we constructed. We provide information about the unique pages, the number of unique links between pages, as well as the number of unique source and target pages in each dataset.

Dataset	Requests	Pages	Links	Sources	Targets
BibSonomy logs	479 471	181 974	248 299	92 842	154 692
BibSonomy logs, min 1 tag	327 060	150 328	208 929	77 125	128 036
BibSonomy link network	-	5 609 774	69 552 262	5 609 774	5 609 774

4.2.4.1 BibSonomy Request Logs

The BibSonomy log files include all HTTP requests to the website (caching is disabled), including common request attributes like *IP address*, *date*, *referer* and *target*, as well as a *session identifier* and a cookie containing the name of the *logged-in user*.

We first restricted the datasets to data that had been created between the start of BibSonomy in 2006 and the end of 2011, since early in 2012 the login mechanism was modified, which introduced significant changes to the logging infrastructure. Out of the over 2.5 billion requests, we additionally filtered out requests to extra resources including CSS, JavaScript, and image files as well as requests from web bots (using a heuristic comparing user agents to those of known bots in various online databases). Due to its high rank in search engines, BibSonomy is a popular target for spammers. Spammers are users who store links to advertisement sites to increase their visibility on the web. BibSonomy uses a learning classifier [125] as well as manual classification by the system’s administrators to detect spam. In our experiments, we only considered direct (i.e., no redirected) valid requests, which have been generated by logged-in nonspammers. Furthermore, both the referer and the target page of a request must be a retrieval page (see Section 3.1.1.2). Table 3.1 lists all considered retrieval pages. For example, the page `/tag/web` is considered a retrieval page, since it issues an explicit query, namely all resources tagged with `web`. On the other hand, `/settings` is not considered as a retrieval page, because it does not query the content database. Although both the landing page as well as search pages show information from the tag assignment table, they do not count as retrieval pages, because the information there was not explicitly requested, as opposed to tag filtering, user requests or document pages. Finally, we only consider those pages which contained any content, i.e., where any tags were shown. This is not the case if, for example, a tag list is filtered by a non-existent tag. After filtering requests the remaining dataset then retains 327 060 log entries, generated by 17 932 nonspammers on 150 328 unique BibSonomy pages. The statistics can be seen in Table 4.12.

4.2.4.2 Human Navigational Paths on BibSonomy

The log dataset unfortunately only contains information about the particular clicks that a user made on BibSonomy, but no direct information about the complete path that she took. Because we have a lot of information about the navigating user, the requested resources as well as the referring pages of a click, we can partially reconstruct how these users actually *navigated* the system along several clicks.

For this, we first group the set of transitions both by the *session ID* and the navigating *user*. We then sort the transition groups by their *date* and then construct a forest of navigation trees by connecting each transition (p_i, p_j) with another transition (p_j, p_k) , which has both the closest click *date* and where there *target page* of the first transition equals the *referrer page* of the second transition. Finally, we consider each path from the root node of such a navigation tree to each of its leaves as a single navigation path. With this method, we were able to create 263 373 paths with a mean length of 2.82 (see Table 4.11).

4.2.5 The BibSonomy Link Graph

Finally, we describe the BibSonomy link graph that we use in Section 5.3. The BibSonomy link graph denotes the potential links that users can follow when browsing on the BibSonomy website, i.e., its user interface.

Most importantly, we have to note that there exists no *explicit* link graph as in Wikipedia, where the link graph is given by the manually inserted links between different articles. A great portion of the BibSonomy link graph is *implicitly* defined by the folksonomy relations between users, resources and tags, as described in Section 3.1.1.2. However, since the BibSonomy user interface provides links that do not exist as folksonomy relations due to usability reasons, the BibSonomy link graph also contains links between related or co-occurring tags, posts with the same resource from different users, or even from users to similar users. All of these links are generated programmatically. As we are mainly interested in navigation on the folksonomy part of BibSonomy, we restricted link generation to only retrieval pages (cf. Table 3.1). Figure 4.2 shows the links that we generated between the retrieval page types for each tag assignment $(USER, TAG, RES) \in Y$. There also exist dynamically created links, for example, from each page to the user page of the currently logged in user. Because we would have to generate a link graph for each user that we analyze, we ignore them in this work. We also do not consider links from tag pages to accumulated tag pages, such as `/tag/web` to `tag/web+science`, since this would result in a exponentially growing link graph, where almost all transitions would never have been visited. The resulting link graph contains 69 552 262 transitions between

4.3 Pretrained Word Embeddings

In order to prove the universality of the proposed methods in Chapter 8 to learn semantic relations in both supervised and unsupervised settings, we applied these methods both

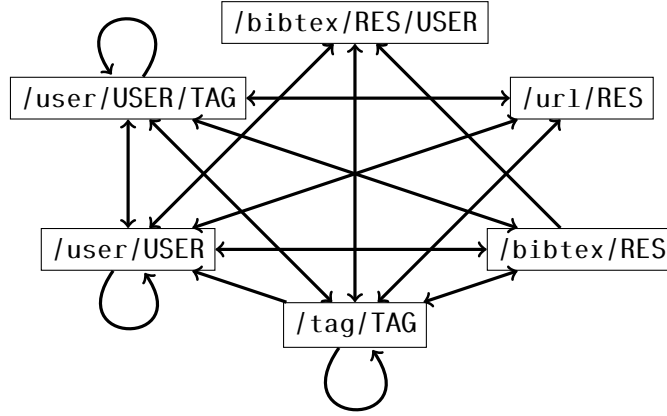


Figure 4.2: The available links between resource pages in the BibSonomy link graph. Loops denote not only that a page can link to itself, but also to other pages of the same type. The BibSonomy link graph contains more edges between different pages than the folksonomy model provides. For example, in a folksonomy, there are no direct links between same entities, which however exist in the BibSonomy link graph.

to our own constructed semantic vector models as well as to pretrained word vectors. This section describes only the pretrained word embedding datasets. We chose the pretrained vector datasets to show an application of the navigational semantics that we investigated in Chapter 7, but also to work with otherwise widely-used datasets, that are also often applied in many NLP tasks. The first two embedding datasets are constructed from Wikipedia article texts and Wikipedia navigation, respectively, while the third set of embeddings has been constructed by combining GloVe vectors with the Google News Word2Vec vectors¹¹ and aligning them with the knowledge base ConceptNet, which has already been mentioned in Section 3.1.4. The algorithms used to create those datasets have already been introduced in Section 3.2.1.2.

4.3.1 WikiGloVe

The authors of the GloVe embedding algorithm [184] trained several datasets of vector embeddings on various text data and made them publicly available.¹² Because it has been demonstrated several times that the textual content of Wikipedia articles can be exploited to calculate semantic relatedness [81, 189], we use the vectors based on Wikipedia as a reference for word embeddings generated from natural language. This dataset consists of 400 000 vectors with dimension 300.

¹¹<https://drive.google.com/file/d/0B7XkCwpI5KDYN1NUTT1SS21pQmM/edit>

¹²<https://nlp.stanford.edu/projects/glove/>

4.3.2 WikiNav

Wulczyn published a set of word embeddings generated from navigation data on the Wikipedia webpage [254] using Word2Vec [156]. Word2Vec was originally intended to be applied on natural language text, though it can also be applied on navigational paths [55]. While technically the generated embeddings represent pages in Wikipedia, most pages also describe a specific concept and we can thus use the generated vectors for both page and semantic word representations. It has been shown that exploiting human navigational paths as a source of semantic relatedness yields meaningful results [175, 210, 249]. The dataset at hand consists of 1 828 514 vectors with 100 dimensions. The vector embeddings have been created from all navigation data in the month of January 2017 and are publicly available.¹³

4.3.3 ConceptNet Numberbatch

Speer, Chin, and Havasi [217] combined sets of GloVe and Word2Vec vector embeddings together with relations from their own semantic network *ConceptNet* using retrofitting [69] to receive vector embeddings which currently pose the state-of-the-art concerning performance in several semantic evaluation settings, such as word relatedness. The produced vector embeddings achieve correlation values of 0.866 with the MEN dataset or 0.828 with the WS-353 dataset. We are using the 16.09 version of the ConceptNet Numberbatch vectors, because these vectors are given as reference in [217], where the authors claim to have posed a new state-of-the-art performance in correlation with human intuition. This dataset consists of 426 572 vectors of size 300. The vectors are freely available for download.¹⁴

4.4 Semantic Relatedness Datasets

In order to achieve a realistic judgment of the quality of semantic relatedness scores of any similarity measure, we need to compare them with human intuition of semantic relatedness. For this, we use datasets with collected scores of human judgment about the semantic relatedness of a set of word pairs, which is the most direct representation of human intuition we can find. These human scores have been gathered by crowdsourcing jobs, i.e., from more or less random people, who are not necessarily linguistic experts. Still, the chosen vocabulary and combination of word pairs have been specifically designed for each crowdsourcing experiment.

In this section, we describe several such datasets with regard to their size, their covered vocabulary and their construction process. Table 4.13 provides an overview over all semantic relatedness datasets.

¹³https://figshare.com/articles/Wikipedia_Vectors/3146878

¹⁴<https://github.com/commonsense/conceptnet-numberbatch/tree/16.09>

Table 4.13: Base statistics for the human intuition similarity datasets. RPP stands for “Ratings per pair” and denotes the number of judgments that each pair received. The final relatedness scores are calculated as the average of the assigned judgments.

Dataset	Pairs	Words	Score Range	RPP
RG65	65	48	0,0–4,0	15/36
MC30	30	39	0,0–4,0	38
WS-353	353	487	0–10	13-16
MTurk	287	499	1–5	23
MEN	3 000	751	0–50	50
SimLex-999	999	1 028	0 – 10	≈50

4.4.1 RG65

The RG65 dataset has been published for a very early judgment of human understanding of semantic *similarity* and context comparison [199].

Origin of Words: The authors proposed a set of 65 word pairs generated from 48 general nouns, which seem to be chosen arbitrarily. It was explicitly stressed that “there was no necessity to study technical vocabulary”, partially, because “using technical vocabulary would raise the practical difficulty of finding enough competent people to serve as judges of synonymy of such words” [199].¹⁵

Data Collection Process: The scores have been collected from two disjunct groups of students. The first group of 15 subjects was asked to judge the degree of “similarity of meaning” of a set of 48 word pairs, later from which 12 pairs were discarded, resulting in 36 word pairs together with human judgment. Each pair has been given a rating between 0,0 and 4,0. The average inter-rater correlation on these 36 pairs amounted to 85%. The second group, consisting of 36 subjects, was then provided with the final 65 word pairs (which also contain the 36 pairs from the first group). Again, each pair has been given a rating between 0,0 and 4,0. This way, the authors were able to ensure that the provided scores were accurate and of a high quality.

Data Quality: The inter-rater correlation between the first and the second group of raters was as high as 99%. Thus, the data can be assumed to be of a sufficient quality to be regarded as a direct and reliable representation of human intuition of semantic similarity. However, the RG65 collection suffers from its small size of only 65 word pairs. Also, since this dataset contains mainly *similar* word pairs, it is not perfectly suited to evaluate semantic *relatedness*, which is a much broader notion of semantic relationship (see Section 3.2.3.1).

¹⁵Technically, this can be seen as an early variant of the knowledge acquisition bottleneck.

4.4.2 MC30

Miller and Charles re-created a part of the RG65 dataset and selected a subset of 30 word pairs consisting of 39 words, taken from the MC30 pairs. The goal was to determine whether students still agree with the semantic similarity scores from MC30. [160]

Origin of Words: The 30 word pairs are to equal parts taken from RG65 as follows: 10 pairs are from the top portion of RG65, 10 pairs from the medium related pairs, and 10 pairs are taken from the least correlated set of pairs in RG65.

Data Collection Process: A group of 38 students from the State University of New York at Oswego were asked to judge the similarity of meaning of each word pair on a 5-point scale from 0 to 4. In order to obtain a notion of different degrees of synonymy, they were shown three characteristic pairs from RG65. Furthermore, the students were free to re-rate the pairs at will.

Data Quality: The reproduced scores exhibit a Pearson correlation of 0.97 and a Spearman rank correlation of 0.95 with the original RG65 scores. From this, it can be concluded that not only the scores' quality is very high, but also that human "estimates remain remarkably stable over more than 25 years" [160].

4.4.3 WS-353

WS-353¹⁶ (WordSimilarity-353) [75] is one of the most widely-used and established datasets of human generated relatedness scores to evaluate semantic relatedness measures. It consists of 353 pairs of English words and names with corresponding relatedness scores between 0 and 10 for each pair.

Origin of Words: Unfortunately, it is not clear how the words were chosen, except that all 30 noun pairs from the MC30 dataset provided by [160] are contained.

Data Collection Process: Two groups of raters were asked to judge the semantic relatedness of all 153 words in this dataset. They had to assign a relatedness value between 0.0 (no relation) and 10.0 (identical meaning), denoting the assumed common sense semantic relatedness between two words. Finally, the total rating per pair was calculated as the mean value of the raters' judgments.

Data Quality: The correlation of scores in WS-353 with RG65 and MC30 is above 0.9, both with the Pearson and Spearman correlation coefficients. Again, the data quality can be judged as very high. The inter-rater correlation of 0.61 however is not as high as with RG65 and MC30.

¹⁶<http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html>

4.4.4 MTurk

The Mechanical Turk dataset is a collection of 287 word pairs and 499 words from New York Times articles [189]. Its main purpose was to provide additional evaluation data in the problem domain of news articles and general vocabulary.

Origin of Words: First, the set of all possible words from the New York Times archive (1863-2004)¹⁷ was intersected with entities from DBPedia. Stop words and words occurring less than 1000 times were also removed. This way, the resulting vocabulary consisted of actual entities instead of verbs or adjectives. Then, the list of possible word pairs is ordered by their point-wise mutual information score. From this list, finally a stratified sample of 280 word pairs was chosen to be judged by humans.

Data Collection Process: From the previously constructed list of 280 word pairs, batches of 50 word pairs were given to workers from Amazon’s Mechanical Turk to rate them on a scale of 1 to 5. Each word pair has been given 10 ratings.

Data Quality: In order to validate the raters, both ten distinct pairs of WS-353 were included in the judgment process, as well as a captcha variant, where the raters were asked to solve a simple math question. Raters with less than 0.5 correlation on the WS-353 pairs or failing the captcha were discarded.

4.4.5 MEN

The MEN Test Collection [43] contains 3 000 word pairs together with human-assigned semantic relatedness judgments, obtained by crowdsourcing using Amazon Mechanical Turk¹⁸. It is thus the largest collection of human judgment about semantic relatedness that we use in our experiments.

Origin of Words: All 3000 word pairs were randomly selected from the WaCkypedia Corpus [17]. Only words that occur more than 700 times are considered. Furthermore, the words have been taken out of a balanced range relatedness levels based on textual semantic scores to avoid picking only unrelated pairs.

Data Collection Process: The relatedness scores have been collected as follows: Raters were given two pairs of words at a time. They were then asked to choose the pair which pair of words was more related. Each pair was rated 50 times, which leads to a score between 0 and 50 for each pair. Each score represents the number of times that a specific word pair has been chosen to be more related than another, unknown word pair. This also implies that no human worker ever gave a judgment about all word pairs.

¹⁷This is only partially available for the public.

¹⁸<http://clic.cimec.unitn.it/~elia.bruni/MEN>

Data Quality: Due to the generation process of this dataset, the inter-rater correlation can not be calculated. However, the authors of [43] additionally judged each of the 3000 word pairs again and achieved a very high correlation with the MEN scores of 0.84, so it is safe to assume that the collected scores provide a valid representation of human judgment. Also, the inter-rater agreement between the authors was $\rho = 0.68$, which is a strong signal on a large number of pairs like in MEN.¹⁹

4.4.6 SimLex-999

The SimLex-999 dataset consists of 999 English word pairs and was specifically designed to contain information about the *semantic similarity* of words [104]. Of those 999 word pairs, 666 are noun pairs, 222 are verb pairs, and 111 finally are pairs of adjectives.

Origin of Words: The 999 word pairs were sampled from the University of South Florida Free Association Database [166]. The authors were especially careful to only sample *similar*, not *associated/related* pairs and also paid attention to include a mix of more concrete and rather abstract words in the final dataset. Finally, they partitioned their dataset according to the Part-of-Speech tags of the words, i.e., if they were nouns, verbs or adjectives (compare above).

Data Collection Process: We only give a short overview of the data collection process, which itself is rather complicated [104]. Hill, Reichart, and Korhonen instructed 500 US-American workers from Mechanical Turk to explicitly judge the *semantic similarity* of all word pairs. Each worker had to rate 119 pairs in groups of 7 pairs on an integer scale of 0 to 6, which they could visualize on a slider. Additionally, they were presented with all word pairs in a group at the same time, so the workers could even compare and later change their judgments to support rating consistency.

Data Quality: To ensure high quality in ratings, the authors inserted checkpoint questions into the rating process to see if the workers “had retained the correct notion of similarity”. The inter-rater agreement was reported as $\rho = 0.67$, which is on a similar level as that of WS-353 with $\rho = 0.61$. In order to validate the instructions given to the raters, Hill, Reichart, and Korhonen then had a sample of 100 pairs from WS-353 judged. Here they found that the human similarity judgments deviated from those in the WS-353 dataset, indicating that the human raters had well understood their instructions to rate *semantic similarity* instead of *relatedness*.

¹⁹The size of the dataset plays an important role in determining statistical significance of correlation scores, see Section 3.2.4.

Chapter 5

Detecting Semantic Influence on Social Media Navigation

5.1 Introduction

By knowing which factors drive human navigation, we can gain further insights into how users make use of information systems. This can subsequently be exploited to e.g., redesign human user interfaces accordingly to support users in finding the desired information [59, 127]. While others have already analyzed human navigation behavior in social media in general [228, 247], we are mainly interested in the *influence of semantic information on navigation behavior*. We assume that if we are able to show the existence of a semantic component in navigation, we can also extract it to gain another source of semantic information. Additionally, since we already know that user pragmatics influence tagging semantics [122], the question arises if users with certain behavioral traits also navigate differently.

In this chapter, we analyze *navigation behaviour* in social media. Our main objective is to determine if semantic information influences navigation behavior in those systems at all and if so, how strong its impact is. Additionally, we are interested in how far users of social tagging systems with certain traits, e.g., their tagging behavior, also exhibit different navigation behavior. In the same way, we assume that navigation on Wikipedia in different settings also shows differences. In order to provide a qualitative judgment for our objectives, we formulate several hypotheses about navigation behavior. For example, we characterize navigation behaviour by random navigation, teleportation, and only navigating on the resources of a single user. We then use the Hyptrails method introduced in Section 3.3.3 to rank those hypotheses by their evidence, which signifies how well a given hypothesis can explain navigation behavior.

We perform the proposed experiments on human navigation trails collected in game and unconstrained settings on Wikipedia (cf. Section 4.2.1 and Section 4.2.2), as well as on the social tagging system BibSonomy (cf. Section 4.2.4). In order to measure the influence of user characteristics on their navigation behavior in BibSonomy, we analyze navigation on navigation data subsets that only contain navigation from those users.

5.2 Wikipedia

In the following, we will analyze different types of navigation on Wikipedia, namely navigation in a game setting (WikiGame and Wikispeedia) and unconstrained, real-world navigation (WikiClickIU and ClickStream). While we mainly seek to achieve a qualitative judgment how strong semantic information impacts navigation behavior of users, we also want to know if users navigate differently in game settings than in unconstrained settings.

Intuitively, we expect that Wikipedia navigation shows a strong influence of semantic information, as pages represent semantic entities, which are linked to other, semantically related entities. We also expect that the goal-oriented navigation in a game setting is influenced by different incentives than unconstrained browsing navigation.

In this section, we first define the hypotheses used to analyze the respective navigation datasets. After this, we first focus separately on analyzing game and unconstrained navigation. Finally, we compare game and unconstrained navigation regarding the strength of the semantic hypothesis.

5.2.1 Hypotheses

Hypotheses about navigation on Wikipedia were mainly presented in three works: Singer et al. analyzed the navigation in the WikiGame in [207], while Becker et al. analyzed Wikispeedia navigation in [20]. Dimitrov et al. examined unconstrained navigation on Wikipedia [61]. However, all of these works used different hypotheses to examine the corresponding navigational data. In this section, we will use a single set of hypotheses which we will test on all datasets. By doing so, we can compare the relative ordering of hypothesis likelihoods with each other and deduce potential differences in navigation behavior between both navigation settings.

We will use the four hypotheses defined in Section 3.3.3.2. Additionally, we define four degree-based hypotheses that are inspired by results from [210]. To mathematically define the hypotheses, we provide their transition functions $\bar{P}(p_i, p_j)$, which are then converted to probability distributions using the notation defined in Equation (3.41). Each hypothesis is also illustrated in Figure 5.1.

5.2.1.1 Standard Hypotheses

As mentioned above, we use several standard hypotheses that we already introduced in Section 3.3.3.2. Those were the self-loop hypothesis (`selfloop`), the uniform hypothesis (`uniform`), the structural hypothesis (`structural`), and the semantic hypothesis (`tfidf`).

As noted in Section 3.3.3.2, the semantic hypothesis depends on vector representations v_i of the semantics of a page p_i to calculate the semantic similarity of pages. We will now define that representation for Wikipedia articles. Generally, we represent a page p_i as a TF-IDF vector v_i^{tfidf} (cf. Equation (3.9)). Following [207], we remove words from the overall vocabulary that occur in more than 80% of all pages in our Wikipedia article corpus and use a sub-linear term frequency scaling in the TF-IDF construction. Finally,

we reduce the dimension of the page vectors v_i^{tfidf} by performing a sparse random projection [136], which preserves Euclidean distance with a guaranteed maximum error, according to the Johnson-Lindenstrauss Lemma [57]. We finally determine the semantic similarity of two pages p_i, p_j by the cosine similarity of their TF-IDF vectors, as shown in the general definition of the semantic hypothesis in Equation (3.47). Figure 5.1d illustrates how the semantic hypothesis assigns weights to page transitions.

5.2.1.2 High and Low Degree Hypotheses

We also introduce new hypotheses to characterize navigation behavior in Wikipedia. In Section 7.2.2.3, we find that a path subset of 30% with mainly low indegree pages yield more useful semantic information than the whole path corpus. Consequently, we formulate hypotheses that assume that users prefer pages that have either *high* or *low in- or outdegree*. This is a more refined assumption than the simple degree hypothesis from [20], where Becker et al. only postulated that users follow high degree pages.

1. The **High Outdegree Hypothesis (outdeg)** assumes that users navigate towards pages with a certain probability relative to their outdegrees.

$$\bar{P}_{outdeg}(p_i, p_j) = \begin{cases} outdeg(p_j), & \text{if } (p_i, p_j) \in E \\ 0, & \text{otherwise} \end{cases} \quad (5.1)$$

The intuition at this point is that users navigate to pages that provide access to a broad range of subsequent pages. This also resembles the intuition from [247] of a “getting-away” or “zoning-out” pattern in the early stages of a path to reach pages that are connected to many regions of the Wikipedia graph.

2. The **High Indegree Hypothesis (indeg)** is similar to the high outdegree hypothesis, but instead assumes navigation towards pages with a high indegree.

$$\bar{P}_{indeg}(p_i, p_j) = \begin{cases} indeg(p_j), & \text{if } (p_i, p_j) \in E \\ 0, & \text{otherwise} \end{cases} \quad (5.2)$$

Mostly, pages with a low indegree cover less general topics and are only relevant for a very small portion of other topics, thus very few pages link to the low indegree page. On the other hand, pages with a high indegree most likely contain broadly relevant content, since many other pages reference to them.

Figure 5.1e provides an illustrative example of the weighting scheme of outdeg for the page Asteroid, which links to Ireland and C.F. Gauss. Under the outdeg hypothesis, we assume that people would navigate to Ireland with a probability of 0.76, since Ireland links to 2307 pages. In contrast, navigation towards C.F. Gauss, would occur only rarely, since the outdegree hypothesis only assigns a probability of 0.01, as the page points only to 86 pages, which pales in relation to Ireland’s 2307 outlinks.

5 Detecting Semantic Influence on Social Media Navigation

For both high degree hypotheses, we also take their counterpart hypotheses into account, where users navigate to the page with a *low* in- or outdegree. With this assumption, we reflect the “homing-in” pattern described in [247], as we assume that pages with fewer in- or outlinks cover rather specific contents.

3. The **Low Outdegree Hypothesis** (`outdeg_low`) can be expressed as follows:

$$\bar{P}_{outdeg_{low}}(p_i, p_j) := \begin{cases} 1 - p_{outdeg}(p_i, p_j) & \text{if } (p_i, p_j) \in E \\ 0 & \text{otherwise} \end{cases} \quad (5.3)$$

Here, $p_{outdeg}(p_i, p_j)$ denotes the transition probability between p_i and p_j as follows from the transition function of high outdegree hypothesis (Equation (5.1)). Intuitively, $\bar{P}_{outdeg_{low}}(p_i, p_j)$ assigns a higher transition probability $p(p_i, p_j)$ to pages with a *low* outdegree.

4. The **Low Indegree Hypothesis** (`indeg_low`) is defined analogously as

$$\bar{P}_{indeg_{low}}(p_i, p_j) := \begin{cases} 1 - p_{indeg}(p_i, p_j) & \text{if } (p_i, p_j) \in E \\ 0 & \text{otherwise} \end{cases} \quad (5.4)$$

5.2.2 Results and Discussion

We evaluate the hypotheses presented above on two types of navigation datasets. The first kind of navigation data has been collected in a game setting, that is, with a pre-determined navigation goal and external incentives, such as winning the navigation game.

5.2.2.1 Game Navigation: WikiGame and Wikispeedia

The analysis of influencing factors, especially a semantic influence, on game navigation serves as a first step towards understanding the influence of semantic information on general human navigation in the web. Previous work by West and Leskovec found that when users in a navigation game navigate towards their target page, they visit pages that are increasingly semantically related to that target page, the nearer the users are to their goals [247]. West, Pineau, and Precup attribute that to the use of an “inner map”, i.e., their knowledge about the semantic connection of Wikipedia pages, that users take advantage of when playing Wikipedia navigation games [249].

Expectations Based on the observations described above, we can expect that users navigate semantically, i.e., their navigation decisions are influenced by their background knowledge about semantic connections between Wikipedia pages. Additionally, we expect that navigation towards highly connected pages also plays a major role in game navigation. We base both assumptions on the analysis by West and Leskovec [247], who provided a first analysis of the navigation behavior in Wikispeedia.

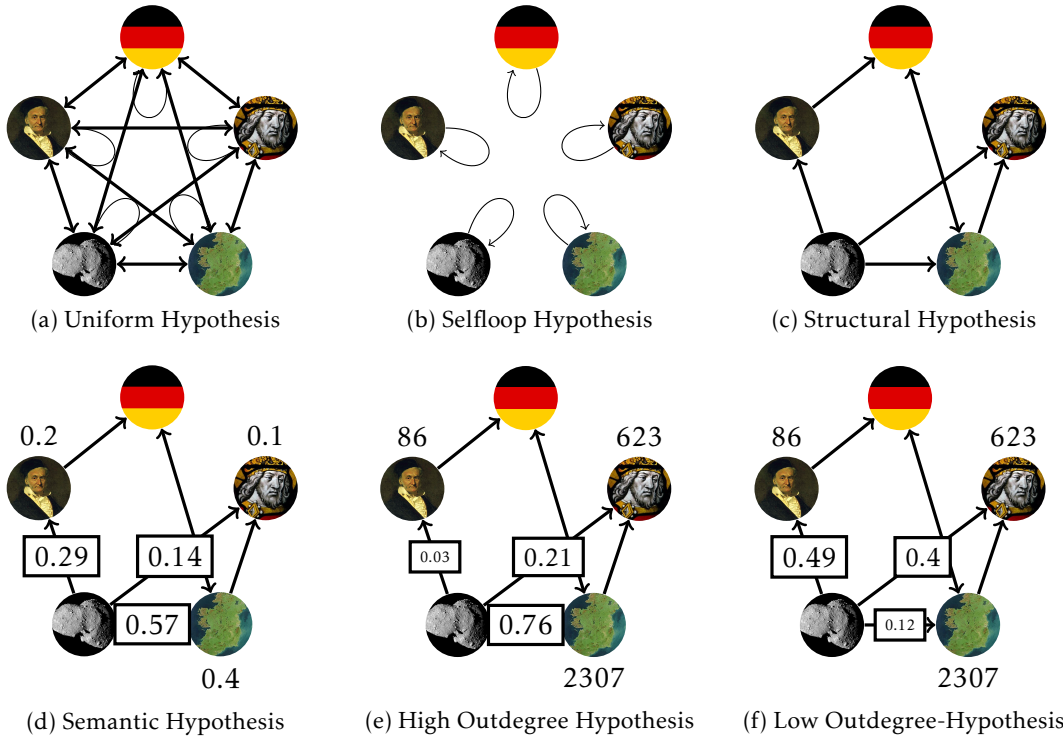


Figure 5.1: Illustrations of the navigation hypotheses on Wikipedia, as defined in Section 5.2.1. A line between two items denotes a hyperlink in Wikipedia. The weight tags on the edges of Figures 5.1d, 5.1e, and 5.1f denote transition probabilities, while the numbers next to the pages in the outdegree-based hypothesis denote the (fictitious) semantic similarities or outdegrees of the respective pages, respectively.

In order to provide a formal characterization of human navigation in game setting, we compute the marginal likelihood of several hypotheses about human navigation (as presented in the preceding section) on both Wikipedia game navigation datasets. Figure 5.2 and Figure 5.3 show the resulting marginal likelihoods for all hypotheses with increasing beliefs.

Results We will first separately interpret the hypothesis results on the WikiGame and on Wikispeedia and then compare the findings on both datasets. Figure 5.2 and Figure 5.3 show the results on logarithmic scales for the WikiGame and Wikispeedia, respectively.

A rather interesting observation is that in the *WikiGame*, the semantic hypothesis (black, solid line, tfidf) does not explain navigation as well as the structure based hypotheses ($\ast\text{deg}$ and structural). This observation is in contrast to the assumption that users in a navigation game would make use of their “inner map” of concepts to

5 Detecting Semantic Influence on Social Media Navigation

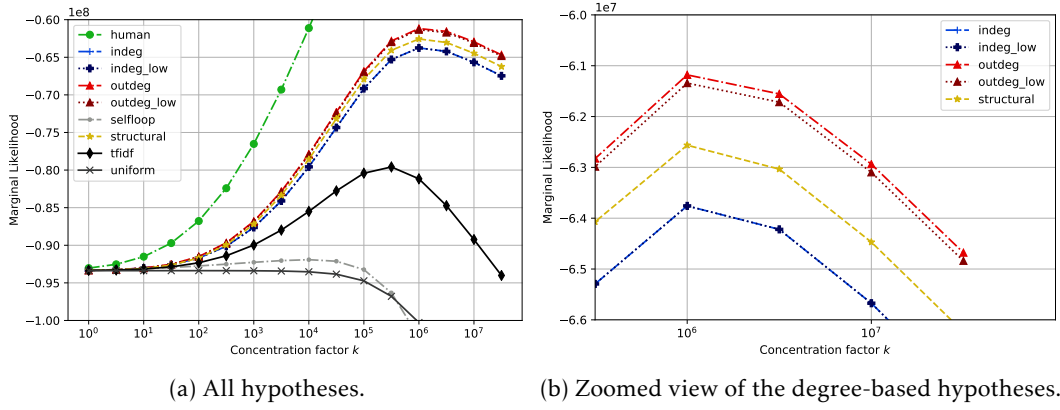


Figure 5.2: Hypothesis-based analysis of game navigation in the WikiGame. Both figures show the marginal likelihoods of the navigational hypotheses on the WikiGame navigation data.

find the target page [249]. We attribute this discrepancy to the fact that users do not necessarily follow the semantically *closest* page, but first of all make use of the link structure to reach their goal. This is especially the case in the first phase of a game, as users tend to quickly navigate towards strongly connected pages [247]: First, users attempt to quickly reach a high outlink page in order to then narrow their search down again for the target page. In the second phase, the so-called “*homing-in*” phase, the page degree decreases as the pages become more specific. At the same time, the users start to select pages according to their semantic relatedness to the target page. This is an intuitive behavior, since most users do not know the Wikipedia link graph structure well enough to directly choose the shortest path to the target page, so they utilize large link hubs (most notably, the “United States of America” page) to have a greater choice of links that potentially bring them nearer to their target. Other potential explanations are that there exist Wikipedia links between semantically unrelated pages or that some links are not semantically obvious. Still, the semantic hypothesis has a higher marginal likelihood to explain navigation in the WikiGame than the two baseline hypotheses (gray lines, *selfloop* and *uniform*). Still, as the evidence of the semantic hypothesis is rising until a certain concentration factor, we can assume that semantic information influences WikiGame navigation, at least to some extent. We can also observe that users tend to navigate more towards high outdegree pages and at the same time low outdegree pages (red lines, *outdeg* and *outdeg_{low}*). While this sounds counterintuitive, it can again be explained because of the two navigation behavior phases in a game setting. However, it seems that structural navigation is generally more important than semantic navigation.

For *Wikispeedia*, we get a very different picture. Here, the semantic hypothesis explains navigation best. We can thus safely assume that navigation behavior in Wikispeedia is heavily influenced by a semantic incentive. We hypothesize that this is due to a largely reduced Wikipedia corpus that provides the basis for Wikispeedia. As this corpus is a

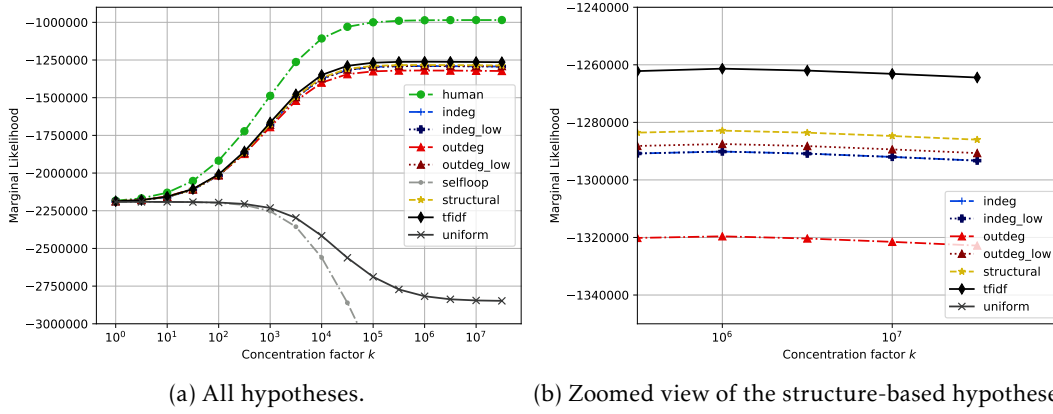


Figure 5.3: Hypothesis-based analysis of game navigation in the Wikispeedia. Both figures show the marginal likelihoods of the navigational hypotheses on the Wikispeedia navigation data.

curated subset of Wikipedia specifically designed for schools, we think that this manual curation effected not only the limited availability of very general real-world pages (like `Table` or `Apple`), but also the number and nature of available links in the dataset. Both the performances of the semantic hypothesis and the low outdegree hypothesis are in line with the observation in [247], that users in their “homing-in phase” increasingly navigate towards semantically more related and at the same time less connected pages. Surprisingly, the “zoning-out” phase seems to be rather weak in Wikispeedia, as the outdegree hypothesis cannot explain navigation as well as the other hypothesis. We can also see that the structural hypothesis (yellow line, `structural`) provides a better explanation than the degree based hypotheses (`indeg` and `outdeg`, both high and low). Again, the low outdegree navigation hypothesis (that is, users navigate towards pages with fewer outlinks) has the highest marginal likelihood of all degree based hypotheses. Finally, all hypotheses exhibit more influence on navigation than the two baseline hypotheses.

In our study, the paths from both WikiGame and Wikispeedia contain transitions that are not represented by actual hyperlinks in the Wikipedia link structure. This is because WikiGame and Wikispeedia only record the succession of actively visited pages, i.e., they leave out the referrer of a request. Additionally, we discarded the backclick information, if it existed (see Section 4.2).

Summary We could show that game navigation is influenced by semantic information. In Wikispeedia, semantic information explains navigation behavior best (Figure 5.3), while in the WikiGame, it still exhibits some influence, but other hypotheses account more for the actual navigation behavior (Figure 5.2). We assume that this is due to the fact that WikiGame contains a lot more data than Wikispeedia. Also, the Wikispeedia dataset is built on a hand-selected collection of Wikipedia pages. Since we know nothing

about its design, we cannot exclude that this influences the semantic navigability of the Wikispeedia graph.

Generally, we can say that for *games on Wikipedia*, navigation according to the low *outdegree* or according to *low indegree* is more important than navigating according to *high indegree*. Considering the two phases in pathfinding proposed by West and Leskovec, this accounts mainly for the “homing-in” phase at the end of game paths.

5.2.2.2 Unconstrained Navigation

By analyzing game navigation on Wikipedia, we have taken a first step towards understanding human navigation behavior in information networks. Still, game navigation is influenced by several external factors. For example Wikipedia pages are rendered differently than on Wikipedia, the underlying Wikipedia corpus is limited to only a subset of pages and the ultimate motivation to click a link is determined by the game setting instead of a potential interest in the contents of a page.

We now want to receive a more realistic impression of how users navigate Wikipedia. For this, we will analyze unconstrained navigation on Wikipedia, that is, real-world navigation on the public Wikipedia without any external constraints. As opposed to game navigation data, information about unconstrained user navigation is more difficult to access or obtain and due to privacy reasons is often accumulated and heavily anonymized.

In this section, we will analyze two navigation datasets with unconstrained navigation. The first dataset contains Wikipedia traffic that originates in the University of Indiana network and is called *WikiClickIU*. It is introduced in detail in Section 4.2.2.2. The second dataset, which is named *ClickStream*, contains accumulated transition data between Wikipedia pages collected on the public Wikipedia servers and has been described in Section 4.2.2.1. Both datasets do not contain any information about the navigating users. For our analyses, we will again use the same hypotheses defined in Section 5.2.1.

Expectations As mentioned before, the paths from the WikiGame and Wikispeedia contain transitions that are not represented by actual hyperlinks in the Wikipedia link structure. For the unconstrained navigation data, we do not have such artifacts, since the unconstrained log data always record both the page where a request originated and where it led. Even if users would return to a page, a new click would then again record the origin of that click. In contrast, the game navigation log data do not record the origin of a click, but only the visited page. Logically, this then induces “wrong” links in game settings when using the back button. We thus expect that the *structural* hypothesis performs much better on unconstrained navigation than on game navigation. Concerning the semantic navigation hypothesis, we expect it to perform better than in a game setting, since users are now free to navigate according to their own interests, not according to a predefined goal page as in a game setting. Finally, Dimitrov et al. found that users on the public Wikipedia move more towards the periphery of the Wikipedia graph, i.e., less connected pages [61]. Consequentially, we also expect the low degree hypotheses to exhibit a higher marginal likelihood than their high degree counterparts.

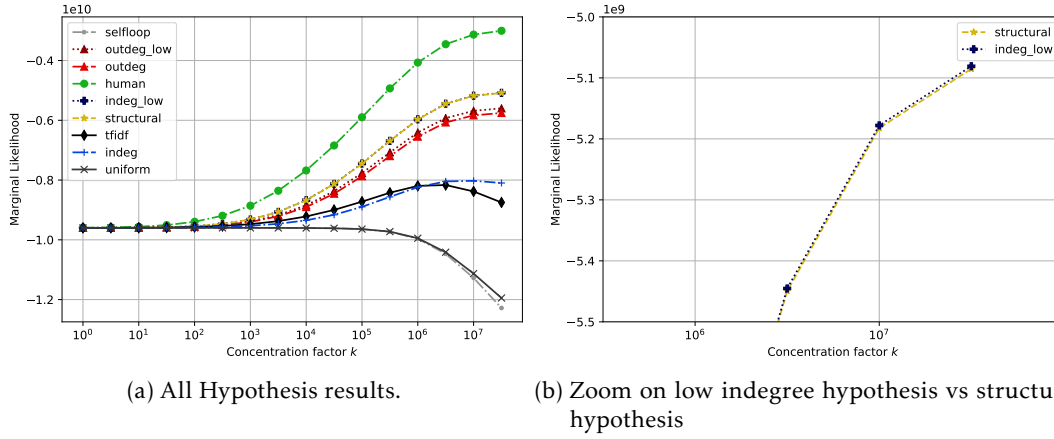


Figure 5.4: Hypothesis-based analysis of unconstrained navigation on the ClickStream data.

Results We will first focus on the results obtained on the *ClickStream* data, which are shown in Figure 5.4. As can be seen, all hypotheses explain navigation better than the two baselines (`selfloop` and `uniform`, grey and darkgrey lines), which indicates that all hypotheses are somehow important. The marginal likelihood of the semantic hypothesis (`tfidf`, black line) shows that real-world navigation on the ClickStream data is to some extent influenced by semantic information. Compared to the indegree hypothesis (`indeg`, blue line), it explains navigation better, but only with a lower concentration factor. As expected, the structural hypothesis has a higher marginal likelihood than almost all other navigation hypotheses. The only exception is the low indegree hypothesis (`indeg_low`, dark blue line), i.e., users often navigate towards pages with a lower indegree than the current page. Figure 5.5b provides a zoomed view of the structural and `index_low` hypotheses so this can be seen better. The `index_low` hypothesis has a slightly higher marginal likelihood than the `structural` hypothesis and thus explains navigation in the ClickStream dataset a bit better. In this light, it is especially interesting that the `indeg` hypothesis, i.e., that users navigate to pages with a high indegree, explains navigation significantly worse than the `index_low` hypothesis. A possible interpretation of this could be that users tend to navigate to more specific instead of more general concepts. A similar observation can be made when comparing results of the `outdeg` and `outdeg_low` hypotheses: Users would rather choose pages with a lower outdegree, i.e., more specific pages, than pages with a high outdegree. Possibly, users do not have to navigate to more general pages as much as they do in a game setting, so they visit connected pages which cover related aspects in more detail, but are consequently less connected.

On the *WikiClickIU* data, we can again see in Figure 5.5 that the structural hypothesis (`structural`, yellow line) is the most influential factor on navigation amongst those that we analyze. With a lower concentration factor k , the semantic hypothesis (`tfidf`, black line) is a close follower, but with increasing k is overtaken by both outdegree

5 Detecting Semantic Influence on Social Media Navigation

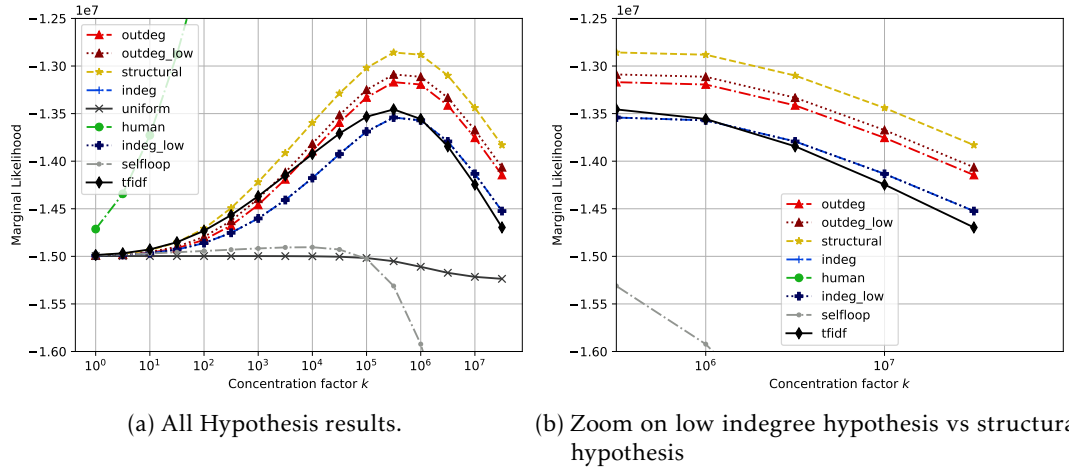


Figure 5.5: Hypothesis-based analysis of unconstrained navigation on the WikiClickIU data.

hypotheses (outdeg and outdeg_low, red and dark red lines). Still, this again shows that semantic information influences real-world navigation, even when restricted to a closed set of users, namely those in the university network of the University of Indiana. Additionally, the semantic hypothesis exhibits a higher marginal likelihood than both indegree hypotheses (indeg and indeg_low, blue and dark blue lines). A potential explanation here is that users at the University of Indiana use Wikipedia as a source of background knowledge concerning only specific topics, e.g., about the state and the history of Indiana. The performance of the outdegree hypothesis (outdeg) compared to the indegree hypothesis (indeg) supports this assumption, as users would rather visit pages with a high outdegree, i.e., a hub page, instead of authority pages with high indegrees. Again we can observe that users also navigate towards the periphery of the Wikipedia graph, since for both the outdegree and indegree hypotheses, it is more likely that users navigate towards low outdegree and low indegree pages instead of highly connected pages (indeg_low, outdeg_low).

Summary On both navigation datasets, the performance of the semantic hypothesis shows that navigation in the real world is influenced by semantic information. This is more obvious on WikiClickIU than on the ClickStream dataset, possibly also because of the different scope of WikiClickIU: WikiClickIU is restricted to users in the University of Indiana’s network. Because of this, we can observe that many of the visited pages have a direct or indirect connection to the university. On the other hand, ClickStream captures navigation from all users that access the public Wikipedia that probably use Wikipedia rather as a one-time retrieval knowledge platform, that is, they access a single article that satisfies their information needs. As already noted by Dimitrov et al. [61], users tend to navigate more towards the periphery of Wikipedia, because the low

degree hypotheses show a higher marginal likelihood than their respective high degree counterparts. Lastly, the structural hypothesis is mostly the best performing hypothesis, as expected.

5.2.3 Conclusion

In this section, we analyzed four datasets with navigation behavior on Wikipedia. Two datasets were collected in a game setting and thus imposed constraints on the possible navigation, namely the WikiGame and the Wikispeedia datasets. Examples of such constraints are restrictions on the possible link network or the exclusion of content from Wikipedia pages that led to pages that are “forbidden” in the respective game. The other two datasets, ClickStream and WikiClickIU contained real-world navigation, i.e., without any external constraints. This means that users were free to navigate to wherever they want on Wikipedia.

We found that *semantic information exerted a notable influence on navigation behavior both in game and unconstrained settings*. Still, it is not the most defining factor in explaining human navigation on Wikipedia, except on Wikispeedia, but also plays a strong role in WikiClickIU navigation. We assumed that on those smaller datasets, the semantic navigation information is more prominent than on the bigger datasets, where it is simply drowned out by other types of navigational behavior.

Across all datasets, we could observe that navigation according to structure is often a good explanation for navigation. This is especially interesting, since in both game datasets, some users also transitioned between pages that were not interlinked. It also became clear that *users in an unconstrained navigation setting are less inclined to navigate towards general pages than in a game setting*, since possibly they are interested in more detailed and related information to their current page, which is found more towards the periphery of the network. This is supported by the larger marginal likelihood of the low degree hypotheses.

Finally, we could again show that *navigation in Wikipedia is oriented towards the periphery of the Wikipedia graph*, i.e., users navigate towards less connected pages. While this has already been noted in [61] for unconstrained navigation in the ClickStream dataset, we showed this for another dataset of unconstrained navigation (WikiClickIU) and on the game navigation datasets (WikiGame and Wikispeedia).

5.3 BibSonomy

In the previous section, we analysed human navigation behavior on Wikipedia. While we gained interesting insights on how humans navigate in game and in unconstrained scenarios and how this navigation is influenced by semantic information, we restricted ourselves only to a single webpage system. In this section, we extend our analysis onto the social tagging system BibSonomy. As mentioned in Chapter 4, BibSonomy is a social tagging system and thus somewhat different than Wikipedia, both in usability and user behaviour. In Wikipedia, users contributed content in the form of encyclopaedic articles. BibSonomy on the other hand stores personalized bookmarks and publications for users,

that annotate those resources with tags. The emerging structure (as introduced in Section 3.1.1) is called a *folksonomy* and serves as the main navigational concept in social tagging systems, providing links between co-occurring entities. Through those links, folksonomies possess an inherently semantic nature. We thus assume that navigation in folksonomies is influenced by such semantic information. Also, studying the general navigation behavior of users in social tagging systems is of great interest. Consequently, different studies have addressed this issue: Heckner, Heilemann, and Wolff conducted a user survey on usage motivation in social tagging systems [100] and Doerfel et al. used log files to study actual navigation behavior of the overall user population through request counts [64].

To characterize navigation on BibSonomy, we will formulate explicit navigation hypotheses that are inspired by the hypotheses in Section 5.2.1. However, we adapted these hypotheses on the unique features of social tagging systems that we described in Section 3.1.1. We also study how the performance of explaining navigation behavior differs on different data subsets, such as navigation grouped by gender, tagging behavior, or long-term experience. In the process, we revisit the aspects described in [64], and extend on their work, providing additional explanations for user intentions during navigation and their comparison. We compare these hypotheses on different subsets of BibSonomy navigation, gathered over the course of 6 years (see Section 4.2 for a description of the dataset). Again, we discuss the influence of semantic information on navigation, but this time, we also include information about the tagging behavior of users. By this, we aim to show that users with different tagging behavior also exhibit different navigation behavior. If that is the case, it supports our assumption that navigation is influenced by semantic information. The content of this section has previously been published in [173].

5.3.1 Hypotheses

Because the website structure of BibSonomy differs to that of Wikipedia, we adapt the hypotheses that we applied on Wikipedia (see Section 5.2.1) to the unique features of social tagging systems. For example, we reformulate the structural hypothesis on Wikipedia to the folksonomy hypothesis for BibSonomy, i.e., we assume that users use the folksonomy structure to navigate the BibSonomy website. Each of these hypotheses represents a basic aspect of navigation in BibSonomy, although they can also be directly applied to model any kind of social tagging system. In this section, we first formulate the adapted hypotheses and additionally describe combinations of those hypotheses. Again, we use the hypothesis notation introduced in Section 3.3.3. The pages are characterized as defined in Section 3.1.1.

5.3.1.1 Standard Hypotheses

We can use some of the standard hypotheses defined in Section 3.3.3.2 as they are and apply them on BibSonomy. The uniform hypothesis (*uniform*) is trivially defined the same way as in Equation (3.44). In this section, we use the page consistent hypothesis

(page), which is equivalent to the self-loop hypothesis from Equation (3.43). As with Wikipedia, we have to define how we represent a page p_i as a vector v_i so we can compute the semantic navigation hypothesis (tfidf). Based on the page's tag cloud, i.e., the set of tags which appear on that page with respective frequencies (see Section 4.2), we again define its representation as a TF-IDF vector v_i^{tfidf} . Here, a page is treated as a document, while its tag cloud is treated as the document's "text". Finally, page similarities are again computed with the cosine similarity measure.

5.3.1.2 Social Tagging System Specific Hypotheses

We also introduce two new hypotheses that make use of the folksonomic structure of social tagging systems, namely the user consistent and category consistent hypotheses.

Category Consistent Hypothesis (cat) Doerfel et al. found that transitions between two pages often occur between pages of the same category, i.e., after a user has visited a *tag* page, the next page is likely to be a *tag* page again [64]. The same holds for *resource* and for *user* pages. The classification of page URLs in one of these categories is given in Table 3.1. The hypothesis therefore states that users will navigate towards pages with the same category, which is a generalization of the previous hypothesis. Figure 5.6d shows a sketch of this hypothesis. Its formal definition is as follows:

$$\bar{P}_{cat}(p_i, p_j) = \begin{cases} 1, & \text{if } cat(p_j) = cat(p_i) \\ 0, & \text{otherwise} \end{cases} \quad (5.5)$$

User Consistent Hypothesis (user) Similar to the category consistent hypothesis, the user hypothesis assumes that in a transition, the target and the source page belong to the same user. Figure 5.6e illustrates the assumed behavior in this hypothesis. The motivating intuition for this hypothesis is that visitors, who are interested in the work of a specific user, will not only read one, but several of her articles and try to further explore her personomy (cf. Section 3.1.1). It is defined as follows:

$$\bar{P}_{user}(p_i, p_j) = \begin{cases} 1, & \text{if } user(p_j) = user(p_i) \\ 0, & \text{otherwise} \end{cases}, \quad (5.6)$$

Folksonomy Hypothesis (folk) As mentioned in Section 3.1.1, the link network of the user interface of BibSonomy is heavily determined by the underlying folksonomy. Each tag-assignment (u, t, r) links the user u with the tag t , the tag t with the resource r and the resource r with the user u and vice versa, respectively. In the *folksonomy hypothesis* (folk), we assume that users follow links that are provided by the folksonomy structure:

$$\bar{P}_{folk}(p_i, p_j) = \begin{cases} 1, & \text{if } p_j \text{ is directly reachable in} \\ & \text{the folksonomy from } p_i \\ 0, & \text{otherwise} \end{cases} \quad (5.7)$$

The folksonomy consistent hypothesis (*folk*) corresponds loosely to the structural hypothesis, which is defined in Equation (3.45). However, the user interface of BibSonomy provides connections to pages that are not covered in the folksonomy to improve navigability. For example, it is possible to navigate from a tag page to another tag page, although folksonomies do not allow direct links between those entities. Also, from each page it is possible to directly navigate to one's own user page, which is obviously not possible in a folksonomy.

We allow only tag-tag connections as the single exception in this hypothesis to deviate from the folksonomy structure. To calculate reachability we construct the page graph from the tag-assignments in the folksonomy dataset described in Section 4.1 and (since they are an integral part of the BibSonomy user interface) we add tag-tag relations, when tags occur at the same posting. Usually, these connections are normally not part of a folksonomy.

5.3.1.3 Combined Hypotheses

In order to investigate possible mutual influences between hypotheses, it is also possible to combine them. We choose a multiplicative combination, which corresponds with the assumption that *both* combined hypotheses must hold. In contrast, an addition would mean that *at least one* hypothesis holds. As we however want to explore the effects of two hypotheses at the same time, we have to multiply two hypotheses. In the following, we motivate and describe certain combinations.

Folksonomy Consistent & Semantic Navigation Hypothesis As described earlier, it is a natural assumption that users utilize the folksonomy structure when navigating a social bookmarking system. If the folksonomy does indeed exhibit notable semantic properties, we should be able to see that adding a semantic component to folksonomic navigation improves the evidence of this hypothesis compared to the bare folksonomy navigation hypothesis. We define the hypothesis as follows:

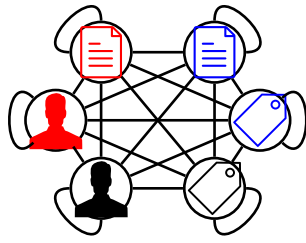
$$\bar{P}_{folk-tfidf}(p_i, p_j) := \bar{P}_{folk}(p_i, p_j) \cdot \bar{P}_{tfidf}(p_i, p_j) \quad (5.8)$$

User Consistent & Semantic Navigation Hypothesis A similar motivation as with folksonomic and semantic navigation arises when we combine user consistent and semantic navigation. Users are normally thematically restricted in their research interests. Navigation in the user's personomy, i.e., all the posts from a user, is expected to show a strong semantic component. This hypothesis is defined as:

$$\bar{P}_{user-tfidf}(p_i, p_j) := \bar{P}_{user}(p_i, p_j) \cdot \bar{P}_{tfidf}(p_i, p_j) \quad (5.9)$$

User Consistent & Folksonomy Navigation Hypothesis The intuition behind combining user consistent and folksonomic navigation is that if users mostly navigate on their own pages [64], we expect them to navigate there according to the folksonomy structure to find their way in their own resources. We define this hypothesis as:

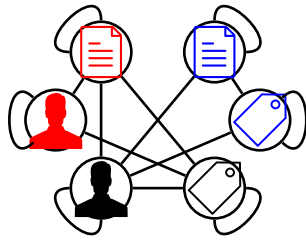
$$\bar{P}_{folk-user}(p_i, p_j) := \bar{P}_{user}(p_i, p_j) \cdot \bar{P}_{folk}(p_i, p_j) \quad (5.10)$$



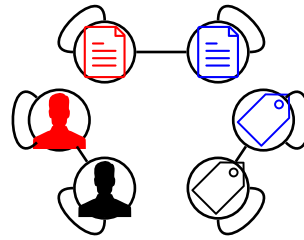
(a) Uniform Hypothesis



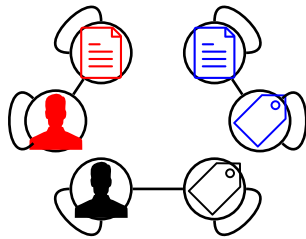
(b) Page Consistent Hypothesis



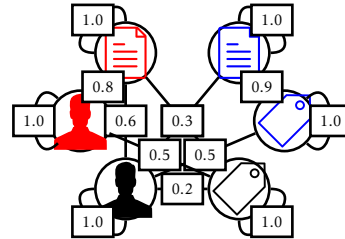
(c) Folksonomy Hypothesis



(d) Category Consistent Hypothesis



(e) User Consistent Hypothesis



(f) Semantic Hypothesis

Figure 5.6: Illustrations of the navigation hypotheses defined in Section 5.3.1. Same colored items depict pages that belong to the same user, while same item symbols denote page categories. A line between two items denotes a hyperlink in BibSonomy. The weight tags in Figure 5.6f denote the cosine similarity of the connected pages.

Table 5.1: Transition statistics of different BibSonomy request log subsets, with respect to user features. We provide the number of unique source states, the number of unique transitions, the amount of all relevant transitions and the subsection where the datasets are used.

subset	source states	links	transition counts	used in
overall	55,129	149,542	327,060	Section 5.3.2.1
inside	37,244	105,222	261,300	Section 5.3.2.2
outside	14,757	28,760	42,193	
male	23,090	61,616	130,988	Section 5.3.2.3
female	5,598	14,413	29,705	
neutral	28,726	73,575	161,830	
shortterm	10,285	21,912	48,221	Section 5.3.2.4
longterm	45,535	126,453	274,302	
lower_trr	30,368	83,268	176,755	Section 5.3.2.5
upper_trr	7,084	15,474	32,517	
lower_ten	3,459	6,959	15,451	
upper_ten	51,542	140,844	307,072	

5.3.2 Results and Discussion

We now evaluate the previously proposed hypotheses on the BibSonomy navigation data. First, the hypotheses are compared on the overall request log dataset. After that, several subsets of the request logs, filtered according to certain user types, are analyzed with the same hypotheses. We expect that there are subsets of the data where some hypotheses perform differently than on the overall dataset. For example, we assume that users with a *categorizing* tagging behavior also navigate BibSonomy in a different way to *describers* (for an explanation of both user types, see Section 3.3). Thus, we investigate different data subsets as listed in Table 5.1. We describe each subset in the corresponding experiment section.¹

5.3.2.1 Overall Request Log Dataset

All of the basic hypotheses explain the observed transitions better than the baseline (*uniform*) to varying degrees. Figure 5.7 indicates that they all introduce at least some structural properties which help to explain the observed transitions.

Besides this fact, there is a clear order of hypotheses: the user consistent hypothesis works best, the semantic and the folksonomy hypotheses are somewhat similarly plausi-

¹Note that while some data subsets are rather small, we still get statistically significant results with regard to the relative ordering of the studied hypotheses. This has been explained in detail in Section 3.3.3.

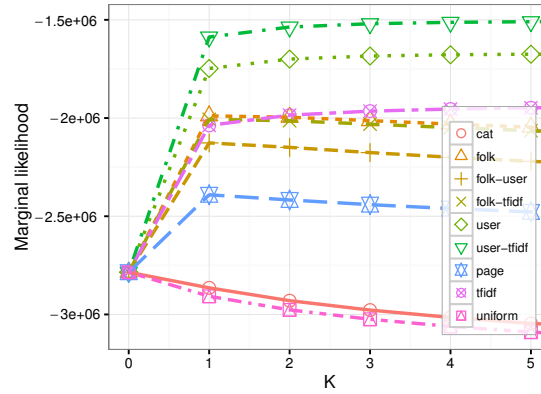


Figure 5.7: Evidence chart for the navigational hypotheses hypotheses on the complete request log dataset. The values of K denote the logarithm to base 10 of the concentration factors, i.e., the results at $K = 3$ were computed using a concentration factor of 10^3 . Of the basic hypotheses, the user consistent hypothesis explains the data best, followed by the semantic and the folksonomy hypotheses. When combining the user consistent hypothesis with a semantic bias, the evidence improves. This indicates that users are actually semantically biased while navigating through resources. In contrast, combining other hypotheses with the folksonomy hypothesis does not yield better explanations for the observed navigation.

ble, followed by the page consistent and the category consistent hypotheses. Many of the observed effects are explainable by the large number of self-transitions in the dataset caused, for example, by pagination (cf. Table 5.2):

1. The page consistent hypothesis strongly improves on the uniform hypothesis.
2. The category consistent hypothesis is more plausible than the the uniform hypothesis, even though it directly contradicts navigation as induced by a folksonomy structure.
3. The user consistent hypothesis as well as semantically induced hypotheses are strongly favored because of self-transitions, which account for roughly 40% of all transitions.

Nevertheless, the user consistent as well as semantically induced hypotheses are also more plausible than the page consistent hypothesis, indicating that their structural properties cover further important factors. That is, the superiority of the user consistent hypothesis indicates that users indeed navigate mostly on their own resources (cf. Table 5.1). In contrast, the valid assumption that this could also be caused by users, who consistently navigate on content from other users, directly contradicts the results from [64], where Doerfel et al. found that users mostly navigate on their own pages. This can also be seen in Table 5.1: navigation inside one's own resources accounts for

Table 5.2: Selected request log statistics. Self transitions are transitions from one page to itself and are mostly induced by pagination effects. Own-transitions are transitions, where the logged in user owns both the referer and the target page in a transition.

Distinct visited pages	103,415
All transitions	327,060
self-transitions	123,452
own-transitions	261,300

roughly 80% of all considered transitions. In fact, we found even another proof for that assumption. The good performance of the semantic hypotheses indicates that semantic similarity of pages (with regard to tags) is a strong explaining factor for navigation on our dataset.

Finally, we consider the *folksonomy hypothesis* which models the navigation we expect in a folksonomy (see Figure 5.6c). It performs similarly well as the semantic hypotheses. We observe that the corresponding evidence curve crosses the semantic hypothesis (TF-IDF) for increasing believe factors K . This indicates that the folksonomy hypothesis covers an important factor of the navigation, but fails to model certain transitions, which are covered by the semantic hypothesis. The fact that the folksonomy hypothesis cannot cover certain transitions is due to navigation outside the folksonomy structure as elaborated in Section 4.2.

When analyzing the *combined hypotheses*, we see that overall, the combination of the user consistent and the semantic hypotheses performs best, indicating that navigation on BibSonomy can mainly be explained by semantic navigation within one’s own resources. Although again this could mean that navigation can also occur on the page subset of another user, we can safely assume that this holds mostly for navigation in one’s own resources. In contrast, combining the folksonomy hypothesis with the semantic hypothesis decreases the observed evidence slightly. Also, combining the folksonomy hypothesis with the user consistent hypothesis decreases the observed evidence dramatically. Both observations indicate that users excessively take advantage of additional navigation features provided by BibSonomy (see Section 3.1.1) when navigating on their own resources.

5.3.2.2 Outside Navigation

Because the results for inside navigation (the source and target state belong to the navigating user) hardly differ from the results on the overall dataset (see Figure 5.7), we only show and discuss the results for outside navigation.

Motivated by the fact that users often navigate on their own pages [64], we investigate whether users behave differently when they are browsing the folksonomy *outside* of their own resources. In particular, we study the transitions where the source as well as the target state do not belong to the browsing user. The results can be seen in Figure 5.8. Here,

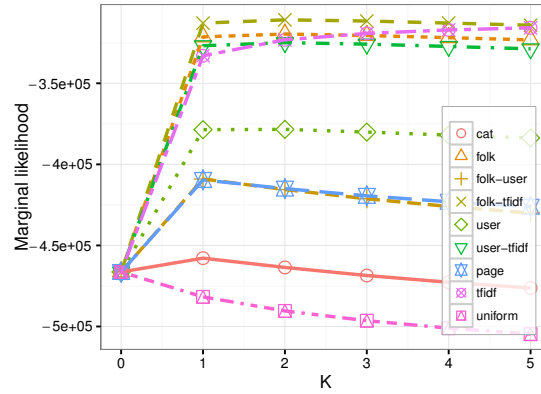


Figure 5.8: Evidence curves for navigation outside of the user’s resources. In contrast to the overall dataset we observe that outside navigation can be explained best by a hypothesis assuming semantic behavior on the folksonomy structure (cf. *folk-tfidf*).

the best explanation for the observed navigation is the *folk-tfidf* hypothesis which considers semantic behavior in combination with the structural properties of the folksonomy. This allows us to conclude that while users do not use the folksonomy structure when accessing their own resources (because they most likely explicitly access known publications), they fall back to the folksonomy structure when browsing resources outside of their scope. Furthermore, the evidence for the user consistent hypothesis drops strongly compared to the other hypotheses, because it is restricted to user consistent navigation *outside* of the browsing user’s resources. This leads us to believe that browsing outside of the own resources is a process aimed at the discovery of new resources which in turn is not bound to the ownership of resources. Additionally, the fact that the plain user consistent hypothesis performs similarly well as the self-transition hypothesis indicates that the observed user-consistent outside navigation is mostly due to pagination effects.

5.3.2.3 User Gender

Since gender bias in online systems is an active research area [107, 237], we also investigate the navigation for different genders. In BibSonomy, users can set their gender explicitly. If no gender was set, we assign the label *neutral*, otherwise, we can distinguish between *male* and *female*. In Figure 5.9 we show the evidence charts for the different gender navigation subsets.

We hardly observe any difference between the genders. There is only a slight difference when considering the semantic hypotheses compared to the folksonomy hypothesis. It seems that the navigation behavior of male users shows a tendency towards following the folksonomy structure whereas the navigation behavior of female and especially neutral users can be better explained using the semantic hypothesis.

5.3.2.4 Usage Continuity

Since we expect users to adapt to systems they are using, we investigate if their navigation behavior changes over time. We divide users into *short-term* and *long-term users*, according to the temporal difference of their first and last request. If the difference is less than half a year, we classify a user as a *short-term user* and as a *long-term user* otherwise. In Figure 5.10, we report the results for short-term users. The results for long-term users are very similar to the results of the overall dataset (cf. Section 5.3.2.1).

When comparing against the overall dataset (or, equivalently, the long-term user group), we observe two conspicuous differences for short-term users. First, the semantic hypothesis performs significantly better when compared to the folksonomy hypothesis. Second, the page-consistent and the folk-user hypotheses are explaining navigation equally well. The former may be explained by the fact that new users are not as tuned to the folksonomy structure as long-term users. Thus, we may actually observe a learning process: the longer users work with the folksonomy, the more they exploit the folksonomy structure in order to navigate their own resources or to discover new ones. The similar evidence curves for the page-consistent and the folk-user hypotheses can be explained by increased pagination effects while exploring the system in combination with the lack of transitions on resources owned by the browsing user. In contrast to outside navigation, here, the lack of transitions on own resources can be explained by the fact that new users have no or a lot less own resources than long-term users.

5.3.2.5 Tagger Classes

In [122] and [176], different types of folksonomy users were characterized by their tagging behavior. [122] define *categorizers and describers* and [176] identify *generalists and specialists*. Categorizers and describers are classified by their *tag-resource-ratio* (or short *trr*, see also Section 3.3). That is, while categorizers use a small set of different tags for a large number of resources indicating elaborate category systems, describers use many different tags indicating a very descriptive approach when tagging. Generalists and specialists can be classified using *tag entropy* (or short *ten*, see also Section 3.3). Where generalists have a high tag entropy, indicating a wide variety of tagged topics with regard to their resources, specialists have a low entropy indicating a very specialized set of topics. For both classes we order users according to their *trr* and *ten* values separately and select the upper and lower 30%, respectively. We observe that categorizers and generalists show very similar evidence curves when compared to the overall navigation dataset. We show the marginal likelihood charts for describers and specialists in Figures 5.11a and 5.11b, respectively.

For both describers and specialists, we see the same tendency as for short-term users: the semantic hypothesis works better compared to the folk hypothesis and the user-consistent hypothesis has a tendency to perform equally well as the folk-user hypothesis.

The tendency towards semantic navigation over structural navigation on the folksonomy structure can most likely be explained by the nature of the tagging types: Specialists can be considered to be interested in rather few abstract topics, implying a more directed

browsing behavior than generalists (whose interests are more varied). Consequently, their navigation is expected to also be more semantically influenced, because of their use of a small, but highly specialized tag subset. As for describers, resources are tagged with more keywords. Thus, for the semantic measure based on TFIDF, calculating the similarity may simply be easier than on the very sparse tagging structures induced by a categorizer's tagging habits. This is because if a categorizer only assigns a low number of tags to her resources, the cosine similarity to another page with only a low number of tags can easily be zero because of the sparsity of the page representation. Consequently, the probability to navigate towards that page is zero. On the other hand, if a describer uses a lot of tags, the chance that the vector representations of two pages share a tag overlap is higher, resulting in a non-zero cosine similarity between both pages and thus a positive transition chance. This in turn again influences the navigation hypothesis by allowing such a transition.

In general, both types, specialists and describers, can be considered to be of a more explorative nature, as can be seen by the relative performance drop of the folk-user hypothesis and/or the increase of evidence for the self-transition hypothesis.

5.3.3 Conclusion

Understanding human navigation in web systems is an important step towards improving the design and usability of web pages. In this section, we analyzed navigational behavior of users in a social tagging system. We presented several hypotheses on navigational patterns and evaluated them on a large weblog dataset of the social tagging system BibSonomy.

Beyond confirming the results from Doerfel et al. [64], that is, that users mainly navigate on their own resources (cf. also Table 5.2), we were able to show that within these resources, navigation follows a semantic bias (cf. Figure 5.8). Also, the semantic hypothesis performs well in general, confirming the semantic component in navigation behavior on BibSonomy.

Furthermore, we studied different navigation subsets and were able to find significant differences in behavior. This includes that even though semantic, user consistent navigation represents a major aspect of the navigational characteristics of BibSonomy, users fall back to the folksonomy structure when browsing outside of their own pages. (cf. Section 5.3.2.2). Also while different genders did not exhibit interesting behavioral deviations (Section 5.3.2.3), short-term users, as well as different tagging types, follow certain behavioral patterns matching their individual characteristics (Section 5.3.2.4 and Section 5.3.2.5). In particular, while it was only hypothesized in prior work [122] that categorizers and describers (as well as generalists and specialists) differ in navigation behavior, we have found specific components of their behavior which differ significantly, thus, indicating that navigation behavior and tagging pragmatics are indeed connected.

Overall, we were able to gain new insights into the underlying processes of navigation in tagging systems, which can be extended and leveraged in the future, for example, by considering new hypotheses, improving navigation experience or extracting semantics.

5.4 Summary

This chapter focused on the characterization of navigation in social media systems, concretely on Wikipedia and BibSonomy. A special focus was given on detecting a semantic influence on the navigation behavior of users.

In the case of Wikipedia, we discerned between game navigation, i.e., navigation in a game context, where users have to reach a given target page as fast as possible, and unconstrained navigation, i.e., navigation without any exterior incentive. We proposed a set of navigational hypotheses and measured their influence on actual navigation data. In the following, we will shortly discuss our findings.

Both on Wikipedia and BibSonomy, navigation is influenced by a semantic component. Although it is not always the strongest factor in navigation, it plays a significant role. While the semantic component best explains navigation in Wikispeedia (cf. Figure 5.2), we can see in Figure 5.3 that up to some point of increasing belief, the semantic hypothesis explains some part of the navigation data. A similar behavior can be observed in unconstrained navigation. Figure 5.5 and Figure 5.4 show that the semantic hypothesis does not suffice to explain navigation, but again explains a part of the navigation behavior. This can motivate further research as to which part of navigation behavior interacts with the semantic hypothesis so that the combination of both hypotheses explains navigation as good as possible. Finally, in BibSonomy, users navigate semantically on resources of the same user (see Figure 5.7), but the semantic hypothesis still explains navigation in other aspects as well.

Generally we can say that, since users navigate at least to some extent according to semantic incentives, we should be able to extract the information generated by human navigation. However, we might be unable to gather *all* contained information, since navigation is also influenced by a lot of other factors, such as the degree (in Wikipedia) or the ownership (in BibSonomy) of the successive page. Methods to extract at least a portion of this semantic component will be presented and discussed in Chapter 7.

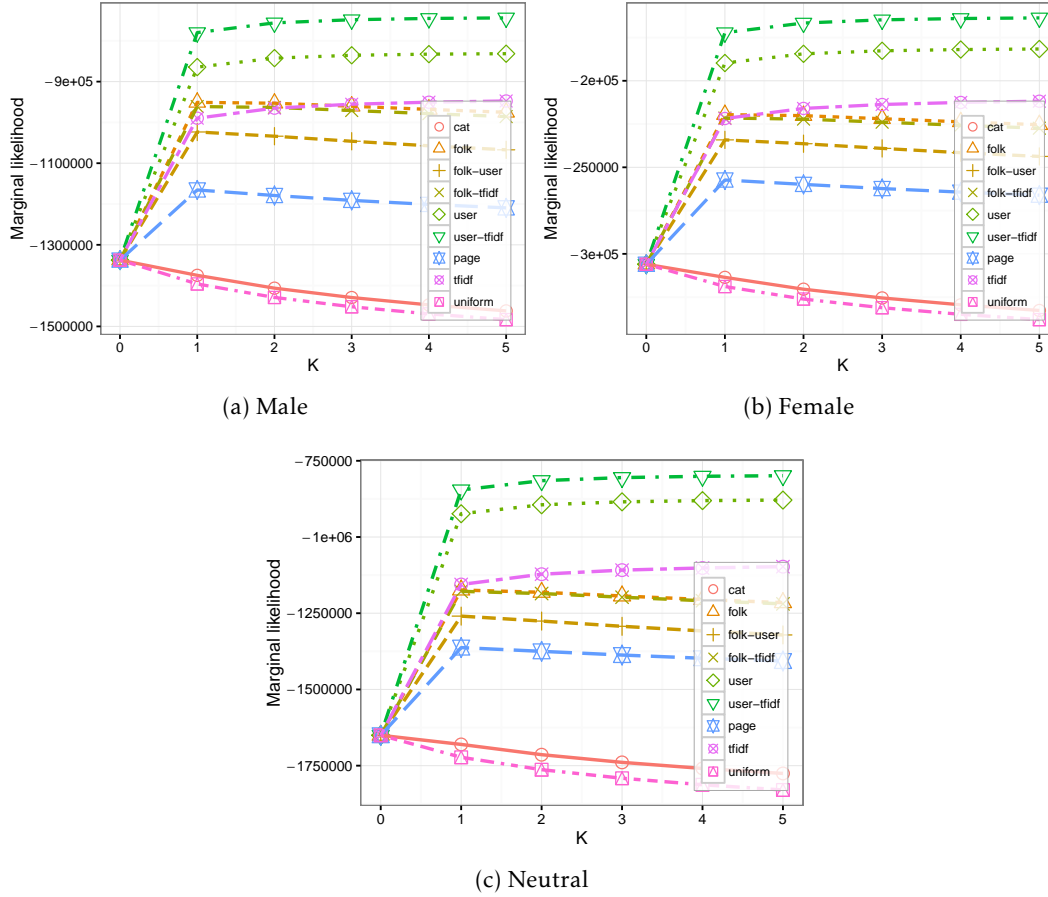


Figure 5.9: Evidence charts of hypotheses split by gender. If a user did not explicitly define his or her gender in BibSonomy, he or she was classified as neutral. While there is a slight difference when considering the folksonomy and the semantic hypothesis (male users seem to navigate using the folksonomy structure a little more than the other user groups), we observe that all hypotheses perform equally for all data subsets.

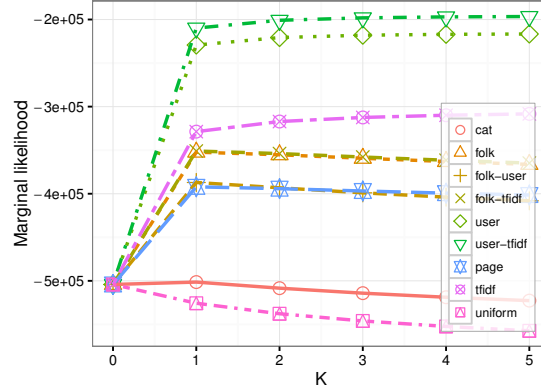


Figure 5.10: Evidence charts for short-term users (\leq half a year according to their first and last request). We observe a stronger performance of the semantic hypothesis compared to the folksonomy hypothesis and see that the self and folk-user hypotheses perform equally well in explaining navigation. We attribute this to the increased browsing aspect of new users.

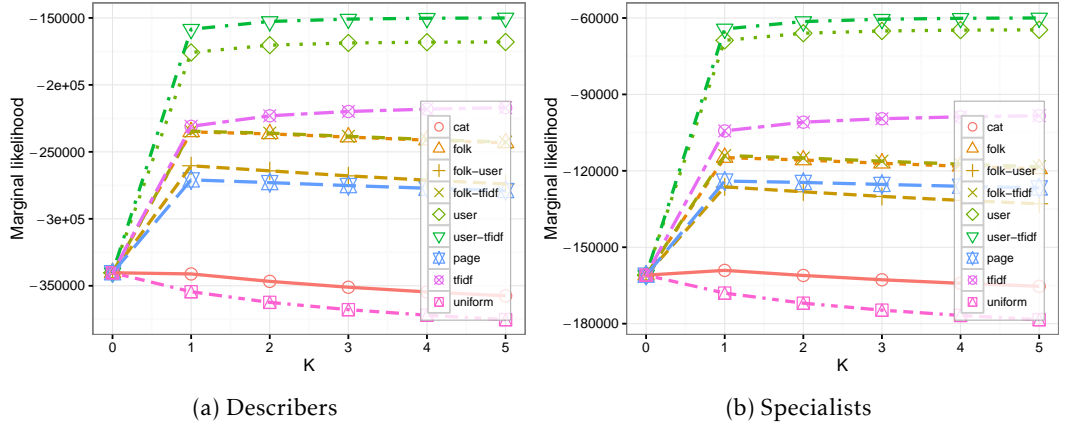


Figure 5.11: Evidence chart for different user types according to tagging behavior: describers and specialists. We see approximately the same characteristics as for short-term users. Again, we attribute this to a more pronounced bias towards browsing instead of explicit access to known resources.

Chapter 6

Capturing Semantics in Social Tagging Data

6.1 Introduction

Social Tagging Systems are built on semantic information. As we have already noted in Section 3.1.1, the assignment of free-form textual annotations or tags to posted resources, for example images or publications, lets semantic structures to emerge [85]. Several authors introduced methods to capture these emergent semantics. For example, Cattuto et al. modeled each tag as a high-dimensional co-occurrence vector [48]. Körner et al. then proceeded to show that there even exist groups of users whose tagging behavior has a positive influence on the emerging semantics in folksonomies [122].

To show that these tagging vectors really captured semantic relatedness, Cattuto et al. grounded them on WordNet using the Jiang-Conrath similarity measure [112]. Budanitsky and Hirst have shown that the Jiang-Conrath measure produces scores that correspond well with human intuition [44]. Consequently, Cattuto et al. argued that the scores produced by Jiang-Conrath generalize well and can thus also serve as a basis to evaluate their tagging vector space model.

However, there exist several issues regarding that evaluation procedure. Most importantly, in [44] the Jiang-Conrath measure was evaluated only on two very small semantic relatedness datasets. Since the evaluation results in [44] showed very high correlation of the Jiang-Conrath measure with human intuition, the first issue should exhibit only a limited impact. However, the second issue needs to be addressed by extending the evaluation on several other, bigger evaluation datasets. If this extended evaluation shows that the Jiang-Conrath measure does in fact not generalize, then results obtained from this evaluation approach should also be reconsidered.

An issue with the tagging vector space model itself, as proposed by Cattuto et al., is that the vectors representing tag possess a very high number of features [48]. While most tag vectors are often sparsely populated, the high number of dimensions increases computational costs, especially in machine learning applications. Additionally, because of the curse of dimensionality, it gets increasingly more difficult to accurately measure semantic similarity, since vector distances tend to converge to a uniform distribution in

high dimensions.

In this chapter, we will approach both problems. We first discuss advantages and shortcomings of two options to evaluate the semantic quality of vector representations in Section 6.2. The first option uses scores produced by the Jiang-Conrath measure as a base for comparison, while the other option directly evaluates semantic similarity scores on human intuition. We find that the scores produced by the Jiang-Conrath measure do not generalize as well as assumed to human intuition of semantic relatedness. Consequently, this necessitates re-evaluation of some results that were obtained using the WordNet-based approach, concretely those from Cattuto et al. [48] and Körner et al. [122]. We will *repeat those experiments from those papers that are based on the Jiang-Conrath measure* in Section 6.3, but this time *evaluate the results on human intuition* instead of the Jiang-Conrath measure. Additionally, we extend the work by Körner et al., which measured the influence of tagging pragmatics on tagging semantics. We do so by introducing *two new classes of tagging pragmatics*. In Section 6.4, use those same classes to show that the task of *tag sense discovery*, i.e., discovering and disambiguating different *meanings* of tag, is also affected by tagging pragmatics. Finally, we also address the issue of sparsity in vector representations of tags by exploring the applicability of *word embedding algorithms on social tagging data* in Section 6.5. We perform parameter studies for all of the applied embedding algorithms and compare results using the previously presented evaluation procedure. Our results show that embeddings produced by the GloVe algorithm [184] using Delicious tagging data achieve competitive results on a series of human evaluation datasets. In conclusion, this makes a strong point for the use of tag embeddings in practical applications.

The contents of this chapter are based on several papers. Section 6.2 and Section 6.3.2 originate from an unpublished technical report from 2015 [171]. The Generalists and Specialists that we use in Section 6.3.3 have been first defined in a publication from 2013 [176]. That same work also contributes the contents of Section 6.4, where we explore the impact of tagging pragmatics on tag sense discovery. Finally, Section 6.5 has been published in 2017 in [174].

6.2 A Critical Examination of WordNet-based Evaluation of Semantic Similarity

In [48], Cattuto et al. not only proposed several ways to model the semantic information contained in tagging data, but also introduced a new way to evaluate these models with regard to their fit to human intuition of semantic relatedness. Concretely, they compared the similarity scores obtained from the tag vectors to those of two WordNet based measures, the Jiang-Conrath semantic similarity measure [112] and the taxonomic shortest path distance measure [105] (cf. Section 3.2.3.3). Both of those rely on the structure of semantic relations in WordNet, which we already introduced in Section 3.1.3. This evaluation approach of not directly evaluating on human intuition is based on the assumption that the Jiang-Conrath measure corresponds extremely well with human intuition, as shown in [44]. In their work, Budanitsky and Hirst evaluated

several WordNet-based semantic similarity measures in two settings, namely on a word similarity task and a malapropism task, i.e., how well these measures are fit to detect spelling errors. The word similarity task is based on comparing the similarity measures directly on human intuition, that is, semantic relatedness datasets as we introduced in Section 4.4. Budanitsky and Hirst used two very small and overlapping semantic relatedness datasets, namely RG65 and MC30, which contain 65 and 30 word pairs, respectively.

Not only due to the small size of both datasets, but also to the fact that MC30 is completely contained in RG65, it is difficult to say how well the WordNet-based similarity measures generalize, even although they achieved very high evaluation scores on both datasets. However, if these measures cannot generalize well, then we also need to re-evaluate approaches that build on semantic similarity scores from WordNet, for example the tag grounding procedure as introduced in [48] which has also been used in [122].

In this section, we re-evaluate some of the WordNet-based semantic similarity measures that were presented in [44] on a wide range of semantic relatedness datasets. In fact, we evaluate the Jiang-Conrath and taxonomic shortest path similarity in the same word similarity setting as in [44] using the MC30 and RG65 semantic relatedness datasets, but also extend the evaluation on other, bigger semantic relatedness datasets. We expected to find a clear signal whether or not these measures are able to generalize on measuring semantic similarity of words and are thus either suitable as a proxy for human intuition or not. Our findings however indicate that the Jiang-Conrath similarity measure only achieves inferior evaluation results on larger semantic relatedness datasets and is thus less suited as a general proxy for human intuition. While the widely used alternative of evaluating directly on human intuition seems the more viable way to obtain valid evaluation results, this method also has some drawbacks. Concretely, we address the issue of the inferior word coverage of human intuition datasets and analyze how the quality of evaluation results depends on the covered vocabulary. To demonstrate how this issue can be overcome, we introduce a new semantic relatedness evaluation dataset specifically collected to measure the semantic relatedness of tag from the BibSonomy folksonomy.

6.2.1 Experimental Setup

This section describes the used datasets and resources as well as the experimental setup.

Resources. We evaluate the Jiang-Conrath semantic similarity measure [112] and the taxonomic shortest path distance [105] on a range of semantic relatedness datasets. This includes the MC30 and RG65 datasets that were used in [44], but also more recently published and bigger datasets, namely WS-353, MEN, MTurk, and SimLex-999 (cf. Section 4.4).

Re-Evaluating the Jiang-Conrath Measure on Human Intuition. The general goal of semantic relatedness measures is to come as close as possible to human intuition,

that is, to depict the generally accepted strength of relation as accurately as possible. Thus, it is crucial that the evaluation scenario sufficiently reflects such human intuition, as anything else would not provide a valid basis for evaluation. In this experiment, we discuss whether each evaluation approach is a suitable representation of human intuition by itself.

A good evaluation method should consistently produce valid and reliable results, regardless of the given topic context. Naturally, this is hardly possible for highly specialized topic contexts, such as scientific or medical vocabulary, as opposed to a general topic context where only general vocabulary is concerned. For example, the meaning of a *paper* in a scientific context, i.e., a publication describing a piece of research, differs from its general meaning, which describes a material that such pieces of research are usually printed on [102]. Especially in such cases it is necessary to quickly extend the evaluation basis to reflect special meanings and relations.

We compare the vocabulary coverage of all tagging datasets on both WordNet and all human intuition datasets on the example of social tagging systems, as these cover a limited range of topics. Then we introduce a specifically created evaluation dataset that targets scientific and technical vocabulary, just as in Delicious and BibSonomy. We show that using this dataset, we can more accurately assess the quality of semantic relatedness information in tagging vectors constructed from these folksonomies.

6.2.2 Results

In this section, we present the results for all experiments described above.

6.2.2.1 Re-Evaluating the Jiang-Conrath Measure on Human Intuition

Because the Jiang-Conrath and shortest path distances have only been evaluated on the RG65 and MC30 datasets, we wanted to extend this evaluation to other evaluation datasets presented in Section 4.1 to assess the quality of the WordNet based measures as a representation of human intuition of semantic relatedness. For each pair of words in the semantic relatedness dataset, we match them each on a concept in WordNet. Then we computed the Jiang-Conrath and taxonomic path distance between these concepts. If a word matched more than one concept, we chose that which maximized the similarity score. Table 6.1 shows correlation scores for the evaluation of the WordNet based measures on human intuition. We report both Pearson (r) and Spearman (ρ) correlation scores, since although using Spearman's correlation coefficient is more intuitive in a semantic relatedness setting, we also want to compare with the Pearson correlation scores reported in by Budanitsky and Hirst [44]. We can see that, while the correlation with RG65 and MC30 is very strong, evaluation performance decreases greatly when using the other evaluation datasets. Most interestingly however is that the Jiang-Conrath similarity measure is indeed a very good fit for the semantic similarity dataset SimLex-999. This means that Jiang-Conrath measure seems to measure rather semantic similarity instead of relatedness, as the correlation with the scores of the semantic relatedness datasets MEN, WS-353, and MTurk are low. Still, since the number of evaluable pairs is

Table 6.1: Evaluation of the WordNet semantic similarity measures Jiang-Conrath (jcn) and taxonomic shortest path distance (path) on different datasets with human intuition of semantic relatedness (MEN, MTurk) and semantic similarity (RG65, MC30). When evaluated on more recent datasets, both measures yield less obvious correlation with human judgment, compared to the scores achieved on RG65 and MC30. The Jiang-Conrath and the shortest path distance show a high correlation with similarity datasets, while there seems to be rather low to no correlation on relatedness datasets. We report both Pearson (r) and Spearman (ρ) correlation scores, since although using Spearman’s correlation coefficient is more intuitive in a semantic relatedness setting, we also want to compare with the Pearson correlation scores reported in by Budanitsky and Hirst [44].

		RG65	MC30	WS-353	MEN	MTurk	SimLex-999
r	jcn	0.853	0.868	0.063	0.047	0.057	0.536
	path	0.784	0.755	0.374	0.378	0.446	0.301
ρ	jcn	0.776	0.826	0.298	0.367	0.364	0.545
	path	0.781	0.724	0.296	0.337	0.340	0.221
matched pairs		65	30	348	2606	243	700

very high for all datasets, these results validate our assumption that the Jiang-Conrath and taxonomic path similarity measures can not generalize as well as thought.

6.2.2.2 Vocabulary Coverage and Extensibility of Semantic Relatedness Datasets

We split this section in two parts. First, we talk about the covered vocabulary in the evaluation datasets in order to see if we can obtain believable evaluation results. After that, we describe how we created a semantic relatedness dataset with vocabulary from BibSonomy, to demonstrate how easily we can obtain domain-specific evaluation data.

Vocabulary Coverage of Evaluation Datasets Table 6.3a shows the overlap of several top-frequency tag subsets from both Delicious and BibSonomy with WordNet. As a trend, we can observe a higher overlap among the more frequent tags. Furthermore, the overlap of Delicious and WordNet is almost the same as the numbers which have been reported by [48], although we use a larger, more recent dataset (cf. Table 6.2). This tells us that over time, the semantic structure and frequency of words in the Delicious folksonomy have been very stable.

Because the results in Table 6.1 are mediocre compared to other papers, which achieve much higher correlation values on WS-353 (see Table 6.3), we investigated the reasons for why BibSonomy and Delicious do not yield more competitive results, although it has been shown that folksonomies exhibit strong regularities in tag usage, which should allow to extract semantic relatedness [85]. In Table 6.3b, we show how many words and

Table 6.2: Overlaps of the top tags from Delicious and BibSonomy with WordNet and WS-353.

# top tags	100	500	1k	5k	10k
Delicious	78%	75%	74%	61%	54%
BibSonomy	61%	59%	56%	45%	38%

(a) Overlap with WordNet.

measure	Delicious					BibSonomy				
	100	500	1k	5k	10k	100	500	1k	5k	10k
word overlap	17	61	99	240	302	16	50	80	193	248
pair overlap	0	18	37	136	194	5	14	26	114	151

(b) Overlap with WS-353

Table 6.3: Spearman correlation scores on WS-353 in other literature. Note that all of these works conduct their experiments on Wikipedia data, e.g., the link network instead of WordNet or tagging data.

paper	reported correlation
WikiRelate [223]	0,55
ESA [81]	0,75
WikiGame paths [210]	0,76
TSA [189]	0,8
Conceptnet Numberbatch [217]	0,828

pairs from WS-353 are contained in different subsets of the most frequent tags of the experiment datasets. For both BibSonomy and Delicious, one must consider the top 5k tags to achieve a sufficient amount of evaluable pairs on which evaluation is reasonable. Also, another big part of the pairs is only matchable when the whole experimental vocabulary is taken into account. In contrast, the vocabulary overlap of BibSonomy and Delicious to WordNet is relatively high (5400 tags for the top 10k Delicious tags and 3800 tags for the top10k BibSonomy tags, see Table 6.3a).

Extension and Evaluation on Domain Specific Vocabulary If it concerns domain specific vocabulary, WordNet cannot easily be extended, since it is a resource hand-crafted by linguistic experts [72]. This makes vocabulary extensions extremely expensive, both in time and necessary experience. On the other hand, one can easily collect a number of tag pairs and have them annotated by humans. In fact, several works have created human intuition datasets for their purposes already in order to obtain datasets with more fitting vocabularies. For example, the WS-353 dataset contains very general

Table 6.4: Comparison of different BibSonomy vocabulary subsets with the WS-353 and the Bib100 vocabularies.

BibSonomy subset	word overlap	pair overlap
WS-353 (353 pairs)		
top 100	18	4
top 1 000	89	36
top 10 000	269	168
Bib100 (100 pairs)		
top 100	14	8
top 1 000	89	70
top 10 000	122	100

words from several topics and thus does not provide a good evaluation basis for news article semantics. Thus, Radinsky et al. created the MTurk dataset from the New York Times corpus, to just name a single example.

Because the vocabulary of the WS-353 and MTurk evaluation datasets does not fit well to the vocabulary of the tagging datasets used in this section, we decided to create a new evaluation dataset with a more fitting vocabulary to the one used in BibSonomy and Delicious.¹ We will now describe the creation process of Bib100.

We randomly selected 122 words from the top 3 000 words of the BibSonomy dataset and combined them into 100 word pairs by hand, according to our own perceived similarity ascending from unrelated to strongly related. In [212], Snow et al. show that non-expert annotations are equally well suited to evaluate NLP tasks. Consequently, we had each pair judged 26 times for semantic relatedness using crowdsourced scores by non-experts. The scores ranged between 0 (no similarity) and 10 (full similarity) on the MicroWorkers platform. Each rater had to judge the whole set of 100 word pairs at once. Raters whose ratings deviated from the remaining inter-rater correlation too much, i.e., with a correlation difference of 0.2, were excluded. We also manually filtered raters who obviously judged wrong scores, e.g., zero relatedness between *web* and *internet*. The average interrater agreement is 0.68. With this dataset, we were better able to judge the semantic quality of word representations constructed from both tagging datasets used in this section. While this heuristic of creating an evaluation dataset is rather basic as for example in comparison with the creation process of SimLex-999 [104], we still obtain a reasonable rating distribution.

In Table 6.4, we can furthermore see how much better Bib100 overlaps with BibSonomy tagging data than WS-353 does. Although we obviously can match all pairs of Bib100 on BibSonomy tagging data, we are also able to match almost double the number of pairs on the top 1000 tags with an equal number of matched words. Table 6.5 finally shows the evaluation scores of the Jiang-Conrath and taxonomic path distance measures

¹<http://www.dmir.org/datasets/bib100>

Table 6.5: Evaluation of the Jiang-Conrath and taxonomic path distance measures on Bib100. We also report the results on WS-353 as comparison. The evaluation results of the two WordNet-based similarity measures are still not very good.

		Bib100	WS-353
r	jc	0.255	0.063
	path	0.337	0.374
ρ	jc	0.389	0.298
	path	0.277	0.296
matched pairs		83	348

on Bib100. In the following sections, concretely in Section 6.3 and Section 6.5, we will also use Bib100 as an evaluation dataset for semantics of tagging data.

6.2.3 Discussion

We will now discuss the results from the preceding experiments: We compare the soundness of the evaluation methods and discuss issues of evaluation on human intuition datasets.

6.2.3.1 Comparison of Evaluation Approaches

While WordNet makes it possible to evaluate hierarchical relations between concepts, it is not explicitly designed to calculate semantic relatedness. Nevertheless, similarity measures, such as the Jiang-Conrath or the path distance measure, exploit the WordNet graph. [44] even found relatively large correlation with human intuition on semantic relatedness by comparing against two relatively small datasets. However, we saw in Table 6.1 that correlations are much smaller when comparing with the larger, more recent evaluation datasets WS-353, MEN, MTurk and Bib100. Since the goal of extracting semantic relations is to find measures that are well aligned with human intuition, our results on WordNet-based metrics are strong evidence against their suitability as a single base for evaluating extracted semantics.

A possibility to still use the evaluation method depicted in Equation (3.30) would be to replace the WordNet-based similarity values d_{wn} by human judgments. However, the $argmax$ component yields different word pairs for each semantic relatedness measure we examine. Thus, it is necessary to collect new human judgements every time we evaluate a new measure. In addition, similarity values should be supported by a minimum number of human ratings. Now, the original approach calls for 10 000 words to average over. Assuming a minimum of 10 judgements per word pair, this would require an overall of 100 000 judgments.

Overall, since WordNet based measures are hardly correlated with human judgements and adopting the evaluation method by Cattuto et al. to use human judgements is

expensive, we argue to evaluate directly on datasets containing human judgements on semantic relatedness using the Spearman correlation coefficient as introduced in Section 3.2.4. While corresponding datasets are smaller in size than the WordNet taxonomy and thus are not covering every possible relation between words, they provide a quicker and more realistic setting for evaluating semantic relatedness measures.

6.2.3.2 Shortcomings of Evaluation on Human Intuition Datasets

After we have argued for the use of datasets covering the actual human intuition on semantic relatedness, we now discuss some issues that must be considered when using that approach. First and foremost, the size of the evaluation datasets is important. The MEN collection is by far the biggest dataset, consisting of 3 000 word pairs. The second largest dataset is WS-353, containing 353 pairs. Compared to the sizes of our experiment datasets, and thus to the number of word pairs for which similarity judgments can be extracted, the evaluation datasets are rather small.

Secondly, human raters exhibit deviations and uncertainty in their judgments. For word-pairs with a mean similarity between 2 and 7,5, the standard deviation of ratings is about 3 and lower otherwise. This could be interpreted as insecurity of raters about the extent of similarity of words, which are neither obviously related nor clearly unrelated. Figure 3.9 shows the mean distribution of the word-pair similarities with smoothed standard deviations in Bib100 and WS-353. Interestingly enough, a very similar rating behavior can be observed for the WS-353 dataset.

Finally, our results show that not only size matters, but for evaluating semantic relatedness measures the evaluation datasets must also contain “the right” vocabulary – i.e., yield a high overlap with the experiment datasets. Table 6.3b shows this overlap between the WS-353 dataset and subsets of the top tags of the experiment datasets Delicious and BibSonomy. It takes the top 5k tags in each dataset to find a reasonably large coverage. Another big part of the vocabulary is contained among second half of the top 10k tags. With the creation of Bib100, we were able to show that it is possible to achieve more competitive results by creating a dataset with a more fitting vocabulary. The suitability of the Bib100 vocabulary for both folksonomies can especially be seen in the high percentage of found word pairs in Table 6.4. Additionally, we will use Bib100 as an evaluation dataset in Section 6.3 and Section 6.5, where we will also see that we can better capture the domain-specific semantics of BibSonomy and Delicious than with WS-353.

Moreover, the competitive results on Bib100 are another argument in favor of using such human intuition datasets to evaluate semantic relatedness: It is very easy to construct a domain-specific dataset by picking a set of representative, frequently used words from the experiment dataset vocabulary, combine them into pairs, and to have them evaluated by humans, e.g., through crowdsourcing. They thus provide a cheap, easy, and fast option to judge semantic relatedness, while yielding a plausible evaluation scenario, as they rely directly on the explicitly expressed human intuition of semantic relatedness.

6.2.4 Conclusion

In this section, we challenged the evaluation approach for semantics in tagging data presented by Cattuto et al. in [48], which grounds relatedness scores on those generated by the WordNet-based Jiang-Conrath measure. We did so by comparing the Cattuto evaluation measure to another standard way of evaluating semantic relatedness measures, which compares artificial scores produced by the relatedness measure to actual human intuition. Our experiments showed that the main assumption for the Cattuto evaluation measure, namely that Jiang-Conrath produces human-like relatedness scores, does not hold at all and thus, this way of evaluating relatedness scores is not as viable as previously assumed, if used as the only evaluation measure. Instead, we should evaluate semantic relatedness scores, regardless of the generating textual content, on human intuition directly. While this evaluation approach also has many drawbacks, such as limited or unfitting vocabulary, it is still the best way to directly evaluate semantic relatedness to ensure that the measure best fits human intuition.

6.3 Re-Evaluating Measures of Semantic Tag Similarity on Human Intuition

We have seen in Section 6.2 that the Jiang-Conrath similarity measure does not always represent human intuition as well as was assumed, and that it cannot generalize well across different semantic relatedness datasets. As there are several works which base their results about measuring semantic similarity on evaluation scores that originated from the Jiang-Conrath measure, their experiments need to be repeated, yet this time evaluated directly on human intuition. While the evaluation on Jiang-Conrath can still be considered as a signal for the quality of the semantic similarity measure, a second evaluation will either support the obtained results and thus make them stronger or can serve as instigation for further research.

In this section, we re-conduct experiments from two works on measuring the semantic similarity of social annotations. Concretely, we focus on those experiments that base their results on scores from the Jiang-Conrath measure. In the first work [48] about modelling semantic information in tagging data, Cattuto et al. evaluated different tag context choices to construct vector representations, according to how well they were able to judge semantic similarity. In this paper, the evaluation on WordNet using the Jiang-Conrath measure was initially proposed. The second work [122] investigated the impact of user pragmatics on the semantic quality of the vector representations. A main implication of their work was that they showed first evidence for a causal link between user pragmatics and tagging semantics. Again, this was shown using the Jiang-Conrath measure as evaluation base.

We find that the results from [48] hold in general, but we can also observe some variations in the qualitative ranking of the tag-tag co-occurrence and tag context similarity measures. When re-examining the results from [122], we first observe that we cannot generalize their results across all considered evaluation datasets, as for each evaluation

dataset, we observe slightly different results. We finally attribute this to the semantic information encoded in these datasets, be it either semantic *relatedness* or semantic *similarity* (cf. Section 3.2.3.1), with the latter being measured when evaluating on the Jiang-Conrath measure (cf. Section 6.2).

We will first describe some preliminaries for the experiments conducted in this section, such as the used folksonomy data and the used methodology to construct tag vectors. Then we will describe the experimental setup as well as the results of the re-evaluated experiments from the papers of Cattuto et al. and Körner et al., for which we also discuss the results in the light of the additional evaluation.

6.3.1 Preliminaries

We use the folksonomy data from Delicious and BibSonomy that we introduced in Section 4.1. Both datasets are restricted to the top 10k tags. They also contain the user information needed to construct the folksonomy subsets based on the tagging pragmatics of users in Section 6.3.3. To construct tag vector representations, we use the context based representations that Cattuto et al. used in [48] and which have been described in Section 3.2.2.1, concretely the tag context, resource context and user context measures, as well as the tag co-occurrence measure.

6.3.2 Distributional Measures of Tag Similarity

In [48], Cattuto et al. provided a systematic characterization and validation of several measures of tag similarity. Some of these measures were based on a distributional representation of a tag's context. We introduced all of these context definitions in Section 3.2.2.1. Cattuto et al. compared the distributional tag representations with regard to their ability to measure semantic relatedness between tags. As evaluation, they used the WordNet-based approach we described in Section 3.2.4. Their results showed that resource context based similarity outperformed the other context similarity variants, followed by tag context similarity, tag co-occurrence similarity, FolkRank similarity and finally user context similarity. Still, all measures yielded better reflection of supposedly human intuition than a random baseline. In the end, they however decided to represent tags by the tag-tag co-occurrence context, since it was easier to construct and computationally more feasible, while at the same time only slightly worse than representation by resource context.

Since human intuition in general shows stable patterns in their assessment of semantic relatedness [75, 160, 199], we assume that a good evaluation method for semantic relatedness of words also produces score rankings that are comparable to other sufficiently good evaluation methods. Concretely, if a proven evaluation method determines that a set of word vectors better reflects human intuition of semantic relatedness than another set of vectors, then other evaluation methods should also yield the same ordering for both vector datasets. We expect at least that the ranking of the context based similarity measures still holds in our experiments, i.e., when evaluating on human intuition instead of WordNet. Additionally, as we extend the experiments on another dataset

Table 6.6: Results for the WordNet-based evaluation method used by Cattuto et al. [48] on the Delicious and BibSonomy datasets. We also report the results from [48] as a reference. We used both the Jiang-Conrath distance (JCN) as well as the taxonomic path distance (Path) as semantic similarity ground truths. The best evaluation scores are marked fat.

	Delicious		BibSonomy		Cattuto et al. [48]	
	JCN	Path	JCN	Path	JCN	Path
tag cosine	11,0	6,5	14,2	8,2	10,1	6,2
res cosine	10,8	6,4	14,0	8,1	9,1	6,2
user cosine	13,8	7,6	14,1	8,3	13,3	7,9
tag co-occ	13,1	7,6	13,7	8,1	12,3	7,4

(namely BibSonomy), we also expect to see a similar picture there.

Experimental Setup

In the following, we will use the context variants to construct distributional representations of tags in the Delicious, and BibSonomy folksonomies. The respective folksonomy datasets have already been introduced in Section 4.1. For each combination of folksonomy data and similarity measure, we then evaluate the semantic relatedness scores produced by the distributional representations both with the WordNet-based evaluation as well as on human intuition. We do so to achieve a comparison with the original evaluation results reported by Cattuto et al. on the one hand, but on the other hand to also see if both evaluation approaches rank the similarity measures analogously.

Results and Discussion

We now present and discuss the results of both evaluation approaches. First we discuss the evaluation results on WordNet, before we look closely at the results of evaluation on human intuition. At the end of this section, we compare them with each other.

Evaluation of Word Vectors on WordNet using the Jiang-Conrath Measure In this experiment, we evaluate our results on WordNet analogously to [48] by comparing our results with the taxonomic path length and Jiang-Conrath distances as described in Equation (3.30). Table 6.6 shows the results of this experiment.

The Delicious results are very similar to the reported results in [48], i.e., tag cosine similarity based on resource and tag context finds closer nearest concepts according to Jiang-Conrath and path distance than cosine similarity based on user context or direct first-order co-occurrence. Similar experiments on BibSonomy have not been reported in [48]. In our investigation, however, we find that the results for BibSonomy show little to no variation between the different context descriptions. We will see in the course of this section that BibSonomy yields a similar ranking of the investigated similarity measures.

6.3 Re-Evaluating Measures of Semantic Tag Similarity on Human Intuition

Table 6.7: Results for evaluation on the human similarity intuition datasets for the Delicious and BibSonomy folksonomy dataset. Best evaluation results are marked fat. We can see that generally, the qualitative performance ordering of the context measures holds, as resource context is almost always the best choice, while user context yields the worst results.

sim	WS-353	MTurk	MEN	Bib100	SimLex-999
tag context	0.452	0.505	0.574	0.636	0.311
res context	0.614	0.662	0.736	0.745	0.478
user context	0.432	0.611	0.583	0.445	0.172
tag co-occ	0.554	0.614	0.718	0.643	0.295
#pairs	202	103	1376	94	398

(a) Delicious

sim	WS-353	MTurk	MEN	Bib100	SimLex-999
tag context	0.384	0.548	0.433	0.626	0.138
res context	0.606	0.668	0.587	0.642	0.171
user context	0.260	0.496	0.381	0.381	-0.030
tag co-occ	0.531	0.675	0.394	0.556	0.104
#pairs	158	62	463	100	213

(b) BibSonomy

Evaluation of Word Vectors on Human Intuition In the following, we evaluate the folksonomy data using the context measures from [48] directly on human intuition of similarity, which is represented by the evaluation datasets from Section 4.1. The evaluation setting has already been described in Section 3.2.4. To provide a wide evaluation base, we use the same human intuition datasets that we already used in Section 6.2, namely WS-353, MTurk, MEN, Bib100, and SimLex-999. We omit the results for RG65 and MC30 because for both datasets, the number of evaluable pairs is always less than 5, from which we cannot infer any valid evaluation scores.

Table 6.7 shows the results for evaluation on human intuition datasets for the different context representations. The performance order as reported in Table 6.6 of the results shown in Table 6.7 is visible to some part, e.g., resource cosine similarity shows a better correlation than all other context representations. Also, the user context representation still yields the worst results. However, the distinction between tag co-occurrence and tag context similarity is not as clearly pronounced as in [48]. In some evaluation settings, tag co-occurrence is better than tag context similarity, while in other settings, it's the other way around.

Conclusion

In this part of this section, we re-examined the results from [48] in the light of a different evaluation approach. While in [48], Cattuto et al. compared different context based vector representations of tags with each other by comparing them to WordNet-based evaluation scores, we changed the evaluation setting to a direct comparison with human intuition. We wanted to see if the results from [48] still hold in a qualitative manner, concretely, if the ranking of the context-based representations with regard to their semantic evaluation scores are still present.

We found that while the first and last ranks consistently stayed the same, namely that resource-based context representations provide the best semantic information, we could not see such a clear distinction between tag-tag co-occurrence and tag context similarity. In [48], tag context similarity clearly outperformed tag-tag co-occurrence, but we sometimes observed the opposite. However, we attribute this to the imperfect vocabulary coverage of each evaluation dataset.

Future work in this section could be to investigate which kind of semantic relatedness is actually evaluated. Since we found in Section 6.2 that the Jiang-Conrath measure rather fits to the SimLex-999 dataset, which is explicitly built to measure semantic similarity, the WordNet-based semantic evaluation approach probably also focusses on determining semantic similarity instead of relatedness. Also, a combined evaluation approach, using both WordNet-based evaluation as well as semantic relatedness datasets would probably be able to combine the best of both worlds and thus provide an extremely helpful tool for the evaluation of semantic information.

6.3.3 Impact of User Pragmatics on Tagging Semantics

After Cattuto et al. have shown in [48] that distributional similarity measures can capture the semantic information in folksonomy data, Körner et al. extended upon this work to investigate potential influencing factors on the quality of the distributional measures. They assumed that “users with certain usage patterns might contribute more to the resulting semantics than others” [122]. In their work, they established a link between *user pragmatics*, i.e., the characterization of why users assign tags, and *tagging semantics*, i.e., the semantic meaning of those tags.

They defined several measures to characterize users either as categorizers or describers, which we described in Section 3.3. This notion of different tagging behavior has already previously been established in [221]. Their results indicate that users with many tags and few resources, i.e., describers, are more useful than others to extract semantic information. As in [48], they base their results on evaluation scores obtained from the WordNet-based evaluation using the Jiang-Conrath measure. Consequently, we also repeat the results from this work and evaluate the resulting vectors on human intuition. Although details may vary, we expect to see a similar pattern of tagging pragmatics influence on the extraction quality.

Experimental Setup

For our evaluation, we use tagging data from Delicious and BibSonomy, each restricted to the top 10k tags. To model the semantic information as vectors, we use the tag context similarity as defined in Equation (3.17), i.e., we count how many posts were both annotated with the same pair of tags. We evaluate these vectors on the MEN, WS-353, MTurk, Bib100, and SimLex-999 semantic relatedness datasets. However, we will not discuss th

To measure the impact of the different user groups, we follow [122] and define several folksonomy partitions. For each metric $m \in \{trr, tpp, vocab, orphan, mdc, mqdc, ten, ssc\}$, we obtained a list L_m of all users $u \in U$ sorted in ascending order according to $m(u)$. All our measures yield low values for categorizers / specialists, while giving high scores to describers / generalists. This means that e.g., the first user in the mean degree centrality list (denoted as $L_{mdc}[1]$) is assumed to be the most extreme specialist, while the last one ($L_{mdc}[k], k = |U|$) is assumed to be the most extreme generalist.

Because we are interested in the minimum amount of users needed to provide a valid basis for disambiguation, we start at both ends of the respectively sorted user list L_m and extract two folksonomy partitions CF_{10}^m and DF_{10}^m based on 10% of the “strongest” categorizers/specialists ($CatSpec_{10}^m = \{L_m[i] \mid i \leq 0.1 \cdot |U|\}$) and describers/generalists ($DescGen_{10}^m = \{L_m[i] \mid i \geq 0.9 \cdot |U|\}$). The sub-folksonomy

$$CF_{10}^m = (CU_{10}^m, CT_{10}^m, CR_{10}^m, CY_{10}^m) \quad (6.1)$$

of \mathbb{F} is then induced by $CatSpec_{10}^m$, i.e., it is obtained by

- by the restricted user set $CU_{10}^m := CatSpec_{10}^m$
- the restricted set of tag assignments $CY_{10}^m := \{(u, t, r) \in Y \mid u \in CatSpec_{10}^m\}$
- the restricted set of tags $CT_{10}^m := \pi_2(CY_{10}^m)^2$
- and the resources tagged by the restricted users $CR_{10}^m := \pi_3(CY_{10}^m)$

The sub-folksonomy DF_{10}^m is determined analogously by $DescGen_{10}^m$. We extracted partitions CF_i^m and DF_i^m for $i = 10, 20, \dots, 100$.

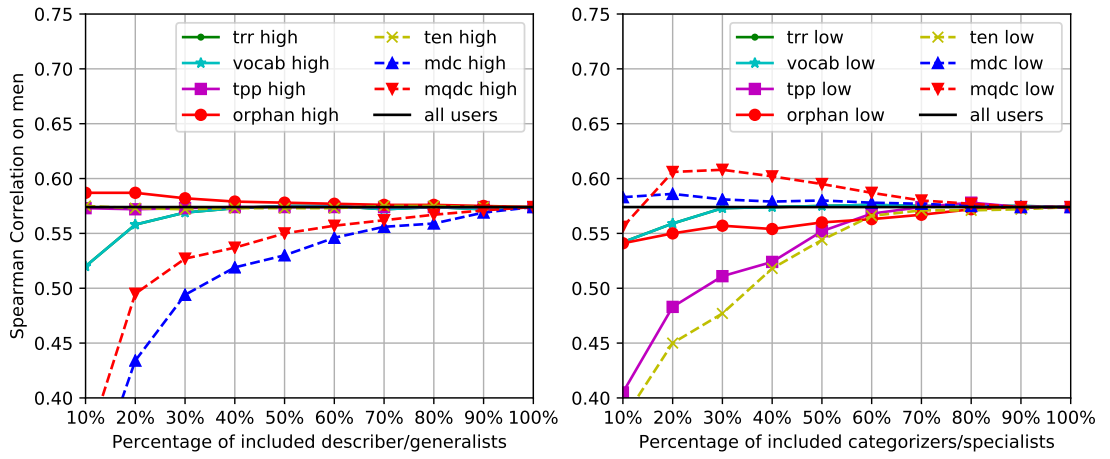
Results and Discussion

We will now describe and discuss the results of re-evaluating the impact of tagging pragmatics on tagging semantics.

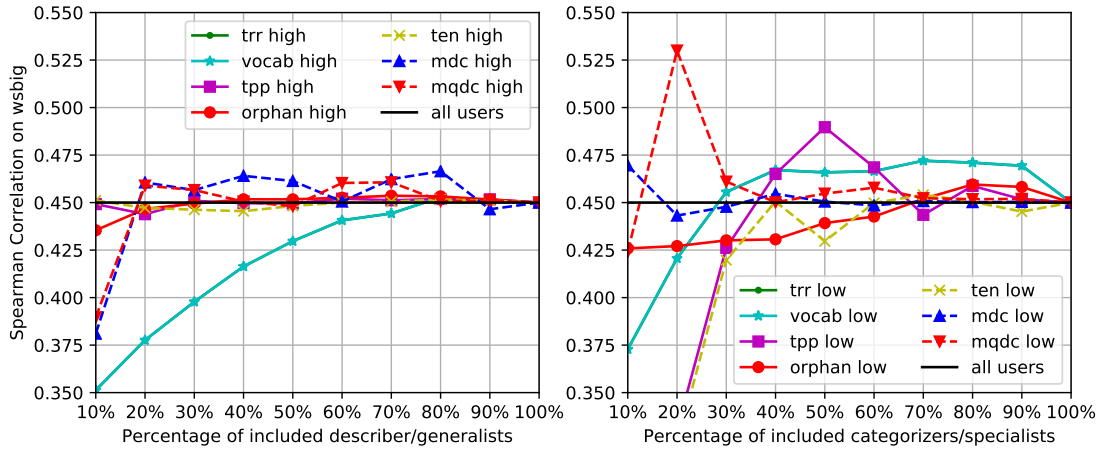
In Figure 6.1a, we can see that *mqdc specialists, more concretely 30% of low mqdc users, generate the best tagging semantics on Delicious and MEN*, where they can achieve a correlation of 0.61. With this, they outperform the baseline score of 0.574, as obtained by all users. The same signal can be seen for *mdc* specialists, although slightly weaker. Furthermore, we can see that *orphan* describers also produce slightly improved semantics,

² π_k is the projection operator on the k -th component of a tuple, i.e., $\pi_2((2, 3, 4)) = 3$

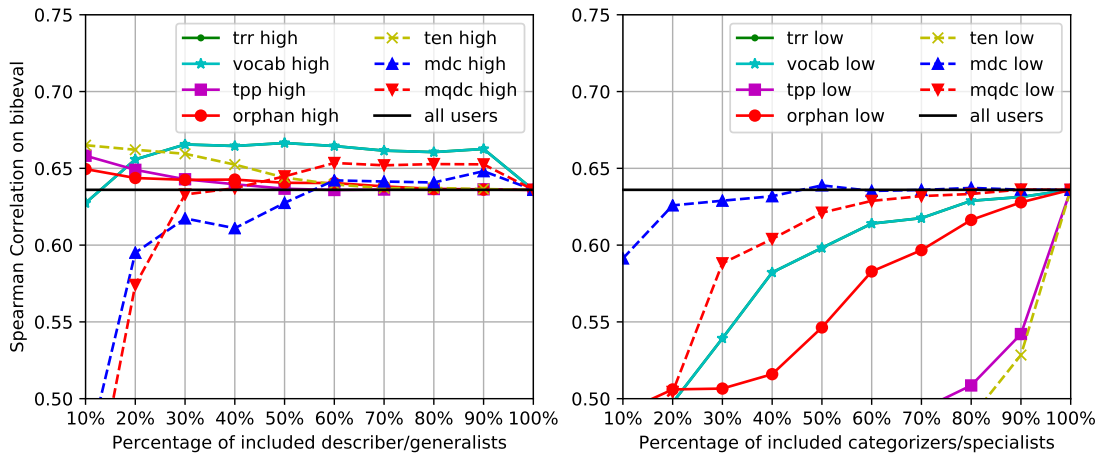
6 Capturing Semantics in Social Tagging Data



(a) Delicious results on MEN.



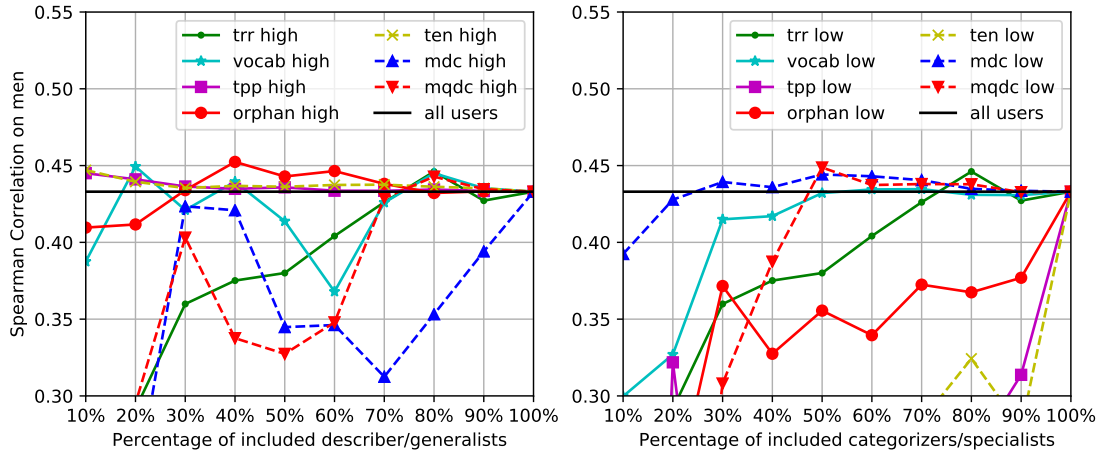
(b) Delicious results on WS-353.



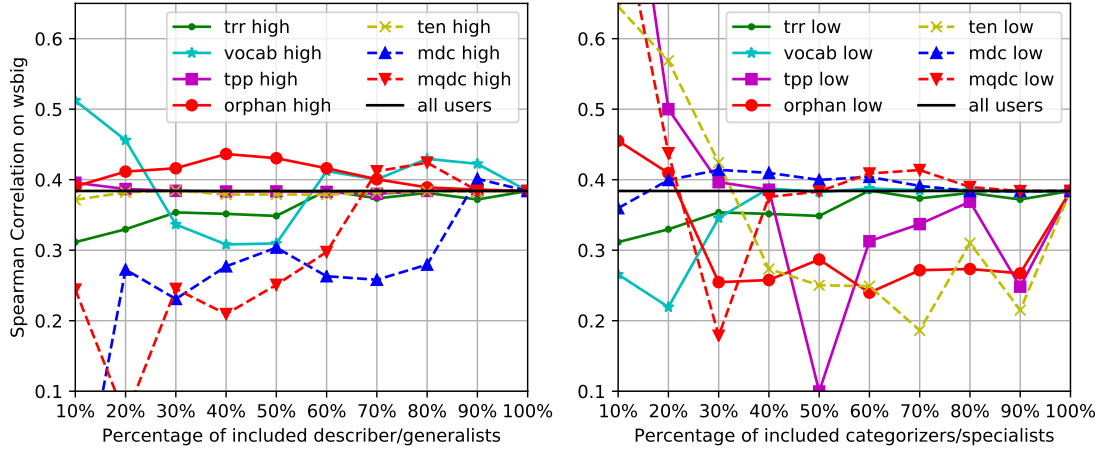
(c) Delicious results on Bib100.

Figure 6.1: Impact of tagging pragmatics on tagging semantics in Delicious. The x axis denotes the percentage of considered users, the y axis shows the Spearman correlation scores on the corresponding semantic relatedness datasets MEN, WS-353, and Bib100. Dashed lines indicate generalists/specialists, solid lines indicate describers/categorizers.

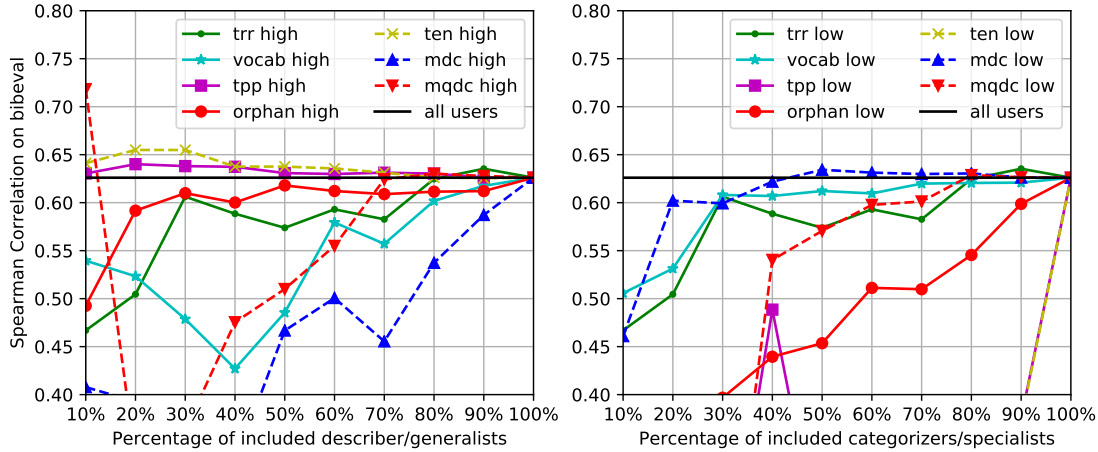
6.3 Re-Evaluating Measures of Semantic Tag Similarity on Human Intuition



(a) BibSonomy results on MEN.



(b) BibSonomy results on WS-353.



(c) BibSonomy results on Bib100.

Figure 6.2: Impact of tagging pragmatics on tagging semantics in BibSonomy. The x axis denotes the percentage of considered users, the y axis shows the Spearman correlation scores on the corresponding semantic relatedness datasets MEN, WS-353, and Bib100. Dashed lines indicate generalists/specialists, solid lines indicate describers/categorizers.

however also with only a weak signal. On WS-353 and Bib100 however, we see different results (cf. Figure 6.1b and Figure 6.1c). On Bib100, describers and generalists generate better semantics, with *vocab* and *trr* describers achieving the best correlation scores. However, also *mqdc* describers yield semantics better than all users on Bib100. This is especially interesting, as it directly contradicts our observation that *mqdc categorizers* yield better semantics on MEN. We will discuss this later.

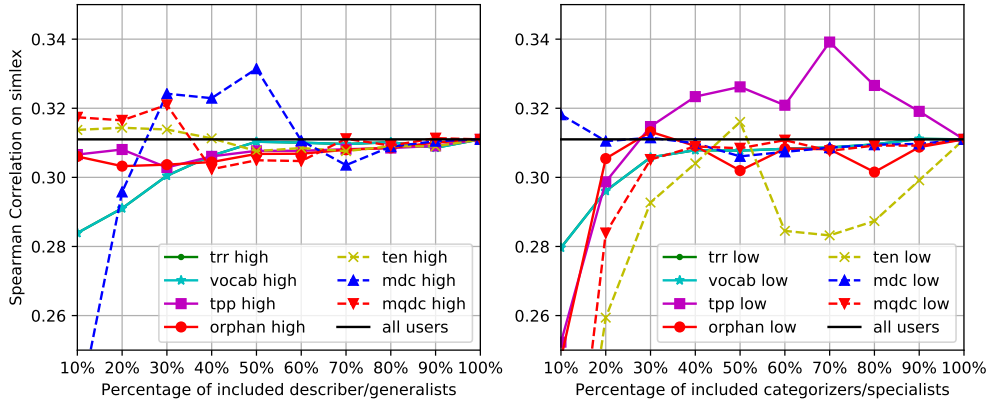
We additionally performed the experiments on the BibSonomy tagging dataset, as we wanted to see if the results from Körner et al. and our results on Delicious can generalize to other social tagging systems that are not too far away in terms of topical coverage. Figure 6.2 shows the results we obtained there. The first thing we can observe is that on MEN, *orphan* describers and *m(q)dc* specialists still yield the most useful semantic information, as we already found on Delicious. Additionally, this time 80% of *trr* categorizers also achieve a correlation above the baseline. On WS-353, we obtain on the one hand similar positive results as on MEN, such as *orphan* describers showing superior performance³ We also see some differences to the evaluation of Delicious on WS-353: First of all, the dominant user group are not *m(q)dc* generalists anymore, but *orphan* describers (like on MEN), but *vocab* categorizers also show decreased performance. Finally, on Bib100, we can roughly see a similar behavior of the evaluation scores as on Delicious, while the describers and generalists behave entirely different. Most notably, while describers/generalists on Delicious all outperformed the baseline after taking a portion of 60% of users into account, now almost all scores generated by describers and generalists are lower than the baseline, with the exception of extreme *tpp* describers and *ten* generalists. Finally, 90% of all *trr* describers also manage to achieve a score above the baseline.

Now when comparing these results to those obtained by Körner et al. [122], we can only see a limited amount of resemblance. While Körner et al. reported that they can achieve the best evaluation performance by taking 90% of *trr* categorizers into account, we can only see this reflected on WS-353, where we still achieve the best correlation scores by considering either 10% of *mqdc* specialists or 50% of *tpp* categorizers. The usefulness of *tpp* describers, as advertised in [122], can also only be confirmed on WS-353, where they achieve extremely similar results to *vocab* describers. On MEN on the other hand, we cannot see either pattern reflected. Instead, we see other regular patterns, such as for example the superiority of *mqdc* specialists. We explain the different results in the different evaluation datasets as follows. While MEN, Bib100, and in parts also WS-353, all aim to evaluate some kind of semantic *relatedness* between words, we might find different user groups that satisfy these conditions. On the other hand, as we noted in Section 6.2, the WordNet-based evaluation is based on comparison with the Jiang-Conrath measure. However, the Jiang-Conrath measure specifically measures semantic *similarity*, which on the other hand is not reflected very much in the evaluation datasets we mentioned above.

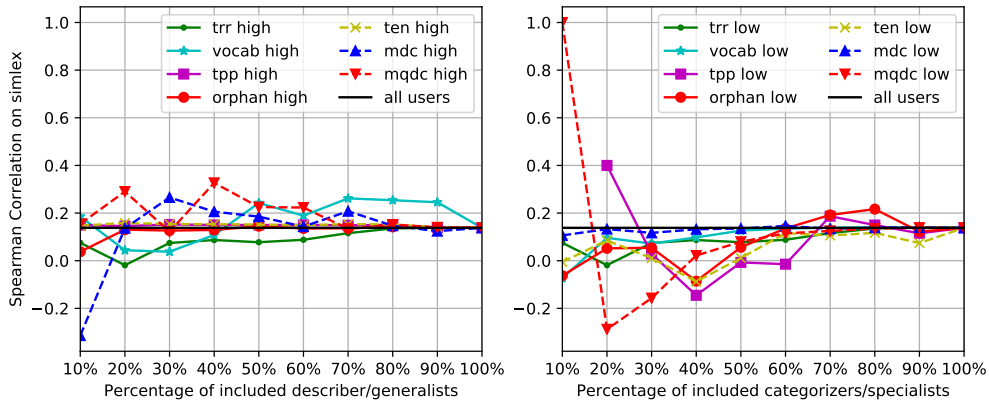
In fact, we also evaluated the produced semantic scores on the SimLex-999 dataset,

³We do not take the WS-353 scores obtained on small categorizer/specialist subsets into account, since we can only cover a small subset of WS-353 pairs there.

6.3 Re-Evaluating Measures of Semantic Tag Similarity on Human Intuition



(a) Delicious results on SimLex-999.



(b) BibSonomy results on SimLex-999.

Figure 6.3: Impact of tagging pragmatics on tagging semantics, evaluated on the semantic similarity evaluation dataset SimLex-999.

which resembles human intuition of semantic similarity, in Figure 6.3. We can see that on Delicious, *tpp* categorizers outperform the semantic baseline by a great margin, which is in line with [122]. However, *vocab* and *trr* describers cannot produce score that are better than that of all users. Still, in [122], these describer groups also only slightly outperformed the baseline score. On BibSonomy, *tpp* categorizers also yield a score slightly above the baseline, but here are outperformed by orphan categorizers. As for describers, we can again see that *m(q)dc* as well as *vocab* describers perform better than others. The *trr* describers again still perform similarly well as the baseline.

Conclusion

In this subsection, we re-evaluated the influence of user groups with a certain tagging behavior on the extraction of semantic information from tagging data. Concretely, we

repeated the experiments by Körner et al. [122], only this time we evaluated the semantic relatedness scores on human intuition instead of using a WordNet-based evaluation. We could show that tagging pragmatics can positively influence tagging semantics, but in many different ways. While in [122] only a large subset of users managed to slightly outperform the baseline of all other users, we found that, depending on the evaluation dataset and the measured type of semantic relatedness, other user groups were more useful. Concretely in relation to the results of [122], we only sometimes saw the same signals that they reported. So for example, while on MEN, we saw that We especially found that different user groups also address a different kind of semantic relatedness, something that [122] did not distinguish.

6.3.4 Summary

After we found in Section 6.2 that the Jiang-Conrath semantic similarity measure does not generalize well, we repeated the experiments of two papers that base their evaluation on the scores of the Jiang-Conrath measure. We found that the results from [48] hold to some extent, i.e., resource context similarity yields the best results, while user context similarity yields the worst results. However, tag context similarity and tag-tag co-occurrence were not as clearly separated as in [48] and sometimes even switched position in the ranking. We also repeated the experiments from [122], where Körner et al. investigated the influence of tagging pragmatics on semantic information of tagging data. While we found that their results are in general still valid, we observed different performances of user pragmatics when evaluating on different evaluation datasets. We especially noticed that some user groups are more useful for measuring semantic *relatedness*, while others benefit the determination of semantic *similarity*.

Overall, we could see that, despite the fact that the Jiang-Conrath measure does not generalize as well as assumed, the impact on the evaluation results from [48] and [122] is limited. Still, we could observe some differences, which validated our assumption that a second evaluation would also bring new and interesting results.

6.4 Influence of Tagging Pragmatics on Tag Sense Discovery in Folksonomies

Until this point, we assumed that each tag represents a single semantic concept. If the contexts of two tags were similar enough, we would call them *synonymous*, i.e., describing the same context. For this, we always aggregated all of the different contexts that tags appeared in (see Section 3.2.2.1). More realistically though, the *current* meaning of a tag, i.e., its tag *sense*, is always determined by its *current* context and does not always describe the same concept. For example, in Figure 6.4, the tag `apple` can refer to a well-known IT company or (more likely, judging from the context) to the fruit hanging from a tree. Because `apple` can have several meanings, it is called *polysemous*. This can introduce biases in measuring tag relatedness, since all the contexts of a tag are always compared with *all* contexts of the other tag, regardless of the current context. While we

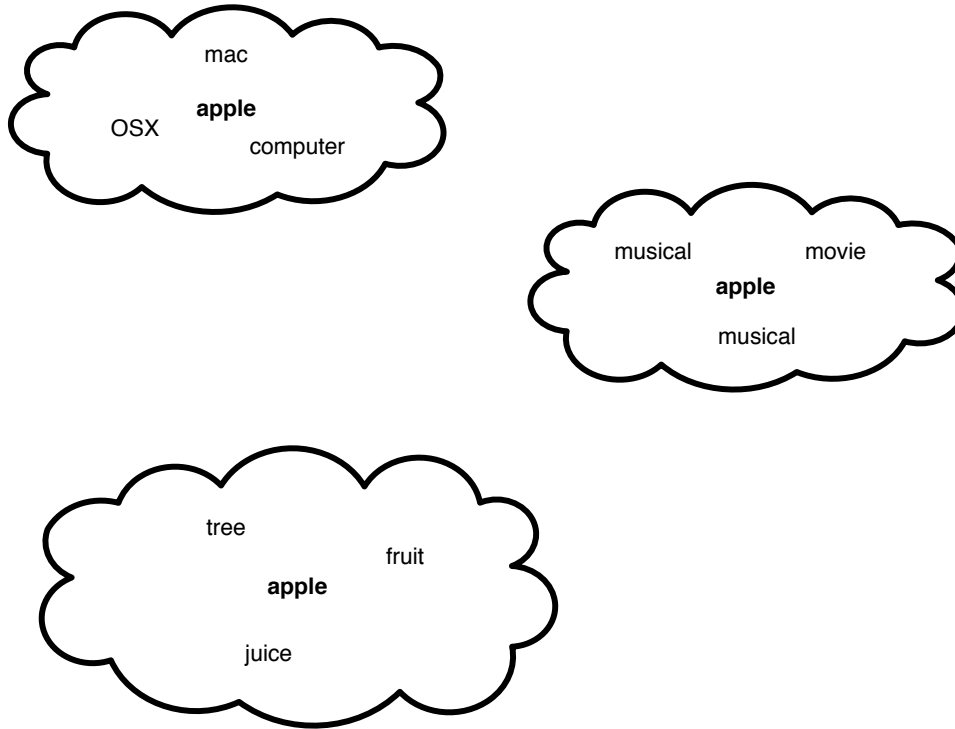


Figure 6.4: Illustration of different word senses for *apple*. Depending on the context, apple can refer to the fruit, to musicals in *The Big Apple* New York, or to the computer manufacturing company.

already learned in [122] as well as in the re-evaluation in Section 6.3.3 that the tagging pragmatics of users, i.e., the way users assign tags, impact tag semantics, we also assume that this is the case with polysemy detection.

In the preceding section, we re-evaluated an approach by Cattuto et al. to capture the semantic information in tagging data and to measure semantic tag relatedness and similarity, with the goal of discovering *synonymous* tags [48]. As an extension, we also reconsidered the work of Körner et al. how users of social tagging systems with different *tagging pragmatics influenced tagging semantics* [122]. We now build upon both works and apply their ideas on the tasks of *tag sense discovery* and *disambiguation*. Concretely, we use the semantic information model from [48] to measure the influence of user pragmatics on the quality of the discovered tag senses. For this, we use the pragmatic user measures that we defined in Section 3.3.1, which partially have been introduced in [123] and used in [122]. Furthermore, the actual sense discovery process is based on a hierarchical clustering approach of the context of a given tag [28]. We will specifically focus on tagging semantics in this section, since in social tagging systems, we already have a notion of user pragmatics, i.e., of categorizers and describers. Thus, this section will focus on *tag sense discovery* and the influence of *tagging pragmatics* in social tagging

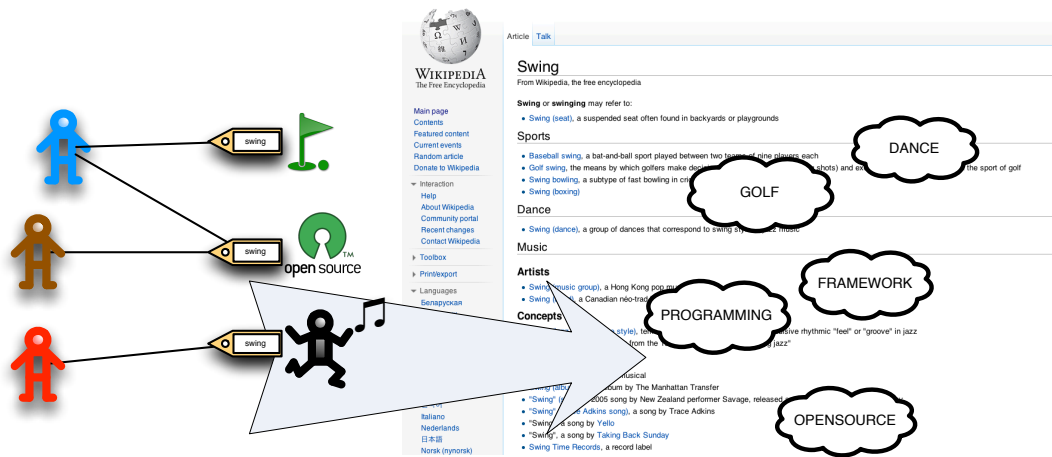


Figure 6.5: Evaluation of discovered tag senses on Wikipedia. A tag sense cluster matches a Wikipedia sense, if the sense cluster and the Wikipedia sense overlap in at least one word.

systems.

6.4.1 Experimental Setup

In order to explore the effects of tagging pragmatics on the ability to discover senses in tags, we set up a series of experiments where we apply the previously introduced method for tag sense discovery. Then we segment the entire folksonomy in several sub-folksonomies based on the pragmatic measures for distinguishing between different types of users and user behavior. Subsequently, we evaluate the performance of different subpopulations on this task. We start by describing how we obtained a “ground truth” for evaluation from Wikipedia, the WikiTagSenses dataset which we described in Section 4.1.5.

Evaluation of Discovered Tag Senses on Wikipedia Clearly, identifying a representative and reliable ground truth dataset which captures (most of) the different senses of a particular tag is a difficult task. While expert-built dictionaries like WordNet⁴ contain descriptions of different word senses, their coverage is limited (e.g., roughly 60% of top Delicious tags are present in WordNet). Furthermore, due to the dynamic nature of social tagging systems, “new” senses might emerge quickly which are not yet covered in the dictionary. For this reason, we have chosen the English version of Wikipedia⁵ as ground truth, as its coverage is higher (89% for BibSonomy, and 85% for Delicious) and we expect the community-driven sense descriptions to be more complete compared to

⁴<http://wordnet.princeton.edu>

⁵<http://en.wikipedia.org>

WordNet. The english Wikipedia provides about 4 million articles and covers a huge range of topics.

Our main source for sense descriptions are *disambiguation pages*, as introduced in Section 4.1.5. For a given term t , we first looked up its disambiguation page, and iterated over all contained bullet list items b_1, \dots, b_{n-1} . Because the first paragraph preceding the bullet list often describes the “standard meaning”, we added it as an additional item b_n . If no disambiguation page was available, we use the first paragraph of the corresponding article as a single sense description. The textual description for each item was then transformed into a bag-of-words representation by (i) splitting it using whitespace as delimiter, and (ii) removing stopwords and t itself. As a result, we obtain for each term t a set of Wikipedia sense descriptions WP_t^1, \dots, WP_t^n , each being essentially of a set of describing terms. We illustrated this in Figure 6.5.

Determining the clustering parameters As described in Section 3.2.5.2, a crucial aspect when using hierarchical agglomerative clustering for sense disambiguation is which parameter to use for “slicing” the dendrogram into distinct clusters. Because we are finally performing a gold-standard based evaluation of our clusters, our goal was to fix a parameter in advance for all experimental conditions which is oriented towards Wikipedia. Varying the distance threshold mainly influences the number of obtained clusters: While a very low threshold leads to many (singleton) clusters, a high value results in a single global cluster. So our goal was to set the threshold in such a way that the resulting distribution of senses comes close to the one found in Wikipedia. More precisely, we varied the threshold in steps of 0.05 between 0.0 and 1.95, recorded the distribution of the numbers of senses q , and compared it via the Kullback-Leibler divergence to the Wikipedia sense distribution p according to $D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$. First, we could observe a decrease of the KL divergence when increasing the distance threshold for both datasets - i.e., the distributions become more similar. After peaking at 0.45 (BibSonomy) and 0.55 (Delicious), the divergence becomes greater again. Based on this observation, we fixed the distance threshold for all hierarchical clustering conditions to 0.45 for BibSonomy and 0.55 for the Delicious dataset.

Evaluation metrics Based on the sense clustering SC_t^1, \dots, SC_t^m obtained for each tag t in each folksonomy partition, we evaluated the “quality” of each clustering by comparison with the corresponding Wikipedia senses WP_t^1, \dots, WP_t^n of t . A crucial question here is at which point a particular clustered sense SC_t^i “matches” a reference sense WP_t^j . We used a simple approach to this end and counted a “hit” when there existed an overlap between both sets, i.e., when $SC_t^i \cap WP_t^j \geq 1$. We refer with $matches(SC_t^1, \dots, SC_t^m)$ to the set of clustered senses which match at least one Wikipedia Sense, and with $matches(WP_t^1, \dots, WP_t^n)$ to those Wikipedia senses which match at least one clustered sense. While this represents only an approximate matching, inspection of a small sample of sense pairs revealed that the approach works reasonably well. Future research might focus on developing and applying more elaborate sense matching approaches. Based on these matches, we computed two measures inspired by precision and recall according

to:

$$precision(\{SC_t^1, \dots, SC_t^m\}, \{WP_t^1, \dots, WP_t^n\}) = \frac{matches(SC_t^1, \dots, SC_t^m)}{m}$$

and

$$recall(\{SC_t^1, \dots, SC_t^m\}, \{WP_t^1, \dots, WP_t^n\}) = \frac{matches(WP_t^1, \dots, WP_t^n)}{n}$$

6.4.2 Results and Discussion

Figure 6.6 depicts the quality obtained for different disambiguation conditions for the Delicious dataset. Along the x -axis of each plot, users are being added, sorted by each pragmatic measure, respectively. This means that the folksonomy partitions are growing towards the size of the full dataset – which is the reason that all lines meet in their rightmost point. The y -axis measures precision and recall as defined above. The black solid line corresponds to the random baseline, in which users were added in random order.

Table 6.8 gives a qualitative perspective of the implications of tagging pragmatics on tag sense discovery. While the senses inferred from extreme describers represent plausible tag sense clusters and accurate matches to Wikipedia senses⁶, the senses inferred from specialists or the entire population of users do not capture the different meanings of the tag *swing*.

Interpretation: When comparing with the baseline, a first observation is that most induced sub-folksonomies based on specialist and categorizer intuitions remain below the random baseline, with increasing quality towards the full dataset condition. This suggests that tagging data produced predominately by categorizers and specialists does not enhance performance of the tag sense discovery task.

For describers and generalists, the situation becomes more interesting: While many partitions based on generalists show a similar behavior and remain below the random baseline, those based on tag entropy (*ten*) and partially those based on mean degree centrality (1st quartile, *mqdc*) perform better, and score higher precision and recall values than the complete dataset. This effect is even more pronounced for partitions based on describers (using *trr*). It suggests that the pragmatics of tagging influence the performance of knowledge acquisition tasks such as tag sense discovery. *But how do the pragmatics influence tag sense discovery in detail?*

Our results offer preliminary *explanations*, identifying the particular types of behavior (such as extreme describers or extreme generalists) that outperform other types of behaviors (such as categorizers or specialists). On a general level, we can explain some ways in which tagging pragmatics influence tag sense discovery. For example, while categorizers and specialists in our experiments seem to negatively effect the ability to discover senses from tags, data produced by describers and generalists has demonstrated a potential to improve performance on this task. On a more specific level, we can observe

⁶<https://en.wikipedia.org/w/index.php?title=Swing&oldid=495762681>

6.4 Influence of Tagging Pragmatics on Tag Sense Discovery in Folksonomies

Table 6.8: Examples of detected senses produced by hierarchical clustering with the optimal distance criterion threshold for the tag **swing** based on different partitions of included users of the Delicious dataset. **Bold** sense labels were successfully matched with the corresponding Wikipedia senses based on Wikipedia disambiguation pages.

folksonomy subsets	sense label	sense tags
Top 10% Describers (trr)	opensource	video opensource tools blog software
	programming	tutorial tips howto reference web programming
	framework	development gui java ui framework library
	swt	swt
	game	game
	dance	dance
	golf	golf
Top 10% Specialists (ten)	ui	icon ui font
	gui	gui
	tutorial	tutorial tutorials
	reference	reference blog cool
	border	border
	trash	trash
	xml	development java xml programming
	test	test
	examples	examples
	mvc	mvc
	focus	focus
	cursor	cursor
	patterns	patterns
All users	gui	gui tutorial java tips library ui programming development
	instructions	callaway instructions cart carts pga clubs balls golf
	equipment	equipment accessories discount used
	courses	courses

that the best performance globally can be found for one of the smallest partitions, i.e., the one induced by 10% of describers. Their annotations (though technically consisting of much less data) seem to provide a better basis for discovering tag senses than the total amount of annotations in the system. One possible explanation lies in the intrinsic behavior of these users: Because their goal is to annotate resources with many descriptive keywords, it may not be surprising that they come closer to what Wikipedia editors do when “describing” word senses.

In order to verify the results on the Delicious dataset, we repeated our analyses on our second dataset (BibSonomy). The observations are consistent across our datasets, as can be seen in Figure 6.7

Discussion of implications: Understanding the ways in which tagging pragmatics influence tasks such as word sense discovery is appealing for several reasons. For example, using this kind knowledge, very large datasets could be reduced to smaller datasets, which exhibit better performance on such tasks. Also, system engineers could provide incentives to stimulate a particular style of tagging (e.g. through tag recommender systems), which may help to foster the emergence of more precise semantic structures.

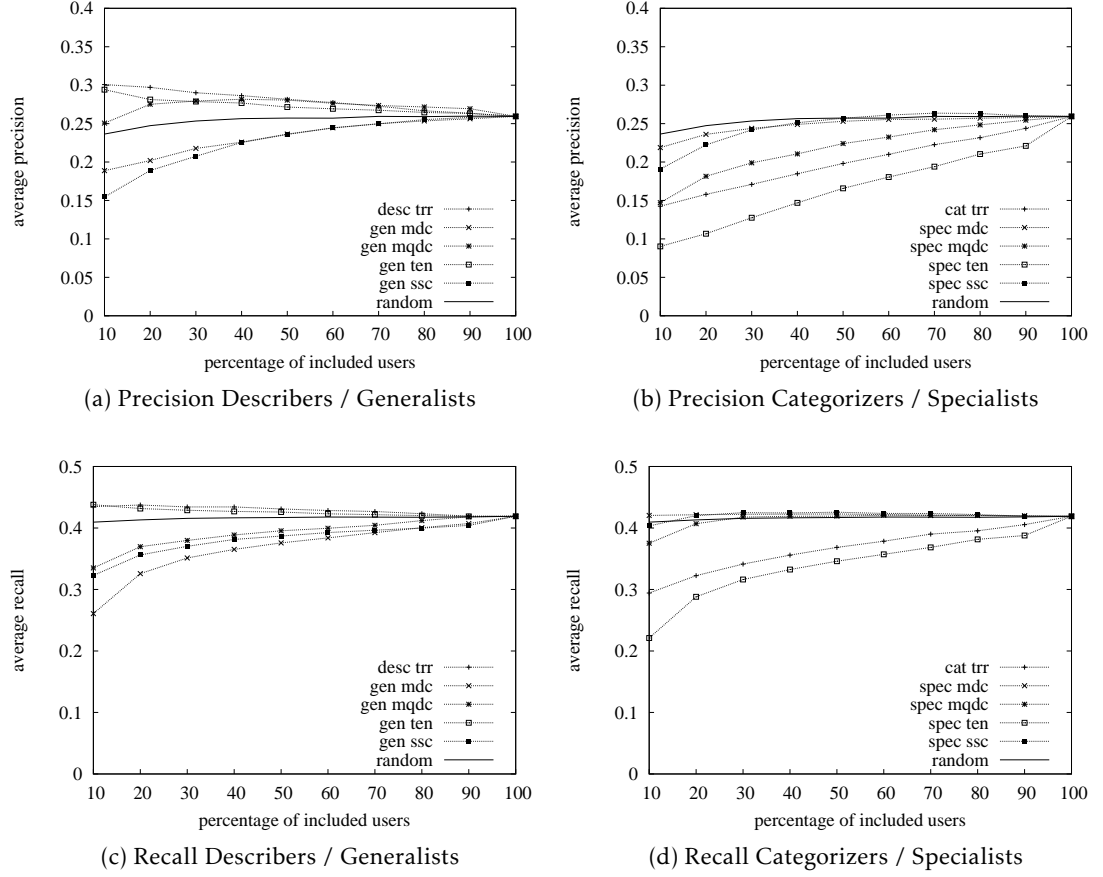


Figure 6.6: Results for the Delicious dataset. The x -axis of each plot corresponds to the percentage of included users, ordered by the different metrics (different lines). The further to the right, the larger are the corresponding folksonomy partitions. The y -axis corresponds to precision / recall as defined in Section 6.4.1. For the case of precision, higher values indicate a higher “correctness” of the discovered senses; for recall, higher values indicate a better “coverage” of Wikipedia senses. The solid line represents the random baseline. Most experimental cases stay close or below the baseline, i.e., they are not particularly well suited for disambiguation; An exception are small partitions consisting of describers (according to *trr*) and generalists (according to *ten* / *mqdc*).

6.4 Influence of Tagging Pragmatics on Tag Sense Discovery in Folksonomies

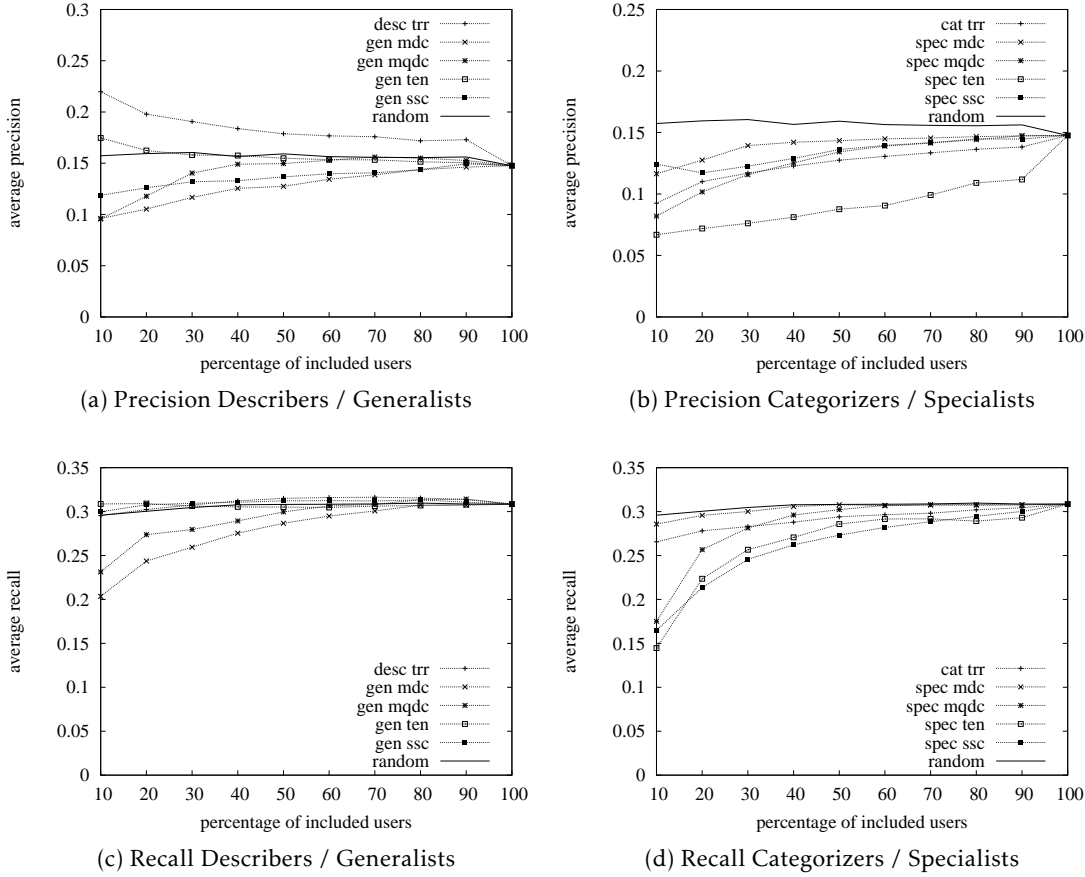


Figure 6.7: Results for the BibSonomy dataset. The x -axis of each plot corresponds to the percentage of included users, ordered by the different metrics (different lines). The further to the right, the larger are the corresponding folksonomy partitions. The y -axis corresponds to precision / recall as defined in Section 6.4.1. For the case of precision, higher values indicate a higher “correctness” of the discovered senses; for recall, higher values indicate a better “coverage” of Wikipedia senses. The solid line represents the random baseline. Most experimental cases stay close or below the baseline, i.e., they are not particularly well suited for disambiguation; An exception are small partitions consisting of descriptors (according to *trr*).

6.4.3 Conclusion

The overall objective of this section was to look for a *signal* - We wanted to explore (i) *whether there is a link between pragmatics and tag sense discovery* and (ii) *if there is, how it might be explained*. Our results provide further evidence that in social annotation systems, knowledge acquisition tasks such as tag sense discovery can not be viewed in isolation

from pragmatic factors, i.e., different kinds of users and user behavior. Our experiments demonstrate that tagging pragmatics can have an influence on the performance of tag sense discovery tasks. Our work also offers explanations, identifying the particular types of behavior (such as *extreme describers* or *extreme generalists*) that outperform other types of behaviors (such as categorizers or specialists). These findings represent an important stepping stone for future, more elaborate tag sense discovery methods that leverage pragmatic factors for improving performance. They also illuminate a way for engineers of social annotation systems to direct or influence user behavior in one or the other way to make their tagging data more amenable to a variety of knowledge acquisition tasks. In conclusion, our work further emphasizes the social-computational nature of social annotation systems, in which semantics emerge out of a combination of social user behavior with algorithmic computation.

6.5 Learning Tag Embeddings

In computational settings, tags are often represented by sparse, high-dimensional vectors [28, 29, 48, 122, 176, 205]. These vector descriptions are based on the co-occurrence of tags as described in Section 3.2.2.1. However, although we have shown in the previous section that tagging data contain meaningful semantics, the correlation of semantic relatedness scores from those vectors with human intuition still leaves room for improvement. One possible reason for this could be the high dimensionality of those representations. What is commonly known as the *curse of dimensionality* (cf. Section 3.2.1.2), impacts the measurement of vector similarity, since with increasing dimension of the feature space, the distance distribution between points approximates a uniform distribution, i.e., the distances between all vectors tend to become equal. Additionally, the high number of dimensions renders many learning algorithms computationally expensive, since they have to fit a large number of parameters. All prior studies rely on high dimensional tagging vectors or reduce the vector space arbitrarily by limiting the dimensionality of the space by a fixed number, e.g., to the top 10 000 tags [28, 122, 176], which in turn decreases the fit of the resulting relatedness scores to human intuition. Another option to decrease the dimensionality of those vector representations is to apply word embedding algorithms.

In this section, we propose to apply *word embedding algorithms* on tagging data to obtain a *dense representation* of tags in a *low-dimensional vector space*. We contribute a thorough exploration of three well-known embedding algorithms on tagging data, with regard to their general applicability and their optimal parameter choices. We first analyze the parameters of each algorithm, before we optimize these settings to produce the best possible word embeddings from tagging data. Then we compare the embeddings of each algorithm with each other as well as with traditional sparse representations by evaluating them on human intuition. We show that all produced embeddings outperform high-dimensional vector representations. We discuss the results in the light of other approaches to measure semantic relatedness and show that we reach competitive results, on par with recent work on extracting semantic information.

The contents of this section have been previously published in [174].

6.5.1 Applicability of Embedding Algorithms on Tagging Data

This section describes the different embedding algorithms that we explored. We already introduced each algorithm in Section 3.2.1.2. In the following, we will thus focus on the applicability and the hyperparameters of each algorithm. This includes a short summary, an enumeration of its parameters for each model and a short discussion how the model can be applied to tagging data.

6.5.1.1 Word2Vec

The most well-known embedding algorithm used in this work is the Word2Vec algorithm [156]. Word2Vec is actually comprised of two algorithms, SkipGram and CBOW (Cumulative Bag of Words).⁷ Word2Vec trains a shallow neural network on sequences of words to predict a word from its context, i.e., from its neighboring words in a given context window. We use the implementation provided in version 3.6.0 of the Gensim framework [194], which is publicly available as a Python package⁸.

Parameterization Word2Vec takes two parameters. The first parameter is the *window size*, which determines the amount of neighboring words in a sequence considered as context from which a word will be predicted. The second parameter denotes the *number of negative samples* that are used in the Noise Contrastive Emulation step of Word2Vec to approximate the gradient in the optimization step.⁹ The last parameter is a *minimum occurrence count* of words to be considered as sufficiently meaningful in a context. We only make use of the first two parameters, as we already filtered our datasets in a preparation step.

Applicability The Word2Vec algorithm normally processes sequential data. However, the sequence of tags normally does not hold any meaning, so this could possibly pose a problem if the window size is chosen too small. In order to be able to apply Word2Vec on tagging data, we grouped the tag assignments into posts and fed the random succession of tags as sentences into the algorithm.

6.5.1.2 GloVe

GloVe is an unsupervised learning algorithm for obtaining vector representations for words on aggregated global word-word co-occurrence statistics in a given corpus [184].

⁷In the course of this work, every time we refer to Word2Vec, we talk about the CBOW algorithm, as is recommended by [156] for bigger datasets.

⁸<https://pypi.org/project/gensim/>

⁹For a more detailed explanation, see <http://ruder.io/word-embeddings-softmax/index.html#noisecontrastiveestimation>

Its main objective was to capture semantic relations such as *king* – *man* + *woman* \approx *queen*. We use the C implementation provided by Pennington, Socher, and Manning.¹⁰

Parameterization The main parameters of the GloVe algorithm are x_{max} and α . x_{max} denotes an influence cutoff for frequent tags while α determines the importance of infrequent tags. According to [184], GloVe worked best for $x_{max} = 100$ and $\alpha = 0.75$. We will choose these as initial values in our experiments.

Applicability Since GloVe depends on co-occurrence counts of words in a corpus, it is very easy to apply on tagging data. For this, we construct the tag-tag-context co-occurrence matrix and can then directly feed it into the algorithm.

6.5.1.3 LINE

The goal of the LINE embedding algorithm is to create graph embeddings where the first- and second-order proximity of nodes are preserved [227]. As introduced in Section 3.2.2.2, the *first-order proximity* in a network is the *local* pairwise proximity of two nodes, i.e., the weight of an edge connecting these two nodes. The *second-order proximity* of two nodes (i, j) in a network is then the similarity between their *first-order neighborhoods*. To produce the LINE embeddings, we use a implementation in C by Tang et al.¹¹

Parameterization LINE takes two different parameters: The amount of edge *samples per step* and the amount of *negative samples* per edge. To decrease complexity of solving the proposed model in [227], the authors employed a noise contrastive estimation approach as proposed by [156] using negative sampling. Furthermore, to avoid high edge weights to outweigh lower weights by letting the gradient explode or vanish, LINE employs a sampling process of edges and then ignoring their weight instead of actually using the edge weights in its objective function.

Applicability Similar to GloVe, this algorithm processes a network with weighted edges, such as a co-occurrence network. Thus, we only have to construct the co-occurrence network from the tagging data and apply LINE on that network.

6.5.1.4 Common Parameters

While each of the following algorithms can be tuned with a set of different parameters, there are some parameters common to all algorithms. First, the *embedding dimension* determines the size of the produced vectors. A higher embedding dimension allows for more degrees of freedom in the expressiveness of the vector, i.e., it can encode more information about word relations. Standard ranges for low dimensions are between 25 and 300.

¹⁰<https://github.com/stanfordnlp/GloVe>

¹¹<https://github.com/tangjianpku/LINE>

Table 6.9: Spearman correlation values for the co-occurrence baseline. For all evaluation datasets, we give the total number of pairs in the original dataset and the number of matched pairs, i.e., where both words were present in the tagging corpus.

	WS-353 (353)	MTurk (287)	MEN (3000)	Bib100 (100)
BibSonomy	0.384 (158)	0.548 (62)	0.433 (463)	0.626 (100)
Delicious	0.450 (202)	0.505 (103)	0.574 (1376)	0.636 (94)
CiteULike	0.186 (139)	0.469 (53)	0.423 (404)	0.270 (87)

Secondly, the *initial learning rate* of an algorithm determines its convergence speed. Fine-tuning that parameter is crucial to receive optimal results, because if chosen badly, the learning process either converges very slowly or might be unable to converge at all.

6.5.2 Experimental Setup

In the following, we describe the baseline and the initial parameter settings for each algorithm. We perform all experiments on the preprocessed tagging datasets introduced in Section 4.1, concretely on Delicious, BibSonomy and CiteULike.

Baseline: Tag-Tag-Context Co-Occurrence Vectors As a baseline, we produced high dimensional co-occurrence counting vectors from all three tagging datasets. As described in Section 3.2, we counted co-occurrence of tags in a tag-tag-context, i.e., the context of a tag was given as the other tags annotated to a given resource by a certain user (cf. Equation (3.17)). Since there is no option to vary the dimension of the tag-tag-context co-occurrence vectors except truncating the vocabulary, we only report the values for a truncated vocabulary of 10,000 tags, which is in line with [48].

From Table 6.9, we can see that vectors built on Delicious data mostly produce the best results and can always match most of the evaluation vocabulary, compared to the other two tagging datasets. At least the second fact is not surprising, as Delicious is by far the biggest tagging dataset. At the same time, the size of Delicious also intuitively accounts for a lower signal-to-noise ratio in the top 10,000 tags. We can thus expect more clear-cut results compared to the BibSonomy and CiteULike datasets.

Parameter Settings for each Algorithm For each of the following algorithms, we conducted the experiments as follows: As initial parameter setting, we used the standard settings that come with the implementation of each algorithm. The corresponding values are given in Table 6.10. We then varied the initial learning rate for each algorithm in the range of 0.01 to 0.1 in steps of 0.01. After that, we varied the embedding dimension on the set of {10, 30, 50, 80, 100, 120, 150, 200}. For Word2Vec and LINE, we now varied the number of negative samples on the set of {2, 5, 8, 12, 15, 20}. For GloVe, we varied $x_{max} \in \{25, 50, \dots, 200\}$ and $\alpha \in \{0.5, 0.55, \dots, 1\}$ simultaneously. Finally, for Word2Vec, we varied the context window size between {1, 3, 5, 8, 10, 13, 16, 20}, while for LINE, we varied the number of samples per step on $\{1, 10, 100, 1000, 10000\} \cdot 10^6$. To rule

Table 6.10: Initial parameter values for each algorithm. If a parameter is not applicable for an algorithm, it is marked with a dash.

Algorithm	learning rate	dimension	samples per step	negative samples	(x_{max}, α)	window size
LINE	0.025	100	$100 \cdot 10^6$	5	-	-
GloVe	0.05	100	-	-	(100, 0.75)	-
Word2vec	0.025	100	-	5	-	5

Table 6.11: Best parameter values for each algorithm for the MEN dataset on Delicious data.

Algorithm	learning rate	dimension	samples per step	negative samples	(x_{max}, α)	window size
LINE	0.1	100	$100 \cdot 10^6$	15	-	-
GloVe	0.1	120	-	-	(100, 0.75)	-
Word2vec	0.1	100	-	20	-	5

out influence of a random embedding initialization, each experiment was performed 10 times and the mean result was reported. After each experiment, we chose the best performing parameter settings on the respective tagging datasets across the four evaluation datasets and used them for all other experiments.

6.5.3 Results

We will now present the evaluation results. We trained each algorithm on the tagging data from Delicious, CiteULike and BibSonomy (cf. Section 4.1) and evaluated them on the WS-353, MEN, MTurk, and Bib100 datasets (cf. Section 4.4) using the Spearman correlation coefficient (see Section 3.2.4). For space reasons, we only report the evaluation results on MEN. For each algorithm, Table 6.11 gives the parameter settings which produced the highest-scoring embeddings. In Figure 6.8, Figure 6.9, and Figure 6.10, we report both the evaluation results of the embeddings for a given parameter as well as the corresponding baselines produced by the high-dimensional vector representations given in Table 6.9. Figure 6.11 finally shows a comparison of the evaluation results for the best configurations of each algorithm.

Word2Vec Although Word2Vec is meant to be applied on sequential data, as opposed to the bag-of-words nature when assigning tags, the generated embeddings yielded better correlation scores with human intuition than their high-dimensional counterparts. However, we did not shuffle the tag sequence in posts, which is left to future work. Figure 6.8a shows that fine-tuning the *initial learning rate* exhibits a great effect on the quality of word embeddings from BibSonomy, with general peak performance at $\alpha = 0.1$, while Delicious data seem unaffected. Increasing the *embedding dimension* only improves the embeddings' semantic content up to a certain point, which is mostly reached at around a very low number of dimensions between 30 and 50 (Figure 6.8b). Anything above that does not notably add to the performance of the embeddings. The number

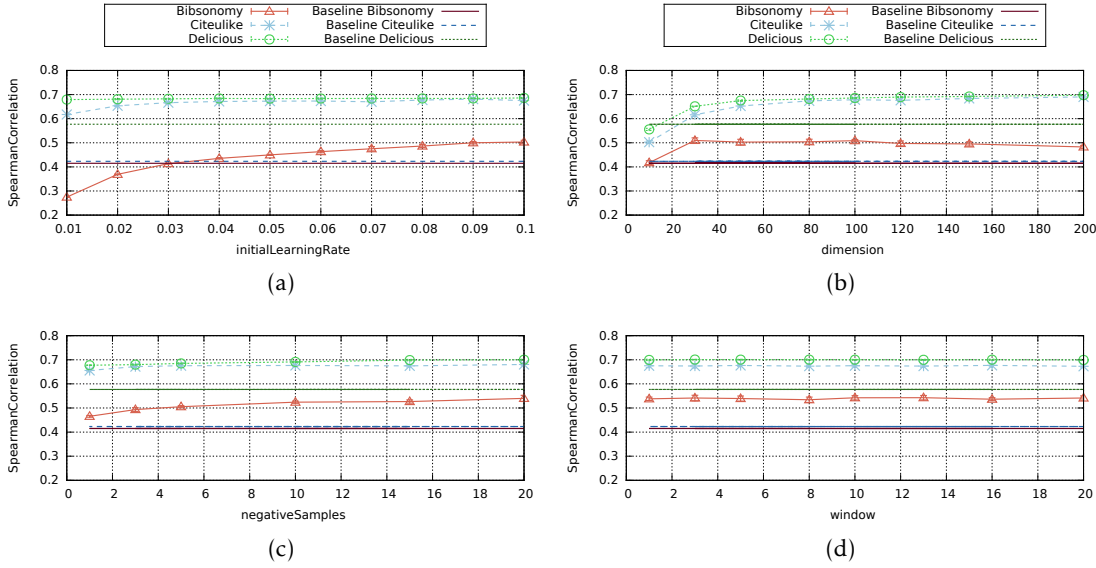


Figure 6.8: Evaluation results for embeddings generated by Word2Vec on MEN. For CiteULike and BibSonomy, tuning the learning rate notably improves results. The embedding dimension seems to be the best for all tagging dataset at 100; afterwards the quality of BibSonomy embeddings decreases again. As reported in [156], a number of around 10 negative samples seems sufficient for small datasets such as BibSonomy, while a lesser number of samples fits for bigger datasets.

of negative samples seems to be sufficient at around 10 samples and even earlier for Delicious and CiteULike (Figure 6.8c). The influence of the *context window size* on the semantic content of the generated embeddings was negligible (Figure 6.8d). In the end, we obtained our best results for Word2Vec on Delicious tagging data with a correlation score of 0.7 with MEN, which is already a major improvement upon the baseline correlation of 0.577 (cf. Table 6.9).

GloVe GloVe generates embeddings from co-occurrence data. As mentioned in Section 6.5.1.2, GloVe is parameterized by the learning rate, the dimension of the generated embeddings as well as by the weighting parameters x_{max} and α , which regulate the importance of low-frequency co-occurrences in the training process. While the *learning rate* does not seem to have great effect on embeddings generated from Delicious data, fine-tuning influences the semantic content of CiteULike and BibSonomy embeddings notably (Figure 6.9a). Mostly, peak performance is reached at an *embedding dimension* around 100 or even earlier, except for Delicious (Figure 6.9b). Furthermore, BibSonomy is quite sensitive to poor choices of x_{max} and α , i.e., if both are chosen too high, performance suffers greatly (Figure 6.9c). However, Delicious and CiteULike seem unaffected by those parameters, at least in our experimental ranges (Figures 6.9d and 6.9e). The

6 Capturing Semantics in Social Tagging Data

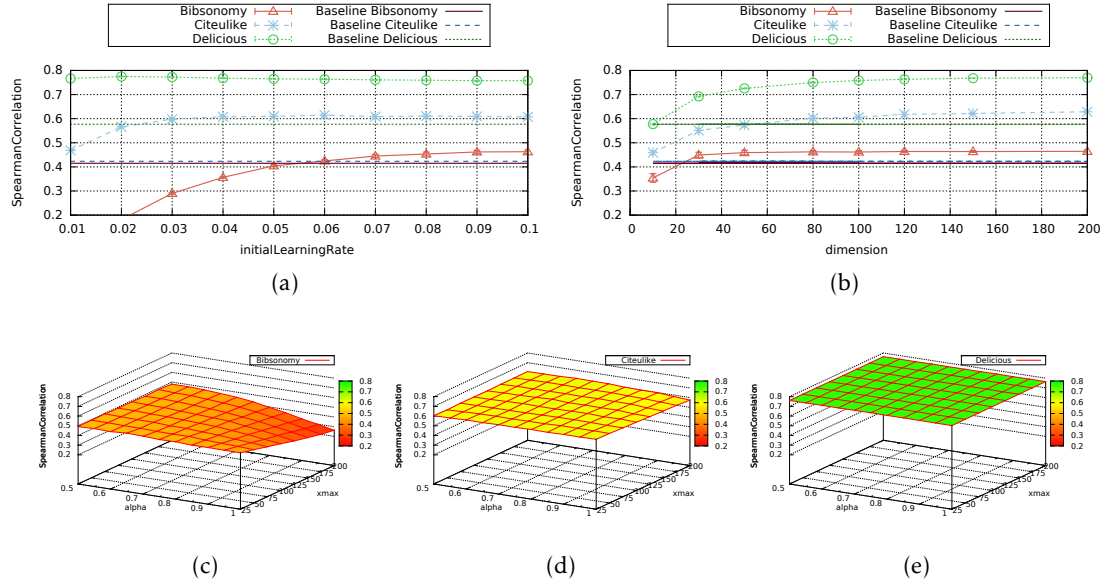


Figure 6.9: Evaluation results for embeddings generated by GloVe on the MEN dataset. The initial learning rate only influences the smaller tagging datasets, while Delicious profits most from increasing dimension. BibSonomy is influenced by a high cutoff x_{max} the most.

GloVe embeddings trained on Delicious achieve the best result on MEN with a score of 0.76 correlation, compared with a score of 0.7 by Word2Vec. Most notably, the GloVe embeddings capture human intuition significantly better than the tag-tag co-occurrence vectors by Cattuto et al., which only achieve a correlation of 0.577. Thus, we can outperform the original result by almost 20% correlation score. In fact, GloVe is the only algorithm that consistently achieves better correlation scores than the co-occurrence counting baseline in all tagging datasets across all evaluation datasets.

LINE LINE generates vertex embeddings from graph data, preserving the first- and second-order proximity between vertices. Its parameters are the initial learning rate, the embedding dimension, the number of negative samples per edge and the number of samples per training step. Its results are given in Figure 6.10. While influence of the *initial learning rate* is visible, it is not as great as with GloVe. Also, the *embedding dimension* gives similar results above 50 and only lets performance suffer if chosen too small. Interestingly enough, the number of *negative samples* seems to have almost no effect on the generated embeddings across all tagging datasets. In contrast, choosing the number of *samples per step* exerts great influence on the resulting embeddings. The best correlation with MEN we can achieve with LINE on Delicious tagging data is slightly below 0.7 (cf. Figure 6.10d), slightly lower than that of Word2Vec.

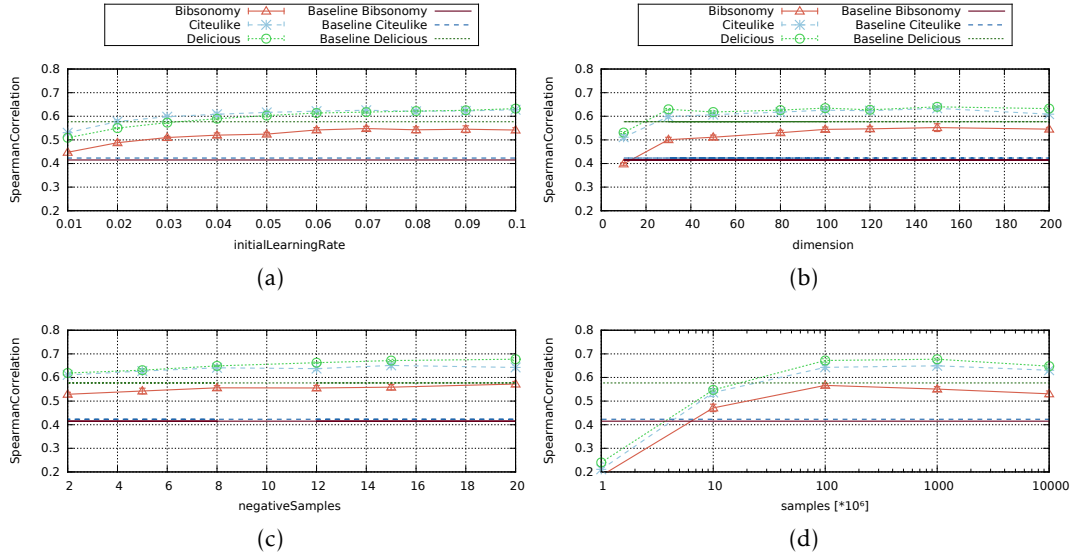


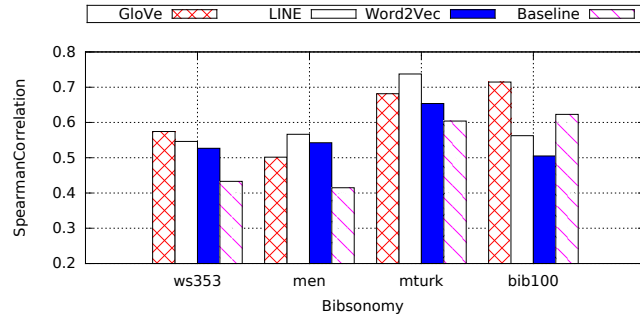
Figure 6.10: Evaluation results for embeddings generated by LINE on the MEN dataset. While the initial learning rate, the embedding dimension and the amount of samples per step exert a notable influence on the evaluation result, increasing the number of negative samples per edge only slightly improves results.

6.5.4 Discussion

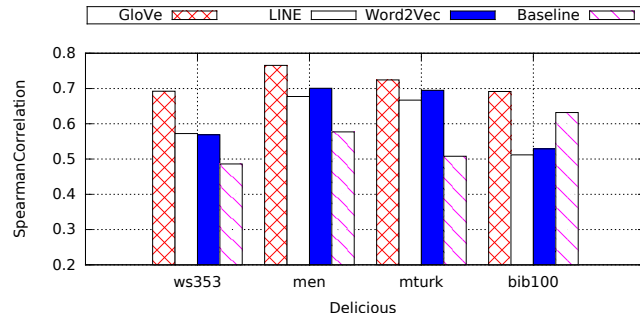
Across all algorithms, fine-tuning the *initial learning rate* greatly improves results for embeddings based on BibSonomy, especially with GloVe. The effect of the *embedding dimension* is much less pronounced across all three embedding algorithms. Peak evaluation performance is often reached with an embedding dimension between 50 and 100 and stays quite stable with increasing dimension. Varying the number of *negative samples* influences evaluation results of BibSonomy, but only at a very high number of 20 negative samples. In contrast, Delicious and CiteULike only show small performance changes already with 3 to 5 samples. Finally, GloVe’s *weighting factors* x_{max} and α negatively influence results on BibSonomy, while barely affecting evaluation performance on Delicious and CiteULike, due to BibSonomy being our smallest tagging dataset with rarely any co-occurrences above a high x_{max} .

In Figure 6.11 we finally compared the best achieved scores from each algorithm on the different tagging data. We can see that all investigated embedding algorithms produce usable and high-quality embeddings from tagging data. Concretely, Although in [48] it was found that tagging data contain high-quality semantic information, the standard high-dimensional vector representation proposed there does not seem to capture this information very well, when evaluated on human judgment (see Table 6.9). In contrast, the generated embeddings seem better suited to capture that information, as they almost always outperform the tag-tag-context based co-occurrence count vectors (Figure 6.11).

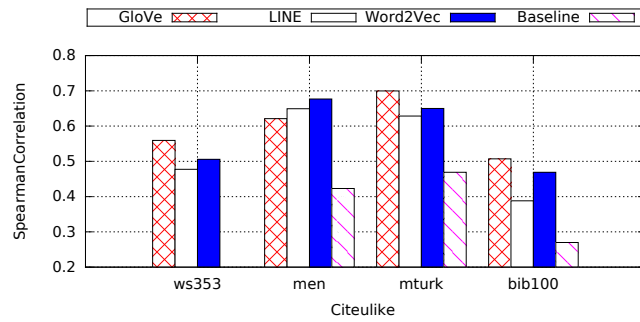
6 Capturing Semantics in Social Tagging Data



(a) BibSonomy



(b) Delicious



(c) CiteULike

Figure 6.11: Evaluation results for embeddings generated by the best parameter settings across the different tagging datasets. GloVe mostly produces the best embeddings and is the only algorithm to always outperform the baseline.

Furthermore, the best result achievable on WS-353 in this work is from Delicious data using the GloVe algorithm of around 0.7 (cf. Figure 6.11b), which is on par with with other well-known works, such as ESA [81], achieving correlation around 0.748, or the work done by Singer et al. on Wikipedia navigation [210] with the highest correlation at 0.76, but generally achieving scores around 0.71 (cf. also Section 7.2.2).

6.5.5 Conclusion

In this section, we explored embedding methods and their applicability on tagging data. We conducted parameter studies for three well-known embedding algorithms in order to achieve the best possible embeddings based on tagging data regarding their fit to human intuition of semantic relatedness.

Our results indicate that i) tagging data provide a viable source to generate high-quality semantic embeddings, even on par with current state-of-the-art methods and ii) that in order to achieve competitive results, it is necessary to choose correct parameters for each algorithm instead of the standard parameters. Overall we bridged the gap between the fact that tagging data yield considerable semantic content and the current state-of-the-art methods to produce high-quality and low-dimensional word embeddings. We expect our results to be of special interest for folksonomy engineers and others working with semantics of tagging data.

Future work includes investigation of the influence of different vector representations on tagging-based real-world applications, such as tag recommendations in social tagging systems, tag sense discovery and ontology learning algorithms.

6.6 Summary

In this chapter, we explored new ways to represent and evaluate semantics from social tagging data as well as the impact on existing research. Additionally, we investigated the influence of tagging pragmatics on another type of semantic information.

We first argued that evaluation directly on human intuition of semantic relatedness yields more valid results than evaluation on WordNet. Since that impacts the results of existing research, we repeated experiments from [48] and [122] that were evaluated using WordNet. We found that in general, their results still hold in a qualitative way. However, we also found some differences between the results that we obtained and what they reported. In addition to this, we showed that user pragmatics also influence the tag sense discovery process in folksonomies. Finally, we explored the applicability of word and graph embedding algorithms on tagging data and how this impacts the representation of semantic information in tagging data. We found that tagging embeddings can in general better capture the semantic information in tagging data.

With these results, we advanced the standard way of representing social tags as vectors and connected it to the current state-of-the-art in semantics. We could also show that social tagging semantics yield even competitive correlation scores with human intuition, thus again supporting the claim that tagging data contain considerable amounts of semantic information. Open problems are for example finding a evaluation routine that

on the one hand can closely resemble human intuition, but also is less limited in the covered vocabulary. Another issue is that although we saw in Section 6.2 that resource context of tags yields by far the most valuable semantics, we cannot straightforwardly apply the tagging algorithms on it as in Section 6.5, since e.g., GloVe depends on a quadratic co-occurrence matrix instead of a variant of the term-document matrix, as defined by the resource context. Finally, it would be interesting to apply the learned tag embeddings to different applied tasks, such as tag sense disambiguation [176] or taxonomy construction [27]. A next step would then be to even develop new embedding-based techniques to perform these tasks.

Chapter 7

Extracting Semantic Relatedness from Social Media Navigation

7.1 Introduction

In Chapter 5 we showed that there is strong evidence for a semantic component in human navigation behavior, both on Wikipedia and BibSonomy. Similarly to how humans express their thoughts by putting words into semantically meaningful sequences, we assume that they navigate information networks by following a semantically meaningful trail of webpages. We thus assume that we can extract the information contained in these sequences. Specifically, we look at navigation information from Wikipedia and BibSonomy.

Extracting semantic information from human navigation is attractive in several ways. First and foremost, users do not have to actively and consciously contribute complex information. As Hargittai and Walejko found, content in the web is actively contributed by about 1% of the total number of consuming users [95]. In contrast to that, navigating a website such as Wikipedia or BibSonomy does not require users to actively contribute content. Still, the way *how and where users navigate*, also contains latent information about their interests and thoughts. Another great advantage of such navigational paths by humans, compared to actively contributed information, is that it can be captured in a very simple way. The only prerequisite is that there is a group of users that navigate a system. Furthermore, many existing methods only work well if the system at hand provides high quality content that can be leveraged for calculating semantic relatedness.

Secondly, webpages in Wikipedia often describe singular concepts, while on BibSonomy, they mostly describe an entity from the folksonomy structure (cf. Section 3.1.2 and Section 3.1.1.2). We thus do not have to tackle issues that are usually encountered in unstructured text such as polysemy, i.e., a word has different meanings, depending on its context. Third, we do not have to filter navigational data as thoroughly as unstructured text. For example, page names cannot misspelled and there are rarely such things as “stopword pages”. Finally, as we will see in the course of this chapter, we need only a fraction of the data to obtain meaningful and valid results when measuring semantic relatedness.

Despite the advantages of human navigation over natural language text, there are also certain drawbacks. For example, users navigate with different intentions. In Section 3.3.2.1, already described that users navigate Wikipedia with different intentions, such as information need or boredom. While in the first case, navigation can be short and less resourceful, since a user stops navigating after having retrieved the desired information, users in the second case do not necessarily show a single topical interest. This could in turn blur the information contained in navigation. Second, although there are rarely “stopword pages”, most systems still provide functionalities to support users in navigation, which however do not carry stable semantic information. Examples of these pages are the landing pages of Wikipedia and BibSonomy, which contain information about recent events or recently posted publications, yet do not show the same content over a long period of time. Third, as navigation data can contain sensitive information, it is hard to obtain large collections of real-world navigation data. Concretely, navigation on the public Wikipedia also contains traces of users navigating to their own user page. Even after discarding information about the logged-in user in the requests, this behavior still allows to make a good guess about the navigating users identity. The same holds for navigation in BibSonomy, where Doerfel et al. showed that users mostly navigate on their own resources [64].

In this chapter, we will extract the semantic information contained in navigation on Wikipedia and BibSonomy. We propose several methods to model the semantic information as vectors, similar to the vector space model described in Section 3.2.1.1. These methods are applied on game and unconstrained navigation data on Wikipedia and unconstrained navigation on BibSonomy. To account for the lack of “real” human navigation data, we propose to generate random walks on both systems, which we will parameterize with different navigational priors. We find that navigation on Wikipedia contains a high amount of semantic information, regardless of the respective setting. On the other hand, we were unable to obtain semantically meaningful tag vectors from BibSonomy navigation data.

We split our analyses into two parts as follows: Analogously to Chapter 5, we first focus on *navigation in Wikipedia*. Here, we begin with introducing two distributional methods to extract semantic information from navigation on Wikipedia. We first apply them on navigation in a game setting and validate our results on data from the WikiGame and Wikispeedia (cf. Section 4.2.1). After this, we apply the same methods to unconstrained navigation on Wikipedia. Here we explore data from the Indiana click dataset [151] (WikiClickIU) as well as navigation collected on the Wikipedia servers (ClickStream, cf. Section 4.2.2). In order to overcome data sparsity of navigational information, we lastly simulate human navigation behavior by performing random walks on the Wikipedia graph. At the end of that section on Wikipedia navigation, we relate the results to each other. After having thoroughly analyzed the semantic information in Wikipedia navigation, we focus on *navigation in the social tagging system BibSonomy* (cf. Section 3.3.2.2). The structure of the second part of this chapter is similar to the first part on Wikipedia navigation. First, we propose two distributional semantic information extraction methods, before we investigate how we can use them to extract semantic information from unconstrained BibSonomy navigation. After this, we experiment with

generating random walks on BibSonomy from several navigation hypotheses which we already used in Section 5.3. Finally, we *summarize the findings* of this chapter.

The content of this chapter is based on several publications. The analysis of semantics in game navigation has been presented in [210]. We extend that work by also performing the experiments on data from Wikispeedia. The subsequent extension to unconstrained navigation was performed in [175]. Lastly, the idea to perform random walks to simulate human navigation was introduced in [55]. However, we extend this work by performing random walks that are parameterized by the navigational hypotheses proposed in Section 5.2.1.

The transfer of the Wikipedia based methodology to extract semantic relatedness from navigational paths on to the BibSonomy social tagging system was described in [172]. All random walk experiments on BibSonomy have however been conducted especially for this thesis and have not been published before.

7.2 Wikipedia

When navigating a set of articles on Wikipedia, users typically need to tap into their intuitions about real-world concepts and the perceived relationships between them in order to progress towards their set of targeted articles. Humans tend to find rather *intuitive* than *short* paths, while in contrast, an automatic algorithm would try to find a shortest path between two concepts that may not be as semantically rich and intuitive as a navigational path conducted by a human. In this section, we want to tap into the semantic richness of human navigational paths on Wikipedia.

Especially for game navigation collected in the Wikipedia navigation game Wikispeedia, West, Pineau, and Precup introduced a semantic distance measure for Wikipedia concepts [249] based on navigation game paths. While this work serves as the inspiration for this chapter, their proposed semantic distance measure suffers from some design choices. Their semantic distance measure $d(a, g)$ can only compute distances between concepts a , which have to occur on a path p with a goal g , and that goal g . This means that $d(a, g)$ is not defined if (i) a has never been visited in any game with target g or if (ii) g was never the goal of a game. This *goal* and *path dependency* of that measure however strongly limits its ability to measure the semantic distance between arbitrary concepts in Wikipedia.

In Section 7.2.1, we will first introduce two approaches to extract semantic information from human navigation on Wikipedia that are both *path* and *goal independent*. This enables us to measure the semantic similarity between any two concepts in Wikipedia, even if they never co-occurred in the same navigation game. We evaluate both measures in game settings, but are also able to evaluate them in *unconstrained* settings, i.e., without a predefined goal. Concretely, we use the same navigation datasets that we already analyzed in Section 5.2. We find that using our model, we can extract high quality semantic information from human navigational paths on Wikipedia, in both game and unconstrained settings alike. However, as the semantic information cannot be extracted as precisely as on game data when using co-occurrence counting, we need to

modify the extraction method. Finally, we investigate different sets of random walks based on different navigational hypotheses to generate human-like navigation with regard to their extractable semantic information. The reason for this is that both game and unconstrained navigation datasets are in some way limited, either their general setting, their size or the dataset design. For example, the ClickStream data only contain accumulated page-to-page transitions over the course of a whole month, and all user and longer path information was discarded entirely.

7.2.1 Methods

To extract semantic information from navigation on Wikipedia, we assume that navigational paths exhibit a similar structure as unstructured text. Based on this, we introduce two distributional methods to represent *pages* as *co-occurrence based vectors*. Both methods are inspired by context based word characterizations as used for text [230] (cf. Section 3.2.1.2). The first variant counts the co-occurrences of pages in a path if they occur near each other. The second variant only notes if two pages co-occurred in a path once or not at all.

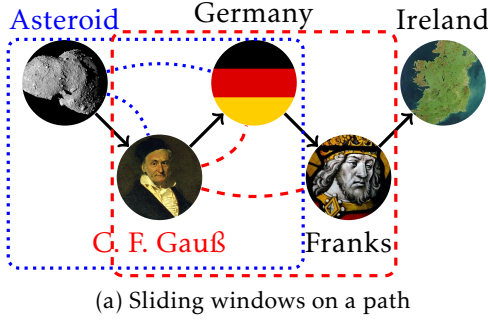
7.2.1.1 Count-based Navigational Semantic Relatedness

The first method to generate distributional semantic vector representations of entities from social media data is to count co-occurrences of those entities in a certain context window. Similar methods to represent entities by their context have already been presented for social tagging data [48] and natural language text [156, 184, 230]. Analogously to this intuition, navigational paths exhibit sequential structures, similar to sentences in natural language text. Thus, the pages in a path can be seen as semantically coherent, i.e., we interpret a navigational path as a type of semantic context, as explained in Section 3.2.2.2.

However, as noted in [247], the semantic coherence is only present for pages not far from each other. Because of this, we further limit the considered context of a page p_i in a path $\mathbf{p} = (p_1, \dots, p_i, \dots, p_n) \in \mathbb{P}$ to only a *window* of size w in front of it, to only capture meaningful relations. As we do this for every page p_i , $i = 1 \dots n$, we call this a *sliding window*. We illustrated this intuition in Figure 7.1a. To construct the vector representation v_i of a page p_i based on the sliding window context, all co-occurrences of pages inside such a window are counted symmetrically. This means that we not only count the forward co-occurrence of $(p_{i,k}, p_{i+1,k})$ in \mathbf{p}_k , but also the reverse occurrence of $(p_{i+1,k}, p_{i,k})$, although this actual navigation step does not necessarily exist.

Analogously to Section 3.2.1.2, the entries $coocc(p_i, p_j)$ of the vector v_i for a page p_i are based on the navigational context of human navigation with a maximum context window size of w . Formally, this can be expressed as follows:

$$coocc(p_i, p_j)_w := \left| \left\{ \mathbf{p} \in \mathbb{P} \mid p_i \text{ and } p_j \text{ are at most } w \text{ clicks away from each other in } \mathbf{p} \right\} \right| \quad (7.1)$$



	Asteroid	C. F. Gauß	Germany	Franks	Ireland
Asteroid	0	1	1	0	0
C. F. Gauß	1	0	1	1	0
Germany	1	1	0	1	1
Franks	0	1	1	0	1
Ireland	0	0	1	1	0

(b) The resulting co-occurrence matrix

Figure 7.1: A sketch of the sliding window approach and the resulting co-occurrence matrix. The blue/red numbers in the matrix correspond to the co-occurrences depicted by the extra edges inside the blue/red windows.

Here, $\mathbb{P} := \{\mathbf{p}_1, \dots, \mathbf{p}_m\}$ is the set of navigational paths $\mathbf{p}_k := (p_{1,k}, \dots, p_{n,k})$ of pages $p_{i,k}$ in the social media system Wikipedia, as introduced in Section 3.1.2. The resulting co-occurrence matrix for Figure 7.1a can be seen in Figure 7.1b.

7.2.1.2 Binary Navigational Semantic Relatedness

A problem of the method presented in the previous subsection is that frequently occurring concepts (such as *United States of America* in the WikiGame dataset) outweigh less frequently occurring concepts. To counter this, the co-occurrence counting approach is modified such that we do not count the actual co-occurrences, but only note that either two concepts were seen inside a context window or not. This approach is called *binarization* and is defined as follows:

$$coocc_{bin}(p_i, p_j)_w := \mathbb{1} \left(\left\{ p \in \mathbb{P} \mid p_i \text{ and } p_j \text{ are at most } w \text{ clicks away from each other in } \mathbf{p} \right\} \right) \quad (7.2)$$

Here, $\mathbb{1}(M)$ is the indicator function, which is 1 if $|M| > 0$ and 0 otherwise. The difference to Equation (7.1) here is that we do not compute the *cardinality* of the set of all paths where p_i and p_j co-occurred, only if that set is empty or not. Another way to express Equation (7.2) is thus

$$coocc_{bin}(p_i, p_j)_w := \begin{cases} 1 & coocc(p_i, p_j)_w > 0 \\ 0 & \text{otherwise} \end{cases} \quad (7.3)$$

Figure 7.2 illustrates the actual difference to the co-occurrence counting method: While the counting method would assign a co-occurrence weight of 2 to the transition *C. F. Gauß - Germany*, the binarization variant only notes that this transition already occurred, so any other occurrence would not influence the result anymore.

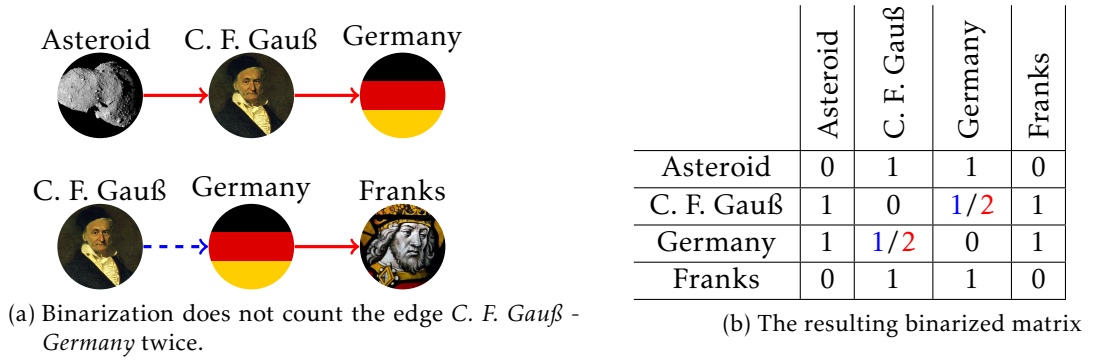


Figure 7.2: A sketch of the binarization approach and the resulting co-occurrence matrix. The red numbers in the matrix correspond to the counted co-occurrences, while the blue numbers denote the binarized result. The black numbers are common to both methods.

7.2.1.3 Evaluation on Human Intuition

To evaluate semantic relatedness, we compare our results to two semantic relatedness dataset, specifically the WS-353 and MEN datasets that were introduced in Section 4.4.3 and Section 4.4.5, respectively. For this, we use the Spearman ranking correlation coefficient ρ to compare the semantic relatedness scores from our model – produced by comparing the model’s vectors using the cosine measure (see Section 3.2.3.2) – with those from human intuition as described in Section 3.2.4. Spearman correlation scores using binarized vectors are denoted by ρ_{bin} .

As Wikipedia articles usually denote semantic concepts instead of single words, we had to map the vocabulary of the evaluation datasets to Wikipedia pages. While for WS-353 we were able to build on previous work by Milne and Witten, who published *WikiSimilarity-353*, a hand-mapped version of WS-353 to Wikipedia articles, we had to match the MEN dataset ourselves. Still, we also had to refurbish the WS-353 mappings.

We will now describe the manual mapping process of MEN and WS-353 to Wikipedia pages. Concretely, we entered each word of the WS-353 and MEN vocabularies as a search query in the Wikipedia search. If we did not find a direct match, we manually disambiguated the results and chose the most probable Wikipedia article. For example, the Wikipedia search for the query *hair* returned an article about the musical *Hair*¹ as well as another article about *Human hair*. The corresponding entry in MEN however was set in relation to *wig*, so we manually mapped the word *hair* to the more general article *Human hair*.²

¹[https://en.wikipedia.org/wiki/Hair_\(musical\)](https://en.wikipedia.org/wiki/Hair_(musical))

²At the date of submission, this is now not valid anymore, since the Wikipedia search directly returns the article for *Hair* for the query *hair*.

7.2.2 Game Navigation

In this section, we extract semantics from two different navigation games on Wikipedia, namely the WikiGame and Wikispeedia. As described in Section 3.3.2.1, the first served only recreational purposes, while the latter has been specifically created to collect navigational data from humans in order to infer semantic distances between concepts [249]. Both Wikispeedia and WikiGame have been the focus of several works analyzing human navigation patterns [20, 207, 228, 246, 247, 249] or have been used to extract semantic relatedness from human navigation [249].

In Section 5.2 we showed that navigation in Wikispeedia was largely determined by a semantic component, while that component only had a lesser influence on navigation in the WikiGame. From these results, it could be expected that Wikispeedia is better suited as a basis to extract semantic relatedness than the WikiGame. In the following, we will however argue that i) the semantic component of human navigation (where the existence has been shown in Section 5.2) can be extracted very well from both datasets (WikiGame and Wikispeedia) and makes a notable difference than compared to the semantics of the pure link network and ii) that there exist specific subsets of navigational paths more suitable than others to extract semantic relatedness from.

7.2.2.1 Contribution of Game Navigation to Semantic Relatedness

To show the usefulness of human navigational paths for calculating semantic relatedness we conduct our experimental steps as described in Section 7.2.1.1. For this, we not only use sliding windows of varying size w but also the principle that all concepts p_i in a path \mathbf{p} can co-occur with all other present concepts p_j in the same path. We denote this as a “none” window size. One can think of the “none” window size as a size that is always exactly as long as the path. Table 7.1 presents the evaluation results for varying window sizes. We report the semantic relatedness scores when evaluated on the WS-353 and MEN datasets (stated in columns *WS-353* and *MEN*) as well as the matchable number of pairs per evaluation dataset (shown right next to the corresponding relatedness score). The exact evaluation procedure is described in Section 3.2.4.1.

A first observation is that the method of letting all concepts in a path co-occur with all other concepts in the path denoted as “none” performs worse than some specific sliding window sizes denoted in the table. This strengthens our assumption that the distance between two concepts in a path is crucial for calculating precise semantic relatedness scores as pointed out in Section 7.2.1.1. Furthermore, we can see that the best accuracy can be achieved using a window size of $w = 3$ or $w = 4$. Hence, letting the surrounding two or three concepts ($w - 1$) given a concept in a path co-occur with the concept seems to be the most precise co-occurrence representation for determining the semantic relatedness between concepts in our corpus of human navigational paths. Interestingly, this observation correlates with the distance often applied in graph based methods for word sense disambiguation, as reported in [164].

To investigate the usefulness of our approach of reporting results obtained from evaluating the scores of all possible evaluation pairs for a specific window size or corpus,

Table 7.1: Semantic relatedness calculated on human navigational paths. Our corpus consists of all available paths in WikiGame and Wikispeedia where different window sizes ($2 \leq k \leq 5$) as well as the principle that all concepts in a path co-occur with all other concepts in the path denoted by “none” were evaluated against the WS-353 and the MEN datasets by calculating the Spearman rank correlation coefficient between the produced rankings of each method and the ones of WS-353/MEN.

k	WS-353 (353)	MEN (3000)	k	WS-353 (353)	MEN (3000)
none	0.649 (299)	0.513 (2159)	none	0.704 (48)	0.560 (208)
2	0.638 (236)	0.573 (1364)	2	0.521 (48)	0.390 (190)
3	0.709 (276)	0.649 (1810)	3	0.715 (48)	0.572 (208)
4	0.718 (287)	0.618 (1992)	4	0.722 (48)	0.570 (208)
5	0.690 (294)	0.587 (2072)	5	0.723 (48)	0.567 (208)

(a) WikiGame
(b) Wikispeedia

we also repeat the experiments by using all word pairs in the evaluation dataset and setting the relatedness scores to zero if we can not cover a pair as this is frequently done in related work (see the last column in Table 7.2). However, this method introduces high negative bias to the results as we observe that not surprisingly, those window sizes or corpora perform better that can simply cover more WS-353 pairs. We also calculate statistical significance tests between the dependent Spearman rank correlation coefficients produced by different window sizes. While the results indicate no statistically significant differences between window sizes 3 to 5, it is clearly visible that we would prefer a window size of 4 over “none” at a significance level of $p < 0.05$ (see Equation (3.29) for the computation of this p -value). Summarized, this evaluation represents a pessimistic evaluation compared to our optimistic one which only evaluates against possible word pairs, as it is hard to judge whether better accuracy is based on more precise calculations of semantic relatedness or simply more well defined term pairs. To further strengthen our evaluation approach we limit the evaluation in Table 7.1 to those pairs available throughout all window sizes (236 pairs) (see fifth column in Table 7.2) and we can observe the exact same trend as our optimistic evaluation approach showed. Finally, we also sample 100 random pairs 100 times and average the results again showing in the fourth column of Table 7.2 that the best accuracy can be achieved using a window size of $w = 3$ or $w = 4$ and making a strong point for our evaluation approach. This agrees with similar observations by [110] when evaluating against different subsets of evaluation pairs that the trend of accuracy always stays the same. Also, [251] pick up on this point as they directly show that as they only include well-defined term pairs to their evaluation, they can achieve the appropriate results.

As the goal of this section is not to achieve the best possible semantic relatedness scores in comparison to related work techniques, but rather to identify whether and if

Table 7.2: Spearman correlation scores with human intuition calculated in a similar fashion as for Table 7.1. This time, we report a variety of different evaluation approaches: (a) “possible pairs” reports the same results as in Table 7.1 and represent our optimistic evaluation, (b) “100 pairs” reports accuracy by sampling 100 word pairs 100 times and averaging the results, (c) corresponds to the accuracy by using only those word pairs that can successfully be determined for all windows sizes and (d) “all pairs” fills in zero semantic relatedness scores for word pairs for which no score can be calculated and represents the pessimistic evaluation. The observations illustrate the usefulness of our proposed “possible pairs” method.

w	WS-353				MEN			
	possible	100	236	all	possible	100	1364	all
none	0.649	0.630	0.632	0.548	0.513	0.511	0.438	0.390
2	0.638	0.633	0.638	0.560	0.573	0.568	0.573	0.507
3	0.709	0.692	0.694	0.586	0.649	0.643	0.591	0.483
4	0.718	0.697	0.695	0.584	0.618	0.621	0.555	0.457
5	0.690	0.690	0.692	0.590	0.587	0.583	0.519	0.434

(a) Spearman correlations for WikiGame data

w	WS-353				MEN			
	possible	30	46	all	possible	100	190	all
none	0.704	0.558	0.665	0.169	0.559	0.566	0.468	0.118
2	0.521	0.387	0.521	0.187	0.390	0.391	0.390	0.143
3	0.715	0.556	0.678	0.169	0.572	0.566	0.486	0.118
4	0.722	0.558	0.686	0.169	0.570	0.570	0.479	0.118
5	0.723	0.542	0.687	0.169	0.567	0.565	0.468	0.118

(b) Spearman correlations for Wikispeedia data

so, human navigational paths can contribute to this task and to find the most appropriate window size and path corpus, we only report results obtained from applying our optimistic evaluation procedure which evaluates the scores of all possible evaluation pairs for a specific corpus. We will also only cover a very small number of pairs later on for our sampling strategies which makes the other evaluation methods not applicable, i.e., only using the same intersection of pairs for all methods would limit the gold standard tremendously (max. 30 pairs) and using all pairs by filling in zeros for missing word pairs would have high negative influences on methods that can only cover a small amount of pairs due to lack of data. This choice is based on abovementioned investigations and observations and gives us a logic way to evaluate our work. Due to

tractability, we focus on window size $w = 3$ for the rest of this section.³

Table 7.2 demonstrates that human navigational paths contain information relevant for calculating semantic relatedness between concepts by exhibiting high quality relatedness evaluated against WS-353 and MEN. We investigate the additional benefit of the paths at hand to several baseline corpora next.

7.2.2.2 Benefit of Human Game Navigation over the Link Network

As our human navigational paths of Wikipedia games are basically subsets of the underlying topological link network, we need to investigate whether the observed effects are based on human intuitions and patterns while navigating or if automatic extractions of paths from the link network can produce similar or even better results. By doing so we can also investigate which role the rich topological link network plays for calculating semantic relatedness on paths. Due to the larger coverage of the Wikipedia link network in the WikiGame, all experiments in this section will only be performed on WikiGame data.

Navigation Baselines In the following, we will briefly describe our navigation baselines which we use to show that human navigation holds more semantic information than artificial navigation.

- **Topological neighbor paths:** A rather simple baseline for comparison consists of artificial sub-paths taken from Wikipedia’s link network limited to concepts available in our WikiGame dataset (see Section 4.2). Given Wikipedia’s topological link graph (limited to WikiGame pages) $\mathbb{W}_{wg} = (V_{wg}, E_{wg})$ with vertices V_{wg} and directed edges $E_{wg} = \{(p_i, p_j) | p_i, p_j \in V_{wg}\}$, we generate all possible paths of length 3, where every page still lies in V_{wg} . This gives us the following set of paths

$$\mathbb{P}_{tb} = \{(p_i, p_j, p_k) | (p_i, p_j), (p_j, p_k) \in E_{wg} \cap V_{wg} \times V_{wg}\} \quad (7.4)$$

The reason for choosing paths with the length 3 for this topological baseline corpus is that we focus on a window size of $w = 3$, i.e., a concept co-occurs with the neighboring $w - 1 = 2$ concepts in a path, throughout this section. Hence, with this corpus of artificial paths we can calculate all possible co-occurrences between concepts in a window of size $w = 3$. This enables us to investigate the influence of the degree of concepts on the results.⁴

- **Permuted WikiGame paths:** To understand how important the underlying link structure is for the task of calculating semantic relatedness on navigational paths

³Note that a window size of $w = 4$ is just by a small, statistically insignificant margin more precise than a window size of $w = 3$ and the reason for only reporting results for $w = 3$ is based on faster runtime and better possibilities for interpreting the results or looking into fingerprints. Nevertheless, we have also conducted further experiments by using a windows size of $w = 4$ which exhibit similar patterns.

⁴Note that again the extracted corpus of paths is a weighted subset of the plain Wikipedia link structure where the weight is influenced by the degree of each page (e.g., a page with a degree of 8 is more likely to get higher co-occurrence counts than a page with an degree of 2.)

and also to explore how much impact the sequence of concepts in a human navigational path has, we create so-called *permuted paths*. In these paths, we are still leaving the position of a concept in a path intact, but swap it with a page on the same position of another path and by doing so we detach the page with preceding and succeeding pages of the path. For a given path $\mathbf{p}_1 = (p_1, \dots, p_n) \in \mathbb{P}$, we randomly choose another path $\mathbf{p}_2 = (p'_1, \dots, p'_m)$ and randomly swap a page p_i in \mathbf{p}_1 with the corresponding page p'_i at the same position i in \mathbf{p}_2 . We receive two new paths $\mathbf{p}'_1 = (p_1, \dots, p'_i, \dots, p_n)$ and $\mathbf{p}'_2 = (p'_1, \dots, p_i, \dots, p'_m)$ where we lose the semantic information around the newly inserted page. Again, we preserve as much structural information as possible of our game paths while randomizing the semantic related information. It is important to note in this scenario, pages might not be linked from their predecessor or to their successor on the underlying Wikipedia topology. These newly created paths are called $\mathbb{P}_{permuted}$ and contain exactly as many paths as \mathbb{P} .

- **Swapped WikiGame paths:** The purpose behind this method is to keep the link structure of Wikipedia intact but to swap out parts of a supposedly meaningful path with parts of another path. Our method works as described in the following: For a given path $\mathbf{p}_1 = (p_1, \dots, p_{i-1}, p_{mid}, p_{i+1}, \dots, p_n)$, we select another path $\mathbf{p}_2 = (p'_1, \dots, p'_{i-1}, p'_{mid}, p'_{i+1}, \dots, p'_m)$, with maybe a different length, but with the property that the middle pages p_{mid} and p'_{mid} are the same page. We cut both paths in half and exchange the back part of p with the one of q in such a way that we receive the new paths $\mathbf{p}'_1 = (p_1, \dots, p_{i-1}, p_{mid}, p'_{i+1}, \dots, p'_m)$, and $\mathbf{p}'_2 = (p'_1, \dots, p'_{i-1}, p'_{mid}, p_{i+1}, \dots, p_n)$. The newly generated paths are called \mathbb{P}_{swap} and contain exactly as many paths as \mathbb{P} .

Results Table 7.3 presents the results using a window size of $w = 3$ with all available WikiGame paths and our baseline corpora as described above. In column *#paths* one can see the number of paths available for each corpus and in column *length* the total accumulated length of all paths in the corpus. Finally, in columns *WS-353* and *MEN* we can see the final *Spearman rank correlation* to the respective evaluation datasets as well as the number of matchable pairs. Further insights from these investigations are discussed next.

Wikipedia topology alone is useful: We know from other semantic analysis methods that the Wikipedia topology alone provides useful information [251]. For confirmation, we evaluated the scores obtained from our *Permuted Wikigame paths corpus*. The corresponding results confirm that we lose semantic preciseness when ignoring the original link and navigation structure. Keeping the original structure intact, but swapping parts of the paths – see *Swapped WikiGame paths* and the corresponding description above – we can see that the original navigation by a user has a high impact on the achieved accuracy, but that we can still achieve reasonable results by leaving the underlying link structure and partly navigational patterns intact.

Human navigation paths improve results: A first observation is that the WikiGame path results outperform the baselines by a relevant margin – for example, it outperforms

Table 7.3: Comparison of semantic relatedness calculations using a window size of $w = 3$ evaluated against WS-353 and MEN on all WikiGame paths with several baseline corpora. We report Spearman correlation scores for co-occurrence counting (ρ) and binarization (ρ_{bin}) semantic information extraction.

Corpus	#paths	WS-353 (#pairs)		MEN (#pairs)	
		ρ	ρ_{bin}	ρ	ρ_{bin}
\mathbb{P}	1,799,015	0.709 (276)	0.707	0.649 (1810)	0.644
\mathbb{P}_{topo}	6,042,578,644	0.659 (308)	0.485	0.594 (2399)	0.393
$\mathbb{P}_{permuted}$	1,799,015	0.381 (292)	0.406	0.090 (2111)	0.101
\mathbb{P}_{swap}	1,799,015	0.668 (273)	0.697	0.633 (1705)	0.636

the best baseline method *Swapped WikiGame paths* by 0.041 on WS-353 (0.709 vs. 0.668). The same holds true for evaluation on MEN. Especially the evaluation results of the binarized semantic relatedness extraction method on the swapped paths support this, as with binarization, the effects of frequently occurring transitions are canceled out, that might otherwise impact semantic information. Although the binarization results of the swapped paths are higher than those of the co-occurrence counting results, they are still outperformed by the binarized evaluation of human navigational paths. When looking at the *Topological neighbor paths corpus*, we can also see that Wikipedia’s inherent link structure already can be used as a powerful resource for calculating semantic relatedness using our methodology.

In order to see how the number of co-occurrences between concepts influences the semantic relatedness, we especially pay attention to the binarization results of the topological neighbor paths, which directly represents the *plain link structure*. We can now see that the accuracy evaluated against WS-353 drops by a significant amount (from 0.659 to 0.485) indicating that the number of co-occurrences between concepts effects our method. However, in all other settings except on the original paths and the topological neighbor paths, the performance of the baselines is actually increased. This means that although we destroyed the inherent structure of the game paths, binarization can still capture more semantic relatedness information than co-occurrence counting on “broken” paths. Still, on the original navigation data, binarization can actually be detrimental to the semantic relatedness extraction process. We can also see the same patterns when evaluating the path baselines on the MEN dataset, even with similar drops and gains in performance.

With this initial exploration, we can conclude that human dynamic navigational paths on Wikipedia can contribute to computing semantic relatedness, but they are based on an already powerful network topology. The weighting provided by users’ choice during navigation exhibits the most precise information for determining semantic relatedness between concepts. Next, we want to identify what kind of navigational paths are most useful for that task.

Table 7.4: Semantics in successful and unsuccessful game paths. ρ and ρ_{bin} denote the Spearman correlation score with WS-353 or MEN of the co-occurrence counting and binarized vector representations. All results are computed using a window size of $w = 3$.

Corpus	#paths	length	WS-353 (#pairs)		MEN (#pairs)	
			ρ	ρ_{bin}	ρ	ρ_{bin}
All Paths	1 799 015	5.98	0.709 (276)	0.707	0.649 (1810)	0.644
Unsuccessful	1 145 934	5.80	0.703 (270)	0.702	0.620 (1688)	0.646
Successful	653 081	6.30	0.631 (230)	0.667	0.592 (1323)	0.627

(a) WikiGame

Corpus	#paths	length	WS-353 (#pairs)		MEN (#pairs)	
			ρ	ρ_{bin}	ρ	ρ_{bin}
All Paths	76 193	5.81	0.715 (47)	0.676	0.572 (208)	0.632
Unsuccessful	24 875	4.68	0.628 (47)	0.621	0.497 (199)	0.569
Successful	51 318	6.36	0.719 (47)	0.683	0.578 (208)	0.639

(b) Wikispeedia

7.2.2.3 Path Selection to Improve Fit to Human Intuition

Human navigational paths in a game setting can be characterized along many dimensions. For example, there exist *successful paths* where users were able to successfully reach the specified target pages, while on *unsuccessful paths* users could not reach their goal. Other path characteristics may mostly move along high degree (vs. low degree) pages. Figure 7.3 shows the distribution of path lengths in all paths (black line), only in successful paths (red line) and only in unsuccessful paths (blue line). Only looking at such path length distributions, we can already see that such distinct path types exhibit different features. We want to explore these differences and investigate their usefulness for the task of calculating semantic relatedness, e.g., investigate whether a subset of only successful paths is more useful than a subset of only unsuccessful paths.

This gives rise to a number of interesting questions related to different navigational paths, such as (a) *Are all navigational paths equally useful for computing semantic relatedness?* and (b) *If some navigational paths are more useful, what are the characteristics of these paths and how can they be exploited?*

To analyze these and other questions, we begin our investigations on the *WikiGame* by taking the corpora of all *successful* and *unsuccessful paths*. Similar to Section 7.2.2.2, we use a window size of $w = 3$ for our co-occurrence calculation and evaluate the relatedness scores against WS-353 and MEN; the results can be seen in Table 7.4a. From that table, we see that a smaller subset of our corpus of all WikiGame paths \mathbb{P} can still perform remarkably well (compare with Table 7.3). Somewhat surprisingly, we see that a

corpus of *unsuccessful paths* performs better than a corpus of *successful paths*. A possible explanation for this behavior is that unsuccessful paths contain the behavior of mostly inexperienced users who try to follow pages whose meanings are very close and hence, remain on a narrow semantic field which may also lose them the game. On the other hand, successful players might navigate through more distant concepts or very central concepts like “United States” which are common strategies for winning a game. Further investigations are necessary in order to explain this behavior in detail, which is not in the scope of this section.

Results on Wikispeedia show a very different result. Here, the number of successful paths is almost double the number of unsuccessful paths, in contrast to the WikiGame data. At the same time, the semantic information generated by successful paths is slightly better than that of the whole path corpus, which also stands in direct contradiction to the results obtained on the WikiGame. In other words, we notice that on Wikispeedia, unsuccessful paths contribute less to the final result than in WikiGame. A possible explanation for this might be that unsuccessful users that do not even reach their goal, also do not minimize the semantic distance toward their target [203]. Additionally, these users often change direction somewhere mid-game, which can be compared to the path swapping experiments on WikiGame and which also resulted in lower scores there (cf. Table 7.3). As to why this seemingly doesn’t affect unsuccessful games on the WikiGame, we can only speculate, as we have no explicit information about backtracking⁵ there.

Regardless of the exact explanation of this behavior, the results suggest that subsets of paths with specific characteristics yield different results. This leads to the idea of investigating whether smaller sets of paths according to specific path characteristics can perform similarly or even more precise in regard to our relatedness calculations on the whole set of paths. In the following section, we will explore this by conducting different path selection experiments. As before, we restrict ourselves to navigation on WikiGame data, since in this case, the selection of smaller subsets of Wikispeedia paths would result in very few selected paths, from which we do not expect valid results.

Characteristics of Paths We introduce several measures $m : \mathbb{P} \rightarrow \mathbb{R}_0^+$ to characterize any path \mathbf{p} in our corpus of paths \mathbb{P} . Each distinct measure makes use of a path characteristic, depending on the visited pages, which actually characterize the path. The resulting measures will be subsequently used in Section 7.2.2.3 to create path selections. In the following, we will elaborate each of the different measures in greater detail. Let $\mathbf{p} \in \mathbb{P}$ be an arbitrary path represented by the sequence of pages (p_1, \dots, p_n) .

- **Length:** We use the length of a path \mathbf{p} , i.e., the number of concepts visited in a path, as a first characteristic:

$$m_{length}(\mathbf{p}) = len(\mathbf{p}). \quad (7.5)$$

Our motivation for taking the length of a path as a characteristic is the notion that longer paths potentially contain more information because of more co-occurrences

⁵Clicking the back button in the browser

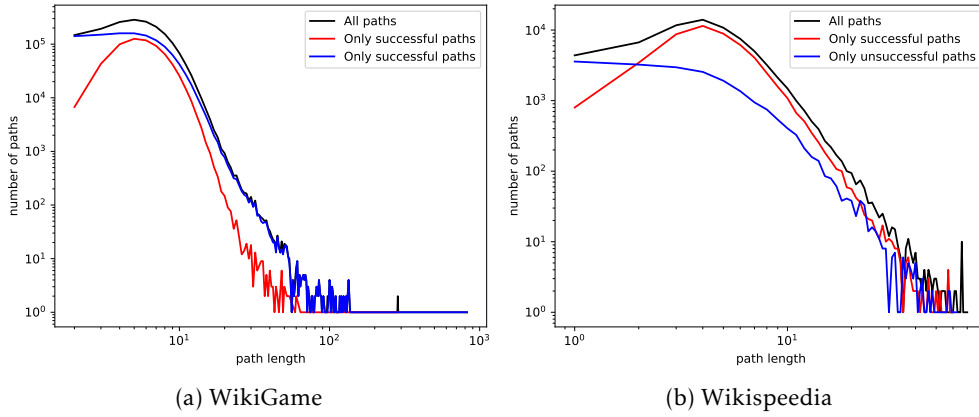


Figure 7.3: Illustration of the distribution of path lengths in all human navigation paths (black dotted line), only in successful paths (red solid line) and in unsuccessful paths (blue dashed line).

between concepts of the paths. Furthermore, we could observe in Figure 7.3 different path length characteristics for different types of WikiGame paths, which is interesting to investigate in greater detail.

- **In- and outdegree:** For a path \mathbf{p} , we determine the in- and outdegree for each concept p_i in \mathbf{p} derived from Wikipedia's complete topological link network. The path itself is characterized by the average in- or outdegree of all the contained concepts. The idea behind this characteristic is to differentiate hubs and strongly connected concepts from dead ends and rather weakly connected concepts. The measure is calculated as ($m_{outdegree}(\mathbf{p})$ is defined analogously):

$$m_{indegree}(\mathbf{p}) = \frac{1}{len(\mathbf{p})} \sum_{k=1}^n indegree(p_k). \quad (7.6)$$

- **Ratio:** This measure represents a ratio of in- and outdegree for each page in a corpus of paths smoothed by the square root of the indegree (see [228]). This characteristic is motivated by the notion that a page with e.g., 200 inlinks and 100 outlinks should be more general than a page with two inlinks and one outlink. This means that we include the semantic *generality* of a page into the measure, since if a node has a higher indegree than another, it is more general than another [120]. If the outdegree for a page is zero, we set the ratio to zero as well. $ratio(p)$ is calculated in the following way for a page p :

$$ratio(p) = \frac{indegree(p)}{outdegree(p)} \cdot \sqrt{indegree(p)}. \quad (7.7)$$

Thus, the value of a path \mathbf{p} is determined by

$$m_{ratio}(\mathbf{p}) = \frac{1}{len(\mathbf{p})} \sum_{k=1}^n ratio(p_k). \quad (7.8)$$

- **TF-IDF:** Interpreting a path as a document and the concepts present in a path as terms, we use the well known *tf-idf* scores (cf. [201]) of each page in a path as a further characteristic. The idea behind this characteristic is that we can identify paths that include many concepts that are very important for the individual path compared to all other paths in the corresponding corpus. Hence, for each path \mathbf{p} , we again take the mean of all *tf-idf* values in the path:

$$m_{tfidf}(\mathbf{p}) = \frac{1}{len(\mathbf{p})} \sum_{k=1}^n tfidf(p_k). \quad (7.9)$$

Path selection strategies Based on the characteristics described above, we now select smaller sets of paths \mathbb{P}_m according to the abovementioned path characteristics m . We investigate whether the relative performance of reduced corpora of paths \mathbb{P}_m , based on the accuracy of our relatedness scores, increases or decreases, compared to the performance of our complete set of paths \mathbb{P} , in analogy to Section 6.3.3 and Section 6.4.

For each characteristic, we calculate ten subsets of increasing size where the tenth subset corresponds to the set of all available WikiGame paths. The sizes of our subsets are calculated by the number of visited pages inside the subset. If we consider the amount of all visited pages $page_sum = \sum_{\mathbf{p} \in \mathbb{P}} len(\mathbf{p})$, a path selection of e.g., 10% does not

necessarily contain $0.1 \cdot |\mathbb{P}|$, but rather $0.1 \cdot page_sum$. More formally, we can express it as follows: Consider an ordered list $L_m = (\mathbf{p}_1, \dots, \mathbf{p}_n)$ of paths, generated by a measure m . A selected subset \mathbb{P}_m^x of size x for measure m can be expressed as:

$$\mathbb{P}_m^x = \left\{ \mathbf{p}_k \mid k = \max \left\{ s \mid \sum_{j=1}^s len(\mathbf{p}_j) \leq \frac{x}{100} \cdot page_sum \right\} \right\}. \quad (7.10)$$

Thus, a potential path selection with very long paths consists of fewer actual paths than a selection with mostly short paths, but both sets contain roughly the *same amount of visited pages*. This renders the selection process more fair than just pure path counting as it enables us to fairly compare two corpora of the same size selected based on different path characteristics. Each selection process generated subsets consist of $x = \{10, 20, \dots, 90\}\%$ of all visited pages. By proceeding with this selection process, the first subset, i.e., the 10% subset, consists of paths with the lowest measures for a corresponding characteristic, e.g., paths with the lowest mean indegree. Furthermore, we also revert the ordered list L_m in order to get a ranking $L_m^{rev} = (\mathbf{p}_n, \dots, \mathbf{p}_1)$ where the small subsets contain paths with higher measures for a specific characteristic, e.g., paths with the highest mean indegree. After the generation of the path ordering lists and the path selection process described above, we run our semantic evaluation for each of these subsets.

Furthermore, we create a baseline for each individual split to learn whether the distinct accuracy results are genuinely dependent on the corresponding path selection process based on several characteristics. We shuffle the corpus of paths independently and randomly ten times in order to remove the original ordering in the complete set of paths. For each of these ten independent shuffles, we extract subsets according to the selection process described above. We end up with ten selections for each subset containing $x = \{10, 20, \dots, 90\}\%$ of the visited pages. Finally, we perform our semantic analysis and evaluate the results accordingly for each selection and subset. We average the results for each subset based on the sum of selections for the corresponding subset and report the results in the following section; we will refer to this baseline as *random baseline*.

Results In Figure 7.4 we present the results obtained from our individual subcorpora of navigational Wikipedia game paths using our selection strategies defined above based on characteristics of paths, or to be precise: characteristics of concepts inside paths averaged for each path. Figure 7.4a and Figure 7.4c illustrate selections where we can achieve a better correlation score evaluated against WS-353 than using random selections of all WikiGame paths, i.e., *random baseline* (black line, ■), while Figure 7.4b and Figure 7.4d show selections performing worse. The horizontal black line with a Spearman rank correlation of 0.709 shows the results achieved when taking a corpus of all WikiGame paths (see also Table 7.1). For all selections we use a window size of $w = 3$ for our co-occurrence and subsequent semantic relatedness calculations. Our key findings are discussed next.

Intelligent path selection improves semantic relatedness. A first observation when looking at Figure 7.4a and Figure 7.4c is that smaller random path selections do not lead to a similar or better accuracy (black line, ■), but that we indeed can find smaller corpora of navigational paths – selected on several characteristics – that perform equally or better than the complete corpus of WikiGame paths (that reaches an accuracy of 0.709 on WS-353). By incrementally adding paths with the lowest average indegree of their concepts, we can achieve the highest Spearman rank correlation with a sub-corpus of only 30% of all WikiGame paths (red line, +) on WS-353 (cf Figure 7.4a). The respective accuracy of 0.760 outperforms the accuracy of the whole WikiGame corpus by about 5% while covering less than a third of all visited pages in the complete corpus. The same thing holds for a minimum of 40% of high tf-idf paths, when evaluating on MEN (see Figure 7.4c), where we reach a maximum correlation score of 0.725 compared to the baseline of all paths of 0.649. Please note that in this case, we evaluate on 1810 instead of 276 word pairs, so this improvement is significant at $p < 0.05$.⁶

Contrarily, we can see in Figure 7.4b that a reverse accumulation of paths, beginning with those having a high average indegree (red line, +), leads to much worse accuracy on WS-353 compared with the random baseline and as well as with the accuracy of

⁶Obviously, a difference of 0.1 in Spearman correlation scores is easier to achieve with only few word pairs. For example, a simple swap of two items in a list of three items causes a greater disturbance in the score, as with a thousand items. Thus, an improvement of 0.1 in correlation score is more significant on many pairs than on fewer pairs. See also Section 3.2.4 for this.

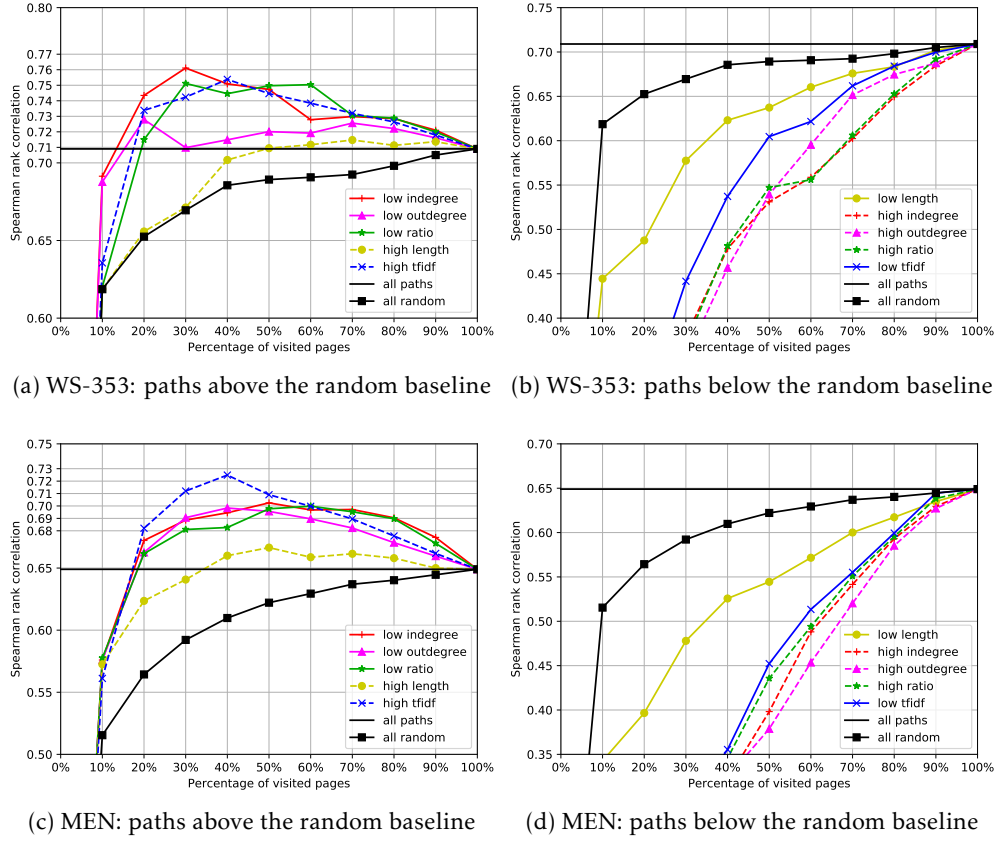


Figure 7.4: Semantic relatedness calculated on different path selections. Larger values on the y-axis correspond to higher Spearman rank correlation with the WS-353 dataset. The black horizontal line depicts the result for the entire set of paths. The figures on the left show selection results with better-than-random performance while the figures on the right show results with worse-than-random performance. In Figure 7.4a, we can see that a small subset of only 30% low indegree paths produces more precise semantics than the whole path corpus \mathcal{P} would (scoring a rank correlation of 0.760 to the WS-353 dataset). On MEN, a subset of 30% high tfidf paths reaches a peak performance of 0.724 correlation. Paths characterized by low in- and outdegree always perform better than a random baseline, while their counterparts, starting from high degrees, perform significantly worse. Similar patterns can be observed when selecting paths according to their tf-idf values.

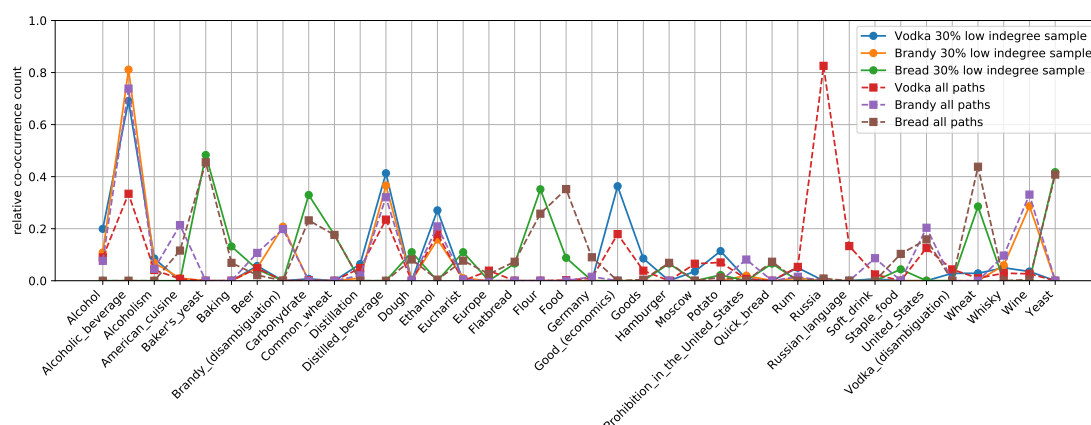


Figure 7.5: Semantic “fingerprints” for the concepts *Vodka*, *Brandy* and *Bread* limited to co-occurring concepts where at least one of all vectors exhibits a co-occurrence count of fifteen to the corresponding concept. All counts are normalized by the L2 norm of the vector and fingerprints for a 30% low indegree selection (solid lines) and the full set of paths (dashed lines) are shown. The 30 % low indegree selection exhibits more fine-grained and precise semantics than the set of all paths.

the complete corpus. Figure 7.4d shows the same behavior of the characteristic subsets when evaluated on MEN. A possible explanation for this is that low indegree pages represent concepts that do not seem to be hubs nor exceptionally abstract concepts in comparison to high indegree pages. Also, high indegree concepts may have much more co-occurrence counts with several other concepts while low indegree concepts may only have co-occurrence connections to a few very specific concepts (even when looking at a window size of $w = 3$). Furthermore, a similar thing holds for high tf-idf paths: Each of the pages in a path are very special for this specific path and thus carry extra meaning. Thus, they also co-occur with only few specific concepts. Hence, the co-occurrence vectors may be sparser, but more precise and this may enable us to calculate more accurate semantic relatedness scores. If we look deeper into the paths included in our selection corpora we can see that paths with the highest average indegree all include the concept *United_States* which is on the one hand, the most central concept in Wikipedia’s topological link network, and on the other hand, also by far the most often navigated concept in our WikiGame paths. Additionally, the tf-idf value of *United_States* is thus very low, which leads to the page not being selected in high tf-idf paths. Hence, this concept co-occurs with many others and is no suitable descriptor for determining the semantic relatedness between concepts while paths with the lowest average indegree contain more variety and also more descriptive co-occurrences. To summarize: *Small selections of low indegree and high tf-idf paths exhibit more fine-grained and precise semantics than the set of all paths.*

To give an example we illustrate in Figure 7.5 the concept co-occurrence vectors for

the concepts *Vodka*, *Brandy* and *Bread* on the one hand, using our best overall performing corpus of 30% low indegree paths (solid lines, ○) and on the other hand, deriving the information from the all path corpus (dashed lines, ■). For visualization purposes the vectors are reduced in dimensionality by only representing co-occurrences to concepts where at least one vector exhibits a count of larger than 15. Furthermore, all counts are normalized by the L2 norm of the complete vector. In Figure 7.5 we can see that the concepts *Alcoholic beverage*, *Distilled beverage* and *Ethanol* exhibit similar peaks for the concepts *Vodka* (blue solid line, ○) and *Brandy* (orange solid line, ○) for the corpus of 30% low indegree paths, while having only few differing peaks, e.g., at *Wine* or *Good (economics)*. We can observe that these common peaks contribute a lot to the high cosine similarity of 0.8043 that we can compute with this subset for the corresponding concept pair. In contrast, we can see that there are only a few similar normalized co-occurrences for the concepts *Vodka* (red dashed line, ■) and *Brandy* (violet dashed line, ■) using the corpus of all paths and that the concept *Russia* exhibits a large diversity regarding the co-occurrence patterns for both concepts negatively influencing the relatedness score resulting in only 0.4205. The co-occurrence vectors for the concept *Bread* show for both corpora – i.e., 30% low indegree paths (green solid line, ○) and all paths (brown dashed line, ■) – no common peaks to both other concepts resulting in extremely low relatedness scores. We can see from this, that our selection of low indegree paths exhibits much more fine-grained patterns for the concept pair *Vodka* and *Brandy* reaching also a higher relatedness score than our corpus of all paths by still keeping low scores for concept pairs, that are not semantically related.

Other degree based selection strategies and corpus based characteristics can also improve accuracy. Similar observations as above can be seen by selecting according to the average outdegree of paths starting with the lowest value depicted in Figure 7.4a and Figure 7.4c (pink line, ▲). Smaller selections can outperform the corpus of all paths, but we can not achieve as good results as with our 30% selection of low indegree paths. Again, the opposite occurs for the reverse selection of paths starting with those having a high outdegree shown in Figure 7.4b and Figure 7.4d (pink line, ▲) – i.e., all selections perform worse than the baseline and the complete corpus. Selections based on the average *ratio* of paths (green line, ✱) not surprisingly show similar patterns as the selection according to in- and out-degree, but indicate that a selection according to the average indegree of paths can achieve higher accuracy than using a combined measure. Selection strategies based on the *tf-idf* values of pages inside paths indicate that we can strongly outperform the baseline and the target accuracy of a corpus of all paths for several sub-corpora incrementally adding paths with a high average *tf-idf* value shown in Figure 7.4a (blue line, ✕). In Figure 7.4c, we can even see that high *tf-idf* paths outperform the baseline and all other path selection subsets by far. Contrary, selecting paths with low *tf-idf* scores never reaches the accuracy of the random baseline as we can see in Figure 7.4b and Figure 7.4d (blue line, ✕). Low average *tf-idf* valued paths exhibit similar patterns than those with a low average indegree. The difference though is that this measure is only corpus dependent and ignores characteristics of the underlying topological link network and this may exhibit advantages for specific scenarios. Finally, we can see from both illustrations in Figure 7.4 that a selection according to the length

of paths (orange line, ○) produces just three sub-corpora of paths – i.e., 70% to 90% selections of longest paths – that can slightly outperform the corpus of all paths.

A combination of successful and unsuccessful paths produces more precise semantics than using unsuccessful paths only. Our initial experiments showed that a corpus of unsuccessful paths outperforms a corpus of successful paths in regard to the accuracy of our semantic relatedness scores (see Table 7.4). Now that we know that a path corpus with lower indegree paths works better one possible reason for the better performance of unsuccessful paths might be that the average indegree of unsuccessful paths is lower as the average indegree of successful paths as we have investigated. However, with the observation that there are more intelligent ways of selecting a corpus of paths accordingly (e.g., by selecting low indegree paths), the question arises if we can furthermore improve the preciseness of semantic relatedness calculation by performing a similar selection just on the corpus of unsuccessful paths. To this end, we use our best performing characteristic measure – namely the *indegree* – and select in the typical way sub-corpora of unsuccessful paths starting with those having the lowest mean indegree. We do the same selection for successful paths to be able to compare both subsets. Again, we accumulate the number of paths in a selection towards the total number of visited pages of the corpus of all paths; we end up with more selections for unsuccessful paths than for successful paths as we have a larger fraction of unsuccessful paths.

In Figure 7.6a we identify that we can outperform the horizontal black solid line indicating the accuracy obtained from a corpus of all WikiGame paths on WS-353. The best results can be achieved by using a 20% split of only unsuccessful paths (blue solid line). While this accuracy of 0.733 outperforms the whole set of all paths, we still get a better result by selecting the whole corpus in a similar fashion as depicted in Figure 7.4a, where we could reach an accuracy of 0.760. When we now look deeper into the subsets of low indegree based selections calculated for the complete dataset, we see that around 25% of the paths inside the best performing 30% low indegree sub-corpus (selected on all paths) are successful paths (see Figure 7.6b). While unsuccessful paths tend to exhibit characteristics that make them more useful for computing semantic relatedness, we find that overall a combination of successful and unsuccessful paths produces the best results. The results also suggest that other characteristics such as the indegree and not success are better suited for selecting good subsets when performed on the whole set of paths.

7.2.2.4 Conclusion

In this section, we investigated human navigational paths on Wikipedia with regard to the contained semantics. We could not only show that these paths can represent a viable source for calculating semantic relatedness between concepts in information networks, but also that semantic relatedness calculated based on human navigational data may be more precise than semantic relatedness computed from Wikipedia's link structure alone. Additionally, we find that not all navigational paths are equally useful. Intelligent selection of navigational paths based on path characteristics can improve accuracy.

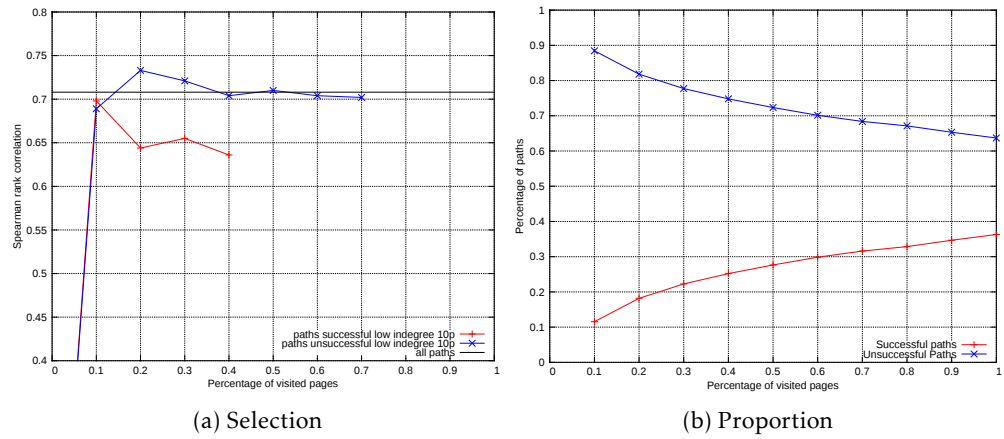


Figure 7.6: Effect of successful / unsuccessful paths: (a) shows successful (red solid line) and unsuccessful paths (blue solid line) selected according to their average indegree starting with low indegree paths and their respective Spearman rank correlation evaluated against WS-353. (b) shows the percentage of successful (red solid line) and unsuccessful paths (blue solid line) for our best performing selection of 30% low indegree paths (see Figure 7.4a). While path selection based on unsuccessful paths performs better than a selection of successful paths, we can see that overall a combination of successful and unsuccessful paths produces the most accurate results.

If we compare our results to those obtained by previous works evaluated on the same full gold standard (see results from some well-known methods in Table 7.3) we can observe that we can match the accuracy of existing methods (our best score ends up at 0.76). Yet, there are obstacles in comparing the results to other methods directly. The main evaluation process of most of the related work remains a black box. Only slight adoptions to the Wikipedia dump used – e.g., by removing low degree concepts as ESA does – can already change the outcome tremendously. As the goal in this section was not to achieve the best performing method but rather detect signals in the data and show the usefulness of our approach we did not directly try to compare us with other works due to abovementioned reasons.

A main limitation of the results obtained in this section is that we focus on human navigational paths derived in game settings. The game design itself may affect the structure of the paths and the resulting semantic relatedness scores. Some possible constraints of the game are: (a) a random choice of start and target pages – hence, users also do target based navigation instead of pure exploration navigation, (b) users have a time constraint while navigating or (c) users tend to evolve strategies in order to win a game that may be counterproductive in terms of specifying semantic relatedness. Contrary, one could argue that real navigation more focuses on the goal of getting as much information as possible. One could also argue that such real human navigational data can even be more useful as humans may take more time for checking the current

page and the next link would be chosen more accurately. They may also navigate on a more semantically narrow path. Nevertheless, the human navigational game paths present an abstraction of real user navigation in information networks and provide a further signal that such data can indeed be very useful for calculating semantic relatedness.

7.2.3 Unconstrained Navigation

When users visit the public Wikipedia page, they do so because they want to find information [209]. While these users do not necessarily have a predefined goal in mind, they might have a rough idea in mind what they are looking for. According to Singer et al., users mostly visit Wikipedia for *shallow* information needs, such as obtaining an overview of some topic or simply for quick fact-checking, but some users also are looking for deeper insights and make heavy use of Wikipedia’s semantic hyperlink structure. The resulting navigational paths are especially interesting, since we assume that users are more likely to click on links that they find interesting when exploring information that they do not have to visit because of navigational constraints.

In Section 7.2.2 we have shown that game navigation on Wikipedia can yield considerable semantic information. A big drawback here is that game navigation is not directly collected from Wikipedia, but recorded in the context of games, which might induce unnatural navigation patterns. In a game setting, users have to navigate towards a goal that is defined by the navigation game. However, the existence of an external goal stands in direct contrast to the *exploring* or *browsing* behavior in an unconstrained setting. Consequently, we now want to investigate if *unconstrained navigation* is a similarly good source of semantic information as game navigation.

As we have shown in Section 5.2, unconstrained navigation on Wikipedia is to some extent explained by a semantic component, although it played a lesser role than in game navigation. However, there are also certain issues that we have to keep in mind. Unconstrained navigation on Wikipedia can almost only be collected on the public Wikipedia. Due to privacy reasons, the published navigation snapshot, the ClickStream dataset (see Section 4.2.2.1), only consists of *page-to-page transitions*, accumulated across all user sessions in a whole month. This also means that there are *no navigational paths*, which could have an impact on the extractable semantic information. While we also investigate the WikiClickIU dataset (see Section 4.2.2.2) which contains slightly more navigation information so we can reconstruct navigational paths, it suffers from a very limited size. Thus, the question arises if *unconstrained navigation contains enough useful information to receive high quality vector representations*.

Concretely, we first want to make the semantic information contained in unconstrained navigation visible. For this, we use the methods introduced in Section 7.2.1 to encode that information in co-occurrence based vector. Again, we evaluate the vectors on human intuition. Second, we want to know if the constructed page vectors of a similar or even higher quality than those constructed from game navigation. Here we show that the *binarization method* (Section 7.2.1.2), is more useful to model the semantic information in unconstrained navigation than co-occurrence counting (Section 7.2.1.1). Finally, we also

investigate if the semantics in unconstrained navigation differ significantly from several heuristic baselines constructed from the plain Wikipedia link network. To validate our results, we perform the semantic information extraction experiments on two datasets containing unconstrained navigation, namely the ClickStream and WikiClickIU datasets (see Section 4.2.2).

We find that unconstrained navigation in the ClickStream data contains significantly more semantic information than the general link network. Compared to game navigation, we do not obtain a clear result which of both yields the more useful semantics. Overall, we can say that unconstrained navigation on Wikipedia is well suited for extracting semantics.

7.2.3.1 Contribution of Unconstrained Navigation to Semantic Relatedness

In the following, we apply both the co-occurrence counting and binarization methods on navigational data in unconstrained settings, namely the ClickStream dataset, which collected page transitions on the public Wikipedia site [255], and the smaller WikiClickIU dataset, which captured navigation from inside the network of the University of Indiana [151].

Experimental Setup. As baseline we use navigation data from the navigation game WikiGame, which has been thoroughly investigated in Section 7.2.2. For a description of those datasets, refer to Section 4.2. We compare the constructed vectors using the cosine measure and use the evaluation methodology proposed in Section 3.2.4. As evaluation datasets, we use WS-353 and MEN, which are described in Section 4.4. The reported results are denominated by $\rho_{dataset}^{\#pairs}$, which describes the Spearman correlation score obtained on a certain number of *pairs* of the denoted *evaluation dataset*. The number of pairs in this case is only given for comparative reasons, since two different datasets also mostly match a different number of word pairs, but almost always share a certain number of *matchable pairs*. Consequently, the given number of pairs denotes this maximum common subset of evaluation pairs for a set of word vector models. Finally, since the ClickStream data only contain *page transition data* instead of navigational paths, we set the window size $w = 2$ throughout all our experiments to make the results comparable.

Results. We will now describe the results shown in Table 7.5.

In Table 7.5a, we can see that *the model of semantic information in both unconstrained navigation datasets does not fit human intuition as good as the model of game semantics, when using the co-occurrence counting approach (Section 7.2.1.1)*. Especially on the smallest common subset of matched evaluation pairs (ρ_{ws}^{107}), ClickStream data yield the worst results, not far behind results from WikiClickIU. Also in the direct comparison of ClickStream and WikiGame (ρ_{ws}^{224}), semantics from unconstrained navigation with a Spearman correlation of 0.506 lags far behind game navigation semantics, where we achieved a rather high correlation with human intuition of 0.688. Only on the maximally

matchable pairs (ρ_{ws}^{max}) can we see that ClickStream produces better semantics than WikiClickIU, however still worse than those that we can extract from WikiGame. On MEN (Table 7.5b), we see roughly the same picture for the co-occurrence counting models, although on the smallest common subset of matchable MEN pairs (ρ_{men}^{468}), ClickStream now shows higher correlation with human intuition than game navigation.

On the other hand, by modelling the semantic information using the binarization approach (Section 7.2.1.2), we can capture the semantic information in unconstrained navigation in the ClickStream data much better. While on WS-353 (Table 7.5a), ClickStream navigation still yields inferior evaluation results to game navigation, the difference is now smaller than when modelling the contained semantic information with co-occurrence counting. For example, while the evaluation scores of the co-occurrence counting models of ClickStream and WikiGame had a difference of $0.688 - 0.506 = 0.182$ on their common evaluation pairs (ρ_{ws}^{224}), the binarization models now only differ slightly in performance with an absolute difference of $0.722 - 0.706 = 0.016$. The same holds for any other comparison of ClickStream and WikiGame. However, while WikiClickIU scores improve a bit, they do not profit from the different modelling approach as ClickStream does.

On MEN, we receive the most interesting results when modelling the semantic information using the binarization approach. While the WikiClickIU based semantics do not show much improvement (ρ_{men}^{468}) and even a decrease in their evaluation score (ρ_{men}^{max}), ClickStream based semantic information can now be modeled in a way that fits much better to human intuition of semantic relatedness. Indeed, binarization models can exploit the semantic information contained in ClickStream navigation even better than the information contained in WikiGame.

Discussion. We now discuss the results presented above.

The first observation is that *binarization can often model the semantic information contained in navigation better than co-occurrence counting*. The score improvements are caused by large entries in the co-occurrence counting vectors that in turn eradicate the impact of less represented information. Concretely, dominant features affect the cosine calculation both in the scalar product as well as the length calculation much stronger than weaker features. This then skews the modelled vectors and the resulting similarity values extremely. Binarization changes the co-occurrence vectors in such a way that the impact of previously extremely dominant features is decreased by a large margin, while the impact of lesser important features is raised relatively, thus taking all features into account more fairly. We briefly explain how the binarization approach concretely affects concept relatedness. We chose four concept pairs of WS-353: (*nature, natural environment*), (*water, liquid*), (*money, bank*) and (*psychology, mind*). The first two concept pairs had their similarity value changed by a large margin when applying binarization, while those of the other two pairs remained almost unaffected. The co-occurrence vectors of the first two pairs contained large, singular peaks in their co-occurrence value distribution, while the vectors of the second two pairs yield a more flat distribution. Such a peak easily becomes the dominant factor in a cosine calculation, thus almost nullifying the impact of the other context, which results in a low (and biased) similarity

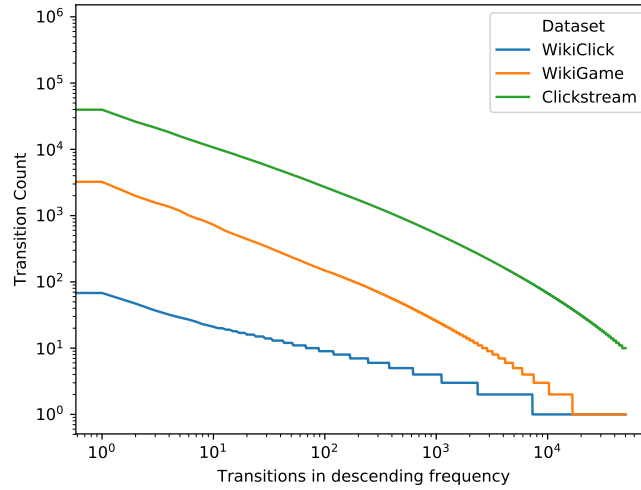


Figure 7.7: Transition count distributions for the ClickStream, WikiClickIU, and WikiGame datasets. For each dataset, we sampled the same stratified amount of transition counts to make the curves comparable. We plotted the amount of transitions in descending order for each dataset. While all distributions follow a power law distribution, ClickStream and WikiGame exhibit a steeper slope than WikiClickIU. Both axes are logarithmically scaled. ClickStream has no transitions that occur less than 10 times by design [255].

value. Binarization erases these peaks and thus decreases the peak dominance, while increasing the impact of the remaining context.

Our second (and main) observation is that *sufficient amounts of unconstrained navigation contain valuable semantic information, even when only considering page-to-page transitions*. One question that we wanted to answer in this section was if unconstrained navigation does contain a rational amount of semantic information. While we already knew that game navigation can be used as a source of semantic relatedness (cf. Section 7.2.2), this was a sensible assumption. However, the game navigation datasets that we investigated there contained complete navigational paths, on which we could show that semantic information is contained mainly in path subsets of length 3 (cf. Table 7.1). However, due to privacy reasons, the ClickStream data only contain accumulated, direct page-to-page transitions. This means that we can only work on path subsets of length 2, i.e., transitions. Additionally, results obtained from modelling these transitions using the co-occurrence counting method defined in Section 7.2.1.1 yielded rather mediocre results, especially in relation to game navigation. However, we attribute this to the different sizes of the datasets, especially to the *transition count distributions*. In Figure 7.7, we can see that while all three transition count distributions follow a power law distribution, the ClickStream distribution is far steeper than that of WikiClickIU. This implicates that the most frequently used transitions in ClickStream also exert a far greater impact on the co-occurrence counting vectors than they could in WikiClickIU. Consequently, the impact

of the frequent transitions in ClickStream is reduced much stronger when modelling the same data using binarization than those in WikiClickIU, which in turn explains the discrepancy in evaluation performance gains. As the slope of the transition count distribution of WikiGame is roughly the same as that of ClickStream, although with a lower γ offset, this also explains why the semantic content of WikiGame navigation can still benefit from being modelled with binarization, however not as strong as that from ClickStream.

7.2.3.2 Benefit of Human Unconstrained Navigation over the Link Network

As another experiment, we compare results on ClickStream against four baselines constructed from the Wikipedia link network and ClickStream. Our goal is to show that user navigation in the form of the selection of source pages as well as corresponding links in ClickStream contains more semantic information than the underlying link network alone. We will first explain the construction procedure of the four baselines. The evaluation results regarding semantic relatedness extraction from the sampling baselines are discussed afterwards.

Experimental Setup. We build four baseline datasets based on WikiLink and ClickStream:

- (i) *full WikiLink network*: The most simple baseline interprets each link in the WikiLink network as a single request. This represents the plain information provided by the static link network.
- (ii) *restricted WikiLink network*: Now we want to know if the selection users make by (not) visiting certain Wikipedia pages contains semantic information. Thus, we build the *restricted* baseline by removing all requests from the binary dataset whose source page is not observed as a source page in ClickStream.

Additionally, we assume that (not) using certain outlinks from a source page also contains semantic information. Thus, we build two sampled datasets based on the *restricted* baseline.

- (iii) *random sampling*: For the *random sampling* baseline, we randomly sample outlinks (not individual requests) without replacement from the *restricted* dataset set until we have the same number of overall observed outlinks as in ClickStream.
- (iv) *distribution sampling*: For the *distribution sampling* approach, we do the same but also keep the number of outlinks for each source page the same as in ClickStream. We generated 10 samplings for each variant, calculated the semantic relatedness performance for each and took the mean of the correlation values.

Results. In Table 7.6, we see that source selection (*restricted*) only shows a small improvement but is based on a considerably smaller dataset (240.1M vs. 494.2M links

Table 7.5: Performance comparison of our datasets on common evaluation pairs. In each column, we give the correlation values ρ_{ds}^x for the denoted evaluation dataset ds and the denoted number of common pairs x . The n/a values are given when we only compared ClickStream and WikiGame on a greater number of pairs than those matchable together with WikiClickIU. “Counting” denotes the co-occurrence counting method described Section 7.2.1.1 and “binary” the binarization approach, as explained in Section 7.2.1.2. It is important to note that the scores are all obtained using a window size of $w = 2$ to make them more comparable, since ClickStream only contains page-to-page transitions instead of longer paths.

Dataset	ρ_{ws}^{224}	ρ_{ws}^{107}	ρ_{ws}^{max}	pairs
ClickStream (counting)	0.506	0.388	0.527	288
ClickStream (binary)	0.706	0.543	0.709	288
WikiGame (counting)	0.688	0.500	0.638	236
WikiGame (binary)	0.722	0.563	0.728	236
WikiClickIU (counting)	n/a	0.398	0.419	120
WikiClickIU (binary)	n/a	0.454	0.458	120

(a) Results on WS-353

Dataset	ρ_{MEN}^{1305}	ρ_{MEN}^{468}	ρ_{MEN}^{max}	pairs
ClickStream (counting)	0.399	0.445	0.370	2906
ClickStream (binary)	0.712	0.494	0.640	2906
WikiGame (counting)	0.581	0.412	0.575	1396
WikiGame (binary)	0.644	0.461	0.639	1396
WikiClickIU (counting)	n/a	0.268	0.225	568
WikiClickIU (binary)	n/a	0.278	0.214	568

(b) Results on MEN

Table 7.6: Sampling results on baseline datasets based on WikiLink and ClickStream. All results are obtained using binarization vectors.

Dataset	ρ_{WS-353}	pairs	ρ_{MEN}	pairs
full WikiLink network	0.625	270	0.548	2137
restricted WikiLink network	0.629	269	0.553	2131
random sampling	0.272	65.8	0.091	362.5
distribution sampling	0.431	156.4	0.384	1011.4
ClickStream (binary)	0.709	288	0.640	2906

and 29M vs. 1.4M source pages) when compared to the *binary* WikiLink dataset⁷. At the same time, we see, that selecting the right amount of links at each source has a strong effect on semantic relatedness (*random vs. distribution sampling*). And finally, we observe that selecting the correct sources as well as the correct links (ClickStream (binary)) results in the best performance for extracting semantic relatedness. Thus, we showed overall that indeed *user behaviour and not merely the underlying link structure is an important factor* for being able to extract semantic relatedness from unconstrained navigation data.

7.2.3.3 Conclusion

The central question of section was if unconstrained navigation on Wikipedia can be used as a source of high quality semantic information. We could show that this is indeed the case, however only if we are able to extract that information in a suitable way.

We compared the evaluation performance of different models to represent semantic information, namely co-occurrence counting and binarization, on two semantic relatedness datasets, WS-353 and MEN. Then we compared these evaluation scores to those obtained from extracting semantic information from game navigation, as well to several baselines based on the static Wikipedia link network.

In this section, we investigated on extracting semantic relatedness from unconstrained navigation on Wikipedia as opposed to game navigation. For this, we applied both the co-occurrence counting and the binarization methods presented in Sections 7.2.1.1 and 7.2.1.2 on two large sets of unconstrained navigation on Wikipedia and compared results with those when applying the same methods on game navigation in the WikiGame.

We found that unconstrained navigation contains a great deal of semantic information by first constructing word vectors and then evaluating them on two standard evaluation datasets. Furthermore, we verified that the transition information from ClickStream data contains a significantly increased amount of semantic intuition compared to the plain link network and adapted baselines. Finally we could show that in certain cases, the vectors constructed by the binarization approach were of even higher quality than those constructed from game navigation.

7.2.4 Random Walks on the Wikipedia Graph

In the preceding two sections, we could show that Wikipedia navigation contains a great amount of semantic information, both in a game (cf. Section 7.2.2) and in an unconstrained setting (cf. Section 7.2.3), although there exist differences in the generating navigation behavior (cf. Section 5.2).

Despite easily collectible and available in great amounts with a lot of side information (cf Section 4.2.1), game navigation is heavily biased by the game setting that the data is collected in. When extracting semantic relatedness from navigation, it is more intuitive

⁷The restriction of WikiLink to ClickStream source pages (*restricted*) should actually contain the same number of matchable evaluation pairs. We attribute this difference to the ever changing nature of Wikipedia.

to exploit unconstrained navigation, i.e., organic navigation on the Wikipedia graph, as it lacks the game bias and represents the natural behavior of people [209]. On the other hand, datasets with unconstrained navigation are only publicly available in a heavily anonymized form, mostly due to privacy reasons (cf. Section 4.2.2). Additionally, available data are either very sparse (such as the WikiClickIU data, cf. Section 4.2.2.2) or represented only by aggregated transition counts, as in the ClickStream dataset (cf. Section 4.2.2.1). While both datasets directly represent natural human navigation on Wikipedia, they still either lack statistical mass or extensive context information. Consequently, we attempt to answer the question *if we can generate random walks to at least alleviate the lack of context in the ClickStream dataset, so we can still extract useful and high-quality semantic information from the generated random walks*. If we can do that, then the issue of insufficient statistical mass obviously is of no concern anymore.

While we cannot correctly reconstruct the actual navigation paths by different users from only transition information, it is still possible to utilize this information in order to simulate human-like navigation by employing different random walk strategies. This has already been attempted in several works: West and Leskovec also explored different strategies to parameterize automatic navigation, i.e., controlled random walks [246]. Similarly, Trattner et al. also proposed different strategies to reach a given *target node* [228], based on decentralized search [119]. However, the goal of both studies was to find the shortest path between two given pages instead of generating additional context to better extract semantic information from navigational paths. In this section, we also *explore different strategies to parameterize the random walk generation process, such that we can approximate the semantic content of human information*.

In the previous sections, we have extracted semantic information from datasets of human navigation on Wikipedia using two different methods. In this section however, we will attempt to overcome design limitations of both datasets, such as the restriction of ClickStream to only page-to-page transitions. Concretely, we exploit the navigational information to augment existing datasets so that we can extract more meaningful semantic information from them by extending the potential context for each page. We achieve this by *generating random walks on the Wikipedia link graph and analyzing their semantic content*. We first perform a qualitative evaluation of this approach by simulating uniform random walks, i.e., completely random navigation according to the link network. After this, we explore different strategies to generate *biased* context from the underlying link network, using navigation hypotheses from Section 5.2.1 to parameterize the random walk process as a simulation of human navigation behavior.

In the following, we first describe our framework for generating random walks on the Wikipedia link graph and look for suitable hyperparameters. Although this drastically reduces the set of usable page concepts, it is in line with the intuition that a great part of pages is visited rarely and does thus not contribute much to the context of more often visited pages. As a next step, we explore the influence of different navigation hypotheses on the semantics of random walks. For this, we make use of the navigational hypotheses used to characterize navigation as presented and discussed in Section 5.2.1.

7.2.4.1 Contribution of Random Walks to Semantic Relatedness

We first analyze the general feasibility of random walks to contain semantic information. After this, we will determine a sufficient minimum number of random walks per node as well as a suitable walk length to stabilize results. We also experiment with different restrictions of the underlying link network by determining the Page Rank of a node and investigate the influence of this restriction on the scores of generated random walks.

Random Walk Generation We will now describe how we generate our random walks. Given a navigation hypothesis \mathcal{H} (e.g., as described in Section 5.2.1) and a corresponding page graph $G = (V, E)$ (cf. Section 3.1.2), it is then possible to perform random walks across G using the transition probabilities from \mathcal{H} to generate pseudo-human navigational paths. In order to capture information about most of the nodes, a fixed number c of random walks is executed for every node in the graph. To restrict the generated paths to a maximum length of l_{max} the random walk is stopped when the path length reaches l_{max} . In each experiment, the next edge to follow in an unfinished path is chosen according to a probability distribution over the weights of the outgoing edges of the current node. Possible distributions are partially based on the navigation hypotheses defined in Section 5.2.1 and are shortly reiterated for convenience later on. Since the investigated graphs are directed it is worth pointing out that only nodes with an out-degree $outdeg(p) > 0$ are used to start random walks, effectively eliminating the possibility of paths with a length of $len(\mathbf{p}) = 0$.

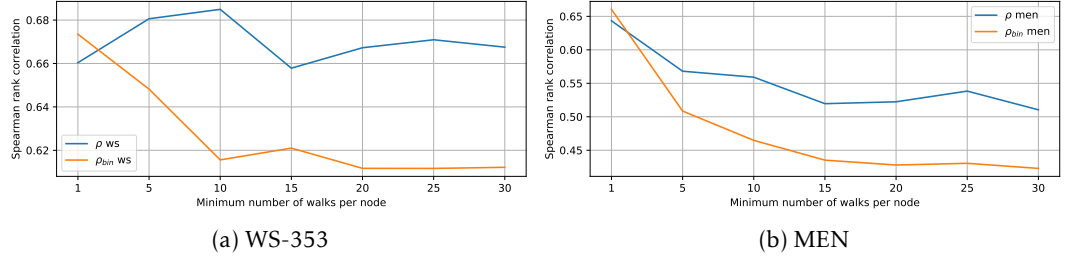
In order to extract semantic information from these paths, we again use the co-occurrence counting and binarization methods described in Section 7.2.1.1 and Section 7.2.1.2. In all experiments, we apply the evaluation scenario as presented in Section 3.2.4 and report the Spearman correlation results with human intuition on WS-353 and MEN (cf. Section 4.4) for co-occurrence counting and binarization as ρ and ρ_{bin} , respectively.

Determining Suitable Hyperparameters In the following, we first conduct parameter studies to determine a mostly well-performing setting for the random walk length and minimum page count, before we experiment with a reduction of the underlying link corpus according to the PageRank of the nodes.

Influence of Number of Walks per Node and Walk Length In the first experiment, we attempt to determine a fitting number of walks started from each node to accumulate enough information to generate meaningful concept vector representations. For this, we again fix the window size to $w = 3$, since we now again work on navigational *paths* instead of *transitions*, and increase the number of walks c for each node in steps of 5 from 1 until 30. The random walks are performed on the Wikipedia link graph from February 2015 (see Section 4.2.3).

Table 7.7 shows that co-occurrence counting largely exhibits the same performance for increasing number of walks, while binary co-occurrence counting instead decreases in performance with increasing c . With the exception of $c = 1$, no apparent dependency

Table 7.7: Spearman correlations achieved with co-occurrence counting and binarization by varying the number of walks started on each node in the network. The window size was set to $w = 3$. ρ denotes the Spearman correlation score when evaluating the co-occurrence counting model. ρ_{bin} gives the evaluation score for the binarization model.



walks/node:		1	5	10	15	20	25	30
WS-353 (306)	ρ	0,660	0,681	0,685	0,658	0,667	0,671	0,668
	ρ_{bin}	0,673	0,648	0,616	0,621	0,612	0,612	0,612
MEN (2943)	ρ	0,643	0,568	0,559	0,520	0,522	0,538	0,510
	ρ_{bin}	0,660	0,509	0,465	0,435	0,428	0,430	0,423

(c) Scores

between the number of walks and the model performance can be observed. We attribute the reason for the slightly different evaluation scores on WS-353 to the randomness in the walk generation process and accept a certain tolerance in these values, as these differences are not statistically significant. On MEN, the difference between 5 and 10 minimum walks per node is not statistically significant either, but the difference between 10 and 15 walks per node is.

Therefore we decided to set the number of walks to $c = 10$ for all further experiments, as this represents an equilibrium between sufficiently much generated information per node as well as mostly a sufficiently good performance.

PageRank Network Reduction With this experiment we want to study how the network size affects the performance of the different semantic vector construction approaches. There are many ways to reduce a network in size but the most intuitive is to only keep the most relevant nodes of a network. To do so, we use the same Wikipedia network as in the previous experiment and compute a ranking for all nodes using Pagerank with 20 iterations. The ranking is then used to create a series of pruned networks using only the top k percent nodes and corresponding edges. Statistics about the obtained networks can be found in Table 7.8 and Table 7.9.

As before we perform a series of random walks on the different networks and evaluate the performance of the two models. Figure 7.8 shows the Spearman correlations for the different networks and approaches. Our results show that the network reduction does

Table 7.8: Overview of all link networks that we use to generate random walks.

dataset	V	E	\emptyset outdeg.
WikiLink-10%	480 150	71 026 160	147,92
WikiLink-20%	960 300	153 751 696	160,11
WikiLink-30%	1 440 450	199 933 460	138,80
WikiLink-40%	1 920 599	231 956 454	120,77
WikiLink-50%	2 400 747	255 128 647	106,27
WikiLink-60%	2 880 883	271 895 395	94,38
WikiLink-70%	3 360 965	285 566 885	84,97
WikiLink-80%	3 841 059	297 133 727	77,36
WikiLink-90%	4 321 158	308 033 742	71,28
WikiLink-full	4 801 501	315 049 408	65,61

Table 7.9: Basic statistics for all random walk datasets generated from Wikipedia link networks. An average path length of less than 21 means that some walks ended with a leaf node, before the max path length could be reached.

dataset	#paths	\emptyset len	\emptyset page freq
Walks-10%	4 800 410	20,90	208,94
Walks-20%	9 601 410	20,92	209,15
Walks-30%	14 402 200	20,92	209,18
Walks-40%	19 202 710	20,92	209,17
Walks-50%	24 001 980	20,92	209,13
Walks-60%	28 800 120	20,91	209,08
Walks-70%	33 595 690	20,91	209,03
Walks-80%	38 389 140	20,91	208,98
Walks-90%	43 178 910	20,91	208,94
Walks-full	47 975 930	20,91	208,92

not negatively affect the model performance for co-occurrence counting and that they perform better than the binarization semantics. However, network size reduction actually has a detrimental effect on the co-occurrence counting scores. The binary counting approach generally performs significantly worse but definitely profits from the network reduction. Overall, a network reduction on the top 50% most relevant nodes according to their PageRank yields a rational trade-off between sufficiently good results across both evaluation datasets and both co-occurrence counting and binarization methods, and simultaneously offering a large enough amount of nodes to not only include high-degree nodes, which (as we already noted in Section 7.2.2.3) do not necessarily yield good semantics in navigational paths.

7.2.4.2 Benefit of Biased Random Walks over Uniform Random Walks

In the following, we explore whether biased random walks contain better semantic information than unbiased random walks. Until this point, we investigated uniformly random navigation on the Wikipedia link network, i.e., the next page in a walk is

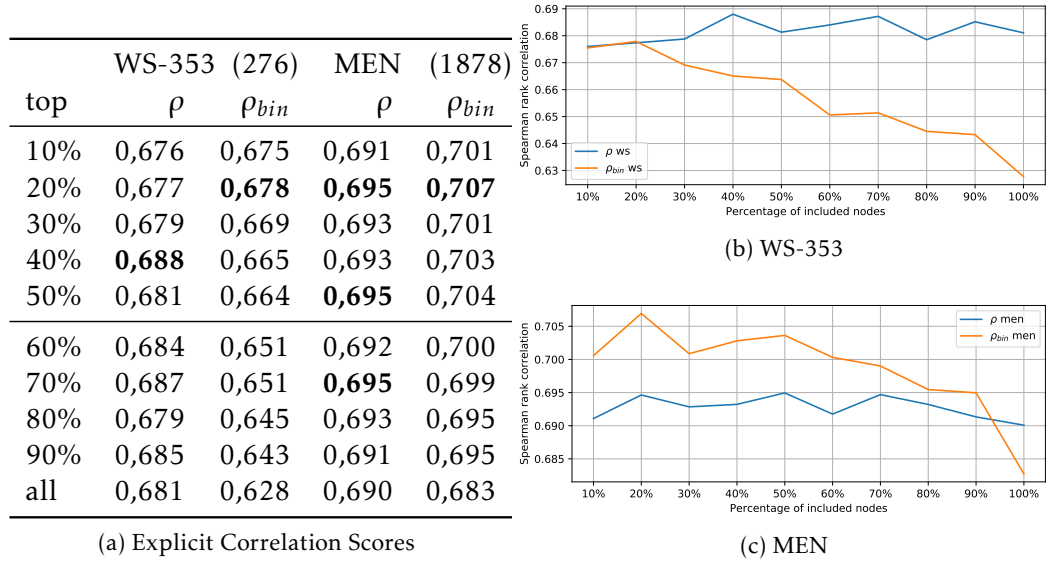


Figure 7.8: Spearman correlations on WS-353 for networks pruned to the top k -percent nodes according to Pagerank. We set the window size $w := 3$. All scores were evaluated on the same set of evaluation data, thus the score at 100% of nodes differs from the score for a min path count of 20 given at Table 7.7.

randomly chosen among the possible outlinks. In order to investigate in how far different navigation strategies exhibit an influence on the extractable semantic information, we will now restrict potential navigation to the networks of WikiGame and ClickStream. We can see in Table 7.10, that the most navigation in these datasets is performed on the top 50% most relevant PageRank nodes, which is also the subset with the best correlation results across both evaluation datasets and both the co-occurrence counting and binarization methods (cf. Figure 7.8). Because of this and since we then can compare our scores to actual human behavior, we conduct all experiments on the WikiGame and ClickStream subnetworks.

The different navigation strategies with which we will experiment are partially parameterized by the navigational hypotheses presented in Section 5.2.1, which we will conveniently reiterate in the following. As baseline strategy we use the *uniform structural* hypothesis.

- The *uniform structural random walk hypothesis* generates uniformly random navigation with respect to the possible links, i.e., users randomly click any link, but still follow the network structure.

We will obviously not focus on the *self-navigation hypothesis* as a parameterization of random walks, since no node would occur in any context of other nodes and thus there would be no semantic relations which we could extract.

We also experimented with more elaborate random walk parameterizations, such as:

Table 7.10: Overlap of the Wikipedia page subsets according to their importance with respect to PageRank with the WikiGame and ClickStream datasets. We furthermore calculated the overlap ClickStream $n > \mu$ with all pages in the ClickStream data where the transition count n was greater than the mean μ of all ClickStream transition counts, with $\mu \approx 80$.

Top	WikiGame	ClickStream	ClickStream $n > \mu$
10%	50.4%	18.7%	30.2%
20%	64.3%	30.7%	43.4%
30%	73.8%	42.3%	55.3%
40%	80.8%	54.3%	67.2%
50%	86.5%	66.4%	77.8%
60%	90.4%	77.0%	85.8%
70%	93.0%	85.6%	91.4%
80%	94.5%	92.2%	95.2%
90%	95.6%	96.3%	97.6%
100%	97.6%	99.7%	99.7%

- the *degree-based hypothesis*, which can be used as an outdegree- or indegree-based hypothesis: The next page would be chosen according to the in- or outdegree of the following page. We consider two variants of the degree-based hypothesis: we either choose a node with a probability *directly* or *indirectly* proportional to its degree.
- the *human-like hypothesis*, where the next page is chosen according to the click probability distribution exhibited by the actual navigation data from the WikiGame or the ClickStream datasets,
- and finally the *semantic hypothesis*, where the article text from each page was transformed into a TF-IDF vector with sublinear TF scaling. To reduce the amount of features, we applied a Sparse Random Projection approach [136], which, according to the Johnson-Lindenstrauss-Lemma, allows us to reduce the vector dimensionality with a guaranteed error on the mutual vector distances of less than 10%. [1] The visitation probability of the next page is then proportional to the semantic relatedness of this page to the current page according to the cosine similarity of the corresponding TF-IDF vectors.

First, we describe the experimental setup. That is, we explain our parameter choices. Second, we present evaluation results and discuss them. We find that while uniform random walks produce semantically useful navigational paths, we increase the semantic quality of those paths by introducing specific biases. Still, not all biases produce useful navigational paths.

Experimental Setup For the random walk generation the length of a random walk is fixed to $l = 20$ transitions, e.g. a walk contains a maximum of 21 nodes. The number of walks per node was set to 10, as mentioned in Section 7.2.4.1. We thus ensure that there is enough context for each node when experimenting with differently sized context windows. A reduction of the network size is not necessary, since navigation on the WikiGame and ClickStream networks happens mostly on the top 50% most relevant PageRank nodes (see Table 7.10). At each step, we first randomly chose the next node connected by a hyperlink to the currently last node in a walk. During a random walk it is possible to reach a node with no neighbours. If that happens, the path cannot be extended anymore, but is still kept in the set of generated paths.

We perform random walks for all the navigation strategies described in the previous subsection and subsequently generate vectors using the cooccurrence counting and binarization methods using window sizes between 2 and 5. Each set of vectors is finally evaluated on the WS-353 and MEN datasets.

Results and Discussion First, we analyze the semantic relatedness scores produced by walks on the WikiGame data and then proceed to investigate the effects of different random walk strategies on the ClickStream network. Please note that the results on WikiGame do not actually represent game navigation, as they are not *goal-oriented*, which is a central feature of navigation in a game setting.

Generally, artificial navigation based on the actual transition counts gathered across all games in the **WikiGame** performs best of all random walks (∇ , pink line). However, human navigation almost always generates the best semantic information (\diamond , gray line). In Figure 7.9, we can see that degree-based navigation directly proportional to the degree of the target node yields the worst results way below the baseline of uniform navigation (blue and orange line, \times and \circ). *If however we revert that notion so that the probability of the next node being chosen is indirectly proportional to the degree (red and purple lines, $+$ and \diamond), the generated random walks yield approximately as much semantic information as the human-like random paths.* When modelling the semantic information in these walks with co-occurrence counting, they are even able to slightly outperform human navigation on a window size of 2, as can be seen in Figure 7.9a and Figure 7.9c. This fits well with the observation that paths with a low mean indegree in the WikiGame yield the best extractable semantics (see Figure 7.4). Especially the evaluation of the count-based vectors based on the random walks also shows that we can transport the optimal window size of 3 or 4 for the extractable semantic information both on WS-353 and MEN. On the binary vectors, we unfortunately cannot reproduce or mirror these results as in the co-occurrence counting vectors. There, we observe a steady decline in performance of all hypothesis-based walks. While this decline can also be seen in human navigation (gray line, \diamond) all random paths yield consistently weaker semantic information than the human navigation baseline. This indicates that human navigation seems to contain factors that cannot be captured by the simple random walk experiments that we performed, *but in some cases, we can approximate its semantic content.* A rather surprising observation is that the semantic hypothesis (brown line, Δ) always produces

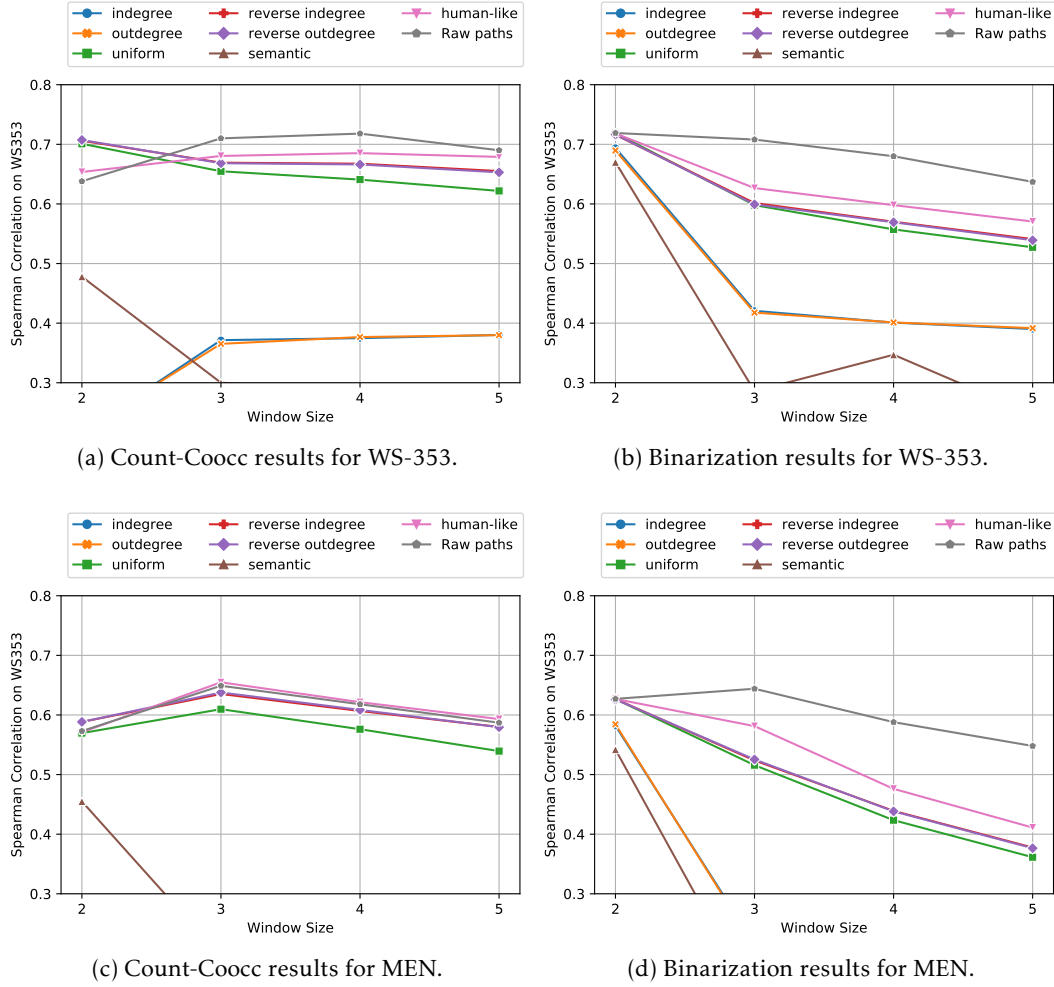


Figure 7.9: Semantic evaluation results for random walks on the WikiGame graph parameterized by different navigation hypotheses. The x axis denotes the window size applied in the semantic information model, while the y axis denotes the Spearman correlation score. We can see that the indegree and outdegree hypotheses consistently generate insufficient semantic information, while the other three hypotheses are able to approximate the semantics in random walks generated from human navigation behavior, at least with regard to its semantic content. However, human navigation itself almost always outperforms all random walks.

inferior results, being almost outperformed by all other hypotheses.

For random walks on the **ClickStream** graph, Figure 7.10 shows that the context provided by only taking simple transitions (i.e., $w = 2$) into account is not sufficient to fully cover the information contained in the ClickStream random walks, although we have shown in Section 7.2.3 that this information also contains notable amounts of semantic relatedness and can be best extracted using the binarized vector construction approach. What is interesting however is that even although we only have information about simple transitions from one page to another and no information about any longer paths taken, we see that *the best extractable semantics are achieved with a window size of 3 to 4, but along random walks with no direct human influence* (see for example Figure 7.10c). Although the random walks parameterized by human behavior (pink line, ∇) often have a slight edge on either uniform (green line, \square) or inverse degree-based random walks (red and purple lines, $+$ and \diamond), the contained semantic information in the generated random walks is not so much better than that of the more artificial hypotheses. A rather surprising fact is that the semantic hypothesis arguably produces the worst random walks for semantic extraction (brown line, \triangle). Only with the co-occurrence count vectors on MEN we see a good correlation score at window size 2. This could indicate that walking to nearby pages that are semantically similar also improves the semantic information contained in the resulting paths. Why however this doesn't replicate with WS-353 correlation we don't know.

Again, as already observed with the random walks on the WikiGame network and on game navigation in Section 7.2.2.3, navigation by high degree (blue and orange lines, \circ and \times) yields less usable semantic information than navigation by low degree, meaning that more specific concepts, i.e., with fewer in- and outlinks, exhibit a greater influence on the granularity of the semantics contained in the generated random walks. Additionally, the uniform hypothesis again yields very good results and often outperforms all other hypotheses, when evaluated with WS-353, indicating that more complex (but still simple) parameterizations do not necessarily result in higher semantic information. In contrast to the binarization results shown in Figure 7.9, the binarization vectors on ClickStream navigation do not always exhibit worse performance. Although they cannot improve upon the results of the co-occurrence counting models, they do not generate much worse semantic information. Again, we attribute this to the slope of the transition count distributions of the walk datasets. Most notably however, we cannot reproduce random walks that are able to beat those scores on a window size of $w = 2$, as we see in Figure 7.10b and Figure 7.10d.

The first question that we wanted to address in this section was if we can generate random walks to make up for the lack of context in the ClickStream dataset. We can see in Figure 7.10 that this is indeed the case, when comparing the evaluation scores of our vector models on window size of $w = 3$. Compared to $w = 2$, which actually is the same as direct page-to-page transitions, we can achieve a great score improvement of at most 0.22 Spearman correlation score (Figure 7.10c). Regarding the question how well we can approximate human navigation with parameterized random walks, our results indicate that on the one hand, we are indeed able to approximate the semantic content in human navigation in some cases by generating random walks according to rather simple

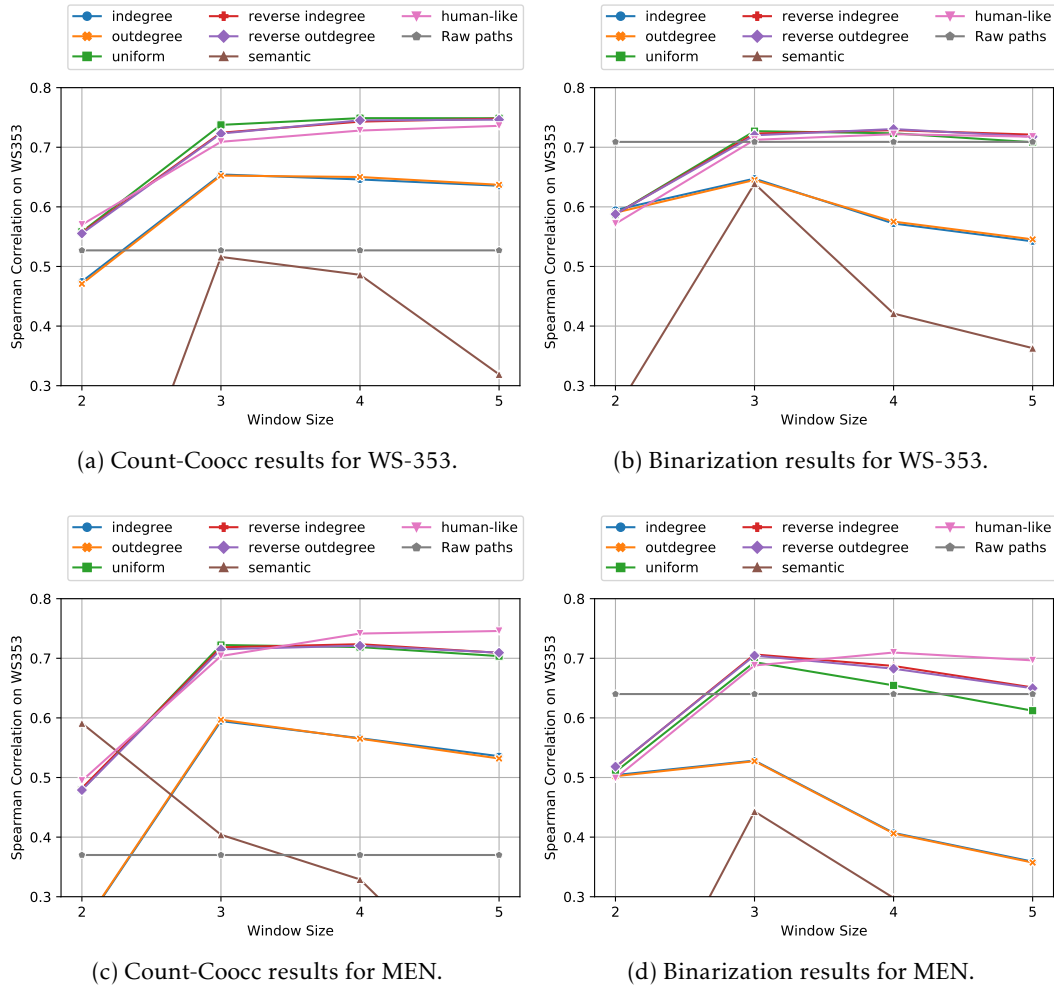


Figure 7.10: Semantic evaluation results for random walks on the ClickStream graph parameterized by different navigation hypotheses. The x axis denotes the window size applied in the semantic information extraction model, while the y axis denotes the evaluation Spearman correlation score. In most cases, the human-like generated navigational paths generate the best semantic information (pink line, ∇). Additionally, we can see that window sizes of length 3 already capture a great deal of semantic information. The *raw path* baseline (gray line, \diamond) here is provided by the scores obtained from page-to-page transitions, i.e., paths of length 2, as given in Table 7.5. Because of this, the baseline scores seem very weak.

hypotheses. On the other hand, it can be seen especially in Figures 7.9b, 7.9d, 7.10b, and 7.10d, that modelling the semantics in random walks using binarization cannot improve scores as much as if we use it to model human navigation itself. In contrast, this is another indicator for our result from Section 7.2.3, that binarization is useful to capture human navigation, especially in the ClickStream dataset.

7.2.4.3 Conclusion

In this section, we generated random walks on several Wikipedia graph networks to make up for lacking datasets with a big enough amount of human navigation and sufficient side information. For this, we compared different approaches to perform random walks on different networks of Wikipedia articles in order to extract semantic relatedness. In that regard, we first studied a suitable parameter setting for the generated random walks to contain sufficient semantic information, before we generated random walks using some of the navigation hypotheses presented in Chapter 5.

Our results show that it is possible to generate random walks using simpler hypotheses instead of human behavior, so that the generated walks can be used to extract semantic relatedness, in some cases even on par with that from real-world human navigation. We confirmed this on two Wikipedia datasets with human navigation, where one is game-based while the other has no navigational constraints, by evaluating the contained semantic relatedness in the generated random walks with two state-of-the-art methods on two large and well-known semantic relatedness evaluation datasets, namely WS-353 and MEN. As we showed in Section 5.2, that humans on Wikipedia tend to follow both reverted indegree and outdegree hypotheses, both in game and unconstrained settings, we can use these hypotheses as a reasonably good proxy for human navigation and thus only need access to the link network itself to receive page vectors which exhibit semantic quality on par with those generated from human navigation. A result that is not so much in line with the results in Section 5.2 is that the semantic hypothesis produces the least useful paths for semantic extraction. We attribute this to the fact that two pages that are semantically similar to a third page do not necessarily have to share any semantic information with each other. However, our results also showed that human navigation, where available, still often outperformed random navigation. This indicates that human navigation still contains behavioral factors that cannot be modeled using simple random walk hypotheses and leaves room for exciting future work.

7.2.5 Summary and Discussion

Wikipedia is a big interlinked collection of knowledge that users navigate to find information [209]. Since both Wikipedia articles [81] as well as the link network of Wikipedia [251] already contain a lot of semantic information, we now searched for ways to extract semantic information from the *usage of these resources*, concretely from *human navigation on Wikipedia*.

7.2.5.1 Summary

To model that semantic information as vectors like in the vector space model from Section 3.2.1.1, we proposed two approaches that are based on co-occurrences of pages in a navigational path.

We started our investigations with a deep analysis of semantics extracted from *game navigation* on the two datasets WikiGame and Wikispeedia (cf. Section 4.2.1). Here we found that not only human navigation in a game setting contains an elevated amount of semantic information compared to the plain link network, but also that there exist subsets of navigational paths that are more suited to extract that semantic information from. Next, we wanted to transfer the semantic extraction methodology to *unconstrained navigation*, i.e., human navigation on Wikipedia without external restrictions as they are encountered in the game setting. Again, we performed our studies on two datasets, namely ClickStream and WikiClickIU (cf. Section 4.2.2). Our results showed that this unconstrained navigation also contained a robust amount of semantic information, however only after modifying the way we modeled that information as vectors, i.e., when using binarization as the modeling technique. Still, we were able to show that both game and unconstrained navigation contain large amounts of semantic information. A big drawback of all navigation datasets that we had access to were their limitations, either their amount or their design. For example, ClickStream only contains accumulated page-to-page transitions instead of full navigational paths like the WikiGame dataset. In Section 7.2.4, we sought to overcome this limitation by *generating random walks* that we parameterized using the navigation hypotheses from Section 5.3.1. Our results showed that our artificially generated paths contained semantic information that in some cases was even on par with human navigation. Additionally, we also could show that artificially extending the ClickStream transitions to longer paths improved the extractable semantic information by a large margin. Despite that, we were unable to consistently surpass the semantic information contained in human navigational paths.

7.2.5.2 Discussion of the Results

We want to include a brief discussion of the relations between the results obtained in this chapter.

Game Navigation vs Unconstrained Navigation. Despite the fundamental differences of game and unconstrained navigation on Wikipedia, in external restrictions as well as the in the way we extract semantic information, both are a good option to extract semantic relatedness. However, since unconstrained navigation is more focussed on retrieving information about concrete instances of abstract concepts, compared to game navigation (cf. Table 3.2), the constructed vector representations might be useful for different applications. For example, while we could use game navigation vectors to construct abstract taxonomies of concepts, we could use unconstrained navigation vectors to improve Wikipedia’s search system. In fact, West, Paranjape, and Leskovec used logs of unconstrained navigation to recommend missing links in Wikipedia [248].

Automatic vs Human Navigation. To overcome potential limitations in data collection and dataset design, we showed that we can generate random walks on Wikipedia that contain an increased amount of semantic information. Still, our results showed that human navigation is superior to automatic walkers, even if we parameterize the walk generation with human transition counts. While one does not have access to longer paths on Wikipedia though, the random walks provide an interesting alternative to still conduct experiments on human-like paths.

Co-Occurrence Count vs Binary Vectors. We showed that both methods are able to extract meaningful semantic information from human navigation, although each works better than the other in different settings. We also identified potential reasons for that, concretely the slope of the Zipfian transition distribution (cf. Figure 7.7). What we did not investigate was a *mixture* of both representations. Although the highly frequently used transitions impact the co-occurrence count model strongly, we assume that the most useful semantic information lies in less used transitions, which in turn are ignored by the binarized model. We expect that for example a logarithmic model or a model with a maximum frequency cutoff would be able to perform well in both navigation settings.

Further considerations. All of these results indicate that there are more factors in human navigation on Wikipedia that need to be explored in order to increase the extracted amount of semantic information. For example, as the WikiGame navigation first showed, we are able to extract meaningful semantics from human navigation. An exciting indicator however also lies in the exploration of unconstrained human navigation on the public Wikipedia. Through the sheer mass of traffic there, we could obtain high quality semantic information which we expect would be very finely grained. Additionally, we could focus more on navigation of different user types there, for example *shallow information seekers*, *deep divers*, or *random surfers*, as were identified by Singer et al. [209]. It would also be interesting if the navigation behavior of these user types still adheres to the patterns observed by West and Leskovec [247]. Again, we could then use these more complex navigation types to generate random walks to become more independent of actual user behavior, which is always a limiting factor, for example due to privacy reasons, as navigation on Wikipedia also includes navigation on one's own pages. Lastly, we could also utilize the vast potential of neural networks to learn latent features of user navigation or navigational semantics.

7.3 BibSonomy

In the previous section, we have investigated different types of human and pseudo-human navigation on the Wikipedia graph in detail in order to extract semantic relatedness from it. We could show that, as the hypothesis-based analysis in Chapter 5 suggested, navigation on the Wikipedia link graph contains considerable amounts of

semantic information. We also proposed and validated several methods to extract this information.

Since we could already show in Section 5.3 that BibSonomy navigation is to a large factor driven by semantic information. In this section, we want to transfer the previously proposed methodology for Wikipedia to the BibSonomy graph (see Section 4.2.5). However, there still are obstacles that need to be overcome. First, the page state space of Wikipedia is limited. A user can only navigate to any other existing page. In BibSonomy, the number of visitable pages is nearly endless, since most of them are generated at runtime. This could lead to a large number of rarely visited pages which do not contribute much information to navigation. Second, BibSonomy users tend to navigate largely on their own pages (cf. Section 5.3.2.2) and revisit the same page very often [64]. This can have severe influence on the usefulness of the extracted semantic information. Finally, we seek semantic information about *tags* instead of the actual pages, as opposed to Wikipedia, where we equated a semantic entity with the page where it is described. Although there are tag pages in BibSonomy which describe that tag through a folksonomy subset, we also would obtain representations of users, resources and mixes thereof. We discuss these and other issues in the remaining section.

Our general goal is to model the semantic information about tags that is contained in human navigation on BibSonomy, as we have shown in Section 5.3. For this, we introduce two adaptations of the semantic information extraction models that we applied on Wikipedia (cf. Section 7.2.1). We subsequently apply them on the human navigational paths dataset on BibSonomy, which we already introduced in Section 4.2.4.2.

In this section, we first describe the method to model semantic information from human navigation on BibSonomy as well as the evaluation procedure and the applied datasets. After this, we perform several experiments to optimize the extraction performance, which include limiting the tag cloud or selecting only specific tags to characterize pages (Section 7.3.2). Finally, in Section 7.3.3 we attempt to simulate human navigation behavior by performing random walks biased by the navigation hypotheses presented in Section 5.3.1. We hope to see, like we have shown for Wikipedia navigation in Section 7.2.4, that there are some hypotheses that are more useful to simulate navigation with high semantic content than others. Lastly, we discuss our results.

7.3.1 Preliminaries

Before we describe our experiments to extract semantic information from navigation on BibSonomy, we will in the following describe the general modelling and evaluation procedures that we applied. We start with the definition of our content-based navigational semantic relatedness measure that we apply on BibSonomy navigation. After this, we state which datasets we used and point to the applied evaluation procedure.

7.3.1.1 Content-based Navigational Semantic Relatedness

Every BibSonomy page describes a subset of the folksonomy, as described in Section 3.1.1.2. We thus also assume that we can model the semantic information rep-

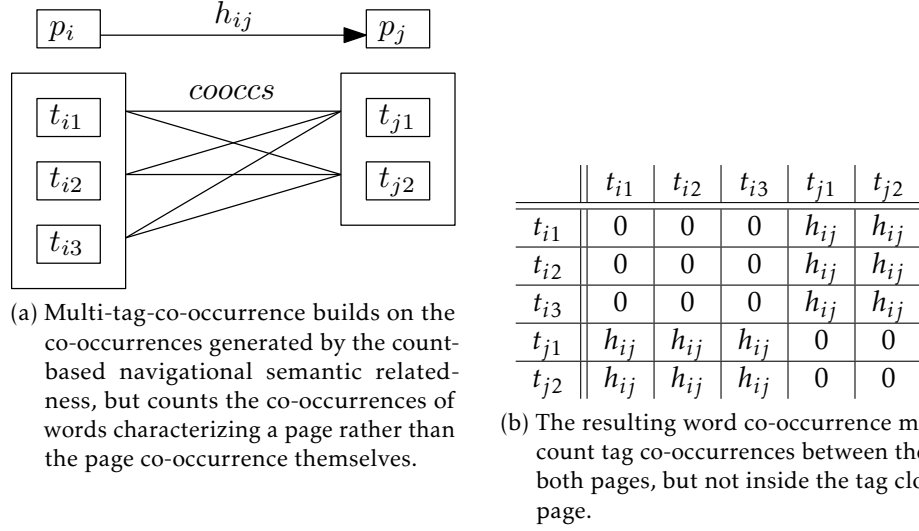


Figure 7.11: Illustration of the Multi-Tag-Co-Occurrence method for a window size of 2. Content-based navigational semantic relatedness builds on counting multi-word co-occurrences of the pages inside a given window. Furthermore, h_{ij} is a possible weight of the transition between p_i and p_j , e.g., a transition count.

resented on that page by its *tag cloud*, i.e., all tags that are shown there. Since the URL of a BibSonomy page is less expressive and precise than that of a Wikipedia page, we think that the respective tag cloud is at least a more verbose description.

The following method is similar to Equation (7.1) in the way that it does not generate *page* vectors based on the navigation context, but actual word vectors based on the content of the context pages.

Based on the count-based navigational semantic relatedness approach to generate co-occurrences from pages inside a certain window of consecutive navigation, the content-based navigational semantic relatedness takes the actual page contents into account and computes co-occurrences between the words on these pages. A sketch of this method is given in Figure 7.11a, the resulting word co-occurrence matrix is shown in Figure 7.11b.

The mathematical formulation of the resulting set of word co-occurrences between two pages p_i and p_j with their corresponding tag clouds T_i and T_j , i.e., the multiset of all shown tags, is as follows: Given two pages p_i and p_j with tag clouds $T_i = \{t_{i1}, \dots, t_{in}\}$ and $T_j = \{t_{j1}, \dots, t_{jm}\}$ of lengths n and m respectively, we count all elements of the *symmetric cartesian product*

$$T_i \times_s T_j := T_i \times T_j \cup T_j \times T_i \setminus \{(t, t) | t \in T_i \cup T_j\} \quad (7.11)$$

as cooccurrences. This adaptation, which we call *multi-tag cooccurrence*, is actually a generalization of the method explained in Section 7.2.1.1, because if we only consider one tag per page, i.e., both tag clouds T_i and T_j have length $|T_i| = n = |T_j| = m = 1$, we end

up with the count-based navigational semantic relatedness.

7.3.1.2 Datasets and Evaluation

We perform all our experiments on navigational paths on BibSonomy, which we introduced in Section 4.2.4.2. To evaluate the vectors that we model from the semantic information in BibSonomy navigation using the method described above, we compute the Spearman correlation of the tag similarity scores with human judgment of semantic relatedness, as described in Section 3.2.4. To represent the human intuition, we use the WS-353 and Bib100 datasets (cf. Section 4.4.3 and Section 6.2, respectively). We did not choose to evaluate on MEN, as it has a comparatively low overlap ratio with tags of BibSonomy of about 12%.

7.3.2 Unconstrained Navigation

Extracting semantic information from navigation on BibSonomy is attractive in some ways. First, in contrast to Wikipedia, usage of BibSonomy is less consumption focused, but to a large part also focused on actually *using* BibSonomy (i.e., posting resources) [64]. Second, we found in Section 5.3 that BibSonomy usage is influenced by a semantic component. Next to this, the semantic nature of tagging systems could additionally influence the semantics of navigation.

Since we could successfully extract semantic information from navigation on Wikipedia, we also expect that BibSonomy navigation yields sufficiently good semantic information that we can extract. For this, we perform several experiments on navigational paths on BibSonomy, including selecting suitable navigation subsets that potentially carry increased semantic information compared to the whole navigation dataset. In the following, we will describe the motivations, setups and progression of each of our experiments, as well as describe the respective results.

7.3.2.1 Page Co-Occurrence Counting Semantics of BibSonomy Navigation

As a first experiment, we follow the same intuition as we did on Wikipedia and assume that a page in BibSonomy represents a single concept only. For example, the page /tag/web denotes the concept *web*, as all resources shown on that page are annotated with the corresponding tag.

we applied the count-based co-occurrence counting and binarization methods as described in Section 7.2.1 to construct context vectors for each page directly on the generated BibSonomy paths. This means, we calculate context vectors for the *pages* we visited and then compare the similarity of the pages using the cosine measure, which is the same as approach as we applied in Section 7.2.

Since in BibSonomy, pages do not fill the role of a concept description as easily in Wikipedia (where we only need the title of the page, because a page normally describes only one concept), we imposed that role of describing a concept on /tag/TAG and /user/USER/TAG pages, i.e., the concept *physics* will be represented by e.g., the BibSonomy page /tag/physics. There are now two possibilities to map a word onto a */TAG

page: We either take only the whole word as the query or we interpret it as a part of the query, e.g., *physics* can be mapped to `/tag/physics` and `/user/USER/physics` (where, of course, we have to cycle through all users which have used this tag) or, more leniently, we also allow partial mappings like `/tag/metaphysics`, which in turn might give us incorrect results such as e.g., `/tag/Samstag` when searching for *tag*, but also allows for a greater variety of possibly correct mappings. If we found several page matches for a given word, e.g., `/tag/physics` and `/user/einstein/physics` and want to know the similarity between *physics* and *science*, we would compute the similarities between all page matches of *physics* and all page matches of *science* and take the maximum similarity over all those match pairs.

The results for both the more narrow, stricter case (`/tag/xyz` is the only tag page match for word *xyz*) as well as the broader, more lenient case (`/tag/vwxyz` also matches *xyz*) can be seen in Figure 7.12. While the lenient matching variant produces higher evaluation scores than the strict variant, we cannot match many words from the evaluation dataset. On those that we can match, we only obtain very weak correlation scores (see Figure 7.12a). Furthermore, increasing the context window does not result in any relevant changes in the evaluation score, either (cf. Figure 7.12b). Binarization also exerts next to no influence on the evaluation results. Finally, the optimistic evaluation indicates a stronger correlation with human intuition than the pessimistic evaluation, but only in the strict matching case. In the lenient case, performance drops strongly, as we obtain almost no correlation ($\rho < 0.1$) anymore.

Because both cases do not yield meaningful results, we investigated possible reasons for this. Since we compare only `*/TAG` pages, where a specific tag was requested, and many of the evaluation words are generally not used often in the audience of BibSonomy, these pages occur very rarely.⁸ The corresponding cooccurrence vectors thus were mostly very sparse, so for most of the WS-353 pairs, we would end up with cosine similarities of 0 between two `*/TAG` vectors and only a few with similarity values greater than 0. Figure 7.12a shows the results for both the standard pessimistic evaluation (include all similarities in the evaluation) as well as for the optimistic evaluation (exclude pairs with similarity of 0), which we have already described above.

Judging from our results, but more so from the statistics around them, we concluded that extracting semantic information from BibSonomy navigation using the page co-occurrence counting method does not yield usable results. While the intuition behind it sounds rational and has been proven in the context of Wikipedia navigation (cf. Section 7.2), we are unable to say if the inferior evaluation scores are due to a lack of navigation data on BibSonomy or due to an imprecise reflection of how semantic information is encoded in BibSonomy navigation.

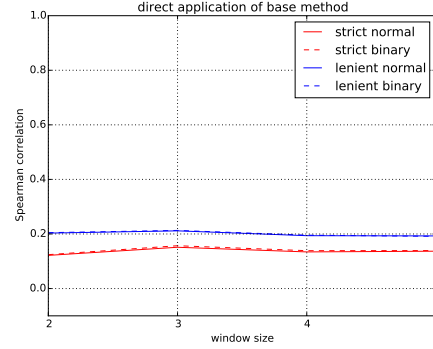
7.3.2.2 Leveraging the Tag Cloud of a BibSonomy Page

Because not every visited page was a `*/TAG` page which can be characterized by the requested tag, we had to somehow induce meaning in other pages like `/user/USER`. As

⁸This only holds for evaluation on WS-353.

approach	ρ	ρ_{bin}	pairs
strict	0.134	0.139	24/156
lenient	0.187	0.182	42/190
strict optimistic	-0.249	0.220	25
lenient optimistic	0.092	0.023	42

(a) Numerical results for the page co-occurrence counting method with window size 4.



(b) Results for the page co-occurrence counting method with pessimistic evaluation.

Figure 7.12: Results for the page co-occurrence counting method on BibSonomy navigation, as described in Section 7.2.1. Here, we considered the URL of a page as a description of its content, similar as in Wikipedia. Because we are interested in tag similarities, we restricted ourselves to `/tag/TAG` pages. We computed both the strict and the lenient variants of matching the page’s tag to a word in our evaluation dataset, as described in Section 7.3.2.1. The strict approach only allows direct tag matching (`/tag/physics` for *physics*), while the lenient approach also allows broader matching (`/tag/astrophysics` for *physics*). Since many of the resulting vectors are extremely sparse, there is a high chance that their similarity score is 0. Optimistic matching means that we consider similarity scores of 0 as lack of data and ignore them in the Spearman correlation computation, while pessimistic matching considers those scores as valid information, i.e., that both tags are semantically dissimilar.

folksonomy navigation strongly depends on the folksonomy metadata, we now leverage the *tag cloud* of a page. Simply using all tags on a page is impractical, since user pages would contribute a huge load of tags, in contrast to resource pages, which usually are annotated with only a few tags. Because of this, we apply several strategies to mitigate this imbalance.

Single Most Frequent Tag. As a first step, we only considered the most frequent tag as a description for a page in BibSonomy, we limit the tag cloud of each page to only a single tag. As mentioned before, this approach then works exactly as described in Section 7.2.1.1.

The results from Table 7.11a show that the single most frequent tag of a page’s tag cloud is not a sufficient representation of the semantic content of that page. This is hardly surprising, as we can gather from the most frequently occurring tags that are

Table 7.11: Evaluation results when characterizing a BibSonomy page by its single most frequently occurring tag. w denotes the window size used in the WikiGame method. ρ and ρ_{bin} are the resulting Spearman correlations for normal and binarized variants. Only a relatively small number of matchable pairs actually yields a non-zero cosine score. Also, with increasing window size, we can match more evaluation pairs, but drop in performance. Additionally, there were 4468 pages that were characterized with the functional tag imported.

k	ρ	ρ_{bin}	pairs		
				tag	pages
				imported	4468
2	0.180	0.203	46/127	wiki	2190
3	0.147	0.188	49/127	web	1278
4	0.118	0.145	51/127	nlp	1232
5	0.104	0.134	52/127	information	1055
				clustering	1038

(a) Evaluation scores on WS-353.

(b) Most frequent tags that occur on at least 1000 pages.

shown in Table 7.11. For example, for 4468 pages, the most frequently occurring tag was imported, which is a functional tag that in most cases does not carry any meaning. Another thing we can see from Table 7.11a is, that with increasing window size, we can match more evaluation pairs, but only at the cost of evaluation performance. Applying binarization improves the fit to human intuition only marginally. We cannot achieve any scores far above a correlation value of 0.2, indicating that there is little to no correlation to human intuition.

Set of n Top Tags. We will now use a greater portion of the tag cloud as a description of the represented concept for each page, because all retrieval pages describe either a user, a tag or a document, which in turn are always describable by tags. This enables a more verbose description of the page’s content. Furthermore, since we now not only consider the co-occurrence of a single tags with another one, but co-occurrences between two tag sets, we will from now on only use the content-based co-occurrence counting approach as defined in Section 7.3.1.1.

We assume that a user has mainly a single interest, a tag mostly has a single meaning and a document describes only a single topic. Although we could naturally exploit the whole set of tags on a page, we chose to assign a limited number of only the most relevant tags. We varied $n := \{1, 10, 20, 50\}$ as possible maximum tag cloud sizes. Subsequently, each page was characterized by the most frequent n tags of its tag cloud. If there were less than n tags assigned to a webpage, we used all of the assigned tags.

Figure 7.13a and Figure 7.13b show the results for the varying tag cloud sizes with window size w varied between 2 and 5 on the WS-353 and Bib100 datasets. First of all, we can see that limiting the size of the tag cloud to the top 10 most frequently occurring

tags positively impacts the binarization evaluation scores on WS-353, however less so on Bib100. While different window sizes only show any impact with a very limited tag cloud of size 1, they do not play any major role later on. Also, tag co-occurrence counting as described in Section 7.3.1.1 yields inferior results to binarization, both on WS-353 and Bib100.

Frequent Tag Weighting Until this point, we used a simple counting approach to context vector construction (see Section 7.3.1.1). This way, we made no difference if we combined two popular tags or two rarely used tags in the symmetric cartesian product. But intuitively, often used tags on a page should receive a higher weight than rarely used ones to underline their importance. Because of this, we did not just *count* a cooccurrence, instead we multiplied both occurrences. Since the TF-IDF measure (cf. Section 3.2.1.1) is a widespread alternative to the first order cooccurrence and also assigns a weight to the use of tags, we also calculated the TF-IDF values for all tags. Here, we interpreted each page as a document.

The results can be seen in Figure 7.13. Judging from Figure 7.13c and Figure 7.13d, it seems that including the frequency of two tags in their containing pages in our semantic information model exhibits barely any impact. In contrast to that, the TF-IDF

7.3.2.3 Removing Potentially Noisy Transitions

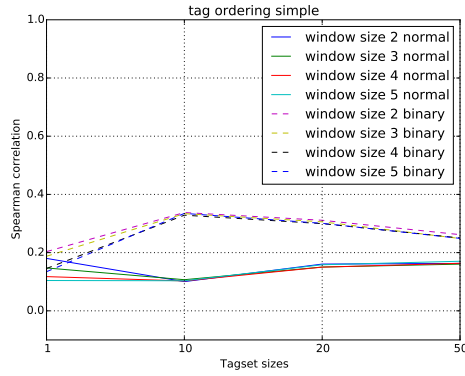
We have seen above that it is difficult to extract semantic information from the whole corpus of BibSonomy navigation data. Potential reasons for this could be that the dataset itself is relatively small, and thus a clear signal for semantics cannot emerge. In the following, we attempt to filter out rarely occurring transitions as well as transitions that introduce a lot of noise, e.g., those that pass /user/USER pages which, by design, contain a lot of general tags.

Core Transitions As the semantics extraction method applied both on WikiGame and ClickStream yielded very good results, as opposed to the WikiClickIU dataset (see Section 7.2.3), we compared these three datasets with respect to size vs usage ratio (see Section 7.2.3). Table 7.12 shows the sizes, usage counts and the size/usage ratio for each dataset, together with the best achievable result of the application of the binarization method described in Section 7.2.1.2 when evaluated on WS-353.

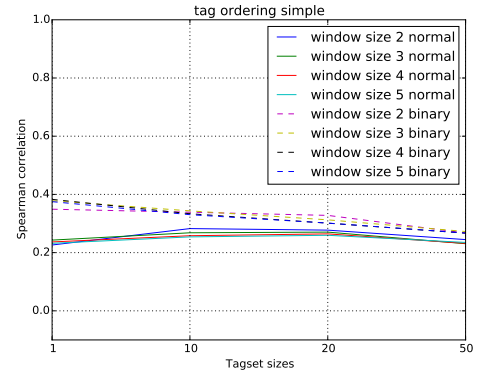
As we can see, the WikiClickIU dataset performs very bad compared to both other datasets. When comparing the features of all three datasets, we can see very quickly that both WikiGame and ClickStream have a very low size/usage ratio, while WikiClickIU yields a very high ratio. Moreover, ClickStream consists of transitions which have been observed at least 10 times. The observation gives rise to the idea that a lower size/usage ratio might also yield more meaningful results (though we didn't test this on the Wikipedia datasets).

Considering this, we limited the BibSonomy request dataset and removed all transitions which occurred less than a predefined threshold to achieve a similar size/usage

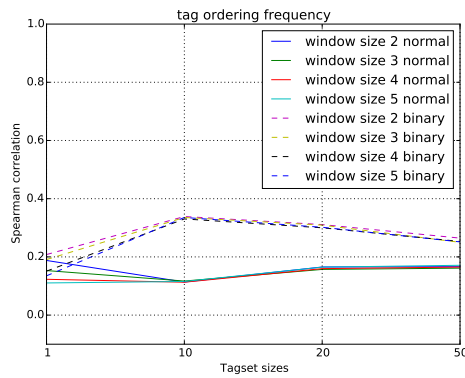
7 Extracting Semantic Relatedness from Social Media Navigation



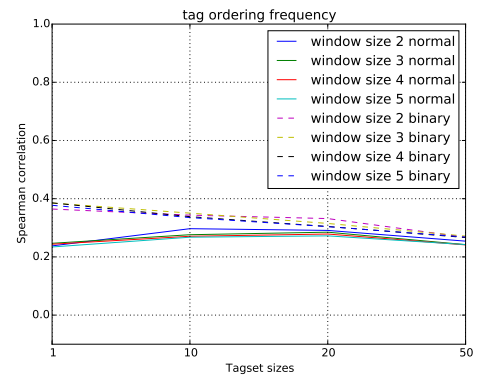
(a) Simple WS-353



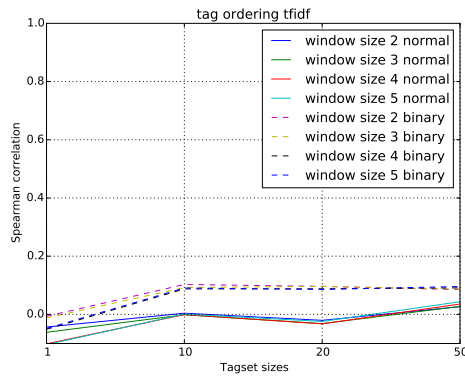
(b) Simple Bib100



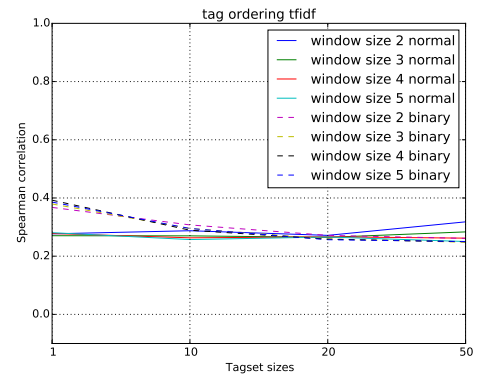
(c) Frequency WS-353



(d) Frequency Bib100



(e) TFIDF WS-353



(f) TFIDF Bib100

Figure 7.13: Results for the tagset weighting experiment. Each picture shows the resulting graphs for varying tagset sizes for the corresponding cooccurrence calculation method, when evaluating the BibSonomy paths on WS-353 and Bib100.

Table 7.12: Comparison of the Wikipedia navigation datasets used in Section 7.2.3, together with their evaluation scores on WS-353. The size denotes the number of unique requests, usage describes the total number of requests made in the dataset. We can see that the lower the size/usage ratio, the higher are the obtained evaluation scores.

	size	usage	size/usage ratio	$\rho_{optimal}$ (pairs)
WikiGame	2.3M	62.5M	0.037	0.728 (236)
ClickStream	14.4M	1 090.2M	0.013	0.709 (288)
WikiClickIU	2.8M	4.0M	0.7	0.458 (120)

ratio or at least a similar effect that fits the idea that low size/usage ratios yield better results. The different dataset sizes can be seen in Figure 7.14a.

From Figure 7.14, we see that judging from the results on WS-353, the restriction to core transitions does not help to improve the extracted semantic information. In contrast, the results on Bib100 suggest that restricting our dataset to transitions with a minimum occurrence of 10 yields the best results throughout all previous experiments. However, Figure 7.14a shows that we can keep only 10 594 unique transitions after filtering, which are however used relatively often, with a relatively low size/usage ratio of 0.054, which is also comparable to that of the WikiGame. Still, the obtained result of a Spearman correlation of 0.408 on Bib100 (cf. Figure 7.14c) is still inferior to that of the folksonomy structure, where we achieved 0.626 on Bib100 (cf. Table 6.7).

Transitions without /user/USER pages Because a /user/USER page is annotated by a big set of tags (see Table 3.1) which are usually spread across several different topics, we tried to exclude those pages to see if the /user/USER pages rather introduce a lot of noise than they are of use. We did so by simply removing all /user/USER pages from the rendered paths, e.g.,

/tag/web \rightarrow /user/hotho \rightarrow /user/hotho/web

became

/tag/web \rightarrow /user/hotho/web.

After removing all /user/USER requests, there are 245 594 paths left with a mean path length of 1.747. We applied our method to these paths with the simple cooccurrence counting method, the tagsize varied between {1, 10, 20, 50} and the window size w varied between 2 and 5.

The results in Figure 7.15 indicate that excluding /user/USER pages indeed improves our fit to human intuition. On the other hand, by removing the user pages from our paths, we lose a lot of transition data, which in turn impacts the validity of our results.

In [48], the authors compared results for calculating semantic similarity based on resource, user and tag level. Since the results for path semantics improved a bit when

7 Extracting Semantic Relatedness from Social Media Navigation

mcnt	size	usage	s/u	paths	avg len
1	181 974	479 471	0.380	263 373	2.821
10	10 594	196 285	0.054	88 675	3.214
15	6 164	166 614	0.037	72 306	3.304
20	4 374	150 600	0.029	63 774	3.361

(a) Data of the core transition experiment. The minimum transition occurrence count *mcnt* is given with the *size/usage ratio (s/u)* as well as the average path length of the resulting paths.

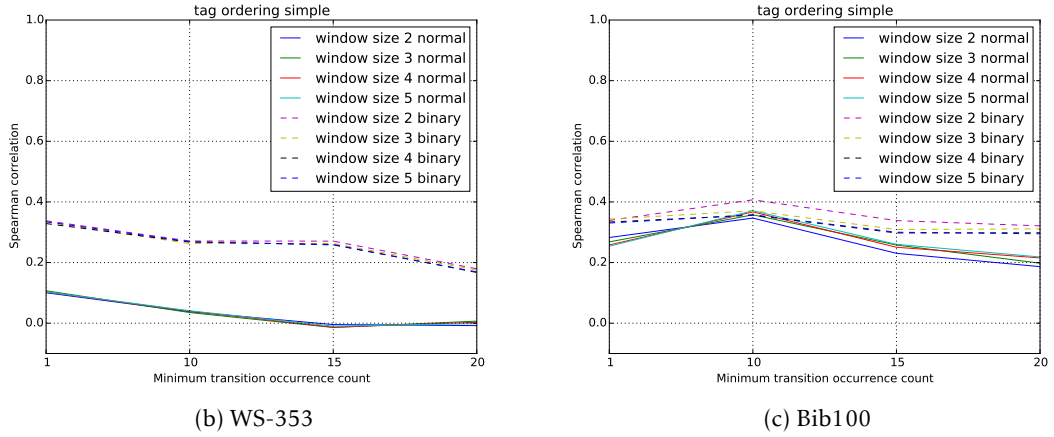


Figure 7.14: Results for the minimum transitions count experiment. The tag cloud size was fixed to 10, since that yielded the best results for the simple cooccurrence counting. Although restricting the transitions to a minimum occurrence count does not improve results on WS-353, we can achieve a rather high evaluation score on Bib100, when we only consider transitions with at least 10 occurrences.

excluding `/user/USER` pages, we could theoretize that the same effect of too many aggregated topics on a user’s personomy⁹ shows an effect here (cf. also Section 6.3.2). Concretely, as the user context tag vectors suffer from a combination of too many semantic topics (after all, the tags were used to describe the resource, not the user), we also assume that this also impacts the quality of the extracted semantics from BibSonomy navigation.

7.3.2.4 Discussion

In this section, we conducted several experiments to extract semantic relatedness information from navigational paths on BibSonomy. Since we achieved remarkable results when extracting such information from navigational paths on Wikipedia, we expected a

⁹All posts from a specific user, see Section 3.1.1.1

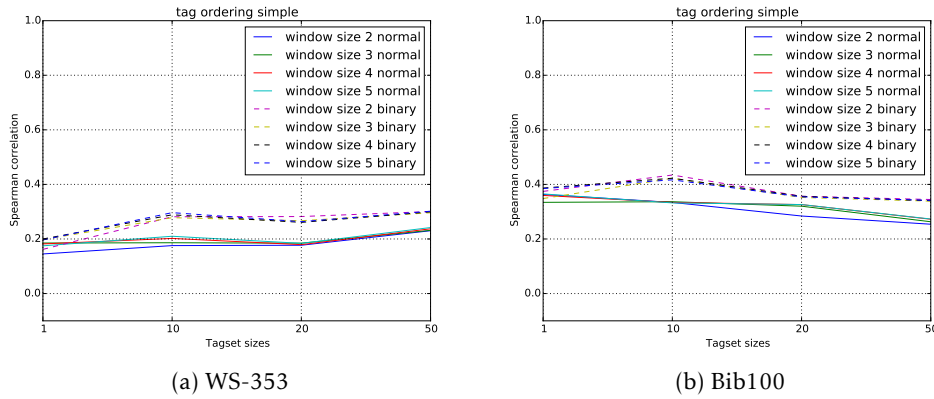


Figure 7.15: Semantic relatedness results from paths when excluding all `/user/USER` pages. Removing the user pages seems to be slightly beneficial for the evaluation scores, but only on Bib100. On WS-353, the scores based on content co-occurrence counting decrease to a correlation score of 0, which indicates no correlation with human judgment at all.

similar outcome on BibSonomy. However, we were unable to achieve a correlation with human judgment on WS-353 as high as those in Section 7.2.

While we can safely assume that this is due to low coverage of word pairs in WS-353, we were surprised to see similarly low scores on the Bib100 dataset, which was specifically constructed for folksonomies (see Section 6.2). A potential explanation would be that the dataset of navigational paths is too small to derive any meaningful results. Another issue is that users navigate mostly on their own pages and then mostly on their `/user/USER` page [64]. As this page contains all tags of a user, it is not very fit to provide a clear description of any potential concept on that page. We found in Section 5.3.2.2 that users in BibSonomy navigate semantically outside their own resources. Because we assume that sparsity is an issue in the analyses of the whole transition dataset in this section, we did not evaluate that subset, since it only contains roughly 42,000 transitions.

In our opinion, the fact that we only have a comparably low number of navigation data on BibSonomy, compared to the potentially huge link graph, strongly impacts the quality and validity of our results. As we could see in Figure 7.14, restricting pages to a minimum occurrence count of 10 already filtered out 170 000 unique transitions.

7.3.2.5 Conclusion

In this section, we extracted semantic information from navigation on BibSonomy. More concretely, we wanted to model the semantics of tags as influenced by navigation instead of characterizing the actual pages, as we did in Wikipedia. After defining the semantic extraction method, we first characterized each page with the most frequent tag on that

page. After this, we increased the amount of tags that we used as a description of the page's content. Additionally, we experimented with different tag selection strategies. After this, we restricted the transitions to those that occur with a minimum count and attempted to remove noise by filtering out `/user/USER` pages. We evaluated all experiments on the WS-353 and Bib100 datasets. Unfortunately, our results were rather mediocre compared to those obtainable on folksonomy data (cf. Section 6.3). In the discussion, we could only assume potential reasons for this. However, we assume that the main reason is the lack of navigation data, both per transitions as well as for the covered pages.

7.3.3 Random Walks on the BibSonomy Graph

Because the results of Section 7.2.4 show that random walks with the right bias can simulate human navigation behavior with regard to the semantic content of navigation, we assume that the same holds for navigation on BibSonomy. Furthermore, we have seen in Section 5.3 that navigation on BibSonomy, similar to navigation on Wikipedia, is influenced by a semantic component. We thus assume that random walks with a certain bias can at least approximate the semantic content of human navigation.

In this section, we will now conduct several random walk experiments on the BibSonomy link graph restricted to those pages which have been visited in the request data from the previous section. For this, we leverage the hypotheses presented in Section 5.3 about user behavior in BibSonomy. We first describe our experimental setup, including the priors that we chose to generate pseudo-human navigational paths. After this, we present and discuss our results.

7.3.3.1 Experimental Setup

Random Navigation Priors In order to generate specifically biased random walks, we make use of several navigation priors, which are partially based on the navigation hypotheses presented in Section 5.3.1. They will all be presented in the following.

We first define simple priors, i.e., priors defined by a single idea. The first two priors (random and folk) can be seen as baselines, while the others are inspired by navigational hypotheses used to analyze navigation on BibSonomy.

- **uniform:** The walker navigates completely random. Teleportation is allowed i.e., it can also reach pages that are not connected via links.
- **folk:** This walker is restricted to the raw folksonomy structure, i.e., the walker cannot follow extra links provided by the BibSonomy webpage. This hypothesis can be seen as a structural baseline.
- **cat:** This walker is based on the category-consistent hypothesis. Navigation is only available to pages in the same category, i.e., from a user page, the walker can only reach other user pages, but no tag pages.

- **own:** Similar to **cat**, this walker can only navigate on pages of the same user, i.e., it cannot cross over to another user or userless pages like `/tag/TAG` pages
- **tag_cos/_tfidf:** Finally, this walker navigates by semantic incentives. This means that pages whose co-occurrence count vectors or TF-IDF representations (for a description of the vector construction, see Section 5.3.1) are highly similar to the current page representation will be reached with a higher probability than less semantically correlated pages.

We did not take the original navigation data as a potential parameterization into account, since the transitions are rather evenly distributed, without clear signs of navigation preference. This can be especially be seen in Figure 7.14a, as only 5% of all transitions occur more than 10 times. The actual navigation would then rather look like the uniform random walker.

Similar to Section 5.3.1, we also use combined priors to model more complex navigation behavior.

- **folk_own:** Here we assume that users will follow the folksonomy structure, but only on their own resources.
- **folk_cos_tfidf:** This walker follows the folksonomy structure with a semantic bias, e.g., given a certain tag page, it will now rather visit a user page of a user more involved in the respectively described topic and which is subsequently semantically higher related than that of another user who rarely uses that tag.
- **own_tag_cos_tfidf:** Finally, we restrict the random walker to the resources owned by the same user, but now bias navigation there in a semantic way.

Random Walk Generation Parameters For the following experiments, we fixed the parameters of our method to extract semantic relatedness from navigational paths on BibSonomy as follows: The maximum tag cloud size for each page is set to 10, i.e., we only look at the 10 most popular tags for each page. The window size varies between 2 and 5. We take the whole navigation dataset into account, i.e., all `/user/USER` pages and every page that occurred at least once. For each page, we generated 10 paths of length 20, so that every page occurred at least 10 times.

Generated Walks For each hypothesis as well as random navigation and true navigation, we created 1 820 080 paths. Depending on the hypothesis, the path length distributions differ, since it is easy to end up in pages with no outlinks (since e.g., they are rendered but do not provide any links to other pages). Table 7.13 gives some statistics about the different paths.

We can see that sometimes, only a fraction of all possible pages is visited, especially on the folksonomy hypothesis. This can mostly be attributed to the fact that a large portion of pages has no outlinks in the folksonomy and thus no paths can be constructed for that particular page. In the case of dead ends in semantic navigation, we argue that

Table 7.13: Statistics about the random paths on BibSonomy generated from the navigation hypotheses presented in Section 5.3.

Dataset	avg. path length	Visited pages	Paths > 1
uniform	20	182 008	1 820 080
cat	19,192	174 270	1 742 700
folk	11,232	99 424	989 060
folk_cos_tfidf	8,314	73 277	720 980
folk_own	11,134	97 079	970 790
own	19,192	174 270	1 742 700
own_tag_cos_tfidf	19,192	174 270	1 742 700
tag_cos	19,192	174 270	1 742 700
tag_cos_tfidf	19,192	174 270	1 742 700
true paths	2,821	181 974	263 373

this is because we end up on pages with either no tags or with links to only uncorrelated tags.

7.3.3.2 Results and Discussion

Now, we will evaluate each path dataset using the multi-tag co-occurrence already used in the previous section from Section 7.3.1.1. We compare the generated scores to the WS-353 and Bib100 datasets.

The general picture that we can see in Table 7.14 is that on Bib100, a) human navigation yields higher semantic scores than `uniform` navigation, indicating a deterministic component in human navigation, and b) the semantic prior generates navigational paths with the highest semantic content (`tag_cos` and `tag_cos_tfidf`). On WS-353, human navigation Interestingly enough, the random walks produced by the pure and folksonomy-based semantic hypotheses perform the semantically most valuable walks with a correlation scores of 0.450 on Bib100 and 0.256 on WS-353, respectively (`tag_cos_tfidf`, `folk_cos_tfidf`). As opposed to the results in Section 5.3, we find that the semantic hypothesis on user-consistent navigation (`own_tag_cos_tfidf`) yields rather bad results, barely above the structural/random navigation hypothesis (`random`). We assume that this is because, although navigation by semantic incentives prefers pages with a high semantic similarity, a user is mostly only interested in a few, limited topics, while navigation on the whole BibSonomy graph can also lead to different, related topics and thus introduce more diverse semantic information. Even more important, by using more complex navigational hypotheses, such as semantic or folksonomic navigation, we can outperform human navigation significantly. On the other hand, restricted navigation such as category-consistent and user-consistent navigation yields worse results than human navigation.

However, comparing all obtained results with those received from taking onl structural data from BibSonomy into account (cf. Table 6.7), we see that the best results we

achieved in this section, a correlation of 0.450 on Bib100 using semantic navigation, is still largely outperformed by all types of context-based vector representations. This indicates that we still need either a better method to extract the semantic information in BibSonomy navigation or more structural information that we can include into our measure.

7.3.3.3 Conclusion

In this section, we generated random walks on the BibSonomy graph influenced by different navigation strategies. The results from Section 7.3.2 for extracting semantic information from human navigation on BibSonomy were less than encouraging due to the small size of the human navigation dataset. We had thus hoped that by artificially generating pseudo-human navigational paths, we can at least find a signal that navigation on BibSonomy is also suitable to extract semantic information. Although the resulting semantic evaluation scores are not very high and leave room for improvement, we are encouraged by the tendency of the results.

We found that random walks biased by a semantic incentive, that is, navigating towards semantically related pages, provide a good base to capture navigational semantic information to a certain extent. Consequently, we assume that since semantic navigation can also partially explain human navigation on BibSonomy, we only need a sufficient amount of navigation information on BibSonomy to capture semantics of a decent quality.

7.4 Summary

In this chapter, we investigated both how to extract semantic information from navigational paths and the quality of that information, i.e., how well it fits to what humans would perceive.

We first proposed a set of co-occurrence counting based methods in order to construct high-dimensional vector representations of words or rather pages in navigational paths. We applied co-occurrence counting and binarization on navigation from Wikipedia with highly promising results, both on game and unconstrained human navigation as well as on random navigation on a set of differently biased link networks. More concretely, we obtained a relatively high correlation with human intuition of 0.709 on WS-353 with the WikiGame data, as well as with unconstrained navigation from ClickStream data on MEN, where we achieved a correlation of 0.64. The results from the random walk experiments showed that we can at least approximate human-like navigation in terms of semantic information, yet we are not able to find paths that provide better results. Methods to determine such paths is left for future work.

Furthermore, we experimented with multi-tag occurrence and its binarization variant on both unconstrained human navigation and biased random navigation on BibSonomy. While the extracted semantics from unconstrained BibSonomy navigation did not produce any good results on WS-353, we could at least find signals for better semantics when evaluating on the Bib100 dataset from Section 6.2. Overall we assume that, if we

had more navigation data on BibSonomy, we could also extract semantics of a higher quality because we then would achieve a more precise representation of human navigation in social tagging systems. We partially tackled this by generating biased random navigation on the BibSonomy subgraph that has also been used in the unconstrained data. Here we could see that semantic navigation, i.e., navigating to semantically more related pages across the whole folksonomy, yielded more precise semantics than random and human navigation. This also strengthens our assumption that when analyzing more navigation data, we could significantly increase the quality of the extracted semantics.

In general, we could show that while navigation on social web systems contains a lot of latent semantic information (see Chapter 5, in this chapter, we made that information visible by constructing different vector space models that capture the contextual semantics produced by human navigation. We could also show that specifically biased random walks can generate paths with considerable amounts of semantic information, superior to completely random navigation and sometimes even human navigation and thus provides a viable alternative to relying on the cumbersome collection user data from web logs.

Table 7.14: Evaluation of all random walks for different window sizes. The scores were evaluated on WS-353 and Bib100. Best results are marked fat.

Hypothesis	ρ				ρ^{bin}			
	2	3	4	5	2	3	4	5
cat	0,238	0,232	0,237	0,234	0,226	0,226	0,222	0,222
folk	0,342	0,305	0,313	0,281	0,255	0,243	0,251	0,254
folk_cos_tfidf	0,348	0,334	0,350	0,338	0,439	0,282	0,280	0,273
folk_own	0,385	0,327	0,333	0,299	0,286	0,253	0,243	0,245
own	0,254	0,256	0,259	0,257	0,217	0,210	0,208	0,206
own_tag_cos_tfidf	0,259	0,246	0,242	0,236	0,209	0,209	0,203	0,205
uniform	0,231	0,224	0,228	0,235	0,235	0,227	0,223	0,222
tag_cos	0,448	0,450	0,452	0,452	0,240	0,228	0,224	0,218
tag_cos_tfidf	0,430	0,433	0,433	0,433	0,229	0,216	0,212	0,210
human navigation	0,309	0,302	0,299	0,297	0,422	0,426	0,416	0,405

(a) Bib100 (all pairs)

Hypothesis	ρ				ρ^{bin}			
	2	3	4	5	2	3	4	5
cat	0,118	0,122	0,123	0,129	0,141	0,131	0,126	0,125
folk	0,140	0,191	0,179	0,181	0,184	0,208	0,206	0,194
folk_cos_tfidf	0,165	0,256	0,262	0,249	0,246	0,226	0,219	0,190
folk_own	0,133	0,220	0,190	0,205	0,133	0,172	0,166	0,163
own	0,050	0,039	0,037	0,038	0,149	0,149	0,141	0,137
own_tag_cos_tfidf	0,120	0,130	0,131	0,134	0,126	0,116	0,110	0,110
uniform	0,118	0,124	0,119	0,120	0,148	0,137	0,131	0,127
tag_cos	0,245	0,235	0,236	0,235	0,185	0,166	0,161	0,155
tag_cos_tfidf	0,200	0,201	0,200	0,198	0,160	0,157	0,150	0,147
human navigation	0,078	0,080	0,080	0,071	0,227	0,227	0,227	0,226

(b) WS-353 (266 pairs)

Chapter 8

Relative Relatedness Learning: Learning Semantic Relatedness from Human Feedback

8.1 Introduction

In the previous chapters, several models have been proposed to extract semantic relatedness information from social media data. These and many other approaches (e.g., [48, 156, 184]) encode semantic information about the context of words in word vectors (cf. Section 3.2.1.1). While many of those models are able to represent a good portion of human intuition about semantic relatedness, they are only able to encode information contained in the underlying corpus. Thus they do not explicitly represent the actual notion of semantic relatedness as expected and employed by humans.

One line of research focussed on exploiting semantic lexicons to post-process word embeddings [69, 163, 250]. A semantic lexicon contains type-level information about words, such as synonymical, hierarchical and paraphrase relations to other words. Examples of semantic lexicon are WordNet [72] and the Paraphrase Database PPDB [181]. However, by capturing synonym relations, works relying on those lexicons cannot capture weaker degrees of semantic relatedness. Information about the *strength* of the degree of semantic relatedness is for example contained in semantic relatedness datasets such as MEN or WS-353, which we introduced in Section 4.4.

As such semantic relatedness datasets provide a very close representation of human intuition (see Section 6.2), a logical approach to make word vectors approximate this human intuition is to incorporate the information contained in these datasets into the word vectors. Often, word embedding datasets are evaluated on such semantic relatedness datasets using the Spearman rank correlation coefficient, which compares the *relative ordering* of word pairs according to their degree of semantic relatedness (see Section 3.2.4.1). That means that it is more important to qualitatively compare the strengths of two relations instead of determining an absolute relatedness score.

Consequently, the central topic of this chapter is a learning approach based on the *relative comparison of semantic relatedness scores* to improve the quality of word vector

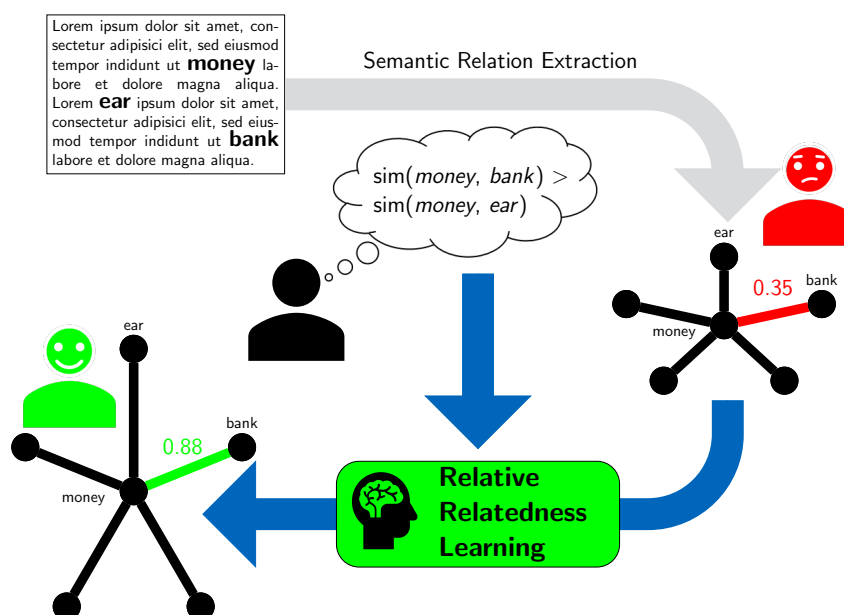


Figure 8.1: A rough sketch of how RRL works. Using human background knowledge about the relative ordering of semantic relations, we retrain a given set of word embeddings to better reflect this background knowledge.

representations. The relatedness scores used for training are obtained from *semantic relatedness datasets with human intuition* such as MEN or WS-353, but can easily be derived from any set of word pairs together with semantic relatedness scores. This approach can be applied on any pretrained word embedding dataset, thus making it universally applicable. We show the effectiveness of our approach in several tasks, such as measuring semantic relatedness between words, sentiment analysis and question answering.

The contents of this chapter are an extended version of the work presented in [169] and [170]. Concretely, we modified the optimization step of the algorithm to apply mini-batch stochastic gradient descent, added comparison with another learning approach from literature and an additional setting in which we evaluate RRL.

8.2 The Relative Relatedness Learning (RRL) Algorithm

In this section, we introduce the Relative Relatedness Learning (RRL) algorithm. First, we motivate RRL by describing how the semantic relatedness of word representations is usually estimated and how such scores are evaluated. We then gradually develop the RRL algorithm that we use to improve the semantics in word embeddings. Figure 8.1 shows a rough sketch of how RRL works.

Table 8.1: Examples of word pairs with human assigned relatedness scores on a scale ranging from 0 to 1. A higher score indicates stronger semantic relatedness.

Word w_1	Word w_2	Semantic Relatedness Score $rel(w_1, w_2)$
Money	Bank	0.85
Alcohol	Chemistry	0.554
Drink	Ear	0.131

8.2.1 Motivation

The degree of semantic relatedness between two words is usually estimated as an arbitrary number on an arbitrary scale. The higher that number for a given pair of words, the more closely related are these words. For example, in Table 8.1, we can see that *Money* and *Bank* are more closely related to each other than *Alcohol* and *Chemistry*, since the former word pair has a higher relatedness score than the latter pair. Such scores about semantic relatedness can be gathered directly from humans using crowdsourcing and thus provide a close, if not the closest reflection of human intuition of semantic relatedness [43, 75] (see also Section 6.2).

Consequently, these scores are also widely used as an evaluation basis for word embeddings [18, 81, 184]: The quality of semantic relatedness information in word representations is usually evaluated on these scores. Using the Spearman rank correlation coefficient (cf. Section 3.2.4.1), it is possible to measure the *monotonic correlation* between the human intuition scores and the semantic relatedness scores produced by word embeddings. This means that it is less important to match the actual *semantic relatedness scores* of word pairs, but their *relative ordering* induced by those scores.

A possible way to improve the relatedness score rankings produced by pretrained word embeddings is to manipulate the corresponding vector space V with regard to a set of *relative constraints*

$$\mathcal{C} := \left\{ \left((w_i, w'_i), (w_j, w'_j) \right) \mid rel(w_i, w'_i) > rel(w_j, w'_j) \right\}. \quad (8.1)$$

For example, this can be done with a linear projection, i.e., a quadratic matrix

$$L \in \mathbb{R}^{n \times n} : V \rightarrow V_L \quad (8.2)$$

between the original vector space V and the manipulated vector space $V_L := \{Lv \mid v \in V\}$, where all vectors $v \in V$ are multiplied with the matrix L . In the following, we propose an algorithm that aims to find such a projection matrix L satisfies the constraints in \mathcal{C} in V_L . Explicitly this means that for two word pairs (w_i, w'_i) and (w_j, w'_j) where $rel(w_i, w'_i) > rel(w_j, w'_j)$, the inequality $\cos(Lv_i, Lv'_i) > \cos(Lv_j, Lv'_j)$ holds. Using a metric learning approach, we learn L implicitly by learning a generalized dot product $\langle x, y \rangle_M := x^T M y$ parameterized by a positive definite and symmetric matrix $M = L^T L$. This way, The

generalized cosine measure is thus defined as

$$\cos_M(v_i, v'_i) := \frac{\langle v_i, v'_i \rangle_M}{\sqrt{\langle v_i, v_i \rangle_M} \sqrt{\langle v'_i, v'_i \rangle_M}} \quad (8.3)$$

To ultimately obtain the desired vector space transformation matrix L , we apply a Cholesky decomposition to the learned matrix M , which can be used to project the original embedding vector space V to the manipulated vector space V_L .

We again emphasize that our model is independent of the algorithm that the pre-trained vectors were created with, as it only relies on a set of vectors and a set of semantic relatedness scores between word pairs.

8.2.2 Optimization Objective

We will now define the optimization objective of the RRL algorithm. It is composed of two parts: The *constraint violation loss*, which penalizes violations of relatedness constraints, and the *regularization loss* to ensure that the transformation matrix M does not degenerate.

Constraint Violation Loss. If a constraint $((w_i, w'_i), (w_j, w'_j)) \in \mathcal{C}$ is satisfied, that is if $\cos_M(v_i, v'_i) > \cos_M(v_j, v'_j)$, it intuitively should have no impact on the optimization objective. However if the constraint is violated, i.e., $\cos_M(v_i, v'_i) < \cos_M(v_j, v'_j)$, we want to penalize this violation and thus increase the loss function. The loss term for a single constraint $(w_i, w'_i), (w_j, w'_j)$ with $rel(w_i, w'_i) > rel(w_j, w'_j)$ can thus be formulated as follows:

$$closs_{(w_i, w'_i, w_j, w'_j)}(M) = \left(\max\{0, \cos_M(v_j, v'_j) - \cos_M(v_i, v'_i)\} \right) \quad (8.4)$$

By summing up the loss terms for all constraints \mathcal{C} , we obtain the following loss function for our constraints:

$$closs_{\mathcal{C}}(M) = \sum closs_{(w_i, w'_i, w_j, w'_j)}(M) \quad (8.5)$$

Regularization Loss. We want to keep M as close as possible to the identity matrix I_n to prevent overfitting to the constraints and keep most of the original embedding structure intact. To additionally ensure that we learn a *bijective* projection, the full rank of M needs to be preserved. For this, we introduce an additional regularization loss term $rloss(M)$ based on the *Bregman divergence* [41]. Bregman divergences are a generalization of for example the Euclidean distance and possess a lot of interesting properties. Concretely, due to their formulation, they guarantee strict convexity. Thus they always approximate a globally optimal point when optimized. The Bregman divergence $D_\phi : \Omega \times \Omega \rightarrow \mathbb{R}$, parameterized with a strictly convex and differentiable function $\phi : \Omega \rightarrow \mathbb{R}$, is defined as the difference between $\phi(x)$ and the first Taylor expansion of ϕ evaluated at y :

$$D_\phi(x, y) := \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle. \quad (8.6)$$

Here, Ω is a closed convex set, such as the convex cone of positive definite matrices in $\mathbb{R}^{n \times n}$. As mentioned above, it is closely related with the Euclidean distance: when choosing $\phi(x) = \|x\|^2$, we obtain the squared Euclidean distance $D_\phi(x, y) = \|x - y\|^2$.

For our algorithm, we follow [126] and choose $\phi(M) := -\log \det(M)$ with positive definite M as parameterization to ensure that M stays positive definite and does not degenerate. M must be positive definite, because otherwise $\langle x, y \rangle_M$ would not be a valid scalar product anymore, which is however crucial for our purposes, as it could not provide a mathematically valid metric anymore. $\log \det M$ denotes the logarithm of M 's determinant and is always defined, since M is positive definite and thus has full rank, i.e., $\det M > 0$. We can now formulate the regularization loss term which minimizes the Bregman matrix divergence $D_\phi(M, I_n)$ between M and I_n :

$$rloss(M) := D_\phi(M, I_n) = \text{tr}(M) - \log \det M - n \quad (8.7)$$

Here, $\text{tr}(M) = \sum_i m_{ii}$ is the *trace* of M , i.e., the sum of all its diagonal elements, while n is the rank of M .

Optimization Objective. Thus, our optimization objective is to minimize the sum of Equation (8.5) and Equation (8.7).

$$loss(M) = closs_{\mathcal{C}}(M) + rloss(M) \quad (8.8)$$

The resulting optimal matrix M then satisfies as many constraints as possible and at the same time stays close to the identity matrix I_n , i.e., attempts to keep most of the word embedding structure intact.

8.2.3 Optimization

To minimize Equation (8.8), we use a projected mini-batch gradient descent approach. This means that we calculate the gradient of Equation (8.8) for a sample of randomly selected relative relatedness constraints, perform the gradient descent step and then project the matrix $M' := M - \nabla loss(M)$ to the cone of positive definite matrices. This last step is important, since it ensures that M still defines a valid scalar product and that M can be decomposed into $M = LL^T$ in a Cholesky decomposition. A matrix M' can easily be projected onto the PSD cone by obtaining its eigendecomposition $M' = U^T \Sigma U$ and setting all negative eigenvalues on the diagonal of Σ to a small $\varepsilon > 0$.

8.3 Experiments and Results

To demonstrate the performance of RRL, we evaluate its impact on measuring semantic relatedness and on a question answering task.

8.3.1 Preliminaries

Before we describe the experiments to evaluate RRL and their corresponding results, we list the used resources and the general evaluation procedure.

Resources. We use different pretrained word embedding datasets, namely 1. Delicious GloVe embeddings from Section 6.5. 2. WikiNav navigational embeddings, like those from Section 7.2.3. 3. WikiGloVe and ConceptNet embeddings (see Section 4.3). The latter three datasets were introduced in Section 4.3.

To train RRL in the word relatedness task, we use the WS-353, MEN, Bib100 and SimLex-999 datasets, which contain human-assigned semantic relatedness scores of several word-word relations (cf. Section 4.4). It is important to note that while WS-353, MEN and Bib100 mainly contain semantic *relatedness* relations, SimLex-999 was specifically designed to contain information about the semantic *similarity* of words. This means that in this case we expect to improve the ability of measuring semantic similarity with the improved word embeddings.

For the question answering task, we train RRL only on the WikiGloVe and ConceptNet embeddings, using MEN.

General Evaluation Procedure. For the Word Relatedness tasks, we compare a set of vectors to human intuition using the Spearman correlation. The exact procedure is described in Section 3.2.4. Additionally, we show that some evaluation score improvements are indeed statistically significant at $p < 0.05$, i.e., the improvements are not caused by artifacts in the training process. To calculate p values, we use Fisher’s z -test. Equation (3.29) describes how to compute these p values, depending on two correlation scores ρ_1 and ρ_2 and the number of samples n , on which those scores were obtained.

8.3.2 Word Relatedness

We evaluate RRL in four settings of a word relatedness task.

First, we explore the *general performance of RRL*. We do so by training RRL on one portion of a semantic relatedness dataset and then compare the performance of the trained vectors and their original versions on the remaining part of the relatedness dataset. Additionally, we investigate the amount of necessary background information to achieve a significant increase in evaluation performance. Our goal in this setting is to show that we can learn semantic relatedness using RRL.

Second, *we compare our approach with a strong baseline* from literature, namely Retrofitting (RF) [69]. As already explained in Section 3.2.5.1, Retrofitting uses taxonomic information to adjust pretrained word embeddings to improve semantic similarity. In this work, we use information from the Paraphrase Database PPDB [181] to train Retrofitting. Note that Retrofitting uses a different source of semantic knowledge to improve for training, so while it is interesting to see how this impacts results, we also expect improvement by *combining* RRL with Retrofitting by training RRL on the retrofitted embeddings. Our goal is on the one hand to show that RRL is able to produce competitive results, and on the other hand to look for a potential synergy effect between both methods.

Third, we want to *investigate the robustness of word embeddings* using RRL. We artificially generate fake semantic relatedness scores by randomly generating new scores for a given set of word pairs. To ensure that these scores do not accidentally correlate with the original human intuition scores, we enforce a very low correlation of $p < 0.0005$ between

the random and the original scores. Here we want to show that i) wrong ratings do not completely collapse the semantics of the trained vectors, which ultimately makes our approach robust for different users with different intuitions of relatedness, and ii) that the promising results of the previous experiments are indeed caused by the successful injection of human intuition into the training process.

Finally, we want to see if the knowledge that we encoded by training RRL on one semantic relatedness dataset is general enough to positively influence evaluation scores on other datasets of semantic relatedness. In the first three experiments, we trained RRL only on a portion of a semantic relatedness dataset, in order to evaluate the trained model on the remaining part. Now, we train RRL on one complete semantic relatedness dataset and evaluate the resulting model on a different dataset. For example, we train on all WS-353 relations, but evaluate the model on the MEN dataset. By this we evaluate if the learned knowledge generalizes from one notion of semantic relatedness (represented by a specific dataset) to another.

8.3.2.1 Integrating Different Levels of Background Knowledge

In this section we investigate how the amount of semantic background information from human intuition datasets used for training influences the quality of the refined word embeddings. To this end, we evaluate various training set sizes extracted from the semantic relatedness datasets introduced above on the different word embedding datasets.

Experimental Setup. For each semantic relatedness dataset, we first randomly sample 20% of all matchable pairs as test set and use the remaining 80% as training data. From the training data, we sample training sets of different sizes (10% - 100%). On each of these subsets of the training data, we train RRL. The resulting trained vectors are evaluated on the previously sampled test data. We repeat this procedure of sampling training and test sets and subsequent training of RRL for 25 times. Finally, we report the mean and the standard deviation of evaluation scores over all 25 repetitions for each training subset. As a baseline, we also report the Spearman correlation score of the untrained vectors on the test sets, i.e., when setting M as the identity matrix I_n in Equation (8.3).

Results. As can be seen from Figure 8.2 and Figure 8.3, we are generally able to *learn semantic relatedness from human intuition using RRL*. However, we need a certain minimum amount of training data to produce significantly improved results.

On the MEN data set (Figure 8.2), we can always outperform the raw baseline by a significant margin ($p < 0.05$) with a minimum amount of 40% sampled training data. On the ConceptNet embeddings, we can even increase the correlation with MEN to 0.88, while the ConceptNet embeddings only reach a correlation score 0.856 on the test data. It is also an interesting observation that our result of 0.88 correlation is also very close to the interannotator agreement of 0.84 reported in [43], which means that (using a

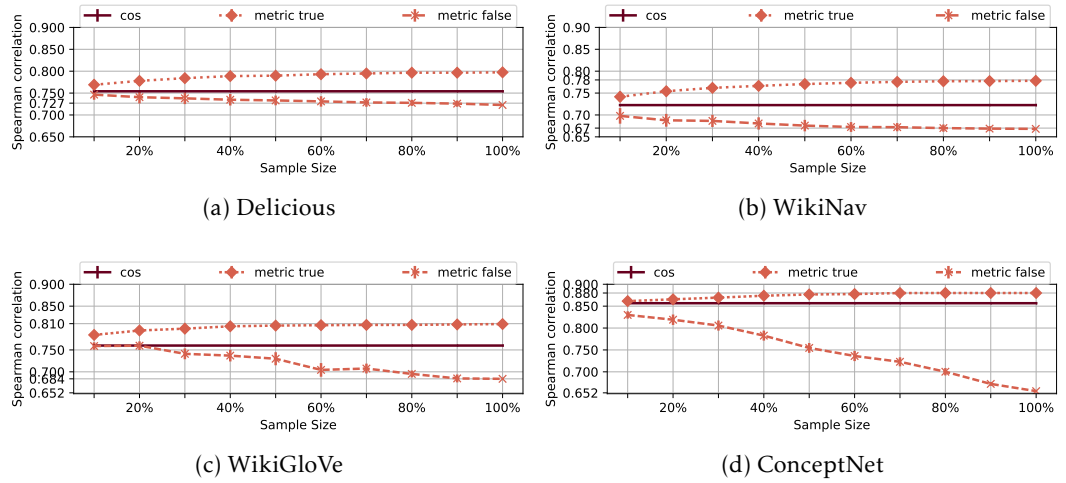


Figure 8.2: Results on different levels of amounts of MEN training data. The dotted lines show the mean Spearman correlations on the test portion of MEN, using the retrained word embeddings, together with the standard deviation of the results (vertical lines at each datapoint). The continuous lines depicts the Spearman score of the untrained word embeddings. Dashed lines denote the scores of the retrained word embeddings, this time using randomly generated relatedness scores.

suitable set of word embeddings) we are able to achieve human-like performance in this setting.

When training RRL using the WS-353 data set (Figure 8.3), evaluation results are unfortunately not as clear-cut as when using MEN. However, it can still be seen that by using RRL, an improvement in correlation score can be achieved on all datasets except Delicious, where we even observe a slight, but insignificant decrease of our results. Still, this improvement is very marginal and probably due to the small size of WS-353 and thus the very small number of items in the test set¹.

8.3.2.2 Comparison and Combination with Retrofitting

To show the usefulness of our approach, we compare ourselves to the Retrofitting algorithm as a baseline, which has already been presented in Section 3.2.5.1. Retrofitting takes a set of word embeddings and a set of similarity constraints into account to fine-tune the embeddings and improve their depiction of semantic *similarity*, as Retrofitting replaces a vector by the mean of its synonyms given by the word lexicon. [69]

Experimental Setup. In this experiment, we train RRL as well as Retrofitting on all four embedding datasets. For RRL, we use WS-353, MEN, Bib100 and SimLex-999

¹ 20% of 353 word pairs are only ≈ 71 pairs

8.3 Experiments and Results

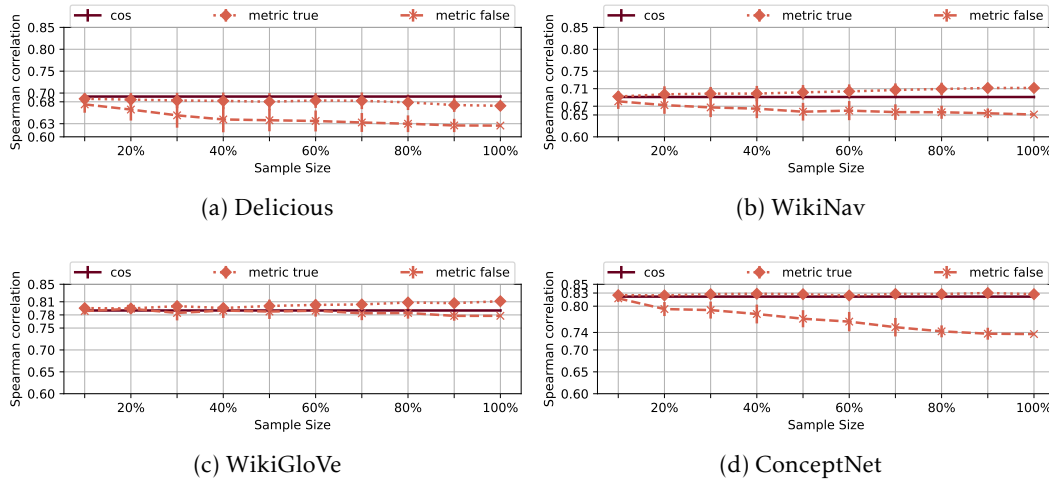


Figure 8.3: Results on different amounts of WS-353 training data using the pretrained embeddings. Injecting true user feedback now only leads to a marginal increase in correlation on the test data for WikiNav, WikiGloVe and ConceptNet, while on Delicious, the correlation even decreases (diamond, dotted). Injecting false user feedback to RRL lets performance decrease significantly and dramatically everywhere except on WikiGloVe (star, dashed). The continuous line serves as baseline, i.e., the standard cosine score on the test data.

for training, concretely we use the full amount of the best set of the sampled training data from the first experiment. To train Retrofitting, we use paraphrase data from the paraphrase database PPDB. Additionally, we train RRL on top of the retrofitted vectors. We evaluate each set of vectors on the corresponding test data from the first experiment.

Results. From Table 8.2, we can see several intriguing results. First of all, both Retrofitting and RRL can increase the fit to human intuition of semantic relatedness on almost all word vector datasets. While in some settings Retrofitting outperforms RRL, there are also cases when training vectors with RRL results in higher correlation with human intuition than when using Retrofitting. The most notable result however is that *combining Retrofitting and RRL almost always yields the highest correlation with human intuition*, but always outperforms the baseline. Only on test sets with very few pairs (e.g., Bib100 or WS-353 on Delicious and WikiNav), this is not the case and Retrofitting shows a clear edge. Another interesting result is that on Delicious, RRL outperforms Retrofitting when training with Bib100 data. We attribute this to the nature of Delicious’ and Bib100’s topic overlap (see also Section 6.3).

Table 8.2: Baselines for word embedding fine tuning experiments. All reported values have been evaluated on the corresponding evaluation dataset splits for each vector dataset. For Retrofitting (RF), we used the PPDB-XL paraphrase lexicon as training data.

		WS-353	MEN	Bib100	SimLex-999
ConceptNet	pairs	67	600	20	200
	raw	0,858	0,871	0,727	0,603
	RF	0,862	0,874	0,748	0,660
	RRL	0,864	0,904	0,734	0,669
	RF+RRL	0,874	0,905	0,738	0,724
WikiGloVe	pairs	67	600	20	200
	raw	0,598	0,754	0,801	0,335
	RF	0,626	0,779	0,881	0,483
	RRL	0,636	0,838	0,799	0,428
	RF+RRL	0,676	0,842	0,831	0,553
WikiNav	pairs	33	246	10	76
	raw	0,708	0,725	0,733	0,406
	RF	0,743	0,751	0,883	0,523
	RRL	0,616	0,762	0,717	0,455
	RF+RRL	0,694	0,781	0,817	0,525
Delicious	pairs	39	275	20	72
	raw	0,741	0,731	0,772	0,255
	RF	0,844	0,784	0,788	0,399
	RRL	0,747	0,750	0,801	0,389
	RF+RRL	0,846	0,817	0,849	0,500

8.3.2.3 Robustness of Pretrained Embeddings

Now we inject false user feedback into RRL to see if we can influence the score in not only a positive, but also a negative direction. We tried to revert the positive effect of injecting human feedback into a semantic relatedness measure.

Experimental Setup. Here we inject wrong semantic relatedness information into our learning process. We artificially generate “wrong” semantic relatedness by randomly generating new scores for a given set of word pairs, which have a very low correlation of $p < 0.0005$ to the original scores. The goal is to show that i) wrong ratings do not completely collapse the relatedness measures, which ultimately makes our approach robust for different users with different intuitions of relatedness, and ii) that the promising results of the previous experiments are indeed caused by the successful injection of user

feedback.

Results. Figures 8.2 and 8.3 show that *wrong semantic information exhibits a large negative influence* on the trained vectors (dotted line, star markers). As in the first experiment, the result decrease on MEN is stronger than on WS-353. We again attribute this to the smaller size of WS-353. Despite the (expected) score decrease, we were unable to completely distort the semantic relatedness scores produced from the resulting vectors. This is most evident in Figure 8.2a, where we almost couldn't induce a negative bias in the Delicious embeddings with the MEN dataset. We assume that the score decrease is mitigated by the inherent semantic content of the embeddings. Overall, this shows that although we can improve the fit of word vectors to human intuition, as shown in the first experiment, we need a high semantic quality of both the word vectors and the latent collected relatedness scores through human feedback.

8.3.2.4 Transporting User Intentions

The previous experiments showed that the integration of a semantic relatedness dataset into a relatedness measure results in higher agreement of the measure with human intuition. By this we evaluate if the learned knowledge generalizes from one notion of semantic relatedness (represented by a specific semantic relatedness dataset) to another.

Experimental Setup. Now, in order to transfer different user intentions across different settings, we trained metrics on one complete relatedness dataset and evaluated them on a different semantic relatedness dataset. For example, training was done using all WS-353 relations but the metric was evaluated on the MEN dataset. We repeat this for each combination of word vectors and pairs of word embedding datasets MEN, WS-353, and SimLex-999.

Results. Results are given in Table 8.3. For each line, its header defines the dataset on which the metric was trained, while the column header is the dataset on which the trained metric was then evaluated. In each cell, the first value denotes the Spearman correlation of the cosine measure with the human relatedness scores in the evaluation dataset. The second value is the Spearman correlation of the relatedness scores calculated from the trained metric with the human relatedness scores in the evaluation dataset. Depending on whether the trained metric increased or decreased correlation with human intuition, we depict upwards or downwards arrows.

Unfortunately, we cannot say that we are able to transport the semantic information contained in one semantic relatedness dataset to another. Almost no scores improve upon the untrained vector baseline score. The only exceptions are when we evaluate a pretrained set of embeddings on the training set. Even there, the scores on the SimLex-999 dataset decrease, when we train RRL on WikiNav and WikiGloVe embeddings. While Delicious and ConceptNet vectors trained on WS-353 can slightly outperform the baseline, we do not consider this a relevant result since the difference is insignificant.

Table 8.3: Results for user intention transport experiments. We trained a metric on all word pairs from the dataset given at the start of each line and evaluated them on the dataset given in the column header. The first value is the Spearman correlation for the cosine measure on the evaluation dataset, the second value is the Spearman value for the trained metric. The arrow denotes if we could transfer relevant information from one dataset to another or not.

Evaluated on Trained on	MEN	SimLex-999	WS-353	Evaluated on Trained on	MEN	SimLex-999	WS-353
raw	0.752	0.336	0.690	raw	0.709	0.340	0.729
MEN	0.811	0.298	0.666	MEN	0.737	0.222	0.571
SimLex-999	0.720	0.389	0.670	SimLex-999	0.655	0.305	0.670
WS-353	0.753	0.337	0.745	WS-353	0.689	0.330	0.732
(a) Delicious				(b) WikiNav			
Evaluated on Trained on	MEN	SimLex-999	WS-353	Evaluated on Trained on	MEN	SimLex-999	WS-353
raw	0.749	0.371	0.609	raw	0.863	0.614	0.828
MEN	0.781	0.175	0.463	MEN	0.915	0.581	0.803
SimLex-999	0.570	0.312	0.444	SimLex-999	0.851	0.728	0.815
WS-353	0.740	0.368	0.831	WS-353	0.862	0.616	0.920
(c) WikiGloVe				(d) ConceptNet			

8.3.3 Question Answering

As an applied evaluation setting, we train a question answering model based on word embeddings. Because we could show that word embeddings adapted using RRL exhibit an increased correlation with human intuition, we also expect them to do better in a task that utilizes pretrained word embeddings. With this, we want to show the usefulness of RRL in other settings than measuring semantic word relatedness.

Experimental Setup. We use the DrQA question answering framework provided by Chen et al. [49], which trains a Machine Reading model on a preprocessed Wikipedia dump and evaluates it on the Stanford Question Answering Dataset *SQuAD* [190]. This dataset consists of over 100,000 question and answer pairs on more than 500 Wikipedia articles. We compare the performance of the trained question answering model when using untrained vectors and vectors trained on the MEN dataset.

In this task, we fit the WikiGloVe and ConceptNet embedding datasets on the complete MEN relatedness scores. The parameter choices for RRL were as follows. The initial learning rate was set to $l = 0.01$, the batch size was set to 5000, and we trained each model for a maximum of 100 epochs. To evaluate how well our embeddings do in the DrQA question answering task, we train a document reader model for 100 epochs for each word embedding dataset and their adapted variants. Except for the number of epochs, we used the default parameter settings of DrQA.

Table 8.4: Evaluation results for the question answering task using DrQA. We report the F1 scores when evaluating the trained DrQA model on the SQuAD development set with 100 epochs. Delicious and WikiNav have not been evaluated in this setting, as the vocabulary (tags and Wikipedia article identifier) is not only semantically, but also structurally different to that of WikiGloVe and ConceptNet.

	raw	RRL
WikiGloVe	74.01	72.92
ConceptNet	73.14	74.37

We report the F1 scores of the trained reader model using the official evaluation script from the authors of SQuAD, evaluated on the SQuAD development set. This is the standard procedure, as the actual test data are not publicly available.

Results. In Table 8.4, we see that training WikiGloVe with MEN decreases performance in the question answering setting. In contrast, the re-trained ConceptNet embeddings improve upon the score of the untrained vectors and even outperform the result of the untrained WikiGloVe vectors. Consequently, this serves as another signal that injecting additional semantic information in word embeddings using RRL improves their usefulness in applications that rely on their semantic quality. However, Chen et al. report a maximum F1 score of 78.8%, which we cannot reach either way. Even the more comparable score of 77.3% they achieve when leaving out re-alignment of their word embeddings is higher than those we obtain. Still, we could show that on ConceptNet, we at least were able to improve the results.

8.4 Discussion

We now interpret and discuss the results presented in Section 8.3 in the order of the experiments.

Integrating Different Amounts of User Feedback. Throughout the whole experiment, we observed that using a higher amount of training data always resulted in improved relatedness scores, if there was a positive effect.

Metrics trained on the MEN dataset yield competitive results, which are even statistically significant at $p < 0.05$ with at least 50% training data. While this might be due to its big size, it might also be due to this dataset was created²: Using crowdsourcing, each human worker was shown two pairs of words and had to determine which of both pairs is more related. The higher a word pair in MEN is rated, the more often it was considered more related than the other one that was given. Our approach exploits

²<https://staff.fnwi.uva.nl/e.bruni/MEN>

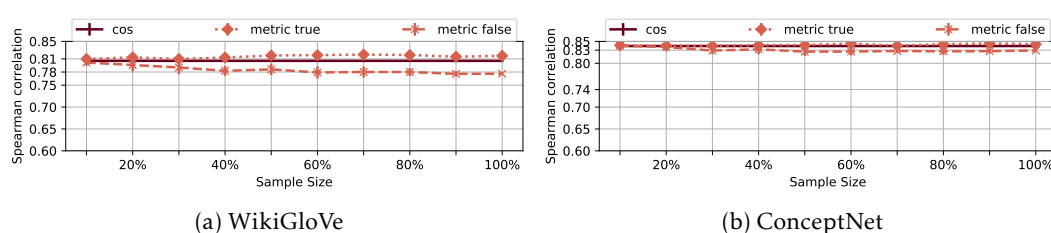


Figure 8.4: Results for training a measure on only 353 equidistant pairs from MEN. For both WikiGloVe and ConceptNet, the changes are only marginal and give the impression that there are not sufficient data to actually train a metric with significant differences. Yet, when utilizing the full MEN dataset, we achieve competitive results, as given in Figure 8.2. This supports the argument that not the content of WS-353 is responsible for the mediocre results shown in Figure 8.3, but rather the small size of WS-353.

very similar constraints for learning. Keeping this explanation in mind, we are able to give a recommendation on how to gather human feedback in order to learn semantic relatedness with our method.

Regarding results on WS-353, we still see that the RRL approach yields improved correlation scores, however not as much as when training metrics on MEN. We wanted to confirm our intuition that the comparatively bad performance of metrics trained on WS-353 is mainly due to size reasons and not because WS-353 contains unusable semantic relatedness information, so we reduced MEN to 353 uniformly distributed word pairs and performed the same training experiments as described in the previous section. The results from this experiment are given in Figure 8.4 and confirm our intuition that the cause of WS-353's bad performance can be attributed to its size, as the downsampled MEN dataset shows similarly marginal, but positive improvements as WS-353, although training on the full MEN dataset achieved remarkable results before.

An interesting pointer for future research lies in the fact that although Delicious and WikiNav share only a small overlap with WS-353, the impact on the trained metric's performance is very different. Here, it is interesting if there are certain subsets of word pair datasets from which training a metric is easier, even although they are small.

Transporting User Intentions. In this experiment, we attempted to see if the learned semantic information contained in one semantic relatedness dataset can also be beneficial for the evaluation on another relatedness dataset. However, our results indicate that this is not the case. Throughout all combinations of training and evaluation dataset, even across all word embedding datasets, we were unable to significantly increase correlation to human judgment. The only obvious exceptions are when training and evaluating on the same dataset. However, on the SimLex-999 dataset, our scores even deteriorated when using the WikiNav and WikiGloVe embeddings, which indicates that RRL cannot use these vectors to encode the information contained in SimLex-999. In how far this is

a signal of a possible (negative) correlation of WikiGloVe, remains a problem for future work.

Robustness. On all four embedding datasets, evaluation performance decreases notably with the MEN dataset, with the worst performance loss on ConceptNet. We observe similar responses on WikiGloVe. These results confirm (again) that word embeddings successfully manage to encode semantic information, and also that we cannot just “unlearn” it. Furthermore, all embedding sets react the most when used with a measure trained on MEN. We attribute this to the same reasons as why MEN is seemingly best suited to learn semantic relations from, i.e., it is constructed in a very similar way to the form of the constraints that the learning algorithm is parameterized with and it can provide a sufficient amount of training data to provide clear results. Another consequence of this is that the promising results of the previous experiments are indeed caused by the successful injection of semantic side information into the relatedness measure.

Comparison and Combination with Retrofitting. Table 8.2 shows that on the one hand, RRL trained with MEN often achieves better results than those of Retrofitting. On the other hand, RRL and Retrofitting can benefit from each other, as the scores of vectors trained with both algorithms often achieve the highest evaluation results. This is a very strong signal, especially since it holds true on 600 word pairs from MEN, which thus renders the score improvements statistically significant. It also shows that Retrofitting and RRL address different kinds of semantic relatedness: While Retrofitting was specifically designed to improve the semantic *similarity* of vector datasets, RRL can address the semantic *relatedness* component.

Question Answering In the question answering task, the results unfortunately are not as clear-cut as they are in the other experiments. While retraining of ConceptNet vectors actually improves the question answering scores, they deteriorate on retrained WikiGloVe vectors. Compared with the significant and large improvements in the word similarity task, this leads to the impression that semantic relatedness does not play a very important part in the question answering task. Another explanation could be that the semantic information obtained when using MEN as training data cannot support the QA task. However, this investigation is left as future work.

8.5 Conclusion

In this section, we presented the RRL approach to learn semantic relatedness from human intuition based on word embeddings. After motivating the algorithm design, we provided a detailed description of the algorithm. We performed several experiments to show the validity and usefulness of RRL. Concretely, we trained RRL on different amounts of training information, compared the algorithm to Retrofitting, tested the robustness of the word embeddings to indirectly obtain a qualitative judgment about the

validity of our results and finally attempted to transfer the semantic knowledge from one semantic relatedness dataset to another. Lastly, we evaluated RRL in a question answering setting.

Our results indicate that RRL can exploit semantic relatedness information from semantic relatedness datasets to more realistically assess human judgment of semantic information, regardless of the underlying embedding dataset. As a special result of our work, we outperformed the current state-of-the-art correlation scores on the MEN dataset and thus set a new record on measuring semantic relatedness on MEN.

Chapter 9

Conclusion and Future Perspectives

In this thesis, we investigated semantics in social media data. After determining the influence of semantics on user behavior in social media systems, we proposed methods to extract this semantic information from traces of user behavior and finally use human intuition to fine-tune the resulting semantic representations. Here we could also quantify the influence of user pragmatics on the extractable semantics in different settings. In this chapter, we will first summarize the contributions and results from each chapter. Finally, we point out open challenges and possible perspectives for future work on detecting signals of semantics in social media data, making these signals visible in the form of vector representations, and fine-tuning these vector representations for both general and specific purposes.

9.1 Summary

This thesis starts with an introduction to the topic of Semantics in the Web in Chapter 1. We start with a motivating example of how semantic technologies can be applied to make computers “understand” plain text and respond accordingly. Then, we show the connection to the Semantic Web and finally describe the covered research topics and contributions of this work. Chapter 2 presents an overview of relevant work and the state-of-the-art in computational semantics. We talk about the extraction of semantics from knowledge bases, folksonomies and Wikipedia, learning and adjusting semantic representations, the mutual influence of semantics and user behavior and potential applications of semantics. To understand the remainder of this work, we introduced the theoretical foundations in Chapter 3. Here, we first described important data structures, before we introduced baseline methods both to extract and learn semantics from knowledge graphs, folksonomy metadata, and Wikipedia. Additionally, we presented two approaches to evaluate semantic relatedness measures, one based on WordNet and the other based on human intuition. Finally, we defined measures for user behavior, both in tagging and navigation contexts. Chapter 4 gives an overview of the used datasets in this thesis. We extract semantics from text-based datasets, such as tagging data, and sense distributions from Wikipedia disambiguation pages. Furthermore, we describe all link-based datasets used to extract semantics. For our learning experiments, we

compare ourselves with pretrained word embedding datasets and finally we evaluate all our relatedness scores on human intuition, which is encoded in ground-truth datasets with crowd-sourced scores.

Chapters 1 to 4 described the necessary background for this work. All subsequent chapters, namely chapter 5 to 8, now describe the contributions of this thesis.

In Chapter 5, we analyzed different navigation datasets with human navigation both in Wikipedia and in BibSonomy. We especially focussed on uncovering a semantic component in navigation on these networks. We could show that in all link networks, semantic navigation plays a certain role in human wayfinding; it is however not as strong as we thought it would be. Nonetheless, our work on BibSonomy could show that users with different tagging behavior also exhibit different navigation behavior.

Chapter 6 extended the thesis of Benz [26] by first challenging the way that tagging data were evaluated until this point. We argued that evaluation on the Jiang-Conrath measure on WordNet would not yield a valid evaluation score, since Jiang-Conrath does not correlate well with human intuition. Furthermore, although evaluation on human intuition datasets is a more valid evaluation approach, they need to contain a fitting vocabulary for a sufficiently interpretable result. We proposed a new evaluation dataset called Bib100. After that, we discussed the applicability of word embedding approaches on tagging data to move tagging semantics on the next level. We found that not only can we show that tagging data indeed contain considerable amounts of semantic information, but we were also able to extract representations with a higher quality than the previous representations from [26].

Subsequently, in Chapter 7, we then proposed several semantic relatedness measures to extract semantic relatedness from human navigation on social media. For this, we investigated Wikipedia game navigation, unconstrained navigation on Wikipedia and random walks on Wikipedia subgraphs. Here, we could show that not only Wikipedia navigation contains a great amount of semantic information, but it can be easily extracted. Furthermore, we also proposed to construct word embeddings from navigation, which boosted the quality of vector representations significantly. We extended the semantic relatedness measures on unconstrained and random navigation on the BibSonomy link network. In order to overcome data sparsity, we explored primed random walks on both navigation datasets and could show that certain primes could even increase the semantic quality of the navigational vector representations significantly.

Finally, Chapter 8 covered both an algorithm to learn semantic relatedness from human intuition and a study on user behavior influence on a word sense disambiguation algorithm in a social tagging system. We performed experiments with different word embedding datasets as well as different competing similarity learning algorithms and could show that our RRL algorithm outperformed the current state-of-the-art. Additionally, we were able to show that the tagging behavior of users indeed influences the word sense disambiguation process in a social tagging system. This is perfectly in line with [122] and [173], where it was also shown that tagging behavior influences other topics in folksonomies as well.

9.2 Open Problems and Future Perspectives

While this thesis made several contributions to further the state-of-the-art, we left several interesting research pathways untouched. In the following, we will point out potential for future work.

Analyzing Semantics in Navigation. In Section 5.2, we have analyzed several datasets of human navigation in Wikipedia. With our analyses, we have barely scratched the surface of what is possible. Potential extensions include the hypothesis-based analysis of navigation of successful and unsuccessful users. Since we found in Section 7.2.2 that successful paths yield less precise semantics, we also expect to see different navigation patterns in either settings. Similar things hold for unconstrained navigation on Wikipedia. Since Wikimedia Research regularly publishes clickstream data, similar to the dataset we used (cf. Section 4.2.2.1), it would also be interesting if navigation behavior is generally consistent through time. Finally, there also exist many options to exploit the semantic information models we created in Section 7.2. One could for example use them to recommend pages that are relevant to a currently browsing user’s interests. Additionally, we could use them to improve the category system of Wikipedia by clustering pages according to their navigational semantic similarity.

Semantics in Social Tagging Systems. By bridging the gap between tagging semantics and word embeddings, we were not only able to capture more of the semantic information in tagging data, but we also connected tagging semantics to today’s world of NLP research. Using these tag embeddings, we could now proceed to apply more sophisticated methods to e.g., embed posts using CNNs [118], improve the detection of hierarchical relations in tagging data [198], and finally even improve upon learning taxonomic relations from tagging data using word embeddings [28].

Learning Semantics. A canonical extension of the RRL algorithm could be to classify or even measure hierarchical relations between words. Here, we could learn a single translation matrix that learns an antisymmetric matrix to characterize the inverse nature of hyponymy and hypernymy. To provide a concrete application for improved word embeddings and word embeddings in general, we propose to investigate how much. Another idea is to learn *translations* between word embeddings, not necessarily generated by different languages. An open problem is to overcome the Out-of-Vocabulary limitation of semantic models. This could for example be solved by a translation matrix between two embedding datasets which projects key vectors from one embedding space to the other and vice versa. A possible application of such a translation algorithm could be to measure social *bias* in word embeddings and effectively erasing it. By training word embeddings both on original and preprocessed text by replacing every gendered form by only one gender, it is possible to train the translation matrix and then to apply this matrix on the gendered embeddings.

List of Notations

1 General Notation

$x \in \mathbb{R}^n$	vector
$M \in \mathbb{R}^{n \times n}$	quadratic matrix

2 Folksonomy Notation

\mathbb{F}	Folksonomy
\mathbb{P}_u	Personomy of user u
U, T, R	Users, Tags and Resources in a Folksonomy
Y	The set of tag assignments in a Folksonomy
P	An unspecified post in a Folksonomy
P_{ur}	A post of user u with the resource r
T_{ur}	The tags assigned to the post P_{ur}
T_u	The tags assigned by user u
R_u	All resources that user u posted

3 Semantics Notation

\mathcal{V}	Vocabulary
$\cos(x, y)$	Cosine similarity measure of two vectors x, y
$\cos_M(x, y)$	Parameterized cosine similarity measure

Evaluation notation	
$\rho(X, Y)$	Spearman rank correlation coefficient of two rankings X and Y
$\rho_n(X, Y)$	Spearman rank correlation coefficient restricted to a certain number of word pairs
$\rho_{bin}(X, Y)$	Spearman rank correlation coefficient of rankings generated by vector binarization

4 Graph and Navigation Notations

p	Page in a social media system
$p^{(u',t',r')}$	a page in the social tagging system link graph
\mathbf{p}	Path of pages, $\mathbf{p} := (p_1, \dots, p_n)$
\mathbb{P}	Set of all navigational paths, $\mathbb{P} := \{\mathbf{p}_1, \dots, \mathbf{p}_m\}$
<hr/>	
$G = (V, E)$	A graph G with vertices $V = \{p_1, \dots, p_n\}$ and edges $E = \{(p_i, p_j)\}$
V_u	All pages \mathbf{p} that “belong” to user u in a Folksonomy
<hr/>	
\mathcal{H}	A navigation hypothesis
$\bar{P}_{hyponame}$	A transition function that defines the hypothesis $\mathcal{H}_{hyponame}$
<hr/>	

Bibliography

- [1] Dimitris Achlioptas. “Database-friendly random projections: Johnson-Lindenstrauss with binary coins”. In: *Journal of Computer and System Sciences* 66.4 (2003). Special Issue on PODS 2001, pp. 671–687.
- [2] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. “A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches”. In: *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL ’09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 19–27.
- [3] Eneko Agirre, Ander Barrena, and Aitor Soroa. *Studying the Wikipedia Hyperlink Graph for Relatedness and Disambiguation*. cite arxiv:1503.01655. 2015.
- [4] Morgan Ames and Mor Naaman. “Why We Tag: Motivations for Annotation in Mobile and Online Media”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’07. New York, NY, USA: ACM, 2007, pp. 971–980.
- [5] Sofia Angeletou. “Semantic Enrichment of Folksonomy Tagspaces.” In: *International Semantic Web Conference*. Vol. 5318. Lecture Notes in Computer Science. Springer, 2008, pp. 889–894.
- [6] Sofia Angeletou, Marta Sabou, and Enrico Motta. “Semantically enriching folksonomies with FLOR”. In: *Proceedings of the CISWeb Workshop, located at the 5th European Semantic Web Conference ESWC 2008*. 2008.
- [7] Francis J Anscombe. “Graphs in statistical analysis”. In: *The American Statistician* 27.1 (1973), pp. 17–21.
- [8] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. “Linear algebraic structure of word senses, with applications to polysemy”. In: *arXiv preprint arXiv:1601.03764* (2016).
- [9] Ching Man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. “Contextualising tags in collaborative tagging systems”. In: *Proceedings of the 20th ACM conference on Hypertext and hypermedia (HT2009)*. New York, NY, USA: ACM, 2009, pp. 251–260.

Bibliography

- [10] Ching-man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. “Web search disambiguation by collaborative tagging”. In: *Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR’08)*. 2008.
- [11] Ching man Au Yeung, Nicholas Gibbins, and Nigel Shadbolt. “Understanding the Semantics of Ambiguous Tags in Folksonomies”. In: *Proceedings of the International Workshop on Emergent Semantics and Ontology Evolution (ESOE2007) at ISWC/ASWC2007, Busan, South Korea, November*. 2007.
- [12] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. “DBpedia: A Nucleus for a Web of Open Data”. In: *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 722–735.
- [13] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*. Vol. 463. ACM press New York, 1999.
- [14] Dzmitry Bahdanau, Tom Bosc, Stanisław Jastrzębski, Edward Grefenstette, Pascal Vincent, and Yoshua Bengio. *Learning to Compute Word Embeddings On the Fly*. cite arxiv:1706.00286. 2017.
- [15] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”. In: *arXiv preprint arXiv:1409.0473* (2014).
- [16] Satanjeev Banerjee and Ted Pedersen. “An adapted Lesk algorithm for word sense disambiguation using WordNet”. In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer. 2002, pp. 136–145.
- [17] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. “The WaCky wide web: a collection of very large linguistically processed web-crawled corpora”. In: *Language Resources and Evaluation* 43.3 (2009), pp. 209–226.
- [18] Marco Baroni, Gerorgiana Dinu, and German Kruszewski. “Don’t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors”. In: *ACL* (2014), pp. 238–247.
- [19] Martin Becker, Kathrin Borchert, Matthias Hirth, Hauke Mewes, Andreas Hotho, and Phuoc Tran-Gia. “MicroTrails: Comparing Hypotheses About Task Selection on a Crowdsourcing Platform”. In: *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*. i-KNOW ’15. New York, NY, USA: ACM, 2015, 10:1–10:8.
- [20] Martin Becker, Florian Lemmerich, Philipp Singer, Markus Strohmaier, and Andreas Hotho. “MixedTrails: Bayesian hypothesis comparison on heterogeneous sequential data”. In: *Data Mining and Knowledge Discovery* 31.5 (2017), pp. 1359–1390.

- [21] Martin Becker, Hauke Mewes, Andreas Hotho, Dimitar Dimitrov, Florian Lemmerich, and Markus Strohmaier. “SparkTrails: A MapReduce Implementation of HypTrails for Comparing Hypotheses About Human Trails.” In: *WWW (Companion Volume)*. ACM, 2016, pp. 17–18.
- [22] Martin Becker, Philipp Singer, Florian Lemmerich, Andreas Hotho, Denis Helic, and Markus Strohmaier. “Photowalking the City: Comparing Hypotheses About Urban Photo Trails on Flickr”. In: *Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings*. Cham: Springer International Publishing, 2015, pp. 227–244.
- [23] Aurélien Bellet, Amaury Habrard, and Marc Sebban. *A Survey on Metric Learning for Feature Vectors and Structured Data*. cite arxiv:1306.6709. 2013.
- [24] Fabricio Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgilio Almeida. “Characterizing User Behavior in Online Social Networks”. In: *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*. IMC ’09. New York, NY, USA: ACM, 2009, pp. 49–62.
- [25] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. “A neural probabilistic language model”. In: *JMLR* (2003), pp. 1137–1155.
- [26] Dominik Benz. “Capturing Emergent Semantics from Social Annotation Systems”. PhD thesis. 2012.
- [27] Dominik Benz, Andreas Hotho, Robert Jäschke, Beate Krause, Folke Mitzlaff, Christoph Schmitz, and Gerd Stumme. “The Social Bookmark and Publication Management System BibSonomy”. In: *VLDB* (2010), pp. 849–875.
- [28] Dominik Benz, Andreas Hotho, Stefan Stützer, and Gerd Stumme. “Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge”. In: *WebSci*. 2010.
- [29] Dominik Benz, Christian Körner, Andreas Hotho, Gerd Stumme, and Markus Strohmaier. “One Tag to Bind Them All : Measuring Term Abstractness in Social Metadata”. In: *Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011)*. Heraklion, Crete, 2011.
- [30] Bettina Berendt. “Using site semantics to analyze, visualize, and support navigation”. In: *Data Mining and Knowledge Discovery* 6.1 (2002), pp. 37–59.
- [31] T. Berners-Lee, J. Hendler, and O. Lassila. “The Semantic Web”. In: *Scientific American* 284.5 (2001), pp. 34–43.
- [32] Ravish Bhagdev, Sam Chapman, Fabio Ciravegna, Vitaveska Lanfranchi, and Daniela Petrelli. “Hybrid search: Effectively combining keywords and semantic searches”. In: *European Semantic Web Conference*. Springer. 2008, pp. 554–568.
- [33] Jiang Bian, Bin Gao, and Tie-Yan Liu. “Knowledge-powered deep learning for word embedding”. In: *ECML/PKDD*. 2014, pp. 132–148.

Bibliography

- [34] Claudio Biancalana, Fabio Gaspiretti, Alessandro Micarelli, and Giuseppe Sansonetti. "Social semantic query expansion". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 4.4 (2013), p. 60.
- [35] Nikos Bikakis, Giorgos Giannopoulos, Theodore Dalamagas, and Timos Sellis. "Integrating keywords and semantics on document annotation and search". In: *On the Move to Meaningful Internet Systems, OTM 2010* (2010), pp. 921–938.
- [36] J Bollen, H Van de Sompel, A Hagberg, L Bettencourt, R Chute, M A Rodriguez, and L Balakireva. "Clickstream data yields high-resolution maps of science". In: *PLoS One* 4.3 (2009).
- [37] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. "Translating embeddings for modeling multi-relational data". In: *Advances in neural information processing systems*. 2013, pp. 2787–2795.
- [38] Angel Borrego and Jenny Fry. "Measuring researchers' use of scholarly information through social bookmarking data: A case study of BibSonomy". In: *Journal of Information Science* 38.3 (2012), pp. 297–308.
- [39] Fabian Both, Steffen Thoma, and Achim Rettinger. "Cross-modal Knowledge Transfer: Improving the Word Embedding of Apple by Looking at Oranges". In: *K-CAP*. ACM, 2017, 18:1–18:8.
- [40] Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. "Adding dense, weighted connections to WordNet". In: *Proceedings of the third international WordNet conference*. 2006, pp. 29–36.
- [41] L.M. Bregman. "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming". In: *USSR Computational Mathematics and Mathematical Physics* 7.3 (1967), pp. 200–217.
- [42] Sergey Brin and Lawrence Page. "The Anatomy of a Large-scale Hypertextual Web Search Engine". In: *Comput. Netw. ISDN Syst.* 30.1-7 (1998), pp. 107–117.
- [43] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. "Multimodal Distributional Semantics". In: *JAIR* (2014).
- [44] Alexander Budanitsky and Graeme Hirst. "Evaluating WordNet-based Measures of Lexical Semantic Relatedness". In: *Computational Linguists* 32.1 (2006), pp. 13–47.
- [45] Alexander Budanitsky and Graeme Hirst. "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures". In: *Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, USA* (2001).
- [46] John A Bullinaria and Joseph P Levy. "Extracting semantic representations from word co-occurrence statistics: A computational study". In: *BRM* 39.3 (2007), pp. 510–526.

- [47] Ciro Cattuto. “Semiotic dynamics in online social communities”. In: *The European Physical Journal C - Particles and Fields* 46 (Aug. 2006), pp. 33–37.
- [48] Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. “Semantic Grounding of Tag Relatedness in Social Bookmarking Systems”. In: *ISWC*. 2008.
- [49] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. “Reading Wikipedia to Answer Open-Domain Questions.” In: *ACL (1)*. Association for Computational Linguistics, 2017, pp. 1870–1879.
- [50] Haochen Chen, Bryan Perozzi, Yifan Hu, and Steven Skiena. *HARP: Hierarchical Representation Learning for Networks*. cite arxiv:1706.07845. 2017.
- [51] Ed H. Chi, Peter Pirolli, Kim Chen, and James Pitkow. “Using Information Scent to Model User Information Needs and Actions and the Web”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '01. New York, NY, USA: ACM, 2001, pp. 490–497.
- [52] G L Ciampaglia, P Shiralkar, L M Rocha, J Bollen, F Menczer, and A Flammini. “Computational Fact Checking from Knowledge Networks”. In: *PLoS One* 10.6 (2015).
- [53] Rudi Cilibrasi and Paul M. B. Vitányi. “The Google Similarity Distance”. In: *CoRR abs/cs/0412098* (2004).
- [54] Philipp Cimiano. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
- [55] Alexander Dallmann, Thomas Niebler, Florian Lemmerich, and Andreas Hotho. “Extracting Semantics from Random Walks on Wikipedia: Comparing Learning and Counting Methods”. In: *Proceedings of the 10th International Conference on Web and Social Media*. AAI, 2016.
- [56] Laurie Damianos, John Griffith, Donna Cuomo, David Hirst, and James Smallwood. “Onomi: Social Bookmarking on a Corporate Intranet”. In: *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*. 2006.
- [57] Sanjoy Dasgupta and Anupam Gupta. “An elementary proof of a theorem of Johnson and Lindenstrauss”. In: *Random Structures Algorithms* 22.1 (2003), pp. 60–65.
- [58] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. “Indexing by latent semantic analysis”. In: *Journal of the American Society for Information Science* 41.6 (1990), pp. 391–407.
- [59] Dimitar Dimitrov, Florian Lemmerich, Fabian Flöck, and Markus Strohmaier. “Query for Architecture, Click through Military: Comparing the Roles of Search and Navigation on Wikipedia”. In: *arXiv preprint arXiv:1805.04022* (2018).
- [60] Dimitar Dimitrov, Philipp Singer, Denis Helic, and Markus Strohmaier. “The Role of Structural Information for Designing Navigational User Interfaces”. In: *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. HT '15. New York, NY, USA: ACM, 2015, pp. 59–68.

Bibliography

- [61] Dimitar Dimitrov, Philipp Singer, Florian Lemmerich, and Markus Strohmaier. “What Makes a Link Successful on Wikipedia?” In: *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2017, pp. 917–926.
- [62] Stephan Doerfel and Robert Jäschke. “An analysis of tag-recommender evaluation procedures.” In: *RecSys*. ACM, 2013, pp. 343–346.
- [63] Stephan Doerfel, Robert Jäschke, and Gerd Stumme. “The Role of Cores in Recommender Benchmarking for Social Bookmarking Systems”. In: *ACM Trans. Intell. Syst. Technol.* 7.3 (2016), 40:1–40:33.
- [64] Stephan Doerfel, Daniel Zoller, Philipp Singer, Thomas Niebler, Andreas Hotho, and Markus Strohmaier. “What Users Actually do in a Social Tagging System: A Study of User Behavior in BibSonomy”. In: *ACM Transactions on the Web* 10.2 (2016), 14:1–14:32.
- [65] Stephan Doerfel, Daniel Zoller, Philipp Singer, Thomas Niebler, Markus Strohmaier, and Andreas Hotho. “How Social is Social Tagging?” In: *Proceedings of the 23rd International World Wide Web Conference*. WWW 2014. ACM, 2014, pp. 251–252.
- [66] Beate Dorow and Dominic Widdows. “Discovering corpus-specific word senses”. In: *Proceedings of the 10th conference on European chapter of the Association for Computational Linguistics*. Vol. 2. EACL ’03. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 79–82.
- [67] Sergio Duarte Torres, Ingmar Weber, and Djoerd Hiemstra. “Analysis of Search and Browsing Behavior of Young Users on the Web”. In: *ACM Trans. Web* 8.2 (2014), 7:1–7:54.
- [68] Lisa Ehrlinger and Wolfram Wöß. “Towards a Definition of Knowledge Graphs.” In: *SEMANTiCS (Posters, Demos, SuCCESS)*. 2016.
- [69] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard H. Hovy, and Noah A. Smith. “Retrofitting Word Vectors to Semantic Lexicons.” In: *CoRR* (2014).
- [70] Manaal Faruqui and Chris Dyer. “Improving vector space word representations using multilingual correlation”. In: Association for Computational Linguistics. 2014.
- [71] Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. *Problems With Evaluation of Word Embeddings Using Word Similarity Tasks*. cite arxiv:1605.02276v1. 2016.
- [72] Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.
- [73] Jun Feng, Minlie Huang, Mingdong Wang, Mantong Zhou, Yu Hao, and Xiaoyan Zhu. “Knowledge Graph Embedding by Flexible Translation.” In: *KR*. 2016, pp. 557–560.
- [74] Edgar C Fieller, Herman O Hartley, and Egon S Pearson. “Tests for rank correlation coefficients. I”. In: *Biometrika* 44.3/4 (1957), pp. 470–481.

- [75] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppín. “Placing Search in Context: the Concept Revisited”. In: WWW. 2001, pp. 116–131.
- [76] J. Firth. *A synopsis of linguistic theory 1930-1955*. Studies in Linguistic Analysis, Philological. Longman, 1957.
- [77] Dayne Freitag, Matthias Blume, John Byrnes, Edmond Chow, Sadik Kapadia, Richard Rohwer, and Zhiqiang Wang. “New Experiments in Distributional Representations of Synonymy”. In: *Proceedings of the Ninth Conference on Computational Natural Language Learning*. CONLL ’05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 25–32.
- [78] Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. “Learning semantic hierarchies via word embeddings”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2014, pp. 1199–1209.
- [79] Wai-Tat Fu, Thomas Kannampallil, Ruogu Kang, and Jibo He. “Semantic Imitation in Social Tagging”. In: *ACM Trans. Comput.-Hum. Interact.* 17.3 (2010), 12:1–12:37.
- [80] Nancy Fulda, Daniel Ricks, Ben Murdoch, and David Wingate. “What Can You Do with a Rock? Affordance Extraction via Word Embeddings.” In: *IJCAI*. ijcai.org, 2017, pp. 1039–1045.
- [81] Evgeniy Gabrilovich and Shaul Markovitch. “Computing semantic relatedness using Wikipedia-based explicit semantic analysis”. In: *IJCAI*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp. 1606–1611.
- [82] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. “PPDB: The Paraphrase Database”. In: *Proceedings of NAACL-HLT*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 758–764.
- [83] Andres Garcia-Silva, Martin Szomszor, Harith Alani, and Oscar Corcho. “Preliminary results in tag disambiguation using dbpedia”. In: *The Fifth International Conference on Knowledge Capture (K-Cap’09) - First International Workshop on Collective Knowledge Capturing and Representation (CKCaR’09)*. Informatica, 2009.
- [84] Jim Giles. “Internet encyclopaedias go head to head”. In: *Nature* 438 (2005), pp. 900–.
- [85] Scott A. Golder and Bernardo A. Huberman. “Usage patterns of collaborative tagging systems”. In: *Journal of Information Science* 32.2 (2006), pp. 198–208.
- [86] Scott Golder and Bernardo A. Huberman. *The Structure of Collaborative Tagging Systems*. 2005.
- [87] Aditya Grover and Jure Leskovec. “node2vec: Scalable Feature Learning for Networks.” In: *KDD*. 2016.
- [88] Nicola Guarino, Daniel Oberle, and Steffen Staab. “What is an ontology?” In: *Handbook on Ontologies*. Springer, 2009, pp. 1–17.

Bibliography

- [89] Ramanathan Guha, Rob McCool, and Eric Miller. "Semantic search". In: *Proceedings of the 12th international conference on World Wide Web*. ACM. 2003, pp. 700–709.
- [90] Joan Guisado-Gamez and Arnau Prat-Perez. "Understanding Graph Structure of Wikipedia for Query Expansion". In: *Proceedings of the GRADES'15*. GRADES'15. New York, NY, USA: ACM, 2015, 6:1–6:6.
- [91] Joan Guisado-Gamez, David Tamayo-Domenech, Jordi Urmeneta, and Josep Lluís Larriba-Pey. "ENRICH: A Query Rewriting Service Powered by Wikipedia Graph Structure". In: *Tenth International AAAI Conference on Web and Social Media*. 2016.
- [92] Yuhang Guo, Wanxiang Che, Ting Liu, and Sheng Li. "A graph-based method for entity linking". In: *Proceedings of 5th International Joint Conference on Natural Language Processing*. 2011, pp. 1010–1018.
- [93] Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. "Large-scale Learning of Word Relatedness with Constraints". In: *KDD*. New York, NY, USA: ACM, 2012, pp. 1406–1414.
- [94] Harry Halpin, Valentin Robu, and Hana Shepherd. "The Complex Dynamics of Collaborative Tagging". In: *Proceedings of the 16th International Conference on World Wide Web*. WWW '07. New York, NY, USA: ACM, 2007, pp. 211–220.
- [95] Eszter Hargittai and Gina Walejko. "THE PARTICIPATION DIVIDE: Content creation and sharing in the digital age". In: *Information, Communication & Society* 11.2 (2008), pp. 239–256.
- [96] Zellig S Harris. "Distributional structure". In: *Word* 10.2-3 (1954), pp. 146–162.
- [97] S. Hassan and R. Mihalcea. "Semantic relatedness using salient semantic analysis". In: *Proceedings of AAAI Conference on Artificial Intelligence*. 2011.
- [98] Samer Hassan and Rada Mihalcea. "Cross-lingual semantic relatedness using encyclopedic knowledge". In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*. Association for Computational Linguistics. 2009, pp. 1192–1201.
- [99] Marti A. Hearst. "Automatic Acquisition of Hyponyms from Large Text Corpora". In: *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*. COLING '92. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992, pp. 539–545.
- [100] M. Heckner, M. Heilemann, and C. Wolff. "Personal Information Management vs. Resource Sharing: Towards a Model of Information Behaviour in Social Tagging Systems". In: *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*. San Jose, CA, USA, 2009.
- [101] Karl Moritz Hermann and Phil Blunsom. "Multilingual models for compositional distributed semantics". In: *arXiv preprint arXiv:1404.4641* (2014).

- [102] Lena Hettinger, Alexander Dallmann, Albin Zehe, Thomas Niebler, and Andreas Hotho. *ClaiRE at SemEval-2018 Task 7: Classification of Relations using Embeddings*. New Orleans, LA, USA, 2018.
- [103] Paul Heymann and Hector Garcia-Molina. *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems*. Technical Report 2006-10. Stanford InfoLab, Apr. 2006.
- [104] Felix Hill, Roi Reichart, and Anna Korhonen. “Simlex-999: Evaluating semantic models with (genuine) similarity estimation”. In: *Computational Linguistics* (2015).
- [105] Graeme Hirst, David St-Onge, et al. “Lexical chains as representations of context for the detection and correction of malapropisms”. In: *WordNet: An electronic lexical database* 305 (1998), pp. 305–332.
- [106] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. “Information Retrieval in Folksonomies: Search and Ranking”. In: *ESWC*. 2006.
- [107] Christoph Hube. “Bias in Wikipedia”. In: *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee. 2017, pp. 717–721.
- [108] Thad Hughes and Daniel Ramage. “Lexical Semantic Relatedness with Random Graph Walks”. In: *EMNLP-CoNLL*. 2007, pp. 581–589.
- [109] John Hutchins. “Retrospect and prospect in computer-based translation”. In: *Machine Translation Summit VII, 13th-17th September* (1999), pp. 30–34.
- [110] Masahiro Ito, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. “Association Thesaurus Construction Methods Based on Link Co-occurrence Analysis for Wikipedia”. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM ’08. New York, NY, USA: ACM, 2008, pp. 817–826.
- [111] Robert Jäschke, Leandro Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. “Tag Recommendations in Folksonomies”. In: *PKDD*. 2007.
- [112] Jay J. Jiang and David W. Conrath. “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy”. In: *cite arxiv:cmp-lg/9709008Comment: 15 pages, Postscript only*. 1997.
- [113] Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. “Generating query substitutions”. In: *WWW ’06: Proceedings of the 15th international conference on World Wide Web*. WWW ’06. New York, NY, USA: ACM, 2006, pp. 387–396.
- [114] Ion Juvina and Herre van Oostendorp. “Modeling Semantic and Structural Knowledge in Web Navigation”. In: *Discourse Processes* 45.4-5 (2008), pp. 346–364.
- [115] Jeon-Hyung Kang and Kristina Lerman. “Leveraging User Diversity to Harvest Knowledge on the Social Web.” In: *SocialCom/PASSAT*. IEEE, 2011, pp. 215–222.
- [116] Robert E. Kass and Adrian E. Raftery. “Bayes Factors”. In: *Journal of the American Statistical Association* 90.430 (1995), pp. 773–795.

Bibliography

- [117] Joo-Kyung Kim, Marie-Catherine de Marneffe, and Eric Fosler-Lussier. “Adjusting word embeddings with semantic intensity orders”. In: *ACL 2016 workshop on Representation Learning for NLP (Repl4NLP)*. 2016, pp. 62–69.
- [118] Yoon Kim. “Convolutional Neural Networks for Sentence Classification”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. 2014, pp. 1746–1751.
- [119] J M Kleinberg. “Navigation in a small world”. In: *Nature* 406.6798 (2000), pp. 845–845.
- [120] Jon M. Kleinberg. “Authoritative Sources in a Hyperlinked Environment”. In: *J. ACM* 46.5 (1999), pp. 604–632.
- [121] Philipp Koehn, Franz Josef Och, and Daniel Marcu. “Statistical phrase-based translation”. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics. 2003, pp. 48–54.
- [122] Christian Körner, Dominik Benz, Markus Strohmaier, Andreas Hotho, and Gerd Stumme. “Stop Thinking, start Tagging - Tag Semantics emerge from Collaborative Verbosity”. In: *WWW*. Raleigh, NC, USA: ACM, 2010.
- [123] Christian Körner, Roman Kern, Hans-Peter Grahsl, and Markus Strohmaier. “Of Categorizers and Describers: An Evaluation of Quantitative Measures for Tagging Motivation”. In: *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*. HT '10. New York, NY, USA: ACM, 2010, pp. 157–166.
- [124] Hideki Kozima. *Computing Lexical Cohesion as a Tool for Text Analysis*. Tech. rep. 1993.
- [125] Beate Krause, Christoph Schmitz, Andreas Hotho, and Gerd Stumme. “The Anti-Social Tagger - Detecting Spam in Social Bookmarking Systems”. In: *AIRWeb '08: Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web*. New York, NY, USA: ACM, 2008, pp. 61–68.
- [126] Brian Kulis, Mátyás A Sustik, and Inderjit S Dhillon. “Low-rank kernel learning with Bregman matrix divergences”. In: *Journal of Machine Learning Research* 10.Feb (2009), pp. 341–376.
- [127] Daniel Lamprecht, Dimitar Dimitrov, Denis Helic, and Markus Strohmaier. “Evaluating and Improving Navigability of Wikipedia: A Comparative Study of Eight Language Editions”. In: *Proceedings of the 12th International Symposium on Open Collaboration*. OpenSym '16. New York, NY, USA: ACM, 2016, 17:1–17:10.
- [128] Daniel Lamprecht, Kristina Lerman, Denis Helic, and Markus Strohmaier. “How the structure of Wikipedia articles influences user navigation”. In: *New Review of Hypermedia and Multimedia* 23.1 (2017), pp. 29–50.
- [129] Quoc V. Le and Tomas Mikolov. *Distributed Representations of Sentences and Documents*. cite arxiv:1405.4053. 2014.

- [130] Kangpyo Lee, Hyunwoo Kim, Hyopil Shin, and Hyoung-Joo Kim. “Tag sense disambiguation for clarifying the vocabulary of social tags”. In: *CSE (4)*. Vol. 4. IEEE. 2009, pp. 729–734.
- [131] Michael Lesk. “Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone”. In: *Proceedings of the 5th annual international conference on Systems documentation*. ACM. 1986, pp. 24–26.
- [132] Jure Leskovec, Kevin J. Lang, and Michael Mahoney. “Empirical comparison of algorithms for network community detection”. In: *Proceedings of the 19th international conference on World wide web*. WWW ’10. New York, NY, USA: ACM, 2010, pp. 631–640.
- [133] Joseph P Levy, John A Bullinaria, and Samantha McCormick. “Semantic vector evaluation and human performance on a new vocabulary MCQ test”. In: *CogSci 2017 Proceedings*. 2017.
- [134] Omer Levy, Yoav Goldberg, and Ido Dagan. “Improving distributional similarity with lessons learned from word embeddings”. In: *TACL* 3 (2015), pp. 211–225.
- [135] Omer Levy, Yoav Goldberg, and Israel Ramat-Gan. “Linguistic Regularities in Sparse and Explicit Word Representations.” In: *CoNLL*. 2014, pp. 171–180.
- [136] Ping Li, Trevor J. Hastie, and Kenneth W. Church. “Very Sparse Random Projections”. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’06. New York, NY, USA: ACM, 2006, pp. 287–296.
- [137] Dekang Lin et al. “An information-theoretic definition of similarity.” In: *Icml*. Vol. 98. 1998. 1998, pp. 296–304.
- [138] Bing Liu. “Sentiment Analysis and Opinion Mining”. In: *Synthesis Lectures on Human Language Technologies* 5.1 (2012), pp. 1–167.
- [139] E. Y. Liu, Z. Guo, X. Zhang, V. Jojic, and W. Wang. “Metric Learning from Relative Comparisons by Minimizing Squared Residual”. In: *ICDM*. 2012, pp. 978–983.
- [140] Shuang Liu, Clement Yu, and Weiyi Meng. “Word Sense Disambiguation in Queries”. In: *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. CIKM ’05. New York, NY, USA: ACM, 2005, pp. 525–532.
- [141] Jared Lorince, Sam Zorowitz, Jaimie Murdock, and Peter M. Todd. “The Wisdom of the Few? “Supertaggers” in Collaborative Tagging Systems”. In: *The Journal of Web Science* 1.1 (2015).
- [142] Hongyin Luo, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. “Online learning of interpretable word embeddings”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 1687–1692.
- [143] Minh-Thang Luong, Richard Socher, and Christopher D Manning. “Better Word Representations with Recursive Neural Networks for Morphology”. In: *CoNLL-2013* (2013), p. 104.

Bibliography

- [144] Thang Luong, Hieu Pham, and Christopher D Manning. “Bilingual Word Representations with Monolingual Quality in Mind.” In: *VS@ HLT-NAACL*. 2015, pp. 151–159.
- [145] Alexander Maedche and Steffen Staab. “Ontology Learning for the Semantic Web”. In: *IEEE Intelligent Systems* 16.2 (2001), pp. 72–79.
- [146] Okumura Manabu and Honda Takeo. “Word sense disambiguation and text segmentation based on lexical cohesion”. In: *Proceedings of the 15th conference on Computational linguistics - Volume 2. COLING '94*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1994, pp. 755–761.
- [147] Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press, 1999.
- [148] Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme. “Evaluating Similarity Measures for Emergent Semantics of Social Tagging”. In: *WWW*. 2009, pp. 641–641.
- [149] Adam Mathes. *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*. <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>. 2004.
- [150] M. Meiss, B. Goncalves, J. Ramasco, A. Flammini, and F. Menczer. “Modeling Traffic on the Web Graph”. In: *Proc. 7th Workshop on Algorithms and Models for the Web Graph (WAW)*. Vol. 6516. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2010, pp. 50–61.
- [151] Mark R. Meiss, Filippo Menczer, Santo Fortunato, Alessandro Flammini, and Alessandro Vespignani. “Ranking Web Sites with Real User Traffic”. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining. WSDM '08*. New York, NY, USA: ACM, 2008, pp. 65–76.
- [152] Mark Meiss, John Duncan, Bruno Gonçalves, José J. Ramasco, and Filippo Menczer. “What’s in a Session: Tracking Individual Behavior on the Web”. In: *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia. HT '09*. New York, NY, USA: ACM, 2009, pp. 173–182.
- [153] Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. ““The sum of all human knowledge”: A systematic review of scholarly research on the content of Wikipedia”. In: *Journal of the Association for Information Science and Technology* 66.2 (2015), pp. 219–245.
- [154] R. Mihalcea. “Using wikipedia for automatic word sense disambiguation”. In: *Proceedings of NAACL HLT*. Vol. 2007. 2007, pp. 196–203.
- [155] Peter Mika. “Ontologies are us: A unified model of social networks and semantics”. In: *Web Semantics: Science, Services and Agents on the World Wide Web* 5.1 (Mar. 2007), pp. 5–15.

- [156] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. “Distributed Representations of Words and Phrases and their Compositionality”. In: *NIPS*. Curran Associates, Inc., 2013, pp. 3111–3119.
- [157] David R. Millen, Meng Yang, Steven Whittaker, and Jonathan Feinberg. “Social bookmarking and exploratory search”. In: *ECSCW 2007*. Springer London, 2007, pp. 21–40.
- [158] David R Millen and Jonathan Feinberg. “Using Social Tagging to Improve Social Navigation”. In: *Workshop on the Social Navigation and Community based Adaptation Technologies*. 2006.
- [159] George A Miller. “WordNet: a lexical database for English”. In: *Communications of the ACM* 38.11 (1995), pp. 39–41.
- [160] George A Miller and Walter G Charles. “Contextual correlates of semantic similarity”. In: *Language and Cognitive Processes* 6.1 (1991), pp. 1–28.
- [161] David Milne and Ian H. Witten. “Learning to Link with Wikipedia”. In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM ’08. New York, NY, USA: ACM, 2008, pp. 509–518.
- [162] Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. “Predicting human brain activity associated with the meanings of nouns”. In: *science* 320.5880 (2008), pp. 1191–1195.
- [163] Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. “Counter-fitting Word Vectors to Linguistic Constraints”. In: *HLT-NAACL*. 2016.
- [164] Roberto Navigli and Mirella Lapata. “An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 32.4 (2010), pp. 678–692.
- [165] Roberto Navigli and Paola Velardi. “An analysis of ontology-based query expansion strategies”. In: *Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia*. 2003, pp. 42–49.
- [166] Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. “The University of South Florida free association, rhyme, and word fragment norms”. In: *Behavior Research Methods, Instruments, & Computers* 36.3 (2004), pp. 402–407.
- [167] Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. “Integrating Distributional Lexical Contrast into Word Embeddings for Antonym–Synonym Distinction”. In: *The 54th Annual Meeting of the Association for Computational Linguistics*. 2016, p. 454.
- [168] Maximilian Nickel and Douwe Kiela. *Poincaré Embeddings for Learning Hierarchical Representations*. cite arxiv:1705.08039. 2017.

Bibliography

- [169] Thomas Niebler, Martin Becker, Christian Pölit, and Andreas Hotho. *Learning Semantic Relatedness From Human Feedback Using Metric Learning*. cite arxiv: 1705.07425. 2017.
- [170] Thomas Niebler, Martin Becker, Christian Pölit, and Andreas Hotho. “Learning Semantic Relatedness from Human Feedback Using Relative Relatedness Learning”. In: *Proceedings of the ISWC 2017*. 2017.
- [171] Thomas Niebler, Martin Becker, Daniel Zoller, Stephan Doerfel, and Andreas Hotho. *Evaluating Emergent Semantics in Folksonomies on Human Intuition*. Tech. rep. 2015.
- [172] Thomas Niebler, Martin Becker, Daniel Zoller, Stephan Doerfel, and Andreas Hotho. *Extracting Semantic Relatedness from Navigation in a Social Tagging System*. Tech. rep. 2015.
- [173] Thomas Niebler, Martin Becker, Daniel Zoller, Stephan Doerfel, and Andreas Hotho. “FolkTrails: Interpreting Navigation Behavior in a Social Tagging System”. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. CIKM ’16. ACM, 2016.
- [174] Thomas Niebler, Luzian Hahn, and Andreas Hotho. “Learning Word Embeddings from Tagging Data: A methodological comparison”. In: *Proceedings of the LWDA*. 2017.
- [175] Thomas Niebler, Daniel Schlör, Martin Becker, and Andreas Hotho. “Extracting Semantics from Unconstrained Navigation on Wikipedia”. In: *KI – Künstliche Intelligenz* 30.2 (2016), pp. 163–168.
- [176] Thomas Niebler, Philipp Singer, Dominik Benz, Christian Körner, Markus Strohmaier, and Andreas Hotho. “How Tagging Pragmatics Influence Tag Sense Discovery in Social Annotation Systems”. In: *ECIR*. Springer, 2013, pp. 86–97.
- [177] Masataka Ono, Makoto Miwa, and Yutaka Sasaki. “Word Embedding-based Antonym Detection using Thesauri and Distributional Information.” In: *HLT-NAACL*. 2015, pp. 984–989.
- [178] Bo Pang, Lillian Lee, et al. “Opinion mining and sentiment analysis”. In: *Foundations and Trends® in Information Retrieval* 2.1–2 (2008), pp. 1–135.
- [179] Patrick Pantel and Dekang Lin. “Discovering word senses from text”. In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Canada, 2002, pp. 613–619.
- [180] Heiko Paulheim. “Knowledge graph refinement: A survey of approaches and evaluation methods”. In: *Semantic web* 8.3 (2017), pp. 489–508.
- [181] Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevich, Benjamin Van Durme, and Chris Callison-Burch. “PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification”. In: 2015.

- [182] Karl Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.
- [183] Karl Pearson. "Note on regression and inheritance in the case of two parents". In: *Proceedings of the Royal Society of London* 58 (1895), pp. 240–242.
- [184] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global Vectors for Word Representation." In: *EMNLP*. Vol. 14. 2014, pp. 1532–1543.
- [185] Bryan Perozzi, Rami Al-Rfou', and Steven Skiena. "DeepWalk: online learning of social representations." In: *KDD*. ACM, 2014, pp. 701–710.
- [186] Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. "Mimicking word embeddings using subword RNNs". In: *arXiv preprint arXiv:1707.06961* (2017).
- [187] Peter Pirolli and Stuart Card. "Information foraging." In: *Psychological review* 106.4 (1999), p. 643.
- [188] Priya Radhakrishnan and Vasudeva Varma. "Extracting semantic knowledge from wikipedia category names". In: *Proceedings of the 2013 workshop on Automated knowledge base construction*. ACM. 2013, pp. 109–114.
- [189] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. "A Word at a Time: Computing Word Relatedness Using Temporal Semantic Analysis". In: *WWW*. New York, NY, USA: ACM, 2011, pp. 337–346.
- [190] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. "Squad: 100,000+ questions for machine comprehension of text". In: *arXiv preprint arXiv:1606.05250* (2016).
- [191] Daniel Ramage, Anna N Rafferty, and Christopher D Manning. "Random walks for text semantic similarity". In: *Proceedings of the 2009 workshop on graph-based methods for natural language processing*. Association for Computational Linguistics. 2009, pp. 23–31.
- [192] Reinhard Rapp. "Word sense discovery based on sense descriptor dissimilarity". In: *Proceedings of the Ninth Machine Translation Summit*. 2003, pp. 315–322.
- [193] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. "Local and Global Algorithms for Disambiguation to Wikipedia". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 1375–1384.
- [194] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 2010, pp. 45–50.
- [195] Philip Resnik. *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*. cite arxiv:cmp-lg/9511007. 1995.

Bibliography

- [196] Petar Ristoski, Stefano Faralli, Simone Paolo Ponzetto, and Heiko Paulheim. “Large-scale Taxonomy Induction Using Entity and Word Embeddings”. In: *Proceedings of the International Conference on Web Intelligence*. WI ’17. New York, NY, USA: ACM, 2017, pp. 81–87.
- [197] Carl R. Rogers. *A way of being*. Boston: Houghton Mifflin Co., 1995.
- [198] Stephen Roller, Douwe Kiela, and Maximilian Nickel. “Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora”. In: *ACL (2)*. Association for Computational Linguistics, 2018, pp. 358–363.
- [199] Herbert Rubenstein and John B. Goodenough. “Contextual correlates of synonymy.” In: *Commun. ACM* 8.10 (1965), pp. 627–633.
- [200] Sebastian Ruder, Ivan Vulic, and Anders Sogaard. *A Survey Of Cross-lingual Word Embedding Models*. cite arxiv:1706.04902. 2017.
- [201] G. Salton, A. Wong, and C. S. Yang. “A Vector Space Model for Automatic Indexing”. In: *Commun. ACM* 18.11 (1975), pp. 613–620.
- [202] Gerald Salton. *The SMART Retrieval System Experiments in Automatic Document Processing*. Prentice Hall, 1971.
- [203] Aju Thalappillil Scaria, Rose Marie Philip, Robert West, and Jure Leskovec. “The last click: why users give up information network navigation.” In: *WSDM*. ACM, 2014, pp. 213–222.
- [204] Tobias Schnabel, Igor Labutov, David M Mimno, and Thorsten Joachims. “Evaluation methods for unsupervised word embeddings.” In: *EMNLP*. 2015, pp. 298–307.
- [205] H. Schütze and J.O. Pedersen. “A cooccurrence-based thesaurus and two applications to information retrieval”. In: *IPM* 33.3 (1997), pp. 307–318.
- [206] Xiance Si and Maosong Sun. “Disambiguating Tags in Blogs”. In: *TSD*. Vol. 5729. Lecture Notes in Computer Science. Springer, 2009, pp. 139–146.
- [207] Philipp Singer, Denis Helic, Andreas Hotho, and Markus Strohmaier. “HypTrails: A Bayesian Approach for Comparing Hypotheses About Human Trails on the Web”. In: *Proceedings of the 24th International Conference on World Wide Web*. WWW ’15. New York, NY, USA: ACM, 2015, pp. 1003–1013.
- [208] Philipp Singer, Denis Helic, Behnam Taraghi, and Markus Strohmaier. “Detecting Memory and Structure in Human Navigation Patterns Using Markov Chain Models of Varying Order”. In: *PLoS ONE* 9.7 (July 2014), e102070.
- [209] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. “Why We Read Wikipedia”. In: *Proceedings of the 26th International Conference on World Wide Web*. WWW ’17. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017, pp. 1591–1600.

- [210] Philipp Singer, Thomas Niebler, Markus Strohmaier, and Andreas Hotho. "Computing Semantic Relatedness from Human Navigational Paths: A Case Study on Wikipedia". In: *IJSWIS* 9.4 (2013), pp. 41–70.
- [211] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. "Content-based image retrieval at the end of the early years". In: *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* 22.12 (2000), pp. 1349–1380.
- [212] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. "Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 254–263.
- [213] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. "Recursive deep models for semantic compositionality over a sentiment treebank". In: *Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013, pp. 1631–1642.
- [214] C. Spearman. "The Proof and Measurement of Association between Two Things". In: *The American Journal of Psychology* 15.1 (1904), pp. 72–101.
- [215] Lucia Specia and Enrico Motta. "Integrating folksonomies with the semantic web". In: *The semantic web: research and applications* (2007), pp. 624–639.
- [216] Robert Speer and Joshua Chin. *An Ensemble Method to Produce High-Quality Word Embeddings*. cite arxiv:1604.01692. 2016.
- [217] Robert Speer, Joshua Chin, and Catherine Havasi. "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge". In: *AAAI*. 2017, pp. 4444–4451.
- [218] Rohini K. Srihari, Zhongfei Zhang, and Aibing Rao. "Intelligent Indexing and Semantic Retrieval of Multimodal Documents". In: *Inf. Retr.* 2.2-3 (2000), pp. 245–275.
- [219] Darko Stanisavljevic, Ilire Hasani-Mavriqi, Elisabeth Lex, Markus Strohmaier, and Denis Helic. "Semantic Stability in Wikipedia." In: *COMPLEX NETWORKS*. Vol. 693. Studies in Computational Intelligence. Springer, 2016, pp. 379–390.
- [220] Andy Stirling. *A General Framework for Analysing Diversity in Science, Technology and Society*. SPRU Electronic Working Paper Series 156. University of Sussex, SPRU - Science and Technology Policy Research, 2007.
- [221] M. Strohmaier, C. Körner, and R. Kern. "Why do users tag? Detecting users' motivation for tagging in social tagging systems". In: *International AAAI Conference on Weblogs and Social Media (ICWSM2010)*, Washington, DC, USA, May 23-26. 2010.
- [222] M Strohmaier, C Körner, and R Kern. "Understanding why users tag: A survey of tagging motivation literature and results from an empirical study". In: *Web Semant* 17.C (2012), pp. 1–11.

Bibliography

- [223] M. Strube and S.P. Ponzetto. “WikiRelate! Computing semantic relatedness using Wikipedia”. In: *Proceedings of the National Conference on Artificial Intelligence*. Vol. 21. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999. 2006, p. 1419.
- [224] Rudi Studer, V Richard Benjamins, Dieter Fensel, et al. “Knowledge engineering: principles and methods”. In: *Data and knowledge engineering* 25.1 (1998), pp. 161–198.
- [225] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. “YAGO: A Core of Semantic Knowledge”. In: WWW. New York, NY, USA: ACM Press, 2007.
- [226] Deborah Talmi and Morris Moscovitch. “Can semantic relatedness explain the enhancement of memory for emotional words?” In: *Mem Cognit* 32.5 (2004), pp. 742–51.
- [227] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. “LINE: Large-scale Information Network Embedding”. In: WWW. New York, NY, USA: ACM, 2015, pp. 1067–1077.
- [228] C. Trattner, D. Helic, P. Singer, and M. Strohmaier. “Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks”. In: *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*. ACM. 2012, p. 14.
- [229] Peter D Turney. “Domain and function: A dual-space model of semantic relations and compositions”. In: *Journal of Artificial Intelligence Research* 44 (2012), pp. 533–585.
- [230] Peter D. Turney and Patrick Pantel. “From Frequency to Meaning: Vector Space Models of Semantics”. In: *J. Artif. Int. Res.* 37.1 (2010), pp. 141–188.
- [231] Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. “Cross-lingual models of word embeddings: An empirical comparison”. In: *arXiv preprint arXiv:1604.00425* (2016).
- [232] David Vallet, Iván Cantador, and Joemon Jose. “Personalizing web search with folksonomy-based user and document profiles”. In: *Advances in information retrieval* (2010), pp. 420–431.
- [233] Olga Vechtomova and Ying Wang. “A study of the effect of term proximity on query expansion”. In: *Journal of Information Science* 32.4 (2006), pp. 324–333.
- [234] Ellen M Voorhees. “Query expansion using lexical-semantic relations”. In: *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag New York, Inc. 1994, pp. 61–69.
- [235] Denny Vrandečić and Markus Krötzsch. “Wikidata: A Free Collaborative Knowledgebase”. In: *Commun. ACM* 57.10 (2014), pp. 78–85.

- [236] Ivan Vulic and Marie-Francine Moens. “Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*. ACL. 2015, pp. 719–725.
- [237] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. “It’s a Man’s Wikipedia? Assessing Gender Inequality in an Online Encyclopedia”. In: *The International AAAI Conference on Web and Social Media (ICWSM2015), Oxford, May 2015*. 2015.
- [238] Claudia Wagner, Philipp Singer, Markus Strohmaier, and Bernardo A Huberman. “Semantic stability in social tagging streams”. In: *Proceedings of the 23rd international conference on World wide web*. ACM. 2014, pp. 735–746.
- [239] Claudia Wagner and Markus Strohmaier. “The Wisdom in Tweetonomies: Acquiring Latent Conceptual Structures from Social Awareness Streams”. In: *Proceedings of the 3rd International Semantic Search Workshop. SEMSEARCH ’10*. New York, NY, USA: ACM, 2010, 6:1–6:10.
- [240] Xuanhui Wang and ChengXiang Zhai. “Mining term association patterns from search logs for effective query reformulation”. In: *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM. 2008, pp. 479–488.
- [241] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. “Knowledge Graph Embedding by Translating on Hyperplanes.” In: *AAAI*. 2014, pp. 1112–1119.
- [242] Joe H. Ward. “Hierarchical Grouping to Optimize an Objective Function”. In: *Journal of the American Statistical Association* 58.301 (1963), pp. 236–244.
- [243] Kilian Q. Weinberger and Lawrence K. Saul. “Distance Metric Learning for Large Margin Nearest Neighbor Classification”. In: *JMLR* (2009), pp. 207–244.
- [244] Joseph Weizenbaum. “ELIZA a Computer Program for the Study of Natural Language Communication Between Man and Machine”. In: *Commun. ACM* 9.1 (1966), pp. 36–45.
- [245] Robert West. “HUMAN NAVIGATION OF INFORMATION NETWORKS”. PhD thesis. STANFORD UNIVERSITY, 2016.
- [246] Robert West and Jure Leskovec. “Automatic Versus Human Navigation in Information Networks.” In: *ICWSM. The AAAI Press*, 2012.
- [247] Robert West and Jure Leskovec. “Human Wayfinding in Information Networks”. In: *Proceedings of the 21st International Conference on World Wide Web. WWW ’12*. New York, NY, USA: ACM, 2012, pp. 619–628.
- [248] Robert West, Ashwin Paranjape, and Jure Leskovec. “Mining Missing Hyperlinks from Human Navigation Traces: A Case Study of Wikipedia”. In: *Proceedings of the 24th International Conference on World Wide Web. WWW ’15*. New York, NY, USA: ACM, 2015, pp. 1242–1252.

Bibliography

- [249] Robert West, Joelle Pineau, and Doina Precup. “Wikispeedia: an online game for inferring semantic distances between concepts”. In: *IJCAI*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2009, pp. 1598–1603.
- [250] John Wieting, Mohit Bansal, Kevin Gimpel, Karen Livescu, and Dan Roth. “From Paraphrase Database to Compositional Paraphrase Model and Back”. In: *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 345–358.
- [251] Ian Witten and David Milne. “An effective, low-cost measure of semantic relatedness obtained from Wikipedia links”. In: *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, AAAI Press, Chicago, USA. 2008, pp. 25–30.
- [252] Xian Wu, Lei Zhang, and Yong Yu. “Exploring Social Annotations for the Semantic Web”. In: *Proceedings of the 15th International Conference on World Wide Web*. WWW ’06. New York, NY, USA: ACM, 2006, pp. 417–426.
- [253] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. “Google’s neural machine translation system: Bridging the gap between human and machine translation”. In: *arXiv preprint arXiv:1609.08144* (2016).
- [254] Ellery Wulczyn. *Wikipedia Navigation Vectors*. May 2016.
- [255] Ellery Wulczyn and Dario Taraborelli. *Wikipedia Clickstream*. Feb. 2015.
- [256] Han Xiao, Minlie Huang, and Xiaoyan Zhu. “TransG: A generative model for knowledge graph embedding”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2016, pp. 2316–2325.
- [257] Eric P Xing, Andrew Y Ng, Michael I Jordan, and Stuart Russell. “Distance metric learning with application to clustering with side-information”. In: *NIPS* (2003), pp. 521–528.
- [258] Majid Yazdani and Andrei Popescu-Belis. “Computing text semantic relatedness using the contents and links of a hypertext encyclopedia”. In: *Artif. Intell.* 194 (2013), pp. 176–202.
- [259] Eric Yeh, Daniel Ramage, Christopher D. Manning, Eneko Agirre, and Aitor Soroa. “WikiWalk: random walks on Wikipedia for semantic relatedness”. In: *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*. TextGraphs-4. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 41–49.
- [260] Mo Yu and Mark Dredze. “Improving Lexical Embeddings with Semantic Knowledge.” In: *ACL* (2). 2014, pp. 545–550.
- [261] Torsten Zesch and Iryna Gurevych. “The More the Better? Assessing the Influence of Wikipedia’s Growth on Semantic Relatedness Measures.” In: *LREC*. 2010.
- [262] Lei Zhang, Xian Wu, and Yong Yu. “Emergent Semantics from Folksonomies: A Quantitative Study”. In: *Journal on Data Semantics VI* (2006).

- [263] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. “Adversarial training for unsupervised bilingual lexicon induction”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2017, pp. 1959–1970.
- [264] Ye Zhang and Byron Wallace. “A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification”. In: *arXiv preprint arXiv:1510.03820* (2015).
- [265] Yu Zhao, Zhiyuan Liu, and Maosong Sun. “Representation Learning for Measuring Entity Relatedness with Rich Information”. In: *Proceedings of the 24th International Conference on Artificial Intelligence. IJCAI’15*. AAAI Press, 2015, pp. 1412–1418.
- [266] Chen Zheng, Zhichun Wang, Rongfang Bie, and Mingquan Zhou. “Learning to Compute Semantic Relatedness Using Knowledge from Wikipedia”. In: *Web Technologies and Applications: 16th Asia-Pacific Web Conference, APWeb 2014, Changsha, China, September 5-7, 2014. Proceedings*. Vol. 8709. Springer. 2014, p. 236.
- [267] Mianwei Zhou, Shenghua Bao, Xian Wu, and Yong Yu. “An Unsupervised Model for Exploring Hierarchical Semantics from Social Annotations”. In: 2008, pp. 680–693.
- [268] Daniel Zoller, Stephan Doerfel, Christian Pölit, and Andreas Hotho. “Leveraging User-Interactions for Time-Aware Tag Recommendations”. In: *Proceedings of the Workshop on Temporal Reasoning in Recommender Systems*. CEUR Workshop Proceedings. 2017.
- [269] Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. “Bilingual word embeddings for phrase-based machine translation”. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, pp. 1393–1398.
- [270] Arkaitz Zubiaga, Victor Fresno, Raquel Martinez, and Alberto P. Garcia-Plaza. “Harnessing Folksonomies to Produce a Social Classification of Resources”. In: *IEEE Trans. on Knowl. and Data Eng.* 25.8 (2013), pp. 1801–1813.
- [271] Arkaitz Zubiaga, Christian Körner, and Markus Strohmaier. “Tags vs Shelves: From Social Tagging to Social Classification”. In: *Proceedings of the 22Nd ACM Conference on Hypertext and Hypermedia*. HT ’11. New York, NY, USA: ACM, 2011, pp. 93–102.