

Dominik Benz

Capturing Emergent Semantics from Social Annotation Systems

Dissertation zur Erlangung des akademischen Grades eines Doktors der
Naturwissenschaften (Dr. rer. nat.)

im Fachgebiet Wissensverarbeitung,
Fachbereich 16 Elektrotechnik/Informatik der Universität Kassel

This work has been accepted by the faculty of electrical engineering / computer science of the University of Kassel as a thesis for acquiring the academic degree of Doktor der Naturwissenschaften (Dr. rer. nat.).

Supervisor: Prof. Dr. Gerd Stumme
Co-Supervisor: Prof. Dr. Philipp Cimiano

Defense Day: 2nd March 2012

Abstract

The ongoing growth of the World Wide Web, catalyzed by the increasing possibility of ubiquitous access via a variety of devices, continues to strengthen its role as our prevalent information and communication medium. However, although tools like search engines facilitate retrieval, the task of finally *making sense* of Web content is still often left to human interpretation. The vision of supporting both humans and machines in such knowledge-based activities led to the development of different systems which allow to structure Web resources by metadata annotations. Interestingly, two major approaches which gained a considerable amount of attention are addressing the problem from nearly opposite directions: On the one hand, the idea of the *Semantic Web* suggests to formalize the knowledge within a particular domain by means of the “top-down” approach of defining ontologies. On the other hand, Social Annotation Systems as part of the so-called Web 2.0 movement implement a “bottom-up” style of categorization using arbitrary keywords.

Experience as well as research in the characteristics of both systems has shown that their strengths and weaknesses seem to be inverse: While Social Annotation suffers from problems like, e. g., ambiguity or lack of precision, ontologies were especially designed to eliminate those. On the contrary, the latter suffer from a knowledge acquisition bottleneck, which is successfully overcome by the large user populations of Social Annotation Systems. Instead of being regarded as competing paradigms, the obvious potential synergies from a combination of both motivated approaches to “bridge the gap” between them. These were fostered by the evidence of *emergent semantics*, i. e., the self-organized evolution of implicit conceptual structures, within Social Annotation data. While several techniques to exploit the emergent patterns were proposed, a systematic analysis – especially regarding paradigms from the field of ontology learning – is still largely missing. This also includes a deeper understanding of the circumstances which affect the evolution processes.

This work aims to address this gap by providing an in-depth study of methods and influencing factors to capture emergent semantics from Social Annotation Systems. We focus hereby on the acquisition of lexical semantics from the

underlying networks of keywords, users and resources. Structured along different ontology learning tasks, we use a methodology of *semantic grounding* to characterize and evaluate the semantic relations captured by different methods. In all cases, our studies are based on datasets from several Social Annotation Systems.

Specifically, we first analyze semantic relatedness among keywords, and identify measures which detect different notions of relatedness. These constitute the input of concept learning algorithms, which focus then on the discovery of synonymous and ambiguous keywords. Hereby, we assess the usefulness of various clustering techniques. As a prerequisite to induce hierarchical relationships, our next step is to study measures which quantify the level of generality of a particular keyword. We find that comparatively simple measures can approximate the generality information encoded in reference taxonomies. These insights are used to inform the final task, namely the creation of concept hierarchies. For this purpose, generality-based algorithms exhibit advantages compared to clustering approaches.

In order to complement the identification of suitable methods to capture semantic structures, we analyze as a next step several factors which influence their emergence. Empirical evidence is provided that the amount of available data plays a crucial role for determining keyword meanings. From a different perspective, we examine pragmatic aspects by considering different annotation patterns among users. Based on a broad distinction between “categorizers” and “describers”, we find that the latter produce more accurate results. This suggests a causal link between pragmatic and semantic aspects of keyword annotation. As a special kind of usage pattern, we then have a look at system abuse and spam. While observing a mixed picture, we suggest that an individual decision should be taken instead of disregarding spammers as a matter of principle.

Finally, we discuss a set of applications which operationalize the results of our studies for enhancing both Social Annotation and semantic systems. These comprise on the one hand tools which foster the emergence of semantics, and on the one hand applications which exploit the socially induced relations to improve, e. g., searching, browsing, or user profiling facilities. In summary, the contributions of this work highlight viable methods and crucial aspects for designing enhanced knowledge-based services of a Social Semantic Web.

Zusammenfassung

Der anhaltende Zuwachs an Inhalten sowie die steigende Verfügbarkeit über verschiedenste Endgeräte festigen die Rolle des World Wide Web als ein zentrales Kommunikations- und Informationsmedium. Trotz der Unterstützung durch beispielsweise Suchmaschinen bleibt dabei die Zuordnung von *Bedeutung* zu Informationsressourcen immer noch weitgehend dem Benutzer überlassen. Die Hoffnung, Menschen und Computer bei diesen wissensbasierten Aktivitäten zu unterstützen, führte zur Entwicklung verschiedener Systeme, die das Strukturieren von Webinhalten über die Annotation mit Metadaten ermöglichen. Interessanterweise finden sich in zwei bekannten Ansätzen solcher Systeme zwei beinahe entgegengesetzte Herangehensweisen an dieses Problem wieder: Einerseits basiert die Vision des *Semantic Web* darauf, Wissensgebiete mittels nach und nach verfeinerter Konzepte innerhalb einer Ontologie zu formalisieren (“top-down”). Auf der anderen Seite zeichnen sich *soziale Verschlagwortungssysteme*, die im Zuge des Web 2.0 Bekanntheit erlangten, durch einen aufbauenden (“bottom-up”) und unkontrollierten Ansatz der Erstellung eines kollaborativen Vokabulars aus.

Die Erfahrungen im Umgang mit beiden Systemen sowie Forschungen in diesem Bereich zeigten auf, dass die individuellen Stärken und Schwächen beider Herangehensweisen in einem inversen Verhältnis zueinander zu stehen scheinen: Unkontrollierte freie Verschlagwortung bringt Probleme wie zum Beispiel Unschärfe oder Mehrdeutigkeiten mit sich, die in formalen Ontologien nicht existieren. Auf der anderen Seite haben letztere das Flaschenhalsproblem der Wissensakquisition, was wiederum für soziale Verschlagwortung aufgrund der hohen Benutzerbeteiligung eine untergeordnete Rolle spielt. Anstatt beide Klassen von Systemen als konkurrierend zu betrachten, wurden viele Methoden im Bereich eines “Brückenschlages” zwischen beiden Ansätzen von den dabei möglichen Synergien inspiriert. Dies wurde noch verstärkt durch Anzeichen von “entstehender Semantik”, die sich in der selbständigen Herausbildung von impliziten konzeptuellen Strukturen innerhalb der Verschlagwortungssysteme zeigte. Obwohl hierbei viele verschiedene Techniken eingesetzt wurden, um diese nutzbar zu machen, wurde bisher keine systematische Analyse – besonders im

Bereich von Ansätzen aus dem Ontologielernen – durchgeführt. Ebenfalls kaum erforscht sind die Umstände, die zur Entstehung der semantischen Strukturen führen.

Das Ziel der vorliegenden Arbeit ist es, diese Lücke zu schliessen und eine detaillierte Untersuchung von Methoden zur Erfassung von entstehender Semantik sowie deren beeinflussende Faktoren innerhalb sozialer Verschlagwortungssysteme durchzuführen. Das Hauptaugenmerk liegt hierbei auf der Akquisition von lexikalischer Semantik, basierend auf der Netzwerkstruktur zwischen Schlagworten, Benutzern und Ressourcen. Innerhalb verschiedener Teilaufgaben aus dem Bereich des Ontologielernens wird dazu eine Methodologie der “semantischen Erdung” eingesetzt, die es erlaubt, von verschiedenen Methoden erzeugte semantische Strukturen zu bewerten und zu charakterisieren. Die zugehörigen Experimente werden jeweils auf Datensätzen aus verschiedenen Verschlagwortungssystemen durchgeführt.

In einem ersten Schritt wird hierzu die semantische Verwandtschaft zwischen Schlagwörtern analysiert. Insbesondere werden Maße identifiziert, die verschiedene Arten von Verwandtschaftsbeziehungen anzeigen. Diese dienen in einem nächsten Schritt als Eingabe für Algorithmen des Konzeptlernens, die synonyme und mehrdeutige Schlagwörter entdecken. Zu diesem Zweck werden auch verschiedene Clustering-Verfahren angewendet und bewertet. Als Vorarbeit zur Extraktion hierarchischer Beziehungen werden dann Maße analysiert, die den Grad der “Allgemeinheit” eines bestimmten Schlagwortes quantifizieren. Ein Ergebnis in diesem Bereich ist, dass vergleichsweise einfache Maße bereits eine gute Approximation der Art von Allgemeinheit erfassen, die auch in Taxonomien enthalten ist. Diese Ergebnisse bilden die Grundlage für den letzten Schritt, nämlich die Erstellung von Konzept-Hierarchien. Hierbei zeigt sich, dass spezielle Verfahren auf Basis der Allgemeinheit von Schlagwörtern Vorteile gegenüber allgemeinen Clustering-Techniken besitzen.

Als Gegenpart zur Bestimmung geeigneter Methoden, um die entstehenden semantischen Strukturen zu erfassen, werden anschliessend Faktoren untersucht, die zu deren Entstehung beitragen. Hierbei wird empirisch gezeigt, dass die Menge an vorhandenen Daten eine zentrale Rolle für das Erfassen der Bedeutung eines Schlagworts spielt. Aus einer anderen Perspektive werden anschliessend pragmatische Aspekte der Annotation untersucht, die sich in verschiedenen Verschlagwortungsmustern innerhalb der Benutzerschaft widerspiegeln. Basierend auf einer groben Einteilung in “Kategorisierer” und “Beschreiber” zeigt sich, dass letztere Gruppe zu präziseren Ergebnissen führt. Dies weist auf eine

kausalen Zusammenhang zwischen pragmatischen und semantischen Aspekten bei der kollaborativen Verschlagwortung hin. Als letztes wird schliesslich der Einfluss einer besonderen und unerwünschten Benutzergruppe untersucht, die Verschlagwortungssysteme für missbräuchliche Zwecke einsetzen. Dies zeigt sich beispielsweise in massenhafter Annotation von werberelevanten oder anstössigen Inhalten durch sogenannte “Spammer”. Da hierbei teilweise widersprüchliche Ergebnisse auftreten, scheint eine individuelle Betrachtung geeigneter als ein kategorischer Ausschluss von Spammern.

Abschliessend werden eine Reihe von Anwendungen vorgestellt, die die Ergebnisse der vorherigen Studien zur Verbesserung von Verschlagwortungssystemen und semantischen Plattformen nutzbar machen. Diese beinhalten einerseits Implementierungen, die das Entstehen von Semantik fördern. Andererseits werden Applikationen beschrieben, die direkten Gebrauch von den gelernten Strukturen machen, um beispielsweise verbesserte Such-, Navigations- oder Personalisierungsmöglichkeiten anzubieten. Zusammenfassend besteht der Beitrag dieser Arbeit darin, gangbare Methoden und zentrale Aspekte für den Entwurf von verbesserten wissensbasierten Anwendungen eines Social Semantic Web aufzuzeigen.

Contents

1. Introduction	1
1.1. Motivation	1
1.2. Research Questions	3
1.3. Outline	6
1.4. Overview of Author's Contributions	8
1. Fundamentals: Social Annotations and the Semantic Web	13
2. Knowledge on the Web	15
2.1. Knowledge Organization	16
2.2. Knowledge Engineering	18
3. Social Annotations	21
3.1. Social Tagging Systems	23
3.1.1. Strengths and Weaknesses	25
3.1.2. Formal Model	27
3.1.3. Induced Networks	28
3.1.4. Tagging System Characteristics	32
3.1.5. Calculating Relevancy	33
3.1.6. Tagging Pragmatics	34
3.1.7. Related Work	35
3.2. Other Forms	36
3.2.1. Weblogs and Microblogging	37
3.2.2. Wikis	38
3.2.3. Question Answering	39
3.2.4. Implicit Annotation within Logsonomies	40
3.3. Summary	42
4. The Semantic Web	43

4.1. Ontologies	49
4.1.1. Formal model	50
4.1.2. Classifying Ontologies	55
4.1.3. Taxonomies and Thesauri	57
4.1.4. Strengths and weaknesses	59
4.2. Derived Measures	61
4.2.1. Semantic Relatedness	62
4.2.2. Semantic Generality	65
4.3. Summary	67
5. From Social Annotations to the Semantic Web	69
5.1. General Aspects	70
5.1.1. Bottom-up vs. Top-down	71
5.1.2. Categorization vs. Classification	74
5.1.3. Comparison of Strengths and Weaknesses	75
5.2. Ontology Learning to capture Emergent Semantics	77
5.2.1. Emergent Semantics	77
5.2.2. Ontology Learning Tasks	78
5.2.3. Comparison to Ontology Learning from other input	79
5.3. State of the Art	83
5.3.1. Comparison dimensions	83
5.3.2. Capturing Semantic Relatedness	86
5.3.3. Learning Concepts	86
5.3.4. Capturing Semantic Generality	89
5.3.5. Learning Concept Hierarchies	89
5.3.6. Learning Attributes, Relations and Axioms	90
5.4. Evaluation Paradigms	91
5.4.1. Lexical Layer	93
5.4.2. Structural Layer	94
5.5. Approach of this dissertation	97
5.6. Summary	99
II. Capturing Emergent Semantics: Data, Methods and Influencing Factors	101
6. Data	103

6.1. Systems and Datasets	104
6.1.1. BibSonomy	104
6.1.2. CiteULike	105
6.1.3. Delicious	106
6.1.4. Flickr	108
6.1.5. AOL Logsonomy	109
6.1.6. Stackoverflow	110
6.2. Gold-standard Ontologies	111
6.2.1. WordNet	112
6.2.2. YAGO	113
6.2.3. DMOZ	113
6.2.4. Wikipedia Category Hierarchy	114
7. Methods	115
7.1. Capturing Semantic Relatedness	116
7.1.1. Relatedness Measures	118
7.1.2. Qualitative Evaluation	121
7.1.3. Evaluation by Semantic Grounding	126
7.1.4. Alternative Aggregation, Weighting and Similarity Ap- proaches	137
7.1.5. Summary	149
7.2. Learning Concepts	150
7.2.1. Synonym Resolution	151
7.2.2. Evaluation by Semantic Grounding	156
7.2.3. Tag Sense Disambiguation	161
7.2.4. Evaluation by Semantic Grounding	165
7.2.5. Summary	170
7.3. Capturing Semantic Generality	172
7.3.1. Generality Measures	173
7.3.2. Evaluation by Semantic Grounding	176
7.3.3. Summary	185
7.4. Learning Concept Hierarchies	185
7.4.1. Clustering Approaches	187
7.4.2. Generality-based Methods	190
7.4.3. Gold-standard based Evaluation	193
7.4.4. Evaluation by Human Assessment	195
7.4.5. Enhancement by Synonym Resolution and Disambiguation	198

7.4.6. Summary	201
7.5. Learning Attributes, Relations and Axioms	202
8. Influencing Factors	203
8.1. Methodology of Influence Assessment	204
8.2. Keyword Properties	205
8.2.1. Frequency	206
8.2.2. Interface characteristics	209
8.2.3. Summary	210
8.3. Tagging Pragmatics	211
8.3.1. Measures of Tagging Pragmatics	212
8.3.2. Influence Assessment	214
8.3.3. Implications	224
8.3.4. Summary	228
8.4. System Abuse and Spam	229
8.4.1. Spam Definition and Detection	229
8.4.2. Influence Assessment	232
8.4.3. Summary	234
III. Applications and Conclusions	237
9. Applications	239
9.1. Enhancing Social Annotation Systems	239
9.1.1. Fostering the Emergence of Semantics	240
9.1.2. Feeding back Semantics	246
9.2. Enhancing Semantic Applications	248
9.3. Summary	250
10. Conclusions	251
10.1. Summary	251
10.2. Contributions and Outlook	254
10.3. Closing Remarks	256
Bibliography	259

List of Figures

1.1. Graphical outline of the narrative of this dissertation.	7
2.1. Top level classes of the Dewey Decimal System (DDC).	17
3.1. Screenshot of the Social Tagging System BibSonomy.	24
3.2. Visualization of the folksonomy data structure.	29
3.3. Overview of folksonomy-derived 1-mode and 2-mode networks.	30
4.1. Structure of the Syntactic Web graph.	44
4.2. Structure of the Semantic Web graph.	45
4.3. The layered architecture of the Semantic Web.	46
4.4. Graphical illustration of an example ontology.	53
4.5. Classification of Ontology languages according to their semantic spectrum.	56
4.6. Excerpt of the Linnaean taxonomy of biological species.	57
4.7. Taxonomy excerpt to exemplify semantic generality of concepts.	64
5.1. Comparison of top-down and bottom-up approaches of annotation.	71
5.2. Ontology learning layer cake.	79
5.3. Visualization of taxonomic similarity measures.	95
7.1. Layer cake model of capturing emergent semantics from Social Annotation data.	116
7.2. Tag co-occurrence fingerprint of five selected tags in the first 30 dimensions of the tag vector space.	123
7.3. Average rank of related tags according to different relatedness measures.	126
7.4. Average semantic distance, measured in WordNet, between a tag and its most closely related one according to different folksonomy-based measures.	129

List of Figures

7.5. Probability distribution for the length of the shortest path between a tag and its most closely related one according to different folksonomy-based measures.	130
7.6. Edge composition of the shortest paths of length 1 and 2 between a tag and its most closely related one.	131
7.7. Probability distribution of the level displacement Δl in the WordNet hierarchy.	132
7.8. Semantic grounding of tag relatedness measures, generalized to different datasets.	135
7.9. Semantic grounding of different types of co-occurrence.	139
7.10. Schematic overview of micro and macro aggregation.	140
7.11. Semantic grounding of different types of aggregation.	142
7.12. Semantic grounding of different types of weighting schemes (tag context relatedness).	144
7.13. Semantic grounding of different types similarity measures.	148
7.14. Example dendrogram obtained from hierarchical agglomerative clustering.	153
7.15. Performance of different clustering methods for the task of synonym resolution.	158
7.16. Parameter variation for two clustering algorithms for tag sense discovery.	168
7.17. Semantic grounding of abstractness measures (direct evaluation).	178
7.18. Semantic grounding of abstractness measures (derived evaluation).	183
7.19. Results of the reference-based evaluation of concept hierarchy learning.	194
7.20. Results of the user-based evaluation of concept hierarchy learning.	197
7.21. Results of the reference-based evaluation of enhanced concept hierarchy learning.	200
8.1. Methodology of influence assessment concerning factors of emergent semantics.	204
8.2. Influence of tag frequency on emergent semantics.	207
8.3. Influence of restrictions to different fractions of popular keywords (CiteULike dataset).	208
8.4. Influence of tag cleaning on emergent semantics.	210
8.5. Distribution of the values obtained by the measures of tagging pragmatics.	216

8.6.	Results of the influence assessment of tagging pragmatics. . . .	221
8.7.	Implications of tagging pragmatics regarding the minimum amount of data required to reach the semantic precision of the complete dataset.	225
8.8.	Implications of tagging pragmatics regarding the improvement of the global semantic precision.	226
8.9.	Influence assessment of spammers on emergent semantics. . . .	233
9.1.	Screenshot of JabRef, including the BibSonomy plugin.	242
9.2.	Screenshot of the Typo3 plugin configuration.	244
9.3.	Screenshot of a publication list created by the Typo3 plugin. . .	245
9.4.	Screenshot of a semantic browsing facility within BibSonomy. . .	248

List of Tables

3.1. Two Types of Taggers according to (Körner et al., 2010).	35
3.2. Overview of other forms of Social Annotation.	41
4.1. Exemplary results of computing semantic relatedness based on WordNet.	63
5.1. Comparison of Categorization and Classification.	73
5.2. Overview on strenghts and weaknesses of Social Annotation and Ontologies.	76
5.3. Dimensions of comparision between “traditional” input of ontology learning algorithms and input from social tagging systems.	82
5.4. Comparison dimensions and possible values to compare methods of making semantics in folksonomies explicit.	84
6.1. Statistics about the BibSonomy dataset.	104
6.2. Statistics about the CiteULike dataset.	106
6.3. Statistics about the BibSonomy dataset.	107
6.4. Statistics about the BibSonomy dataset.	108
6.5. Statistics about the AOL logsonomy dataset.	109
6.6. Statistics about the Stackoverflow dataset.	110
6.7. Statistical properties of the gold-standard datasets.	112
7.1. Examples of the most related tags for co-occurrence, distributional and folkrank relatedness.	122
7.2. Overlap between the most closely related tags according to different relatedness measures.	125
7.3. Coverage of Delicious tags within WordNet.	128
7.4. Average overlap of the 10 most related tags according to different aggregation schemes with the 10 most frequently co-occurring tags.	143
7.5. Optimal parameters and cluster sizes for synonym detection on different datasets.	160

List of Tables

7.6. Example synonym sets created by hierarchical clustering. . . .	161
7.7. Optimal sense clusters according to different clustering algorithms.	169
7.8. Manual evaluation of tag sense discovery for the Delicious dataset.	170
7.9. Examples of discovered senses of selected keywords.	171
7.10. Statistical properties of the term graphs derived from the Delicious dataset.	176
7.11. Results from the user study on judging semantic generality. . .	181
7.12. Accuracy of the taxonomy-derived abstractness measures. . . .	182
7.13. Statistical properties of the tag-tag-networks used for concept learning.	191
7.14. Statistical properties of induced concept hierarchies.	192
7.15. Lexical overlap among concepts present in the learned and reference taxonomies.	192
8.1. Statistics for the Delicious dataset variants used for the influence assessment of tagging pragmatics.	215
8.2. Statistical properties of selected folksonomy partitions.	222
8.3. Comparison of spammer and non-spammer data within BibSonomy.	232
8.4. Examples of better semantic relations obtained from spam data.	234

Abbreviations, Formatting Conventions, Mathematical Notation

Abbreviations

DDC	Dewey <u>D</u> ecimal <u>S</u> ystem
DMOZ	<u>D</u> irectory <u>M</u> ozilla
FCA	<u>F</u> ormal <u>C</u> oncept <u>A</u> nalysis
HTML	<u>H</u> ypertext <u>M</u> arkup <u>L</u> anguage
IBM	<u>I</u> nternational <u>B</u> usiness <u>M</u> achines
ID	<u>I</u> dentifier
IT	<u>I</u> nformation <u>T</u> echnology
KO	<u>K</u> nowledge <u>O</u> rganization
KOS	<u>K</u> nowledge <u>O</u> rganization <u>S</u> ystem
LSA	<u>L</u> atent <u>S</u> emantic <u>A</u> nalysis
LSI	<u>L</u> atent <u>S</u> emantic <u>I</u> ndexing
NLP	<u>N</u> atural <u>L</u> anguage <u>P</u> rocessing
ODP	<u>O</u> pen <u>D</u> irectory <u>P</u> roject
OL	<u>O</u> ntology <u>L</u> earning
OWL	<u>W</u> eb <u>O</u> ntology <u>L</u> anguage
PLSI	<u>P</u> robabilistic <u>L</u> atent <u>S</u> emantic <u>I</u> ndexing
RDF(S)	<u>R</u> esource <u>D</u> escription <u>F</u> ramework (<u>S</u> chema)
SMM	<u>S</u> eparable <u>M</u> ixture <u>M</u> odel
SNA	<u>S</u> ocial <u>N</u> etwork <u>A</u> nalysis
SPARQL	SPARQL Protocol and RDF Query Language
URI	<u>U</u> niform <u>R</u> esource <u>I</u> dentifier
URL	<u>U</u> niform <u>R</u> esource <u>L</u> ocator
WWW	<u>W</u> orld <u>W</u> ide <u>W</u> eb
XML	<u>E</u> xtensible <u>M</u> arkup <u>L</u> anguage

Formatting Conventions

CONCEPT	Concept identifiers
<i>keyword</i>	Keywords used for Social Annotation
label	Concept labels

Mathematical Notation

$A \cap B$	set intersection
$A \cup B$	set union
$A \setminus B$	set difference
$A \times B$	Cartesian product
$ A $	set cardinality
\emptyset	the empty set
$\pi_i(x_1, \dots, x_n)$.	projection on i^{th} dimension of a tuple
$D(p \parallel q)$	Kullback-Leibler divergence
$f : A \rightarrow B$	function f from values of A to values of B
\mathbb{R}	the set of real numbers
\mathbb{R}^+	the set of positive real numbers (including zero)
$P(A)$	probability of A
$P(A B)$	conditional probability of A given B
$\ v\ _p$	p -norm of vector v
$\log x$	natural logarithm of x
$x \cdot y$	dot product between x and y
$\operatorname{argmax}_x f$	argument x which maximizes function f
$\operatorname{argmin}_x f$	argument x which minimizes function f
$\min f$	minimum value of f
$\max f$	maximum value of f
$\lceil x \rceil$	ceiling of $x \in \mathbb{R}$ (i. e., the smallest integer greater or equal than x)

Chapter 1.

Introduction

1.1. Motivation

The ability to acquire, memorize and process knowledge is an integral part of human intelligence. Throughout the evolutionary and cultural development of mankind, the access to relevant knowledge within a given context has always been an “asset” and a competitive advantage. With the development of computer systems and especially their interconnection by the Internet, the resulting digital ecosystem evolved to an important source of information for large user populations. Especially the invention of the World Wide Web by Tim Berners-Lee around 1990 and its global distribution of immensely growing amounts of digital information resources have fostered this development. However, the massive amounts of data are hereby a mixed blessing: While relevant content is available in abundance, the question how to identify and make it accessible is still a difficult issue.

Because a cognitive strategy of humans to handle large amounts of information is to use abstractions into interrelated *concepts* (Anderson, 2001), an obvious approach is to apply the same principle to the content of the Web by implementing facilities to *structure* and *organize* its information resources. While some existing tools (like, e. g., corporate taxonomies or even personal bookmark folders) work well within individual or organizational boundaries, they do not necessarily scale to the requirements of the dynamic and fast-growing Web environment. Nevertheless, within the history of the Web, several organization systems (often including some form of *annotation* to categorize or classify a particular resource) were proposed and implemented – some of them by authoritative institutions like the W3C¹, others “emerged” as successful applications by innovative individuals, groups or companies.

¹<http://www.w3.org/>

Among the latter, especially two approaches have gained a considerable amount of attention. Interestingly, they are approaching the solution of the problem from nearly opposite directions: On the one hand, the vision of the *Semantic Web* coined by Tim Berners-Lee (Berners-Lee et al., 2001) suggested to formalize the knowledge within a particular domain by means of a “top-down” approach of defining ontologies. Based on an annotation of Web resources using defined classes and relations, Semantic Web agents would be able to perform “intelligent” tasks of information integration and processing. Despite this appealing vision, impeding factors like the knowledge acquisition bottleneck (Cullen and Bryman, 1988) have hindered the mass adoption of such semantic applications on the Web. On the other hand, based on the user-centric paradigms of the so-called *Web 2.0* movement, an alternative way of annotating Web resources in a “bottom-up” manner quickly gained huge popularity. It was implemented within Social Annotation Systems, which allowed their users to mark up different kinds of (possibly shared) resources like, e. g., websites, videos or pictures by using arbitrary keywords. The simplicity and immediate usefulness of these platforms effectively engaged millions of humans in the process of “producing” metadata – an accomplishment that the Semantic Web had never reached before. However, the disadvantages of missing structure and conventions within the uncontrolled vocabularies also quickly became visible, e. g., in retrieval difficulties related to varying or imprecise annotations.

Having observed the obvious *inverse* relation among the strengths and weaknesses of both approaches – i. e., Social Annotation suffer exactly from the problems that ontologies were designed to eliminate, but tackle on the other hand successfully the knowledge acquisition and annotation bottleneck – the idea to combine the best from both worlds by methods to “*bridge the gap*” (Hotho and Hoser, 2007) between the Social and the Semantic Web was picked up by researchers from different communities. Activities in this direction were catalyzed by studies which provided evidence for *emergent semantics* within the user-created Social Annotation vocabularies. The latter became visible, e. g., in a “*nascent consensus*” (Golder and Huberman, 2006) of keyword usage, which might lead to the “crystallization” of concepts without any external control or influence. While several methods to exploit the emergent structures were proposed, a systematic analysis of the applicability of different paradigms especially from the field of ontology learning is still largely missing. This also includes a deeper understanding of the factors which foster the emergence processes.

This dissertation aims to address this gap by providing an in-depth study of concepts, methods and influencing factors of capturing emergent semantics from Social Annotation Systems. It has its roots in the research field of ontology learning, i. e., the (semi-)automatic acquisition of a conceptual domain model from data (Cimiano, 2006; Maedche, 2002). The ultimate goal hereby is to contribute to a synergetic “intermediation” between the social and the semantic paradigm of knowledge organization, intended to pave the way towards augmented knowledge-based services of a next generation Social Semantic Web.

1.2. Research Questions

Because the notion of “capturing emergent semantics from Social Annotation Systems” is rather general, we aim to specify now more precisely the research questions addressed within this dissertation. For this purpose, we will separately treat each main part of this expression above, clarify its understanding for the context of this dissertation and state the related research questions. We will hereby refer to the corpora of Social Annotation Systems as Social Annotation data.

Emergent Semantics: As will be elaborated in greater detail in Section 5.2.1, the study of emergent semantics is concerned with the evolution of decentralized semantic structures. Because Social Annotation Systems are typically made up of three kinds of constituents, namely keywords, users and resources (see Section 3.1.2), it is necessary to clarify the objects of investigation. In other words, when interpreting structures as relations among objects, we need to specify *between which kinds of objects* the relations of interest hold. While there is work on “collective semantics” including emerging relations between users, keywords and resources (Au Yeung, 2009), the focus of this dissertation is solely the evolution of *keyword structures*. The rationale behind this decision is that we want to compare the different paradigms of Social Annotations and the Semantic Web for structuring information resources – without blurring the focus by taking into account, e. g., the detection of user communities. Because ontologies as core components of the Semantic Web have in most cases a lexical layer as well, this approach allows us to contrast the *top-down* induced relations with those emerging in a *bottom-up* manner. Hereby we are finally dealing with the “meaning” of words; hence, this dissertation also has contact points to the

fields of lexical semantics and lexical acquisition (Widdows and Dorow, 2002). So strictly speaking, when we talk about “emergent semantics” in the context of this dissertation, we primarily refer to emergent *keyword semantics*.

Another important question is how the “meaning” of a particular keyword or relation is represented. While there is work on “mapping” keywords to existing semantic resources (Angeletou, 2008), our primary goal is to capture the meaning of a keyword *relative* to the other keywords in the system. An exemplary outcome could be, e. g., that two keywords mean the same thing, i. e., denote the same concept – without the need to be able to explicitly map this concept to an existing one in some external ontology. This is a crucial point, because we expect one of the largest benefits from capturing emergent semantics actually in the *discovery* of previously unknown concepts and relations. However, in order to tune the instruments for this purpose, a prerequisite is an understanding to which extent *existing* structures (e. g., relations defined in reference ontologies) can be reproduced based on Social Annotation data.

Specifically, our interest in the emergence of semantic structures among keywords in Social Annotation Systems can be broken down into the following research questions:

- Is it possible to derive keyword relations from Social Annotation data which correspond to term relations defined within existing semantic resources?
- Are there methods to differentiate the “nature” of emerging keyword relations, i. e., is it possible to find out which *semantic* relation (as defined in existing semantic resources) is captured by a particular type of keyword relation?
- Which factors influence the emergence of keyword semantics, i. e., which influences have a positive (or negative) impact on the similarity of the emerging relations to existing ones within semantic resources?

Capturing: Although there exist also explicit relations among keywords within Social Annotation Systems (e. g., direct co-occurrence), there is evidence for a much richer underlying *implicit* structure (Cattuto et al., 2007). This implies that even when we know that there *exist* meaningful latent relations, the crucial question is how to obtain access to them, or in other words, how to make them explicit. Because this task is also typically addressed in the field

of ontology learning (e. g., by capturing the implicit relations present within natural language text or semi-structured documents), one might assume that methods stemming from this field can be useful for the purpose of analyzing Social Annotation data as well. These methods borrow usually from several research areas, like natural language processing (NLP), machine learning or data mining. However, due to the rich network structure of Social Annotation Systems, promising candidates also exist in related fields like social network analysis (SNA), graph theory or information retrieval. Regarding these fields, another research question addressed within this dissertation is the following:

- Is it possible to apply methods from the fields of ontology learning, knowledge acquisition and related disciplines to Social Annotation data?
- What are the core characteristics which differentiate Social Annotation data from other kinds of input to ontology learning or knowledge acquisition algorithms?

Social Annotation Systems: Besides the question which kind of relations among which objects present in Social Annotation Systems (i. e., users, keywords and resources) should be targeted, another important aspect is which part of the data is considered for inferring those relations. This corresponds to the question which exact *data source* to use as an input for the chosen algorithms. While there exist approaches which require user-specified relations (Plangprasopchok et al., 2010), or analyze the textual content of keywords (Tatu and Moldovan, 2010) or resources (Brooks and Montanez, 2006), the main focus of this dissertation is to exploit solely the tripartite network structure of users, keywords and resources (as well as derived networks, which will be introduced in Section 3.1.3). Hereby we actually ignore all additional information based on the “content” of keywords, users or resources; strictly speaking, most of our methods are actually “blind” for the latter, and address objects of all three kinds basically just by an arbitrary identifier. While hereby surely useful information is left aside, the strong benefit of our approach is its universality: While, e. g., methods based on the analysis of the resources themselves are mostly only applicable to textual content, operating solely on the graph structure retains the possibility to analyze all kinds of resource types. This leads to the following research questions:

- Is it possible to deduce semantic keyword relationships solely based on the network structure of Social Annotation Systems?

- Are there particular networks which are especially well-suited to derive a certain semantic keyword relation from?
- What are the general properties of the different kinds of networks regarding the amount and type of keyword semantics which they encode?

The following section outlines the major steps which will be considered to thoroughly address the aforementioned research questions.

1.3. Outline

Briefly spoken, the overall structure is intended to familiarize the reader first with the social and the semantic approach to Knowledge Organization, and then to examine all relevant aspects for “mediating” between both by means of capturing emergent semantics. More specifically, this comprises the steps explained in the following.

In Chapter 2, some open problems of working with knowledge on the Web will be highlighted, and Knowledge Organization and Knowledge Engineering will be introduced as two approaches to address those. After that, Chapters 3 and 4 provide an in-depth explanation of the relevant paradigms and concepts in the field of Social Annotation and the Semantic Web. Hereby formal models for both are introduced, and special attention is given to the discussion of their respective strengths and weaknesses. Chapter 5 finally introduces the core topic of this dissertation, namely the field of bridging the gap between Social Annotations and the Semantic Web. After presenting general issues which need to be considered hereby, ontology learning as a bridging methodology is introduced. In the sequel, the state of the art and further relevant work in this direction are discussed, based on a set of comparison dimensions. The latter will then be used to concretize again the specific focus of the approach of this dissertation.

The second and main part then starts by a presentation of the Social Annotation Systems, whose data is the object of investigation, along with summary of the reference semantic datasets used for evaluation purposes (Chapter 6). The core methodological contribution is then contained in Chapter 7: Broadly structured along tasks derived from ontology learning, we will systematically study methods which unveil different kinds of semantic relations among keywords. The most basic one is semantic relatedness, which will be addressed

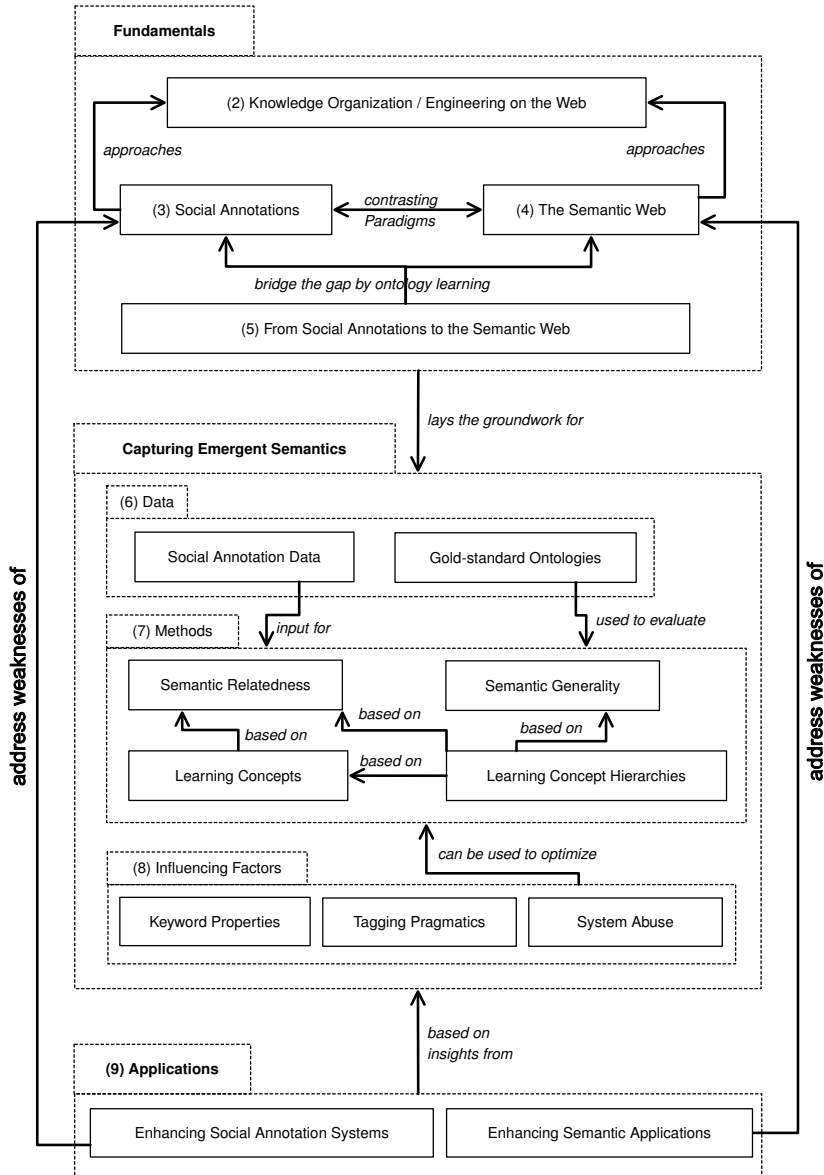


Figure 1.1.: Graphical outline of the narrative of this dissertation.

first. Based on insights gained hereby, the next step is to induce concepts by discovering synonymous and polysemous keywords. As a prerequisite for inducing hierarchical relationships, the subsequent section introduces measures of semantic keyword generality. The latter will finally serve as an input for methods targeted towards building concept hierarchies. Next, in Chapter 8 we will shift perspective and take a look at the factors which have an impact on the emergence of semantics. These belong mainly to three classes, namely (i) properties of the keywords themselves, (ii) tagging pragmatics, i. e., how and why people are annotating, and (iii) the malicious activities of spammer users.

Chapter 9 then outlines a set of existing and envisioned applications, which build on the insights gained in the previous methodological chapter. These comprise on the one hand systems which enhance Social Annotation Systems by (i) fostering the emergence of semantics and (ii) feeding back the learned semantics, and on the other hand external semantic systems, which can benefit from the derived semantic relations. Chapter 10 then summarizes the contributions of this dissertation, and gives an outlook to interesting future research directions.

Figure 1.1 summarizes in a graphical way the outline of this dissertation, especially highlighting the relations among the aforementioned chapters. It is intended to serve as a reader's guide for the remainder of this work.

1.4. Overview of Author's Contributions

Large parts of this work were created during collaborations with colleagues and other researchers. Several approaches and results have also appeared in previous publications by the author, together with collaborators. This section clarifies the author's original contributions, and relates each publication to its corresponding chapter within this thesis. All other parts of this thesis which are not explicitly mentioned below are the sole work of the author.

Chapter 7.1

- Dominik Benz, Beate Krause, Praveen Kumar, Andreas Hotho, and Gerd Stumme. Characterizing semantic relatedness of search query terms. In *Proceedings of the 1st Workshop on Explorative Analytics of Information Networks (EIN2009)*, Bled, Slovenia, September 2009.

- Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th International World Wide Web Conference (WWW2009)*, pages 641–641, April 2009.
- Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In Amit P. Sheth, Steffen Staab, Mike Dean, Massimo Paolucci, Diana Maynard, Timothy W. Finin, and Krishnaprasad Thirunarayan, editors, *Proceedings of the 7th International Semantic Web Conference (ISWC2008)*, volume 5318 of *LNAI*, pages 615–631, Heidelberg, 2008. Springer.
- Dominik Benz, Marko Grobelnik, Andreas Hotho, Robert Jäschke, Dunja Mladenic, Vito D. P. Servedio, Sergej Sizov, and Martin Szomszor. Analyzing tag semantics across collaborative tagging systems. In Harith Alani, Steffen Staab, and Gerd Stumme, editors, *Proceedings of the Dagstuhl Seminar on Social Web Communities*, number 08391, 2008.

The analysis around semantic grounding of tag relatedness was done in close collaboration with Ciro Cattuto and Andreas Hotho; both of them contributed original ideas hereby. More precisely, Ciro Cattuto suggested the distributional measures of relatedness (especially the tag context relatedness), the methods of qualitative analysis, and implemented parts of the path analysis within WordNet. The idea of semantic grounding, and the computation and analysis of FolkRank on the Delicious dataset can be attributed to Andreas Hotho. The author implemented the distributional measures, and generalized the results to other datasets. The idea of alternative aggregation schemes (namely macro and collaborative aggregation) stems from Filippo Mencer and Ben Markines. The alternative weighting schemes and similarity measures are based on the author's ideas and implementations. Most similarity computations were performed using a framework implemented within the scope of a Master thesis by Tobias Gunkel, under the guidance of the author. The analysis of logsonomy data was done together with Beate Krause, who compiled the data and contributed to the discussion of the results.

Chapter 7.2

The baseline approach of grouping synonym keywords resulted from a master thesis by Stefan Stützer, under the guidance of Andreas Hotho and the author. The same holds for the idea of using hierarchical agglomerative clustering for tag sense disambiguation.

Chapter 7.3

- Dominik Benz, Christian Körner, Andreas Hotho, Gerd Stumme, and Markus Strohmaier. One tag to bind them all : Measuring term abstractness in social metadata. In Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Jeff Pan, and Pieter De Leenheer, editors, *Proceedings of the 8th Extended Semantic Web Conference (ESWC2011)*, Heraklion, Crete, May 2011.

The graph-based generality measure were partially implemented by members of the Knowledge Management Institute, Graz University of Technology (Christian Körner, Markus Strohmaier). They also implemented the web platform for the user study. The design of the user-based evaluation was done by the them and the author together. The explanation of the results stems from extensive discussions with them.

Chapter 7.4

- Markus Strohmaier, Denis Helic, Dominik Benz, Christian Körner, and Roman Kern. Evaluation of folksonomy induction algorithms. *Transactions on Intelligent Systems and Technology*, 2011.
- 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 *Proceedings of the 2nd Web Science Conference (WebSci2010)*, Raleigh, NC, USA, 2010.
- Dominik Benz and Andreas Hotho. Position paper: Ontology learning from folksonomies. In Alexander Hinneburg, editor, *Workshop Proceedings of Lernen - Wissensentdeckung - Adaptivität (LWA 2007)*, pages 109–112. Martin-Luther-Universität Halle-Wittenberg, September 2007.

All works concerning the creation of the concept hierarchies (such as the implementation of the clustering and generality-based algorithms, and the pre-processing of the datasets) was done by members of the Knowledge Management Institute, Graz University of Technology (Denis Helic, Roman Kern, Christian Körner and Markus Strohmaier). The author contributed the semantic evaluation of the results, as well as the design and evaluation of the involved user study. The optimized version of the generality-based algorithm as well as the inclusion of disambiguated keywords and synonyms in the learning process was developed during the master thesis by Stefan Stützer, under the guidance of Andreas Hotho and the author.

Chapter 8.3

- Christian Körner, Dominik Benz, Markus Strohmaier, Andreas Hotho, and Gerd Stumme. Stop thinking, start tagging - tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th International World Wide Web Conference (WWW2010)*, Raleigh, NC, USA, April 2010. ACM.

The initiative to analyze semantic implications of tagging pragmatics originates from members of the Knowledge Management Institute, Graz University of Technology (Christian Körner, Markus Strohmaier), who also developed and implemented the measures of tagging pragmatics. The experimental design of assessing the influence of pragmatic factors on emergent semantics was developed to equal parts by them and the author.

Part I.

Fundamentals: Social Annotations and the Semantic Web

Chapter 2.

Knowledge on the Web

As technology and Web access is pervading more and more areas of our personal and professional lives, the role of the World Wide Web as a central communication and information medium is becoming more and more prevalent. However, actually *finding* and *making use* of its contained knowledge can still be difficult. While web search engines like Google¹ allow for keyword-based retrieval, the identification of central concepts and their interrelation within a particular domain of interest is still largely left to human interpretation. To use a more pointed formulation, the role of the Web as a “knowledge repository” still consists to a large extent of how its users actually process and make sense of its contents. However, its sheer size and complexity demands for theoretical and practical solutions to support humans such knowledge-based tasks.

Historically, the establishment of such solutions has been the driving force behind a large number of activities in different research areas. An important representative among them is Artificial Intelligence, which has been concerned since its early days with formalisms and mechanisms to operationalize so-called *Knowledge-based Systems* (KBS) (Studer et al., 1998). The goal hereby was to develop tools which are able to make “intelligent” decisions within a certain domain, coming close to (or even outreaching) human expertise. The discipline of constructing such systems became known as *Knowledge Engineering*.

From a different point of view, a core interest in the field of Library and Information Science were methods and activities to organize repositories of knowledge resources. “Resources” are hereby understood in a general sense as artifacts (e. g., books or web pages) which encode knowledge in a particular format. Efforts in this direction were hereby subsumed under the term *Knowledge Organization Systems* (KOS) (Weller, 2010). A typical example are library indexing schemes, which are intended to structure the available content and hence facilitate its efficient access.

¹<http://www.google.com>

It is clear that the requirements of “web scale” presented new challenges and opportunities for both Knowledge Engineering and Knowledge Organization. Despite both are concerned with making knowledge on the Web accessible, they are focusing hereby on orthogonal aspects: While the primary goal of Knowledge Engineering is to represent this knowledge such that it can be understood and processed by *machines*, the focus of a Knowledge Organization scheme lies more in facilitating its access by *humans*. However, it is clear that both targets are not necessarily mutually exclusive, because a particular form of knowledge representation may be understandable for both humans and machines. In any case, both disciplines benefit from a meaningful modeling and structuring of the concepts and relations within a domain of interest – i. e., in other words, of a precise *semantic* representation.

Because a core topic of this dissertation is how such a representation can be derived from emerging patterns within a particular kind of web applications (namely Social Annotation Systems, which will be introduced in Chapter 3), it clearly touches both mentioned disciplines. As our ultimate goal is to derive a *formal* representation, which can be understood by intelligent agents of the *Semantic Web* (which will be introduced in Chapter 4), we see a stronger relationship to the field of Knowledge Engineering. However, because some concepts and terminology from Knowledge Organization are relevant and useful, we will borrow those when appropriate. We will start by giving a brief overview on the relevant aspects of Knowledge Organization, and continue with a summary of Knowledge Engineering.

2.1. Knowledge Organization

As stated above, within the cultural development of mankind, institutions were developed intended to collect and aggregate knowledge resources in order to keep track with the ever-growing amount of knowledge resources – libraries are an early example. Because a cognitive strategy of humans to cope with large amounts of information is to structure it using categories and concepts (Anderson, 2001), the latter were fundamentally concerned with methods and activities to organize the aggregated repositories. In the field of library and information science, these efforts are subsumed under the term *Knowledge Organization* (KO). Its narrow meaning is defined as follows:

“[...] Knowledge Organization (KO) is about activities such as docu-

000 Computer science, information and general works
100 Philosophy and psychology
200 Religion
300 Social sciences
400 Language
500 Science
600 Technology and applied Science
700 Arts and recreation
800 Literature
900 History, geography, and biography

Figure 2.1.: Top level classes of the Dewey Decimal System (DDC).

ment description, indexing and classification performed by libraries, bibliographical databases, archives and other kinds of 'memory institutions' by librarians, archivists, information specialists, subject specialists as well as by computer algorithms and laymen.”

(Hjorland, 2008, spelling corrected)

Its implementation takes mainly place within *Knowledge Organization Systems* (KOS), for which the following is a generic definition stemming from library science:

“The term knowledge organization systems is intended to encompass all types of schemes for organizing information and promoting knowledge management.”

(Hodge, 2000)

Classical examples are hereby nomenclatures, thesauri and classification systems (Weller, 2010, p.21 ff). The core idea of the latter is to subdivide a given domain of interest in a hierarchical fashion into classes and subclasses. A prominent example is the *Dewey Decimal Classification* (DDC); Figure 2.1 shows its ten top level classes. Please note that each class can also be represented by a so-called *notation* (a numeric value ranging from 0 to 999 for the case of DDC), which allows the language-independent assignment of resources. One of the main purposes of such a cataloguing activity is to facilitate efficient browsing, search and retrieval within large resource collections.

Applied to the context of the World Wide Web, one could interpret the approach of KO as viewing the Web as a “library” of resources, for which a suitable organization scheme needs to be defined. However, it is clear that the different nature of the Web requires novel paradigms to this end. Within its evolution, several organization systems were proposed and implemented – some of them by authoritative persons or institutions like the W3C², others “emerged” as successful applications by innovative individuals, groups or companies. For the scope of this thesis, we will adapt the above definition of KOS from (Hodge, 2000) as follows:

Definition 2.1 *A Knowledge Organization System is a system which allows to structure information resources by annotating them with metadata which indicates their membership in classes or categories.*

We are primarily introducing this notion to have a common terminology for the “social” (Chapter 3) and the “semantic” (Chapter 4) approach of annotating resources on the World Wide Web. In the next section, we will relate both to the discipline of Knowledge Engineering.

2.2. Knowledge Engineering

As stated above, the discipline of Knowledge Engineering (KE) is fundamentally concerned with designing, building and maintaining Knowledge-Based Systems. While this activity had historically been viewed as a *transfer* process (Hayes-Roth et al., 1983) from human expertise into a program, the current consensus is more a *modeling* view:

“Building a KBS means building a computer model with the aim of realizing problem-solving capabilities comparable to a domain expert.”
(Studer et al., 1998)

When we talk about Knowledge Engineering from the World Wide Web, we mean primarily mechanisms and formalisms to model the knowledge present within its vast amount of information resources. Hereby, certain knowledge organization schemes may be used for representation purposes (e. g., a taxonomy of website topics). Hence, we interpret these as an aspect of the general process of Knowledge Engineering from the Web.

²<http://www.w3.org/>

Within the following two chapters, two approaches of knowledge organization systems which are relevant to Knowledge Engineering from the Web will be presented – namely Social Annotations and the Semantic Web. Hereby, especially their respective strengths and weaknesses will be highlighted. In the subsequent chapter, the idea to provide an augmented kind of knowledge organization by combining “the best of the two worlds” will be presented, and the state of the art will be discussed.

Chapter 3.

Social Annotations

Since its invention around 1990 by Tim Berners Lee, the World Wide Web has undergone a tremendous development through various phases. A first major step towards mass adoption was the introduction of graphical web browsers, among which Mosaic (NCSA, 2011) became especially popular.¹ According to its developers, this breakthrough was mainly due to “*features – like icons, bookmarks, a more attractive interface, and pictures – that made the software easy to use and appealing to ‘non-geeks’*” (NCSA, 2011). Despite the Web was subsequently populated by an immensely growing amount of users, the boundary between information consumers (i. e., mostly people browsing the Web) and providers (i. e., mainly website and content authors) remained clearly observable.

Several years later, this dichotomy should be blurred within the course of a set of developments being summarized under the term “Web 2.0”, coming along with a retrospective naming of the prior stage of the WWW as “Web 1.0”. The variety of novel applications, techniques and principles makes it hard to formulate a precise definition what exactly were the distinguishing features of this second stage; however, a common characteristic of many involved applications is their participatory nature and a user-centered design. This allowed end-users to contribute and collaborate in a highly interactive manner, making them in fact an integral part of what became subsequently known as the “Social Web”. While the notation “Web 2.0” may suggest an update of a technical specification, Tim Berners Lee himself pointed out in an interview² that this was not the case, and that instead these novel applications were based on existing web

¹The importance of Mosaic in the early history of the WWW is reflected in the title of the Proceedings of the Second World Wide Web Conference: “Mosaic and the Web” (cf. <http://web.archive.org/web/20050306013919/archive.ncsa.uiuc.edu/SDG/IT94/IT94Info-old.html>)

²<http://www.ibm.com/developerworks/podcast/dwi/cm-int082206.txt>

technologies. Despite that, an important point is that the distinction between information consumers and providers could no longer be drawn at a comparable level of clarity.

A further notable feature of many social web platforms was that they greatly simplified the process of *annotating* digital objects like websites or multimedia items. Annotation could hereby be performed in various ways, e. g., by classifying, voting, editing or rating, and the targeted objects could be created by others or by the annotator himself. In any case, this feature turned out to be immediately useful for the users. As an example, uploading and annotating a holiday picture to Flickr³ has the immediate benefit of having this picture stored on a central server, being able to access it from anywhere; in addition, it is easy to discover further interesting pictures by other users which have used the same keywords for annotation. These and other advantages have in fact engaged millions⁴ of end-users in the process of annotating web resources.

Although Social Annotation yields different kinds of benefits, its function as a means of organizing information has gained particular interest (Halpin et al., 2007). Hereby especially *Social Tagging Systems* (Hammond et al., 2005), which allow the annotation of various kinds of resources with arbitrary keywords or *tags*, were seen as a bottom-up categorization alternative to or even a replacement (Shirky, 2005) for more formalized classification approaches like taxonomies, controlled vocabularies or ontologies (the latter will be described in detail in Chapter 4). While the idea of non-expert manual indexing systems using an uncontrolled vocabulary was not fundamentally new, their implementation on the Web using intuitive interfaces and immediate feedback mechanisms finally leveraged mass adoption (Voss, 2007). Early analyses like (Hotho et al., 2006a) described the resulting complex networks of users, tags and resources as lightweight conceptual structures.

In this chapter, a systematic characterization of Social Annotation as an approach of organizing information will be given. The first and main section is concerned with *Social Tagging Systems*, whose data is the main focus of the analyses presented within this thesis. After a description of their core functionalities, a formal model is given and derived structures as well as specialized ranking algorithms are described. The section closes with a discussion of different kinds of tagging motivations. The second section introduces further kinds of Social

³<http://www.flickr.com>

⁴The factual number of users for several systems can be found in Section 6.1.

Annotation Systems which can be found among Web 2.0 applications. The chapter closes with a summarization of the most important characteristics, intended to lead over to the complementary Semantic Web approach described in the subsequent Chapter 4.

3.1. Social Tagging Systems

As stated above, among all Social Annotation variants, the collaborative assignment of keywords or *tags* to different kinds of resources has gained a large amount of attention, especially in research fields interested in different aspects of organizing information. The basic principle of these so-called *Social Tagging Systems*⁵ is to allow registered users the comfortable maintenance of a resource collection (e. g., videos on YouTube⁶, images on Flickr⁷, URLs on Delicious⁸ or bibliographic references on BibSonomy⁹) within a centralized online location. In order to ease later browsing and retrieval, the users can assign a set of keywords or *tags* to each resource. A key characteristics hereby is that the choice of tags is not limited to a predefined vocabulary; instead, any arbitrary¹⁰ character sequence can be selected. Figure 3.1 shows an exemplary screenshot of the social bookmark and publication sharing system *BibSonomy*, highlighting the occurrence of the three major constituents of social tagging systems – i. e., tags, users and resources.

An typical use case for the BibSonomy system is the collaborative maintenance of a bibliography during a research project as exemplified in the following scenario: During his PhD thesis, a researcher discovers an interesting paper about folksonomy analysis. Then he can create a new publication entry¹¹ and assign, e. g., the following keywords to it: *folksonomy*, *analysis*, *toread*, *thesis* and *2010*.¹² Immediately after storing the entry, the interlinked struc-

⁵also often called *Social Bookmarking Systems* or *Collaborative Tagging Systems*

⁶<http://www.youtube.com>

⁷<http://www.flickr.com>

⁸<http://www.delicious.com>

⁹<http://www.bibsonomy.org>

¹⁰A slight natural exception is that a system-defined character used as a keyword delimiter cannot be part of a tag. Common delimiters are a whitespace character (e. g., in BibSonomy or Delicious) or a comma (e. g., in Flickr).

¹¹BibSonomy supports several ways to enter the bibliographic details, ranging from manual input over copying from other users to extraction techniques from online digital libraries.

¹²Please note that we will use this formatting (namely a slanted monospace font) to refer

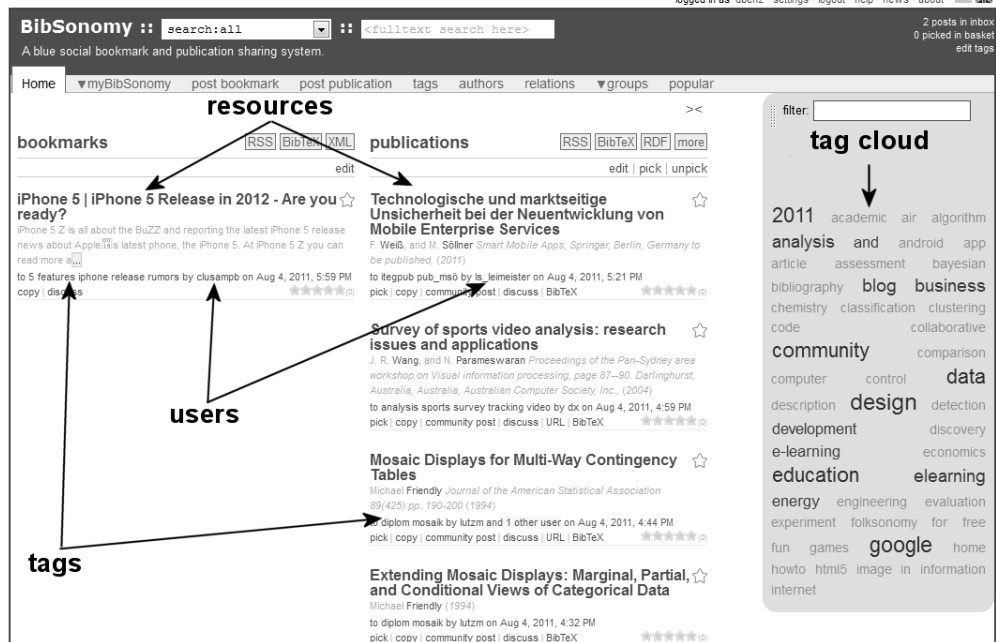


Figure 3.1.: Screenshot of the Social Tagging System BibSonomy. It allows its users the maintenance of a collection of two types of resources, namely URLs (i. e., bookmarks, left column) and bibliographic references (right column).

ture allows him to quickly explore (i) which other users in the system have the same paper in their collection, (ii) which other resources in the system were tagged with the same tags and (iii) which tags other users have used to annotate the paper he just discovered. These immediate feedback mechanisms are offering directly perceivable added values like the discovery of (i) other users with similar research interests and (ii) further potentially interesting publications. The latter combined with the possibility to browse the folksonomy structure along all dimensions (i. e., via selecting users, tags or resources) facilitates especially the serendipitous discovery of relevant items which one did not explicitly search

to keywords stemming from Social Annotation processes throughout this thesis. Further formatting conventions (namely for concepts and concept labels) are introduced in Section 4.1.1.

for. During further tagging activities of our exemplary researcher and other users, the system is able to aggregate tag assignments into *tag clouds* (see also Figure 3.1). These typically depict the tag usage frequencies by different font sizes or color shades, leading to a intuitive description which can be grasped quickly by a user.

Despite their obvious usefulness which was documented by a growing popularity among Web users, systematic research in the nature of these systems showed up inherent strengths and weaknesses, which will be discussed in the following subsection.

3.1.1. Strengths and Weaknesses

With the aforementioned advent of the so-called Web 2.0, existing knowledge organization techniques like taxonomies or ontologies could hardly cope with the masses of user-generated content. The participatory nature of many of the new applications demanded for open and dynamic categorization approaches, which could be handled by end-users themselves in an effortless manner. The growing popularity of Social Tagging Systems can be attributed to their fulfilment of many of these requirements:

“The mass amateurization of Web publishing makes the mass amateurization of cataloguing a forced move. Folksonomies are a trade-off between traditional structured centralized classification and no classification or metadata at all. And they are the best we actually have.”
(Quintarelli, 2005)

More specifically, the following characteristics of Social Tagging were seen as superior compared to existing categorization approaches (Quintarelli, 2005; Mathes, 2004; Sinha, 2005):

- **Inclusiveness and Adaptivity:** Because each tag added by each user is comprised in the global vocabulary without any kind of filtering or centralized control, the resulting structure completely includes all personal viewpoints and preferences. Due to this continuous inclusion process, folksonomies are also adapting quickly to vocabulary changes, contrasting the typically slow process of terminology adaptation in, e. g., controlled vocabularies.

- **Serendipity:** The highly interlinked structure of folksonomies together with their navigability along all dimensions (i. e., by selecting users, tags or resources) makes them an excellent candidate for serendipitous discovery of interesting content which one did not explicitly search for.
- **Low Cognitive Effort:** The uncontrolled nature of the tagging process relieves users from the burden of having to select the “correct” categorization among a set of predefined choices. Furthermore, no special domain expertise or knowledge of domain concepts and their interrelation is required for participation.
- **Immediate usefulness and feedback:** As described earlier in this chapter, Social Tagging Systems directly offer added values: Even after the first contribution, a new user gets immediate feedback on which tags other potentially interesting users have applied and which other relevant resources exist in the system.
- **Scalability:** Especially when a domain is growing, the maintenance of a categorization scheme is an expensive task. Social Tagging Systems have the potential to tackle this problem, because they attract large user populations due to their aforementioned benefits. Or to use the formulation by Shirky (2005): “*The only group that can categorize everything is everybody.*”.

To summarize, Social Tagging seemed to be a pragmatic answer to the question of how to organize the masses of user-generated content within the Social Web. Despite that, most authors agreed that tagging is not the “silver bullet” of knowledge organization – interestingly, it turned out that folksonomies suffered mainly from problems that more formalized approaches were designed to eliminate. Typical arguments in this direction found in the literature are (Mathes, 2004; Quintarelli, 2005; Golder and Huberman, 2006):

- **Low precision:** The uncontrolled nature of the tagging process opens the door for all kinds of noisy annotations, including misspellings, idiosyncratic terms, or even erroneous descriptions due to missing expertise.
- **Lack of tag structure:** For some tasks like tag-based information retrieval, the flat structure of the tag space can prove problematic, because,

e. g., missing hierarchical tag relationships make broadening or narrowing of the retrieval scope difficult.

- **Synonymy and Polysemy:** Because tags can be considered as words, language phenomena like synonymy are becoming problematic in folksonomies when users are naming the same concept by different terms. On the other hand, a given tag may have more than one meaning. Both has a detrimental effect especially on information retrieval tasks.
- **Varying basic levels:** Another problem is rooted in cognitive aspects of categorization. Depending among others on their expertise in a given domain, different users will use terms with varying levels of specificity to describe a given object. Faced with, e. g., a picture of a wildcat, an animal expert might annotate it with *wildcat* or even *felis silvestris*, while an ordinary person would probably use *cat*. These conflicting basic levels lower the usefulness of too specific or too general tags for particular user groups.
- **Limited Retrieval possibilities:** While one of the core strengths lies in the serendipitous discovery of relevant content by browsing activities, searching for specific resources especially via tag-based retrieval is much more difficult, mostly due to the aforementioned problems.

Throughout this dissertation, possible solutions for some of these shortcomings (mainly those related to the lack of structure as well as synonymy and polysemy) will be presented. In order to lay the groundwork for the presentation of these approaches, the following subsection introduces a formal model of folksonomies.

3.1.2. Formal Model

As described in the previous sections, the three main constituents of social tagging systems are users, tags and resources. The assignment of tags to resources by users is given by a ternary relation. For the context of this thesis, we will stick to the following formalization taken from (Hotho et al., 2006b):

Definition 3.1 A folksonomy is a tuple $\mathbb{F} := (U, T, R, Y, \prec)$ where

- U , T , and R are finite sets, whose elements are called users, tags and resources, resp.,

- Y is a ternary relation between them, i. e., $Y \subseteq U \times T \times R$, called assignments, and
- \prec is a user-specific subtag/supertag-relation, i. e., $\prec \subseteq U \times T \times T$.

Definition 3.2 The personomy \mathbb{P}_u of a given user $u \in U$ is the restriction of \mathbb{F} to u , i. e., $\mathbb{P}_u := (T_u, R_u, I_u, \prec_u)$ with $I_u := \{(t, r) \in T \times R \mid (u, t, r) \in Y\}$, $T_u := \pi_1(I_u)$, $R_u := \pi_2(I_u)$, and $\prec_u := \{(t_1, t_2) \in T \times T \mid (u, t_1, t_2) \in \prec\}$.

Users are commonly identified by their user name in the system and tags can be arbitrary character sequences. Different systems support different kinds of resources; however, these are usually identified by a mechanism like URLs independent of their type. This data model is underlying all social tagging systems presented in Section 6.1. Among those, the implementation of the user-specific tag relations \prec differs: While Delicious allows the creation of tag sets called *bundles*, BibSonomy enables users to create directed tag *relations*. Other systems do not offer this feature at all. In such a case or when the relations are irrelevant for the research question under consideration, we will set $\prec = \emptyset$ and regard a folksonomy for simplicity reasons as a quadruple $\mathbb{F} := (U, T, R, Y)$. In Formal Concept Analysis (Ganter and Wille, 1999) this structure is known as a *triadic context* (Lehmann and Wille, 1995). This data structure can alternatively be regarded as a tripartite undirected hypergraph $G = (V, E)$, whose set of nodes is defined by the disjoint union of the sets of users, tags and resources $V = U \dot{\cup} T \dot{\cup} R$ connected by the set of hyperedges $E = \{\{u, t, r\} \mid (u, t, r) \in Y\}$. Figure 3.2 shows an exemplary visualization of a small hypothetical folksonomy.

For convenience reasons, we will refer to a *post* as a triple (u, T_{ur}, r) with $u \in U$, $r \in R$, whereby $T_{ur} := \{t \in T \mid (u, t, r) \in Y\}$ is a non-empty set of tags. In other words, a *post* corresponds to all data which accrues during the annotation of a given resource r by a given user u with a set of tags T_{ur} .

3.1.3. Induced Networks

When researchers started to analyze folksonomies, it turned out that several existing algorithms and techniques were not well developed for the three-mode nature of the folksonomy graph structure. But because the hypergraph also induces other kinds of networks, often two-mode and especially one-mode views on the data were used.

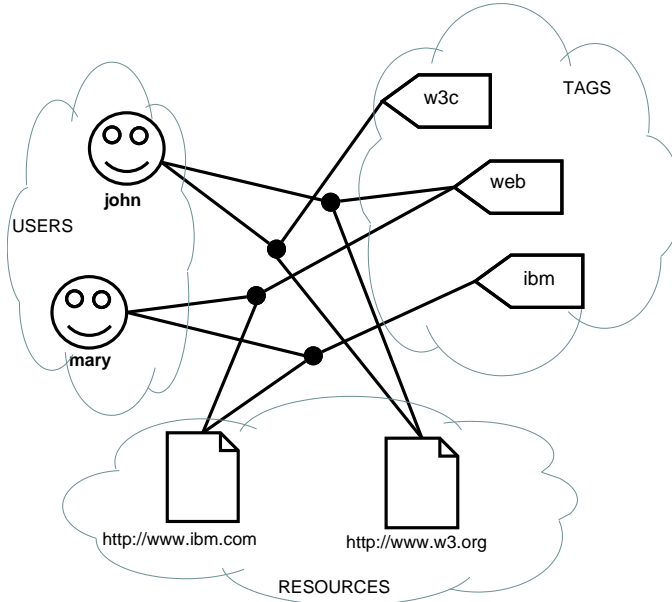


Figure 3.2.: Visualization of the folksonomy data structure.

Two-mode networks: Having users, tags and resources as three folksonomy constituents at hand, it is consequently possible to derive $\binom{3}{2} = 3$ undirected two-mode networks, namely (i) user-tag, (ii) tag-resource and (iii) user-resource networks. Mika (2005) denoted these as actor-concept (AC), concept-object (CO) and actor-instance (AI) graphs, respectively. Figure 3.3 shows a graphical overview. As an example, the edges of the tag-resource network $TR = (T \dot{\cup} R, E_{tr})$ along with an edge weighting function $w_{tr}: E_{tr} \rightarrow \mathbb{R}$ for a given folksonomy \mathbb{F} can be constructed according to

$$E_{tr} = \{(t, r) \in T \times R \mid \exists u: (t, u, r) \in Y\}$$

$$w_{tr}((t, r)) = |\{u: (t, u, r) \in Y\}|$$

This means an edge between a tag t and a resource r is weighted by the number of users u who have used t to annotate r . The user-resource and user-tag networks can be constructed analogously. However, these have seldom been analyzed as such, but were further processed to derive one-mode networks (Mika, 2005; Au Yeung et al., 2009a).

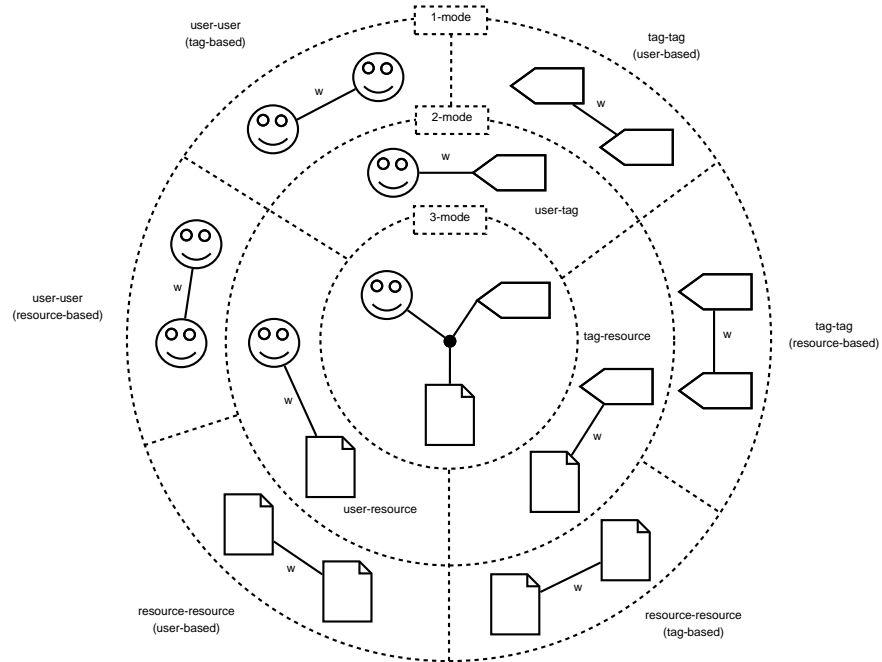


Figure 3.3.: Schematic overview of 2-mode and 1-mode networks which can be derived from the 3-mode folksonomy data.

One-mode networks: For many applications addressing the aforementioned weaknesses of folksonomies (see Section 3.1.1), one-mode networks are of special interest – e. g., user-user networks for community detection, tag-tag networks for inferring semantic relations, or resource-resource networks for item recommendations. Figure 3.3 depicts how six canonical types of one-mode networks can be derived¹³ from underlying two-mode representations. In each case, the connections between the nodes are based on the principle of *co-occurrence* – i. e., the common occurrence of items within the context of tag assignments. As an example, in the tag-based resource-resource network, the connection strength between two resources is growing with the number of times they “co-occurred” by being annotated with the same tag. Because especially tag-tag networks are encoding a lot of relevant information to overcome folksonomy shortcomings like synonymy, polysemy or the lack of tag structure, these will be treated

¹³A more formal description of this process is found in the next paragraph.

separately in the following paragraph.

Tag networks: Because tags are used by humans to annotate resources, it is justified to expect that the common occurrence of tags encodes information about the semantic relation among them. For this reason, different kinds of *tag co-occurrence networks* were employed for the purpose of analyzing tag semantics (Hotho et al., 2006a; Schmitz, 2006; Begelman et al., 2006). The general formulation of these graphs is $\mathbb{T} = (T, E)$, whereby the edges $E \subseteq T \times T$ can be weighted by the function $w: E \rightarrow \mathbb{R}$. There exist three basic kinds of co-occurrence, which differ mainly in the definition of the respective co-occurrence context and the weighting function:

- *Resource-based:* Two tags co-occur when they were used by one or more users to annotate the same resource:
 $w_{res}((t_1, t_2)) = |\{r : (t_1, u, r) \in Y \wedge (t_2, u', r) \in Y\}|.$
- *User-based:* Two tags co-occur when a single user has used both of them to annotate one or more of his resources:
 $w_{user}((t_1, t_2)) = |\{u : (t_1, u, r) \in Y \wedge (t_2, u, r') \in Y\}|.$
- *Post-based:* Two tags co-occur when a single user has used both of them to annotate one of his resources:
 $w_{post}((t_1, t_2)) = |\{(u, r) : (t_1, u, r) \in Y \wedge (t_2, u, r) \in Y\}|.$

These result in the three networks $\mathbb{T}_{res} = (T, E_{res})$, $\mathbb{T}_{user} = (T, E_{user})$ and $\mathbb{T}_{post} = (T, E_{post})$. In all cases, the set of edges is constructed according to $(t_1, t_2) \in E_i \Leftrightarrow w_i((t_1, t_2)) > 0, i \in \{res, user, post\}$. Please note that the adjacency matrix of \mathbb{T}_{res} can also be obtained by multiplying the adjacency matrix of the tag-resource network with its inverse, as described by (Mika, 2005). The same holds analogously for \mathbb{T}_{user} and the user-tag network. It is obvious that post-based co-occurrence is a restriction of resource-based co-occurrence which excludes cross-user co-occurrences, i. e., it holds that $E_{post} \subseteq E_{res}$.

While co-occurrence itself can be interpreted as a kind of relatedness measure among tags, more elaborate measures were used in the literature (see Heymann and Garcia-Molina (2006) and Section 7.1 of this work) to create *tag relatedness networks*, often based on a vector representation of tags within different contexts. A systematic in-depth analysis of which kind of semantics is captured by the individual measures is a core contribution of this thesis and can be found in Section 7.1.

3.1.4. Tagging System Characteristics

While the formal data model presented in Section 3.1.2 is valid for all Social Tagging Systems, the concrete implementations differ significantly. Hence, individual design choices and system characteristics may have a strong effect on the resulting folksonomy structure. Based on a set of key design dimensions proposed by (Marlow et al., 2006), the following aspects can be expected to have an influence:

Broad vs. narrow folksonomies: An early distinction made by (Vander Wal, 2005) has its origin in tagging permissions. While *narrow* folksonomies like Flickr allow only the content creator himself to annotate, objects within *broad* folksonomies like Delicious can be tagged by the whole folksonomy population. This implies that the number of posts in narrow folksonomies is equal to the number of resources. A direct consequence is that resource-based and post-based co-occurrence are essentially identical in such a case.

Tagging Support: The process of annotation itself can be supported in different ways. A first question is if the user is exposed during his tagging activities to the tags other users have used to annotate the object under consideration, possibly causing imitation effects. Other systems go one step further and offer personalized *tag recommendations* (Hotho et al., 2008).

Type of object: The sharing of different kinds of objects (e. g., videos, images, websites, publications, ...) possibly also affects the tagging process. When a user is exposed during the tagging process to textual resource content, one can assume, e. g., a bias towards using words occurring in the resource title. The success of title-based tag recommendations (Lipczak et al., 2009) provides empirical evidence for this assumption.

Spam Detection: Due to their popularity, social tagging systems are also an attractive goal for malicious user activities like link promotion or the distribution of inappropriate content. As an example, (Krause et al., 2008b) reported that among the 20 000 users of BibSonomy, 18 500 were identified manually as spammers, responsible for 90 % of all posted bookmarks. With further system growth, automatic spam detection methods were a forced move. Though it is hard to assess which spam prevention techniques are employed by the different

system administrators, the existence of spam within folksonomy data is a necessary consideration in order to avoid strongly biased results. A study which examines the influence of spammers on emergent semantics can be found in Section 8.4.

A further question which is independent from individual system characteristics is how to harness existing information retrieval techniques to calculate relevancy among objects within a folksonomy. The next section discusses possible solutions for this purpose.

3.1.5. Calculating Relevancy

Although the optimization of information retrieval techniques on folksonomies is not a core topic of this dissertation, the notion of *relevancy* among objects (e. g., search terms and websites for the case of web search) is a valuable source for the analysis of implicit relations. The PageRank algorithm (Brin and Page, 1998) is a popular and successful example of a web search ranking algorithm. It reflects the idea that a web page is important if there are many pages linking to it, and if those pages are important themselves. Faced with a given keyword query, all matching results are ranked by their importance. As it is based upon the Web graph (whose vertices are websites, and hyperlinks correspond to directed edges among them), it cannot be applied directly to the tripartite data structure of Social Tagging Systems.

However, (Hotho et al., 2006b) showed how its principle can be employed for folksonomies: A resource which is tagged with important tags by important users becomes important itself. The same holds, symmetrically, for tags and users. The core idea hereby is that by modifying the weights for a given object in the random surfer vector, the so-called *FolkRank* can compute a ranked list of relevant other objects within the folksonomy. Simply spoken, giving a high weight to one or more objects o_1, \dots, o_n (which can be an arbitrary combination of tags, users and resources) corresponds to submitting a “query”, which yields the most relevant resources (i. e., again tags, users and resources) relative to o_1, \dots, o_n . Of special interest for this dissertation is hereby FolkRank’s ability to compute a tag-specific ranking – i. e., when “querying” with a given tag t , then the most relevant tags to t are found. Section 7.1 presents an analysis of the semantic implications of the obtained relevancy relation.

More specifically, FolkRank considers a folksonomy (U, T, R, Y) as an undirected graph $(U \cup T \cup R, E)$ with $E := \{\{u, t\}, \{u, r\}, \{t, r\} \mid (u, t, r) \in Y\}$.

For a given object o , it computes in this graph the usual PageRank (Brin and Page, 1998) with a high weight for o in the random surfer vector. Then, the resulting vector is compared to the case of PageRank without random surfer (which equals the simple edge count, as the graph is undirected). In this way the winners (and losers) are computed that arise when giving preference to a specific object in the random surfer vector. The objects that, for a given object o , obtain the highest FolkRank are considered to be the most relevant in relation to o . Hotho et al. (2006b) provides a detailed description of the algorithm.

3.1.6. Tagging Pragmatics

Another interest which the research community has developed is concerned with usage patterns of tagging, such as why and how users tag. Early work like (Golder and Huberman, 2006; Marlow et al., 2006) provided first evidence for different usage patterns among users. Further work suggested that tag usage and motivations vary across different tagging systems (Heckner et al., 2009; Hammond et al., 2005). It was also shown that even within the same tagging system, strong differences of tagging motivation between individual users can be observed (Körner, 2009). These observations led to the formulation of the hypothesis that the emergent properties of tags in tagging systems – and their usefulness for different tasks – are influenced by pragmatic aspects of tagging (Heckner et al., 2009). For the context of this dissertation, especially the question to which extent tagging pragmatics are influencing emergent tag semantics will be addressed (cf. Section 8.3).

Previous work such as (Marlow et al., 2006; Hammond et al., 2005; Heckner et al., 2009) and especially (Körner et al., 2010) suggests that a distinction between at least two types of user motivations for tagging is interesting: On one hand, users can be motivated by categorization (in the following called *categorizers*). These users view tagging as a means to categorize resources according to some (shared or personal) high-level conceptualizations. They typically use a rather elaborated tag set to construct and maintain a navigational aid to the resources for later browsing. On the other hand, users who are motivated by description (so called *describers*) view tagging as a means to accurately and precisely describe resources. These users tag because they want to produce annotations that are useful for later searching and retrieval. Developing a personal, consistent ontology to navigate to their resources is not

Table 3.1.: Two Types of Taggers according to (Körner et al., 2010).

	<i>Categorizer</i>	<i>Describer</i>
<i>Goal of Tagging</i>	later browsing	later retrieval
<i>Change of Tag Vocabulary</i>	costly	cheap
<i>Size of Tag Vocabulary</i>	limited	open
<i>Tags</i>	subjective	objective

their goal. Table 3.1 gives an overview of characteristics of the two different types of users, based on (Körner et al., 2010).

While these two types make an ideal distinction, tagging in the real world is likely to be motivated by a combination of both. A user might maintain a few categories while pursuing a description approach for the majority of resources and vice versa, or additional categories might be introduced over time. Second, the distinction between categorizers and describers is purely based on tag usage patterns, and not related to tag semantics. One implication of that is that it would be perfectly plausible for the same tag (for example *java*) to be used by both describers and categorizers, and serve both functions at the same time – for different users. In other words, the same tag might be used as a category or a descriptive label. Hence tagging pragmatics represent an additional perspective on folksonomical data. However, as will be shown later within this thesis, knowledge about the users' motivation for tagging can be useful for the optimizing methods to capture emergent tag semantics (cf. Section 8.3).

3.1.7. Related Work

While the description of related work in the field of capturing emergent semantics is found in Section 5.3, in the following section further relevant work will be briefly covered in order to provide the reader with a sound overall picture of folksonomy research. While Social Tagging entered the scientific discourse roughly around the year 2004 via newsgroups, weblog posts or mailing lists (e. g., (Mathes, 2004; Shirky, 2005)), a first systematic and meanwhile famous¹⁴ analysis was performed by (Golder and Huberman, 2006). The authors discussed

¹⁴cited by more than 700 papers according to <http://scholar.google.de/scholar?cites=14807028176669359586>, retrieved on 2011/08/08

basic differences between taxonomies and folksonomies, and identified different functions of tags. Hammond et al. (2005); Lund et al. (2005) analyzed general aspects of Social Bookmarking, including reasons for participation and community building. Sinha (2005) mentioned a smaller cognitive effort as a main success factor for Social Tagging. Further works focused on comparing Social Annotations with existing methods of (i) professional subject indexing (Macgregor and McCulloch, 2006; Voss, 2007; Veres, 2006; Heymann et al., 2010; Lin et al., 2006), (ii) automatic keyword extraction (Al-Khalifa and Davis, 2006; Mishne, 2006; Chirita et al., 2007) and (iii) paid annotation (Heymann et al., 2010). Another research direction came up with generative models intended to simulate the tagging process (Cattuto, 2006; Dellschaft and Staab, 2008; Halpin et al., 2007). (Millen et al., 2005) reported on experiences of using social bookmarking tools within an enterprise context, while (Ramage et al., 2009) explored the value of social tags as an information source to cluster websites. Researchers were also interested in the statistical and network properties of evolving folksonomies; (Shen and Wu, 2005) reported small world and scale free characteristics, while (Cattuto et al., 2007) adapted standard network measures like characteristic path length and clustering coefficients to the tripartite folksonomy structure. Further addressed questions were how experts could be identified within a folksonomy (Au Yeung et al., 2009b) and recommendation processes could be designed to assist the user in choosing tags (Siersdorfer and Sizov, 2009; Hotho et al., 2008; Wu et al., 2009).

3.2. Other Forms

The popularity of Social Tagging Systems along with the availability of tagging datasets catalyzed a large amount of research activities around folksonomies. Though the role of keywords is less central and less explicit in most cases, other forms of Social Annotations exist as well. The paradigms of the “Web 2.0” itself as well as the increasing spread of mobile devices connected to the Internet led to the situation that a growing amount of everyday professional and leisure activities became digitally observable. Two brands of popular applications are hereby *Weblog Systems* and *Wikis*, which both facilitate the easy publishing and sharing of textual content. *Microblogging* services like Twitter¹⁵ are a more recent development which produces massive amounts of user-generated markup.

¹⁵<http://www.twitter.com>

Additionally, *clicklogs* of traditional web search engines have been interpreted as an implicit annotation of web resources as well. Lastly, keywords as such are used in some *Question Answering* portals like Stackoverflow¹⁶ to categorize questions.

Because the core topic of this dissertation is the analysis of keyword-based approaches, we will use mainly “pure” folksonomy datasets. However, in order to show up the applicability of the studied methods to other kinds of data, these will be complemented by (i) a clicklog dataset and (ii) a dataset derived from a Question Answering portal (see Section 6). In this way, our intention is to highlight promising analogies. We excluded Weblog and Wiki data, because an in-depth analysis of the different characteristics is beyond the scope of this work. So the main purpose of the following section is to provide the reader with pointers towards possible extensions of the methods and techniques proposed in this dissertation.

3.2.1. Weblogs and Microblogging

While the focus of Social Tagging as a Social Web application lies mainly in the field of organizing and categorizing resources, *Web logs* or short *blogs* became popular because they greatly alleviated the process of web publishing for amateurs. Blogs are an interactive kind of website where individuals create new content in a personal or organizational context, typically on a regular basis. These entries (which can consist of different media types like, e. g., text, images or videos) can then be again referenced and commented by visitors. Most platforms also allow an annotation of posts by tags. However, these may typically be assigned by the post creator. Hence, coming back to the discrimination introduced in Section 3.1.4, a web log system can be interpreted as a narrow folksonomy: All contributing authors form the set of users, the blog posts they produce correspond to the set of resources, and the keywords they assign to them can be viewed as the set of tags.

While the tag assignment within Social Bookmarking Systems is mainly driven by the purpose of making one’s *own* resource collection accessible *for oneself* (or as (Golder and Huberman, 2006) formulated: “*users bookmark primarily for their own benefit, not for the collective good*”), the assignment of tags to blog posts probably shows some more extrinsic aspects: Because

¹⁶<http://www.stackoverflow.com>

the blog entries themselves are usually targeted towards a public audience, their annotation with keywords or labels¹⁷ can be expected to serve publicly oriented purposes as well. Among them are (i) summarization of post content, (ii) navigational support for browsing the set of posts, and (iii) the attraction of possibly interested readers via blog indexing services like Technorati¹⁸. Tag disambiguation approaches like (Si and Sun, 2009) have furthermore exploited the fact that textual content is often present within resources, which is not necessarily the case for Social Tagging Systems. Despite that, it is justified to consider keyword annotation of web logs as a special case of Social Tagging as described in Section 3.1.

A further specialized kind of web logs are *microblogging* services like Twitter¹⁹. These restrict the resource content which can be published to short text messages (constrained to e. g., 140 characters). Within these plain text snippets, simple markup conventions are used: “RT” stands for *retweet*, a word-of-mouth like mechanism used to spread messages; references to other user names are indicated by a leading @-character; and finally, a leading #-character turns individual words into so-called *hashtags*. According to (Kwak et al., 2010), the latter are “[...] a convention among Twitter users to create and follow a thread of discussion”. In an analysis by (Wagner and Strohmaier, 2010), indication of context is described as a main function of hashtags. While tripartite data models of microblogging similar to folksonomies have been proposed (Wagner and Strohmaier, 2010), it was also mentioned that “*social awareness streams*” from Twitter have a much more dynamic nature and a different type of complexity. The latter makes it difficult to directly apply methods originating from the analysis of Social Tagging Systems to microblogging data.

3.2.2. Wikis

Wikis are websites which allow easy editing and interlinking via a web browser, typically using a simplified markup language.²⁰ Wikipedia²¹ is probably the most well-known example of a community-created encyclopedia based on this technology. The idea of making its content available as structured information

¹⁷The popular web log platform *Blogspot* <http://www.blogspot.com> denotes tags as *labels*.

¹⁸<http://www.technorati.com>

¹⁹<http://www.twitter.com>

²⁰<http://en.wikipedia.org/wiki/Wiki>

²¹<http://www.wikipedia.org>

is the driving force behind the DBPedia project²². Hereby mainly pattern matching techniques are applied based among others on the article’s textual content and its structuring by templates (Auer and Lehmann, 2007; Lehmann et al., 2009).

From the viewpoint of resource annotation, another mechanism is interesting: Within Wikipedia, articles can be assigned to a *category system* by adding a “category tag” (e.g., [[Category:Science]]) which points to a category page (Voss, 2006). The latter can themselves be assigned to “super-categories”, leading to a polyhierarchical structure whose elements can be created, modified and deleted by every Wikipedia contributor. While the semantics of category relations is not explicitly defined, (Ponzetto and Strube, 2007) presented an approach to identify subsumption pairs among them using network properties and lexico-syntactic matching. In contrast to Social Tagging, the users have to agree on the set of categories for a given article. These properties position the Wikipedia category annotation system “*somewhere between indexing with a controlled vocabulary and free keywords*” (Voss, 2007).

3.2.3. Question Answering

Another participatory brand of web applications which has existed before the advent of the Web 2.0 are *Question Answering* portals (Q&A). Their basic principle is that users can ask questions on various topics, hoping to get an answer from an expert in the respective field. The motivation for answering is usually driven by some form of reputation score. Social Annotations come hereby into play for the purpose of categorizing questions: As an example, on the Q&A platform *Stackoverflow*²³, users have to assign at least one and at most five tags to each question asked. This process is hence more controlled, compared to the open paradigm of typical Social Tagging Systems. Furthermore, often there exist guidelines on which tags should be chosen (see Section 6.1.6). The choice of tags can also be expected to be extrinsically motivated, as a “good” tag choice will heighten the chance of a good answer by a suitable expert.

Hence, the annotation of questions by keywords represents a slightly more restricted variant of Social Annotation. Despite that, considering questions as resources, they fit nicely into the folksonomy model, and will hence be included in the later studies.

²²<http://www.dbpedia.org>

²³<http://www.stackoverflow.com>

3.2.4. Implicit Annotation within Logsonomies

A common characteristics of all aforementioned variants of Social Annotations is that users *explicitly* choose suitable metadata and attach it to resources. While not being a genuine Web 2.0 phenomenon, search engines on the other hand come along with a more *implicit* form of annotation. Within their log files containing queries and clicks of users, a folksonomy-like relation between users, query terms and a resources is induced when a user clicks on a specific URL after submitting a query. The resulting structure of this process, previously called *logsonomy* (Krause et al., 2008a), is a tripartite graph of a set of users, queries and clicked URLs with hyperedges, each connecting one query, one clicked URL and one specific user. Previous work (Krause et al., 2008a) revealed that the latter exhibits structural similarities to folksonomy graphs, e. g., small world properties, a power law distribution of tags and users, and a similar co-occurrence behavior of tags.

Analogous to the folksonomy model described in Section 3.1.2, a logsonomy can be more formally defined as:

Definition 3.3 A logsonomy is a tuple $\mathbb{L} = (U, T, R, Y)$ whereby

- U is the set of users of the search engine.
- T is the set of query terms contained in the queries the users gave to the search engine,
- R is the set of URLs which have been clicked by the search engine users.

Y is a subset of $U \times T \times R$. It contains a tuple (u, t, r) whenever user u clicked on resource r of a result set after having submitted the query term t (eventually with other terms).

Although logsonomies show similarities to folksonomies, (Jäschke et al., 2008b) also mentions some differences: First, there is a bias towards clicking top-ranked search results. Apart from that, “erroneous” clicks are introduced when a user is not satisfied when inspecting a particular URL and returns to the result list. Furthermore, because the applied techniques are not disclosed by the search engine operators, one can not be sure if, e. g., query expansion or reduction affects the relation between search terms and the results. Finally the process of

Table 3.2.: Overview of other forms of Social Annotation.

	<i>annotators</i>	<i>resources</i>	<i>vocabulary</i>	<i>purpose</i>
<i>blogs</i>	blog authors	blog posts	open unstructured set of keywords	organizing and categorizing posts, attraction of readers
<i>Twitter</i>	tweet authors	tweets	open unstructured set of hashtags	creating and following a thread of discussion
<i>Wikipedia</i>	Wikipedia contributors	Wikipedia articles	open polyhierarchically structured set of categories	organizing and categorizing Wikipedia articles
<i>Log-sonomies</i>	search engine users	URLs within search results	open unstructured set of query terms	n/a (implicit annotation)
<i>Social Tagging</i>	resource authors (narrow), resource collectors (broad)	various resource types	open unstructured set of tags	organizing and categorizing personal resource collection
<i>Q&A</i>	question authors	questions	open unstructured set of tags, sometimes with tagging guidelines	categorize questions, assign question to suitable experts

splitting queries and relating individual search terms to results may change the intended meaning.

Apart from that, logsonomies can still be seen as form of knowledge organization which occurs as a by-product of users' information retrieval activities.

3.3. Summary

The objective of this chapter was to familiarize the reader with Social Annotations as a means of knowledge organization, with a special focus on the characteristics and formal properties of Social Tagging Systems. Table 3.2 briefly gives an overview of the different types of Social Annotations which were introduced within this chapter. In summary, their underlying principle of indexing information resources is not fundamentally new and has existed for several years in the fields of library or information science. However, somehow similar to how the first intuitive browsers like Mosaic helped to turn the World Wide Web from a technology enthusiast playground into a mass phenomenon, Social Annotation Systems pioneered to transfer the process of information indexing from specialized domains and trained experts to a general and public audience. Despite its weaknesses, an important benefit is that Social Tagging allows to observe and learn from how users perceive and organize content (Wetzker, 2010, p. 184). Using the insights gained hereby, user-created annotations could play the role of a *bridging technology* towards more intelligent and adaptive information retrieval and management tools like the ones comprised in the vision of a *Semantic Web* described by Tim Berners-Lee (Berners-Lee et al., 2001). The following chapter introduces the core ideas of the Semantic Web as a complementary approach of organizing information resources.

Chapter 4.

The Semantic Web

Although the information on World Wide Web is essentially contained within a giant network of computers, its primary clients which access, extract, interpret and maintain information are still human users (Maedche, 2002). This is rooted in its original design as a medium of information exchange among users – using Tim Berners-Lee’s words: “*Web 1.0 was all about connecting people*”¹. Consequently, most of the communication via this novel medium was encoded in the same way as humans were used to communicate to each other, namely via natural language. This is why the Web 1.0 is also sometimes referred to as the *Syntactic Web* (Breitman et al., 2007), where computers are mainly responsible for information storage and presentation, but the semantic interpretation and hence its organization is delegated to humans. Figure 4.1 depicts the underlying structure of the Syntactic Web graph, indicating that very few machine-processable information is present.² But with the ever-growing amount and availability of online resources, the complexity of this task is growing towards becoming practically unfeasible. As an example, though search engines are of great help in finding relevant content, going through possibly lengthy result lists and judging if an element satisfies the current information need remains a tedious task. To summarize, large parts of organizing the knowledge contained in the syntactic web are left to its users.

A natural question which arises is to which extent these “intelligent” tasks can be delegated to computers. Tim Berners-Lee himself coined in a revolutionary article (Berners-Lee et al., 2001) the vision of a *Semantic Web*, in which Web resources are annotated with machine-readable metadata, enabling software agents to process them in a meaningful way. A common semantic layer defines hereby the meaning of the symbols used for exchanging information. Within

¹<http://www.ibm.com/developerworks/podcast/dwi/cm-int082206.txt>

²The illustrations of Figures 4.1 and 4.2 were inspired by <http://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#%281%29>, retrieved on 2011/08/18.

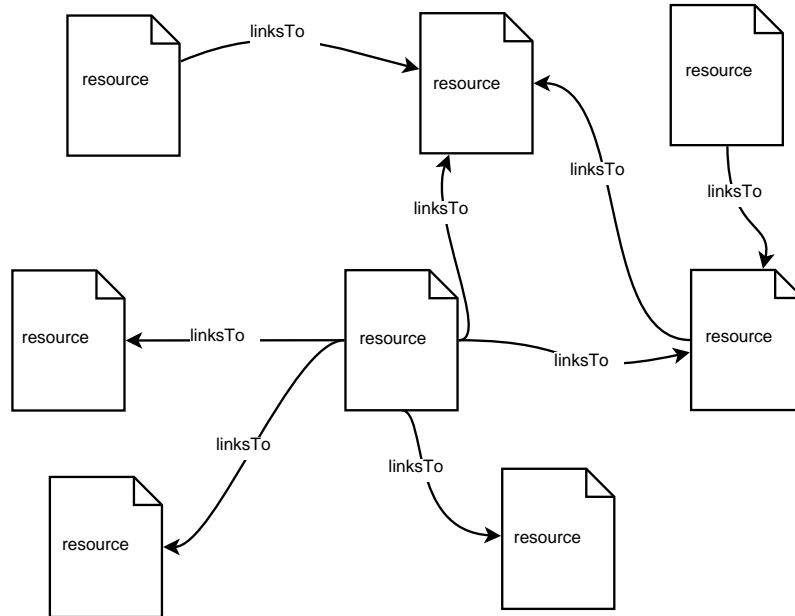


Figure 4.1.: Structure of the Syntactic Web Graph. From a conceptual point of view, all resources (e. g., websites) are equal, and the hyperlinks between them are also all of the same unspecific type.

this layer, the relevant concepts of a given domain as well as their relationship are specified by means of *ontologies* (these will be introduced in detail in Section 4.1). While the idea of facilitating Web information access by standardized categorization has been central to a number of approaches stemming from the research field of Artificial Antelligence (Breitman et al., 2007), James Hendler as another founding father of the Semantic Web points out that decentralization should be one of its core aspects:

“Instead of a few large, complex, consistent ontologies that great numbers of users share, I see a great number of small ontological components consisting largely of pointers to each other.”

(Hendler, 2001)

Figure 4.2 illustrates how this vision leads to a different *semantic* Web graph. It depicts the same Web resources as found in the Syntactic Web graph (see

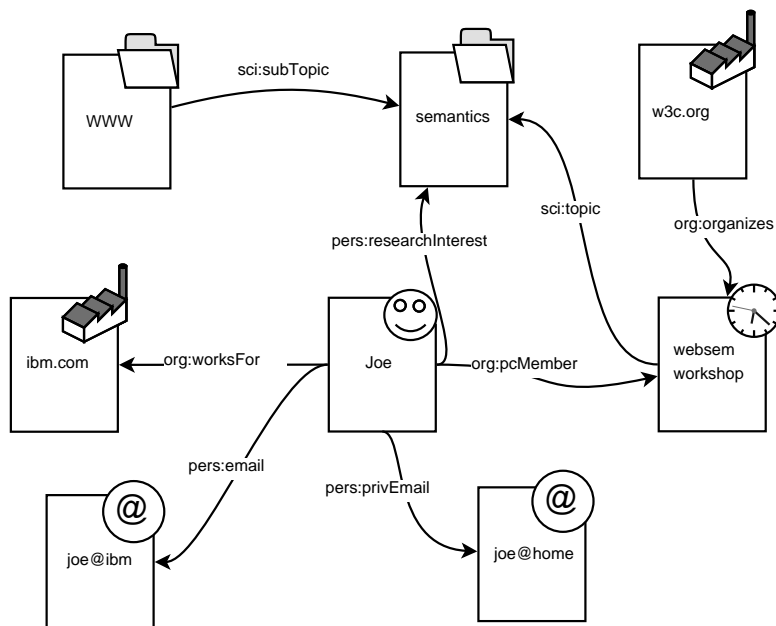


Figure 4.2.: Structure of the Semantic Web Graph. The novelty compared to the Semantic Web graph (see Figure 4.1) is that the resources as well as the links between them are annotated with metadata indicating their type. The label prefixes (*org*, *pers* and *sci*) indicate different ontologies, in which the (link) semantics is defined.

Figure 4.1). However, these are annotated using machine-readable metadata which allows to define different types of resources (highlighted as symbols in the upper right corner of each resource) and links (indicated by the link labels). Hereby it is important to notice that

- the different resources may stem from different data sources (e. g., those labelled *WWW* and *semantics* from an online taxonomy of research topics, and those labelled *w3c.org* and *ibm.com* from a business directory) and
- the semantics of the resource and link types (indicated by the label prefixes *org*, *pers* and *sci*) may be defined within different ontologies (e. g., *pers* could be an ontology describing personal attributes of users, while *org* could model business-related activities).

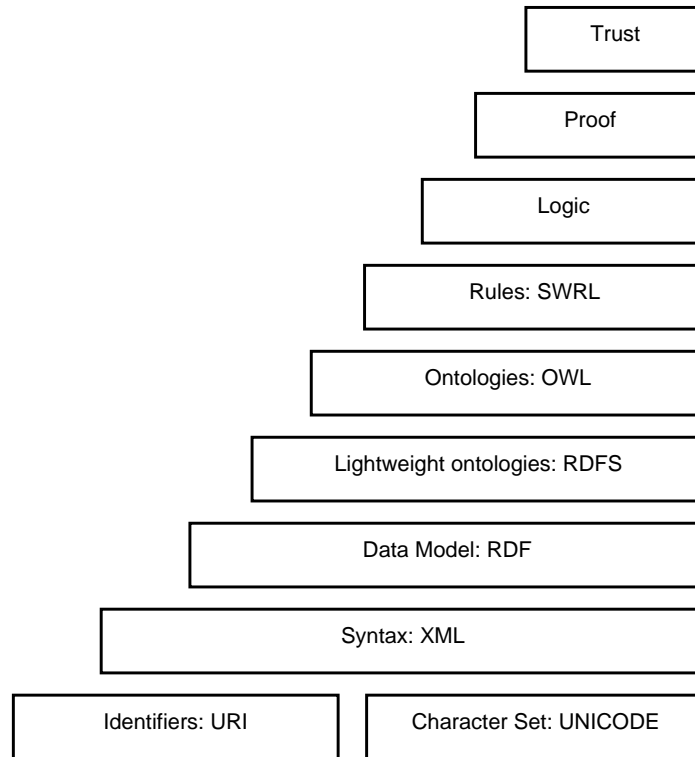


Figure 4.3.: Proposed layered architecture of the Semantic Web.

In other words, the Semantic Web does not intend to be a replacement for the Syntactic Web, but rather an extension based upon the existing infrastructure. This is also reflected in the proposed layered architecture (Berners-Lee et al., 2001) of the Semantic Web (see Figure 4.3), whose lower layers are effectively based upon existing Web technologies. For the scope of this dissertation, primarily the paradigm of knowledge organization implemented in the “Ontologies” layer is relevant. For this reason, the further technical details as well as the concrete languages used in its current implementation will be just briefly covered in the next paragraph, while ontologies will be introduced in detail in the subsequent section 4.1.

Characters, Identifiers, Syntax: Starting from the bottom, a very first prerequisite for the exchange of information with computers is that all participating parties encode and decode the binary message content in the same manner; **Unicode**³ is a standard serving this purpose. Organized in groups, planes, rows and cells, its goal is to provide a unique representation for each character used within any real-world language. Once this is accomplished, the next issue is to be able to assign unique identifiers to objects in order to be able to “talk” about the same object. This is accomplished by **Uniform Resource Identifiers** (URI), which provide “*a simple and extensible means for identifying a resource*”⁴. A commonly used subset of those are *Uniform Resource Locators* (URLs) like `http://www.kde.cs.uni-kassel.de`, which additionally encode the location where the resource can be accessed. For the case of textual resources, the **Extensible Markup Language** (XML) defines a set of syntactical rules to add additional information (called *markup*) to the resource content. This markup adds structure to text documents, which facilitates their automated and standardized processing. Up to this level, all technologies are purely syntactical and represent in fact well-established Web technologies.

Data Model, Lightweight Ontologies: While XML allows to add markup to the content of web resources, the **Resource Description Framework** (RDF) provides syntax and data model to represent additional information about resources themselves. Its basic principle is to denote this information in the form of *triples*, consisting of a subject, a predicate and an object, each being identified by a URI as described in the previous paragraph. An example is to express the affiliation of a person: Let’s say Joe works for IBM. This would be expressed by the following triple:

`http://ibm.com/joe` `http://jobs.org/worksFor` `http://ibm.com`

Please note that the above is a simplified notation – others (especially XML-based variants) exist as well. This triple can be interpreted as a directed labelled edge between two nodes (subject and object) in a resource graph. Coming back to Figure 4.2, the reader may notice that the depicted Semantic Web graph is an alternative representation of a set of triples. So far this corresponds to a further syntactic convention; however, RDF and its extension **RDF Schema**

³<http://unicode.org>

⁴<http://www.ietf.org/rfc/rfc3986.txt>

(RDFS) offer special vocabulary elements whose semantics is defined formally by means of a model theory⁵. As an example, the RDFS vocabulary allows to define classes, class hierarchies, membership within classes and the domain and range of properties. In this way, RDF and RDFS allow to specify lightweight ontologies.

Ontologies, Rules: The limited expressivity of RDFS naturally restricts its ability to model more complex knowledge domains. To facilitate that, the **Ontology Web Language** (OWL) adds further RDF vocabulary elements with defined formal semantics. As an example, it allows to define new classes by intersecting, combining or restricting existing classes. Because an enhanced expressivity needs to be traded off against a greater complexity of reasoning over the captured knowledge, OWL is available in three expressivity levels (OWL Lite \subset OWL DL \subset OWL Full). OWL DL is based on the logical formalism of Description Logics, which are decidable fragments of First Order Logic (FOL). This allows to perform inference tasks like *instance checking* (is a given individual an instance of a given class), *subsumption* (is a given concept subsumed by another concept) or *consistency* (are there contradictions within a set of statements) efficiently. As stated above, the underlying knowledge representation paradigm of ontologies will be discussed in depth in Section 4.1.

While the focus of RDFS and OWL as ontology languages is to represent knowledge, rule languages like the **Semantic Web Rule Language** (SWRL) are designed to formulate rules by which new facts can be synthesized from existing ones (Breitman et al., 2007, p.105). These rules can be, e. g., useful for defining views over ontologies or mappings between heterogeneous data sources.

Other layers: At the time of writing of this dissertation, the top three layers of the Semantic Web architecture were not realized at a comparable level to the lower ones. The idea behind the **Logic** layer is to provide a unifying logic formalism for ontologies and rules; the **Proof** layer is intended to allow the truth assessment of statements by applying rules. **Trust** finally takes furthermore the origin of statements into account to assess their trustworthiness.

Two further layers which are typically included, but were left out intentionally because they are not directly relevant within the scope of this dissertation are (i) Query languages like **SPARQL** (which offer the possibility to issue

⁵W3C Recommendation on RDF Semantics: <http://www.w3.org/TR/rdf-mt/>

queries over ontologies) and (ii) protocols of **cryptography** (which are used for authentication and authorization purposes).

After this brief introduction of the Semantic Web as a whole, the following section is concerned with ontologies as its underlying mechanism of knowledge representation.

4.1. Ontologies

Having its roots in philosophy, the term “ontology” has been used and adapted by various disciplines, and hence it is hard to formulate a generally agreed definition across all communities. Etymologically, the term stems from the greek word “*ontologia*”, composed from “*ontos*” (being) and “*logos*” (word). As a philosophical discipline, Ontology is concerned with the *science of being*. As a branch of metaphysics, a central goal of this discipline was to reason about category systems which account for a certain vision of the world (Breitman et al., 2007, p.17). While the idea of *abstraction* dates back to Platon, his student Aristotle shaped notions like *category* or *subsumption* (Cimiano, 2006) and proposed the first known category system. The latter was commented by the Greek philosopher Porphyry, who arranged the proposed categories in a tree diagram, intended to serve as a basis to classify things. In other words, the “Tree of Porphyry” can be interpreted as one of the first examples of a knowledge organization scheme.

Although the underlying motivation of establishing categories and properties to describe the world is similar, ontologies⁶ in the context of computer science are not necessarily concerned with capturing the “nature of existence” as a whole. Instead, an ontology is primarily understood as “*a formal, explicit specification of a shared conceptualisation.*” (Studer et al., 1998). The authors explain further:

“A ‘conceptualisation’ refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. ‘Explicit’ means that the type of concepts used, and the constraints on their use are explicitly defined. [...] ‘Formal’ refers to the fact that the ontology should be machine readable,

⁶Please note that we stick to the notation “Ontology” (with upper case O and without a plural) to denote the philosophical discipline, and to “ontology” for the use in the field of computer science.

which excludes natural language. ‘Shared’ reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group.” (Studer et al., 1998)

One can observe that the understanding of Ontology as a scientific discipline is replaced by regarding ontologies as resources which describe the conceptual model of a particular domain of interest by a language which specifies its relevant concepts (or classes) and relations. In line with (Corcho and Gomez-Perez, 2000), throughout the remainder of this dissertation, the terms *concept* and *class* will be used synonymously. In the previous chapter, two possible ontology languages with a well-defined syntax and formal semantics were briefly introduced, namely RDFS and OWL. While several others were proposed within the Semantic Web movement (cf. (Gomez-Perez and Corcho, 2002) for an overview), their individual characteristics and differences are not directly relevant for the scope of this dissertation. Instead, we will stick to a more mathematical notion of ontologies based on (Cimiano, 2006) and (Maedche, 2002), which is intended to paint a clearer picture of the underlying paradigm of knowledge organization. This formal model will be introduced in the following section.

4.1.1. Formal model

The following definition is an adaption from the definitions given by (Cimiano, 2006) and (Maedche, 2002) for our purposes:

Definition 4.1 *An ontology is a structure*

$$\mathcal{O} := (C, \leq_C, \text{root}_C, R, \sigma_R, \leq_R)$$

whereby

- C and R are two disjoint sets whose elements are called concept identifiers and relation identifiers, respectively,
- $\leq_C \subseteq C \times C$ is a partial order called concept hierarchy or taxonomy,
- $\text{root}_C \in C$ is a designated root element of the taxonomy \leq_C , i. e., $\forall c \in C : c \leq_C \text{root}_C$,
- $\sigma_R : R \rightarrow C \times C$ is a function called relation signature, and

- $\leq_R \subseteq R \times R$ is a partial order called relation hierarchy, whereby $(r_1, r_2) \in \leq_R$ implies $\pi_i(\sigma_R(r_1)) \leq_C \pi_i(\sigma_R(r_2)), i \in \{1, 2\}$.

For simplicity reasons, we define two projection functions $C(\mathcal{O}) = \pi_1(\mathcal{O})$ and $R(\mathcal{O}) = \pi_2(\mathcal{O})$ on set of concept and relation identifiers, respectively.

If $(c_1, c_2) \in \leq_C$ (or $c_1 \leq_C c_2$) we say that c_1 is a *subconcept* of c_2 , and c_2 is a *superconcept* of c_1 . If this is the case, and there exists no c_3 with $c_1 \leq_C c_3 \leq_C c_2$ then we denote c_1 as a *direct subconcept* of c_2 and accordingly c_2 as a *direct superconcept* of c_1 . Sub- and superrelations and their direct variants are defined analogously. Please note that the above definition differs from (Cimiano, 2006) insofar the relation signature only allows binary relations. This is of course a restriction, as in principle also higher-order relations are possible; but because these are not directly relevant for the scope of this dissertation, we will stick to binary relations only.

For simplicity reasons, we will refer to the concept identifiers and relation identifiers as *concepts* and *relations*, respectively. Because these do not necessarily need to be natural language items, often a *lexicon* is assigned to an ontology, which provides textual concept and relation labels. It is defined as follows:

Definition 4.2 A lexicon for an ontology $\mathcal{O} := (C, \leq_C, \text{root}_C, R, \sigma_R)$ is a structure

$$\mathcal{L} := (L_C, L_R, \text{Ref}_C, \text{Ref}_R)$$

whereby

- L_C and L_R are non-empty sets, whose elements are called lexical entries for concept and relations, respectively,
- $\text{Ref}_C \subseteq L_C \times C$ is a relation called lexical reference for concepts and
- $\text{Ref}_R \subseteq L_R \times R$ is a relation called lexical reference for relations.

Based on Ref_C , we define for $l \in L_C$

$$\text{Ref}_C(l) := \{c \in C : (l, c) \in \text{Ref}_C\}$$

and for $c \in C$:

$$Ref_C^{-1}(c) := \{l \in L_C : (l, c) \in Ref_C\}$$

Both functions also have a counterpart for sets L and C , respectively: $Ref_C(L) = \bigcup_{l \in L} Ref_C(l)$ and $Ref_C^{-1}(C) = \bigcup_{c \in C} Ref_C^{-1}(c)$. Ref_R and Ref_R^{-1} are defined analogously.

Similar to the convention for keywords from Social Annotation (see Section 3.1, we will use a special formatting (namely a monospace font and small capitals) when we refer to *concepts* (or classes), and another one (namely a monospace font) to refer to their *lexical reference(s)* in order to ensure a clear legibility and differentiation within running text. As an example, the concept **COMPANY** could be lexically represented by the terms **company**, **firm** or **enterprise**. As a quick reminder, if the latter term would be used in the context of Social Annotation, we would format it as **enterprise**.

While ontologies capture the conceptualization of a particular domain, and the associated lexicon provides lexical labels for concepts and relations, a *knowledge base* is used to hold information about instances of concepts and relations. In other words, the ontology itself contains (mostly) *intensional* definitions, while the knowledge base comprises (mostly) *extensional* parts, which correspond to a concrete state.

Definition 4.3 A knowledge base for an ontology $\mathcal{O} := (C, \leq_C, root_C, R, \sigma_R)$ is a structure

$$\mathcal{KB} := (I, \iota_C, \iota_R)$$

whereby

- I is a set of instance identifiers or short instances,
- $\iota_C : C \rightarrow 2^I$ is a function called concept instantiation and
- $\iota_R : R \rightarrow 2^{I \times I}$ is a function called relation instantiation.

Similar to concepts and relations, the instances do not necessarily need to be identified by natural language entities, but can be referred to by an arbitrary identifier. In order to assign textual labels to instances, usually an *instance lexicon* is defined. Although instance labels are playing a subordinate role in

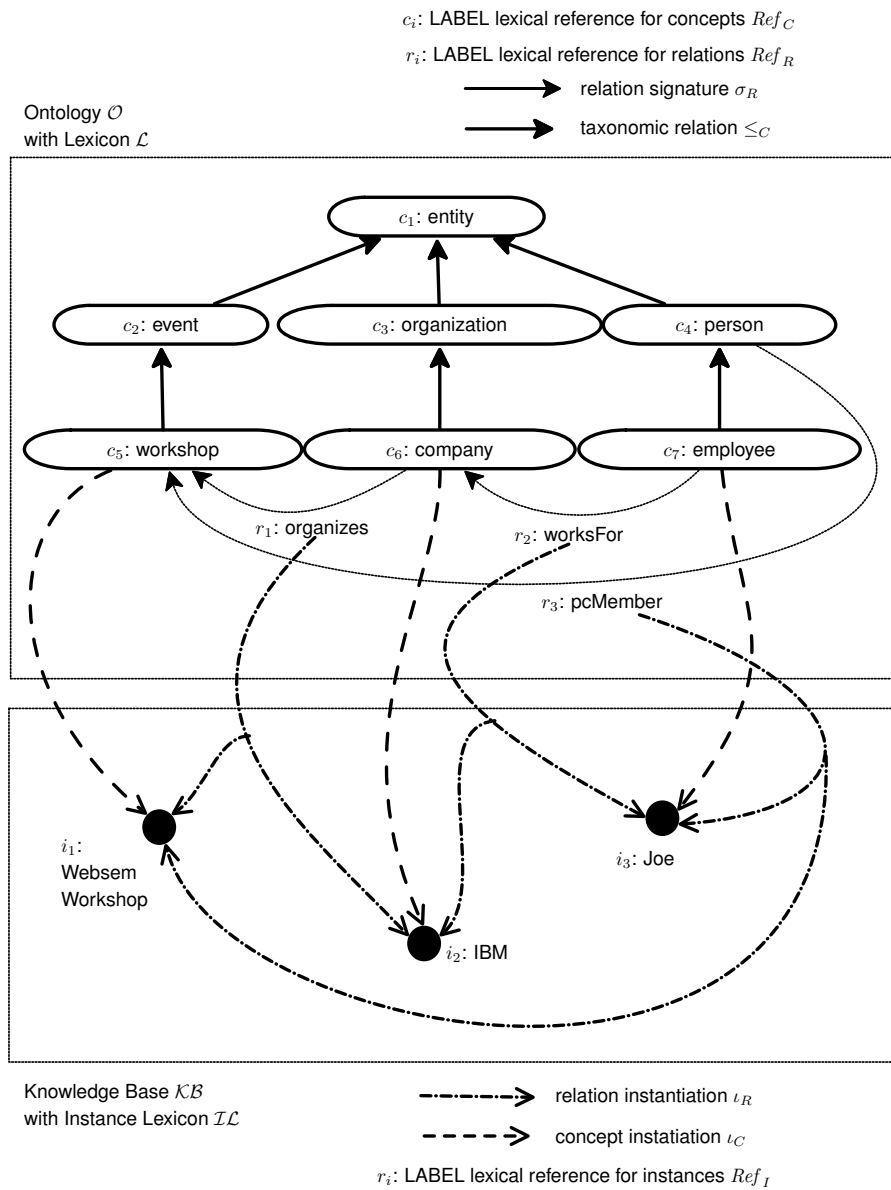


Figure 4.4.: Example ontology \mathcal{O} , including a lexicon \mathcal{L} and an associated knowledge base \mathcal{KB} with an instance lexicon \mathcal{IL} .

the context of this dissertation, we introduce the following definition for the sake of completeness:

Definition 4.4 An instance lexicon for a knowledge base $\mathcal{KB} := (I, \iota_C, \iota_R)$ is a structure

$$\mathcal{IL} := (L_I, Ref_I)$$

whereby L_I is a non-empty set, whose entries are called lexical entries for instances and $Ref_I \subseteq L_I \times I$ is a relation called lexical reference for instances.

Figure 4.4 displays graphically an example ontology \mathcal{O} , together with a lexicon \mathcal{L} and an associated knowledge base \mathcal{KB} including an instance lexicon \mathcal{IL} . The ontology models a small excerpt of concepts and relations possibly present in a scientific or business domain. Its formal representation is the following:

ontology \mathcal{O} :

$$\begin{aligned} C &= \{c_1, c_2, c_3, c_4, c_5, c_6, c_7\} \\ \leq_c &= \{(c_2, c_1), (c_3, c_1), (c_4, c_1), (c_5, c_2), (c_6, c_3), (c_7, c_4), \\ &\quad (c_5, c_1), (c_6, c_1), (c_7, c_1)\} \\ root_c &= c_1 \\ R &= \{r_1, r_2, r_3\} \\ \sigma_R : &\quad \sigma_R(r_1) = (c_6, c_5), \sigma_R(r_2) = (c_7, c_6), \sigma_R(r_3) = (c_4, c_5) \end{aligned}$$

lexicon \mathcal{L} :

$$\begin{aligned} L_C &= \{entity, event, organization, person, workshop, company, \\ &\quad employee\} \\ L_R &= \{organizes, worksFor, pcMember\} \\ Ref_C &= \{(entity, c_1), (event, c_2), (organization, c_3), (person, c_4), \\ &\quad (workshop, c_5), (company, c_6), (employee, c_7)\} \\ Ref_R &= \{(organizes, r_1), (worksFor, r_2), (pcMember, r_3)\} \end{aligned}$$

knowledge base \mathcal{KB} :

$$\begin{aligned} I &= \{i_1, i_2, i_3\} \\ \iota_C : \quad &\iota_C(c_5) = i_1, \iota_C(c_6) = i_2, \iota_C(c_7) = i_3 \\ \iota_R : \quad &\iota_R(r_1) = (i_2, i_1), \iota_R(r_2) = (i_3, i_2), \iota_R(r_3) = (i_3, i_1) \end{aligned}$$

instance lexicon \mathcal{LI} :

$$\begin{aligned} L_I &= \{WebsemWorkshop, IBM, Joe\} \\ Ref_I &= \{(WebsemWorkshop, i_1), (IBM, i_2), (Joe, i_3)\} \end{aligned}$$

4.1.2. Classifying Ontologies

The formal model introduced in the previous section is a generic one and is hence valid for all kinds of ontologies. However, due to great number and variety of ontologies, it would be illusory to treat all of them in the same manner. For the purpose of distinguishing among them, several dimensions of comparisons were proposed in the literature. The following summary is largely based on (Breitman et al., 2007, p.26ff).

Semantic Spectrum: A first differentiation can be made by the “complexity” of semantics which can be captured within an ontology (Uschold and Grüninger, 2004; McGuinness, 2003). Essentially, ontologies serve the purpose of assigning meaning to terms, whereby different ontology languages offer different possibilities and levels of expressiveness. The comparison of these languages based on their degree of formality as done in Figure 4.5 illustrates that these lead to a continuum of ontology types. Starting from the left, one can find languages which merely allow the definition of terms, offering little or no support to specify their meaning. On the opposite end of the scale, logical languages are present which are capable of formulating strictly formal logical theories (Guarino et al., 2009). When moving from left to right, the complexity and amount of meaning which can be captured is growing, together with the level of formality. However, the more complex meaning constructs can be formulated, the more complex becomes reasoning over the captured knowledge. Description Logics offer a good trade-off to this end; this is why they form the basis of many Semantic Web ontologies. Though it is difficult to draw a clear border for the criterion of

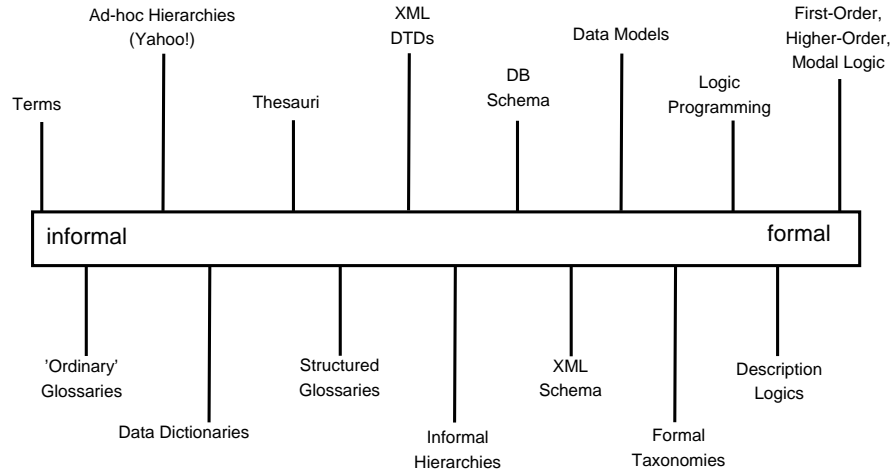


Figure 4.5.: Classification of Ontology languages according to (Guarino et al., 2009) by to their semantic spectrum.

a “formal” specification, in practice these two rightmost categories are usually considered as such.

Generality: Another distinction introduced by (Guarino, 1998) is based on the generality of the captured knowledge. Starting from the most general category, *Upper Level ontologies* are intended to model generic and domain-independent concepts such as space, time or action. The main purpose of these ontologies is to serve as a basis for interoperability in large user communities and for refinement in other ontologies. On the next level of generality, one can find *Domain ontologies* and *task ontologies*. These are specializing the concepts present in top-level ontologies focused on a specific domain or generic tasks or activities. Finally, the most specific kind are *application ontologies*, whose goal is to describe the vocabulary relevant to a particular application in the context of a given domain or task.

Further proposed ontology classification schemes were based on the type of information represented (Gómez-Pérez et al., 2004), the focus on formal, terminological or prototypical representation (Biemann, 2005) and others (Maedche, 2002; Omelayenko, 2001). But for the purpose of laying the groundwork for understanding which kind of ontologies are suitable and feasible to represent

III AMPHIBIA.			IV PISCES.			
Corpus nudum, vel squamosum. Dentes molares nulli: reliqui semper. Pinnae nullae.			Corpus apodum, pinnis veris instructum, nudum, vel squamosum.			
SERPENTIA	Tortu.	Corpus quadrupedum, caudatum, testa munitum.	Tortu. testularia. • • • terrestria. • • • marina.	Thrichechus.	Dentes in utraque maxilla. Dorsum impenne.	Masatus f. Panna mar.
	Rana.	Corpus quadrupedum, cauda desiliatum, squamea creta.	Bufo. Rana arborea. • • • aquaticae. • • • Carolinae.	Catodon.	Dentes in inferiore maxilla. Dorsum impenne.	Cet. Fittule in rostro. <i>Art.</i> Cete <i>Clasf.</i>
	Lacerta.	Corpus quadrupedum, caudatum, squamosum.	Crocodilus. Alligator. Cordylus. Draco volans. Scincus. Salamandra aq. • • • terrestria. Chamaeleo. Seps. Sternobi <i>Arg.</i>	Monodon.	Dens in superiore max. 4. Dorsum impenne.	Monoceros. <i>Unicornus.</i>
	Anguis.	Corpus apodum, teres, squamosum.	Salamandra aq. • • • terrestria. Chamaeleo. Seps. Sternobi <i>Arg.</i> Vipera. Cuculla. Aphis. Caudifoss. Cobras de Cabelo. Anguis <i>Africana</i> . Cenchris. Natrix. Hydrus.	Balena.	Dentes in sup. max. cornes. Dorsum saepius impenne.	B. Groenland. B. Finfich. S. Maxill. inf. latiore. <i>Art.</i>
				Delphinus.	Dentes in utraque maxilla. Dorsum pinnatum.	Orcha. Delphinus. Phocena.
				Raja.	Foramina branch. utriusq. 5. Corpus depectum.	Raja clav. sfp. lev. etc. Squasmo-Raja. Alaveta. Pinnaca mar. Aquila. Torpedo. Bos <i>Fos.</i>
				Squalus.	Foram. branch. utriusq. 5. Corpus oblongum.	Lamia. Galeas. Catulus. Vulpes mar. Zygma. Squatina. Centrina. Prilla.
				Acipenser.	Foram. branch. utriusq. 7. Os edentul. tubulatum.	Sturio. Hufo. <i>Ichthyocolla.</i>
				Petro-myzon.	Foram. branch. utriusq. 7. Corpus bipenne.	Encephthalmus. Lampetra. Mulleta.

Figure 4.6.: Excerpt of the Linnaean taxonomy of biological species. Picture is taken from http://psnt.net/blog/wp-content/uploads/2010/04/Linnaeus_-_Regnum_Animale_1735.jpg

emergent semantics from Social Annotation Systems, the aforementioned categorization dimensions suffice. As will be explained in Section 5.2, these will mainly be domain and application ontologies. Because taxonomies as the conceptual backbone are playing hereby an important role, these will be discussed in greater detail in the next chapter, together with thesauri as a closely related approach.

4.1.3. Taxonomies and Thesauri

Similar to the case of “Ontology” and “ontologies”, *Taxonomy* as a scientific discipline is concerned with all aspects of classification, i. e., the organization of objects by an assignment into a (mostly hierarchical) arrangement of classes. Literally translated from its Greek origin, it means “method of arrangement”. A *taxonomy* on the other hand corresponds to a particular classification of objects belonging to a certain domain of interest (e. g., the taxonomy of mammals). Historically, taxonomies were in fact first used in biological contexts to classify living things – a famous example is the Linnaean taxonomy (Figure 4.6 depicts an excerpt). While being used in the sequel within a broad variety of domains, the underlying principle remained the same: Starting in a top-down manner from a “root” class (denoted by $root_C$ in the ontology model found in Section 4.1.1)

which contains all objects, these are subsequently divided into subclasses based on common properties. In a strict sense, the subclass / superclass relationship corresponds to a refinement in the sense that objects belonging to the subclass (e. g., **MAMMALS**) exhibit the same properties as the ones from the superclass (e. g., **VERTEBRATES**), plus additional ones which allows to distinguish them from the latter. This conceptual generalization-specialization semantics of the taxonomic relation hence corresponds to concept subsumption, also denoted as *is-a* or *type-of* relationship. From a linguistic point of view, this relation corresponds to hyponymy. And with respect to instances, the taxonomic relation implies a directed instance inheritance from subclasses to superclasses: Sticking to the example, each object which is an instance of the class **MAMMALS** also is an instance of the class **VERTEBRATE**. All these properties are characteristic of *formal* taxonomies (cf. Figure 4.5), which are sometimes denoted to exhibit “strong semantics” (Breitman et al., 2007).

Apart from that, there exist also taxonomies whose hierarchy encode a less strict *parent-child* relationship. Although this informal use does not necessarily correspond to a generalization relationship, the aggregation of related concepts under a common class can still be useful for organizing resources. An example is the class hierarchy

COMPUTERS > INTERNET > SEARCHING

found on the user-created Web directory *DMOZ*⁷. The reader will instantly notice that **SEARCHING** is not a “type of **INTERNET**”, and **INTERNET** is not a “type of **COMPUTERS**”. However, a possible observation is that the Internet plays an important role for a large number of Computer users, which makes it an important “sub-aspect” of Computers as such. The same holds for the relationship between **INTERNET** and **SEARCHING**. The precise semantics of this relationship may be hard to capture and vary across and even within such informal taxonomies; one could find here *part-of*, *cause-effect*, *association* or *localization* meanings, just to name a few (Breitman et al., 2007). An analysis by (Veres, 2006) related these mixed semantics to so-called *Wierzbicka categories*, which are strictly speaking not taxonomic as such, but are often based, e. g., on a common function (e. g., **WEAPONS**) or a functional or origin collocation (e. g., **FURNITURE**, **GARBAGE**). The authors identified many examples of Wierzbicka categories within DMOZ. On the semantic spectrum from Figure 4.5, such structures can be found on a range between ad-hoc hierarchies and informal

⁷<http://www.dmoz.org/Computers/Internet/Searching/>, retrieved on 2011/08/25

taxonomies.

A *thesaurus* can basically be regarded as an extension of a taxonomy by a limited set of well-defined additional relationships among classes. The underlying taxonomy can hereby adhere to formal or informal semantics. These hierarchical relationships are usually encoded by prefixes like *BT* (broader than) and *NT* (narrower than). Apart from these, one can typically find (i) *associative* relations *RT* (related term) and (ii) *Equivalence* relations *UF* (used for) to denote synonym or quasi-synonym relations. A further information which is encoded hereby are “preferred” terms, which indicate a standard way to refer to a particular class.

Finally, the question remains how one can distinguish taxonomies and thesauri from ontologies. The following three criteria are mentioned by (Breitman et al., 2007):

- *Formal is-a hierarchy*: While taxonomies and thesauri may contain informal hierarchical relationships, an ontology must strictly adhere to the generalization-specialization semantics of the taxonomic relation.
- *Ambiguity-free interpretation of relationships*: The fuzziness present, e. g., in associative thesaurus relationships is not allowed within an ontology, where the semantics of relationships need to be clearly and unambiguously defined.
- *Vocabulary*: The vocabulary used to specify an ontology needs to be controlled and finite, but extensible.

4.1.4. Strengths and weaknesses

While taxonomies and thesauri are classical approaches to knowledge organization and have existed since a long time, the sheer amount and the dynamics of Web knowledge resources make their straightforward deployment difficult. Ontologies with their enhanced expressiveness have been envisioned by the Semantic Web community to serve as a better basis for structuring Web contents. This vision is mainly nourished by the following strengths of ontologies as knowledge organization formalism⁸:

⁸The enumeration of strengths and weaknesses is based on (Benz, 2007)

- **High precision:** Because domain experts usually craft an ontology using the high degree of expressiveness available in many ontology languages, classes and relations can be precisely specified.
- **Avoidance of ambiguity:** The usage of an ontology specification language with a controlled vocabulary and formally defined semantics makes it possible to avoid ambiguity of concepts or relations.
- **Creation of context:** The (hierarchical) relation structure among classes provides a considerable amount of context, which can be of great help to users interacting with specific classes or instances.
- **Transferability:** Knowledge captured in ontologies is usually meant to be stable over time and different contexts. Together with the usage of a “lingua franca”, ontologically represented knowledge is hence transferable over cultural, temporal and language barriers.

Despite these obvious advantages, to the time of writing of this dissertation, ontologies have not yet fully found their way into all critical Web applications like search engines or knowledge management systems. The following reasons are named in the literature as impeding factors for their widespread adoption:

- **Required Expertise:** The establishment and maintenance of an ontology requires both domain expertise and thorough understanding of ontological techniques. This makes the whole process expensive and thus problematic for large-scale deployment.
- **Metadata annotation bottleneck:** Even when a stable ontology is set up, the sheer mass of information resources on the Web makes their exhaustive annotation with metadata a very time-consuming and practically often unfeasible task.
- **Inflexibility:** With the precision of class and relation definitions, the degree of rigidity increases, as each modification of the scheme needs the review of a central authority. This makes an ontological scheme inflexible with respect to fast changing organization needs.
- **Creator bias:** It has been argued that the establishment and maintenance of an ontology is influenced by subjectivity and cultural background of the

involved people (sometimes called *cataloguers*) (Mathes, 2004). Especially (Shirky, 2005) argued in addition that “*there is no perfect organization scheme*” due to context errors. The author furthermore pointed out that the establishment of an ontological organization scheme requires the cataloguer to *mind-read* the target audience in order to come up with a scheme everybody will agree on. The last problem mentioned is that *future-telling* is necessary to reach a scheme that will be stable over time.

The alert reader will have noticed that these strengths and weaknesses are in a way inverse to those of Social Annotations as presented in Section 3.1.1: While, e. g., folksonomies suffer from problems which ontologies were explicitly designed to eliminate (e. g., ambiguity or lack of precision), their ability to scale and involve large user populations is an advantage hardly reachable by current Semantic Web approaches. Section 5.1 will detail further on the potential benefits and synergies from a combination of both approaches.

4.2. Derived Measures

Besides the explicitly encoded information about classes and their relations within taxonomies, thesauri or ontologies, their internal structure can be exploited to derive further information which is present in a more implicit manner. As an example, even if the two classes **TABLE** and **CHAIR** are not connected by a direct relationship, it is still possible to infer an indirect one, e. g., by observing that they are both subsumed under the class **INTERIOR** in the underlying taxonomy. Measures which are based (among others) on this kind of information are denoted to capture *semantic relatedness* among concepts. These will be described in Section 4.2.1.

Another aspect which is encoded mainly within the taxonomic relation is the notion of *semantic generality*. Intuitively, when humans are asked which of the classes **COMPUTER** or **NOTEBOOK** is more general, most people will probably judge **COMPUTER** as more general. With high probability, this information is encoded directly within a technological taxonomy. On the contrary, when asked if **NOTEBOOK** or **TEXT PROCESSING SOFTWARE** is more general, most people might choose **NOTEBOOK**, despite these two concepts will probably not be related by a (strict) taxonomic relation. Measures of semantic generality have received much less attention compared to measures of semantic relatedness; however, Section 4.2.2 provides a formal definition and summarizes existing approaches

in order to lay the groundwork for the methods of making emergent semantics explicit presented in Chapter 7.

4.2.1. Semantic Relatedness

Measuring the degree of semantic relatedness among concepts has its roots in the field of natural language processing. In this context, tasks like word sense disambiguation, text summarization or spell checking are applications which benefit strongly from relatedness measures. While the notion of “semantic similarity” has also been used partially, we stick to the term “relatedness”, because Budanitsky and Hirst (2006) pointed out that similarity can be considered as a special case of relatedness. Essentially, measures of semantic relatedness assign a relatedness score to pairs of concepts, as can be seen from the following definition:

Definition 4.5 *A Semantic Relatedness Measure based on a set of concepts C is a function*

$$\rho_C : C \times C \rightarrow \mathbb{R}^+$$

For two concepts $c_1, c_2 \in C$ we denote $\rho_C(c_1, c_2)$ as the semantic relatedness of c_1 and c_2 . The higher the value of $\rho_C(c_1, c_2)$ is, the stronger is their semantic relatedness. The possible values are restricted by a defined maximum value \max_{ρ_C} , i. e.,

$$\forall c_1, c_2 \in C : \rho_C(c_1, c_2) \leq \max_{\rho_C}.$$

Furthermore it must hold that all concepts are maximally semantically related to themselves:

$$\forall c \in C : \rho_C(c, c) = \max_{\rho_C}$$

If two concepts c_1, c_2 are semantically unrelated, their value of semantic relatedness is $\rho_C(c_1, c_2) = 0$. Sometimes, the inverse concept of *semantic distance* is used. It differs from measures of semantic relatedness insofar that a *smaller* distance value corresponds to a higher degree of relatedness. In any case, a measure of semantic distance $\rho_C^{-1}(c_1, c_2)$ can be transformed into a measure of relatedness by the a simple inversion according to:

Table 4.1.: Exemplary results of computing semantic relatedness based on WordNet. All concepts are described by are English nouns, and are used in the following senses (taken from WordNet):

TABLE: “a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs”

DESK: “a piece of furniture with a writing surface and usually drawers or other compartments”

CHAIR: “a seat for one person, with a support for the back”

GARDEN: “a plot of ground where plants are cultivated”

SOFTWARE: “written programs or procedures or rules and associated documentation pertaining to the operation of a computer system and that are stored in read/write memory”

<i>concept pair</i>	<i>short. path</i>	<i>Hirst St-Onge</i>	<i>Leacock Chodorow</i>	<i>Lin</i>	<i>Resnik</i>	<i>Jiang Conrath</i>
table – desk	0.5	4	2.99	0.93	7.41	0.91
table – chair	0.25	5	2.3	0.81	6.19	0.34
table – garden	0.07	0	1.12	0.14	1.17	0.07
table – software	0.06	0	0.91	0	0	0.05

$$\rho_C(c_1, c_2) = \begin{cases} \frac{1}{\rho_C^{-1}(c_1, c_2)} & \rho_C^{-1}(c_1, c_2) > 0 \\ \max_{\rho_C} & \text{otherwise} \end{cases}$$

Analogously, a measure of semantic distance can be transformed in to a measure of semantic relatedness. As a basis for computation, often WordNet as a structured lexicon of the English language is used (see Section 6.2.1). Table 4.1 contains exemplary results for a set of word pairs, whose semantic relatedness was captured by different measures⁹. Most of the measures take into account graph-based measures of distance between the two concepts in WordNet’s taxonomy; some of them (e. g., Resnik, Jiang/Conrath) also take into account information-theoretic aspects like the information content of a concept, estimated by its occurrence frequency in a reference corpus. See (Budanitsky and Hirst, 2006) for a detailed explanation of the measures. Budanitsky and Hirst

⁹The values were computed using Ted Pedersen’s WordNet::Similarity library (<http://search.cpan.org/dist/WordNet-Similarity/>)

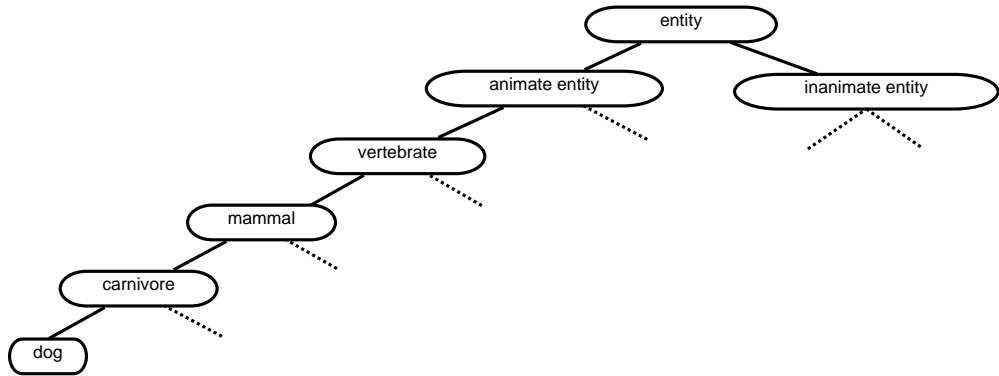


Figure 4.7.: Taxonomy excerpt to exemplify semantic generality of concepts.

also evaluated these measures in the context of a spell-checking application, and found that the Jiang-Conrath distance (which is transformed to a relatedness measure in the rightmost column of Table 4.1) performed best.

Finally, there exist also measures of relatedness which are not based on strongly structured resources like taxonomies, thesauri or ontologies, but on less controlled ones like text corpora (Mohammad and Hirst, 2006), Wikipedia (Strube and Ponzetto, 2006) or even information retrieved via web search engines (Cilibrasi and Vitanyi, 2007). These capture often *co-occurrence* or *distributional* similarity among words. From a linguistic point of view, these two families of measures focus on orthogonal aspects of structural semiotics (de Saussure, 1916; Chandler, 2007). The co-occurrence measures address the so-called syntagmatic relation, where words are considered related if they occur in the same part of text. The contextual measures address the paradigmatic relation (originally called associative relation by Saussure), where words are considered related if they can replace one another without affecting the structure of the sentence. The latter also adopt the *distributional hypothesis* (Firth, 1957; Harris, 1968), which states that words found in similar contexts tend to be semantically similar. The relation between distributional similarity and semantic relatedness is also discussed by (Budanitsky and Hirst, 2006).

4.2.2. Semantic Generality

In a similar way as taxonomies form the backbone of many approaches to compute semantic relatedness, their generalization/specialization semantics also allows to make assessments about the *semantic generality* of a concept. If two concepts are taxonomically related, i. e., $(c_1, c_2) \in \leq_C$, this typically implies that c_2 is *more general* than c_1 . This is the case, e. g., for the two concepts **MAMMAL** and **VERTEBRATE** in the sample taxonomy excerpt shown in Figure 4.7. However, one could argue that this taxonomy encodes further generality information besides the explicitly encoded one: Intuitively, most people would probably agree that **INANIMATE ENTITY** is a more general concept than **DOG**, even though they are not connected by a direct taxonomic relation.

From a completely different point of view, the question of which factors determine the generality of natural language terms has been addressed by researchers coming from the areas of Linguistics and Psychology. The psychologist Paivio et al. (1968) published in 1968 a list of 925 nouns along with human concreteness rankings; an extended list was published by Clark and Paivio (2004). Kammann and Streeter (1971) compared two definitions of word abstractness in a psychological study, namely imagery and the number of subordinate words, and concluded that both capture basically independent dimensions. Allen and Wu (2010) identified the generality of texts with the help of a set of “reference terms”, whose generality level is known. They also showed up a correlation between a word’s generality and its depth in the WordNet hierarchy. In their work they developed statistics from analysis of word frequency and the comparison to a set of reference terms. In (Zhang, 1998), Zhang makes an attempt to distinguish the four linguistic concepts fuzziness, vagueness, generality and ambiguity.

To clarify the meaning of semantic generality for the context of this dissertation, we define the following:

Definition 4.6 A concept generality measure $\sqsubseteq^{\mathcal{O}}$ based upon an ontology \mathcal{O} is a partial order among the concepts C present in \mathcal{O} , i. e.,

$$\sqsubseteq^{\mathcal{O}} \subseteq C(\mathcal{O}) \times C(\mathcal{O})$$

If $(c_1, c_2) \in \sqsubseteq^{\mathcal{O}}$ (or $c_1 \sqsubseteq^{\mathcal{O}} c_2$) we say that c_2 is equally or more general than c_1 . $\sqsubseteq^{\mathcal{O}}$ is reflexive

$$\forall c \in C(\mathcal{O}) : (c, c) \in \sqsubseteq^{\mathcal{O}},$$

antisymmetric

$$\forall c_1, c_2 \in C, c_1 \neq c_2 : (c_1, c_2) \in \sqsubseteq^{\mathcal{O}} \Rightarrow (c_2, c_1) \notin \sqsubseteq^{\mathcal{O}}$$

and transitive:

$$\forall c_1, c_2, c_3 \in C(\mathcal{O}) : (c_1, c_2) \in \sqsubseteq^{\mathcal{O}} \wedge (c_2, c_3) \in \sqsubseteq^{\mathcal{O}} \Rightarrow (c_1, c_3) \in \sqsubseteq^{\mathcal{O}}$$

In the literature, implementations and approaches to measure semantic generality (besides via the direct taxonomic relationship) are scarcely found. As mentioned above, the depth in the WordNet taxonomy has shown to be an indicator of concept generality (Allen and Wu, 2010); another intuition comes from (Kammann and Streeter, 1971), who stated that “*the abstractness of a word or a concept is determined by the number of subordinate words it embraces[...]*”. These two approaches are examples of *generality ranking functions*

$$\gamma_C : C(\mathcal{O}) \rightarrow \mathbb{R}^+$$

which assign a real value to a concept (e. g., the length of the shortest path from a given concept to the taxonomy root, or the size of the taxonomy subgraph rooted a given concept). It is clear that a ranking function γ_C induces a concept generality measure according to

$$(c_1, c_2) \in \sqsubseteq^{\mathcal{O}} \Leftrightarrow \gamma_C(c_1) \leq \gamma_C(c_2)$$

The resulting measure will be denoted as $\sqsubseteq_{\gamma_C}^{\mathcal{O}}$. Please note that all generality measures based on real-value ranking functions are by construction total orders, but this is not mandatory. Later in Section 7.3.1 we will analyze which kinds of ranking functions come close to what humans actually perceive as semantically “general”. The next section will summarize the main aspects of the Semantic Web which were discussed so far.

4.3. Summary

This chapter was intended to familiarize the reader with the concepts and paradigms of knowledge organization within the Semantic Web. After a brief introduction in its layered architecture, which consists essentially in an extension of the existing Web infrastructure, the approach of formalizing knowledge by defining classes and relations among them within an ontology was introduced. A formal model of ontologies was given, together with the introduction of lexicons (containing lexical information about concepts) and knowledge bases (containing instances of classes and relations). As a next step, two dimensions to differentiate ontologies were presented, namely their semantic spectrum and their level of generality. Additionally, ontologies were discriminated from the related concepts of taxonomies and thesauri, with a special focus on formal and informal taxonomic relationships. As a next step, strengths and weaknesses of the “ontological” way of knowledge organization were discussed. Finally, semantic relatedness as well as semantic generality were introduced as sources of additional information which can be derived from knowledge captured within an ontology.

Having introduced two “global players” of knowledge organization on the Web so far (namely Social Annotations and the Semantic Web), a justified preliminary conclusion is that none of them alone can be seen so far as the ultimate solution to bring order into the masses of Web content. In the literature, these two classes of approaches are often not seen as competing, but rather than “*flip-sides of the same coin*” (Mika, 2005). The next chapter will detail on the idea of creating synergies by “combining the best from both worlds” by establishing connections between them.

Chapter 5.

From Social Annotations to the Semantic Web

Since the idea of the Semantic Web was initiated by the visionary article of Tim Berners-Lee in 2001 (Berners-Lee et al., 2001), it has attracted great interest in both academic and corporate contexts. This is reflected in a number of successful conference series¹ and company foundations². Despite that, its outreach has not yet pervaded the experience of large user populations on the Web at a degree comparable with major applications like keyword-based search engines, online social networks or even Social Annotation Systems. In Section 4.1.4, high entry barriers and inflexibility were named among potential reasons which hampered mass adoption of semantic applications for knowledge organization purposes.

On the other hand, Social Annotation Systems with their highly flexible and easy-to-use usage characteristics attracted millions of Web users within short periods of time.³ While early advocates of social tagging like (Shirky, 2005) interpreted this as evidence that the ontological approach was “overrated”, the following years showed that the deficiencies like ambiguity or lack of precision of Social Annotation Systems (cf. Section 3.1.1) effectively hampered interoperability and retrieval mechanisms – which are two crucial aspects of knowledge

¹As an example, since 2004 there exists the *European Semantic Web Conference* (later *Extended Semantic Web Conference*) ESWC (cf. <http://www.eswc2011.org/content/history>, retrieved on 2011/08/27). Since 2002, there exists the *International Semantic Web Conference* ISWC (cf. <http://iswc.semanticweb.org/>, retrieved 2011/08/27).

²As an example, the *Semantic Web Company* www.semantic-web.at provides professional services related to the Semantic Web; *Ontoprise* <http://www.ontoprise.de> offers among others professional knowledge management solutions based on semantic technologies.

³Roughly during their third year after foundation, Delicious reported to have tripled their number of users to one million, see <http://blog.delicious.com/blog/2006/09/million.html>, retrieved on 2011/08/27. Roughly 4 years after the foundation of Flickr, it was reported that 3 billions of images were shared over this portal, see <http://blog.flickr.net/en/2008/11/03/3-billion/>, retrieved on 2011/08/27.

organization systems. In his keynote speech at ESWC 2011⁴, James Hendler retrospectively criticized that “*Tagging has largely failed to meet its promise*”.

Apart from their individual criticisms, the idea of combining the best from both worlds by “*bridging the gap*” (Hotho and Hoser, 2007) was seen as a promising direction towards an augmented kind of knowledge organization for the Web, sometimes denoted as *Web 3.0* (Hengartner and Meier, 2010). While approaches in this direction concern a variety of applications like Wikis or search engines, the focus of this dissertation is how to establish connections between Social Annotations and ontologies as the “heart” of knowledge representation in the Semantic Web.

In the first section of this chapter, a motivation for this combination will be given by covering general aspects which show up potential synergies. In the following section, the presence of *emergent semantics* within social annotation systems is reported, and methods of *ontology learning* are described as potential mechanisms to make the emergent semantic structures explicit. Hereby also a comparison to ontology learning from textual data sources is included. In the context of this section, the precise research problem addressed within this dissertation will be formulated. In the sequel, the state of the art in related approaches will be described, structured along the different tasks involved. Because a crucial question hereby is how to assess the quality of the learned semantic structures, the chapter closes with an overview of evaluation methods.

5.1. General Aspects

Within the previous chapters, a groundwork for the understanding of the two concepts of Social Annotations and the Semantic Web as approaches of Knowledge Organization on the Web was laid. Based on that, this section aims to highlight the core differences, intended to show up aspects of synergies which potentially arise from merging both paradigms. The first dimension is the comparison of bottom-up and top-down structuring, including an identification of explicit and implicit concept representations. Hereby, the necessity to differentiate between classes and categories becomes visible, and will be treated in the subsequent sections. In order to recapitulate the “inverse” relation of strengths and weaknesses of Social Annotations and ontologies, a brief summarization and comparison is given as a last general aspect.

⁴<http://www.slideshare.net/jahendler/why-the-semantic-web-will-never-work>

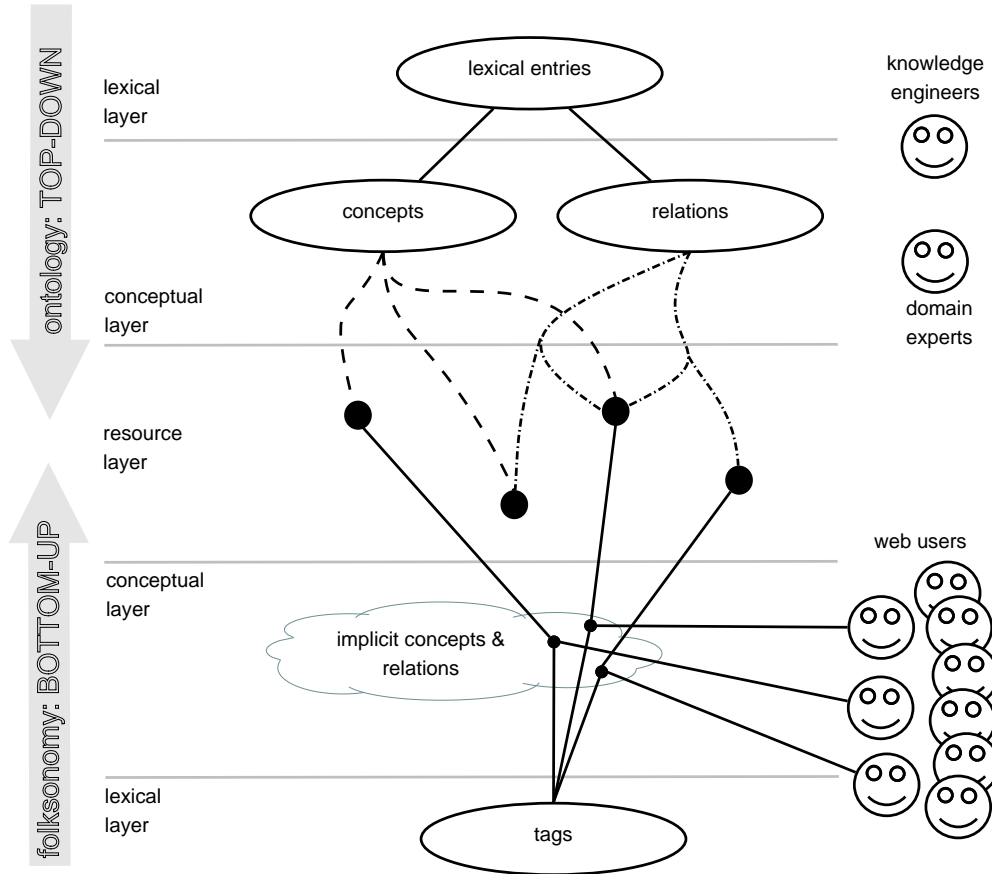


Figure 5.1.: Comparison of top-down and bottom-up approaches of annotation. Each black circle in the resource layer corresponds to a Web resource (e. g., a website). The upper and lower half of the figure visualize the top-down and bottom-up approach of (semantic) metadata annotation, respectively.

5.1.1. Bottom-up vs. Top-down

A core difference in the paradigms of social and semantic annotation has often been seen in the “direction” in which they approach the structuring of knowledge resources (Eda et al., 2009; Quintarelli et al., 2007; Zhang et al., 2006).

Because ontologies need to be constructed *before* resources can be annotated with the defined concepts, the Semantic Web has been regarded as a *top-down* methodology. This view also stems from the hierarchical decomposition of a domain of interest within taxonomies, starting from a root class at the top which subsumes all other classes. It was pointed out (Shirky, 2005) that this works especially well in domains which are characterized by a relatively small corpus of stable and restricted entities. Because the identification of relevant classes and relations usually requires a sufficient degree of domain knowledge, this task is typically performed by users with the corresponding qualification. In addition, the semantic technologies themselves require at least a basic understanding of the underlying concepts and tools. Both leads to a *knowledge acquisition* and *annotation bottleneck*, because the comparatively high requirements limit the number of potential knowledge engineers and annotators. The summary of this top-down approach is depicted in the upper half of Figure 5.1. It is important to notice hereby that the meaning of the terms present in the lexical layer (topmost part of Figure 5.1) is specified by *explicit* concepts and relations within the conceptual layer.

Social Annotations on the other hand do not require a well-structured and predefined vocabulary. Their simplicity is rooted in the possibility that users can assign arbitrary keywords in an uncontrolled manner to resources of their choice. The meaning of these tags is not formally specified by explicit concepts (as it is the case for ontologies), but requires human interpretation. As an example, if a user assigns the keyword *paper* to a resource, only the interpretation based on a sufficient amount of context (e. g., the contents of the resource, or other keywords used by this user) allows possibly to judge whether it is about a scientific paper or paper as a material. So at first sight, the inherent uncontrolled nature of free tagging seems to give rise to a rather chaotic overall structure. However, analyses of tagging networks like (Golder and Huberman, 2006) interestingly reported stable patterns and a “*nascent consensus*” on appropriate descriptions. This phenomenon was attributed to *emergent semantics* (described in greater detail in Section 5.2.1), mainly due to imitation effects and shared background knowledge among participants. So in summary, it seems to be the case that *implicit* concepts and relations are present in the aggregated annotation data, which form a kind of conceptual layer (see Figure 5.1). Because the semantics “crystallizes” *after* the annotation with lexical items, this opposite direction (compared to the ontological approach) has been denoted as a *bottom-up* methodology.

Table 5.1.: Comparison of Categorization and Classification as methods of organizing information.

<i>Aspect</i>	<i>Categorization</i>	<i>Classification</i>
<i>Process</i>	Creative synthesis of entities, based on context or perceived similarity	Systematic arrangement of entities based on analysis of necessary and sufficient characteristics
<i>Boundaries</i>	Non-binding group membership, “fuzzy” boundaries	Mutually-exclusive and non-overlapping classes, fixed boundaries
<i>Membership</i>	flexible, based on generalized knowledge and / or immediate context	rigorous, binary class membership (yes / no) based on class intension
<i>Criteria for Assignment</i>	context-independent and context-dependent	predetermined guidelines or principles
<i>Typicality</i>	graded structure, individual members can be ranked by typicality	ungraded structure, all members equally representative
<i>Structure</i>	clusters of entities, may form hierarchical structure	hierarchical structure of fixed classes

Figure 5.1 contrasts the top-down and bottom-up approach paradigm visually. One apparent potential synergy of a combined approach would be to make the implicit concepts within social annotations explicit. In this way, still large user populations would participate in the annotation process, alleviating the knowledge acquisition bottleneck. In addition, the explicit concepts would facilitate interoperability and resolve ambiguity. In order to achieve this goal, methods are required which detect categorization patterns and deduce potential emergent concepts or classes. Because this transition from categories to classes needs to cross systemic differences between both approaches, the following section contrasts their underlying organizational structures.

5.1.2. Categorization vs. Classification

The schema by which Social Annotation Systems and ontologies structure a set of information resources differs fundamentally. While Social Annotation is usually said to produce *categories* (Shirky, 2005) of associated items, an ontology consists of precisely specified *classes* (recall from the definition of the ontology model introduced in Section 4.1.1 that the term *classes* is used interchangeably with *concepts*). The corresponding processes are called *categorization* and *classification*. Reinforced by both being essentially mechanisms for organizing information, (Jacob, 2004) reports within the literature several occurrences of the misconception that categorization and classification are in fact synonymous. Coming back to the ontology model introduced in Section 4.1.1, please recall that the term *classes* is used interchangeably with *concepts*. However, there is a fundamental difference in their underlying structure and their paradigms of organization. The following systemic properties which differentiate them from each other were identified (Jacob, 2004): (i) process, (ii) boundaries, (iii) membership, (iv) criteria for assignment, (v) typicality and (vi) structure. Table 5.1 summarizes the comparison. As will become apparent during the following explanation of the dimensions, Social Annotation resembles more closely to a categorization approach, while ontologies belong more to the classification paradigm. Both will be used as illustrative examples during explanation. If not stated otherwise, the comparison is based on (Jacob, 2004), and literal quotations within are taken from there.

The **process** of categorization is concerned with “*dividing the world into groups of entities whose members are in some way similar to each other*”. As an example, the assignment of, e. g., tags within a tagging system is grouping Web resources sharing some kind of commonality. This generally unsystematic process is mainly based on individually perceived suitability of a given tag for the resource under consideration, which may be driven by the current context, personal goals or individual experience. Ontologies on the other hand stand for a much more systematic approach, because the definition of concepts or classes is typically based on a thorough analysis of object characteristics within a domain of interest.

Because the context may play an important role for category membership, the **boundaries** of categories tend to be “fluent” or fuzzy in the sense that a given object may be contained under certain contextual circumstances, but not under others. Hence, category **membership** is flexible, and the **criteria**

for assignment may vary. Because ontologies are intended to capture context-independent transferable knowledge, their class boundaries are fixed, class membership is rigorous (i. e., true or false) and the assignment of objects to classes requires predetermined principles. Though mutual exclusiveness is not a hard requirement for ontology classes as such, the idea of a (formal) taxonomy remains to assign an object to a single class.

Another implication of the binary class membership principle of ontologies is that all class members share the same level of *typicality* – i. e., there is no internal structure within a class. Coming back to the ontology example presented in Section 4.1.1 and the **COMPANY** class, this means that there is no “most typical” company. On the contrary, such distinctions could be reflected, e. g., in the popularity of resources: If the website of IBM is tagged very often with the keyword *company*, then one could hypothesize that IBM is a more prototypical example of a company than, e. g., a seldom tagged local web design company.

As a last and most important criterion, which is influenced by the aforementioned ones, the hierarchical **structure** of well-defined classification systems provides rich context and a cognitive scaffolding. The encoded class relationships make it a stable medium for the “*accumulation, storage and communication of information*”. Categorizations like found in Social Annotation Systems exhibit a more ephemeral nature, i. e., their meaning may be short-lived and subject to change. This implies an increased adaptivity, but hampers their use as a persistent medium.

So in summary, categories are characterized by lightweight, flexibility and a kind of “plasticity”, while classes are more static and formal groupings. It is clear that each has its right to exist, being particularly suited for different tasks. However, an especially appealing synergetic vision is to observe the process of category formation until a certain degree of stability is reached, and then to try to “transform” stable categories into classes.

5.1.3. Comparison of Strengths and Weaknesses

In the previous Sections 3.1.1 and 4.1.4, the respective advantages and disadvantages of Social Annotations and ontologies were explained in detail. In order to summarize them for comparison and to point out their “inverse” relation, Table 5.2 contains a condensed representation of the most important aspects. Note that the notation of “+” and “–” within the table corresponds to relative

Table 5.2.: Overview on strengths and weaknesses of Social Annotation and Ontologies.

<i>Aspect</i>	<i>Ontologies</i>	<i>Social Annotation</i>
precision	+	–
consistency	+	–
ambiguity	+	–
interoperability	+	–
creation of context	+	–
support for retrieval	+	–
requirements	–	+
flexibility, adaptivity	–	+
inclusiveness	–	+
scalability	–	+

advantages and disadvantages when contrasting both approaches. , e. g., a “+” sign under the aspect of ambiguity means that within ontologies, ambiguity is a smaller problem than within Social Annotation Systems, because the different meanings are explicitly represented. Similarly, a “–” in the requirements aspect means that ontologies have a disadvantage here, because they impose higher requirements on contribution than Social Annotation Systems. Please refer to Sections 3.1.1 and 4.1.4 for a detailed explanation of the aspects.

The essence of this comparison is the following: Social Annotation systems suffer on the one hand side exactly from problems that ontologies were designed to eliminate (e. g., ambiguity, lack of precision, low interoperability); but on the other hand, their scalability and flexibility let them achieve a level of dissemination and widespread use hardly reachable for purely ontological approaches. To use the formulation of (Mathes, 2004), “*A folksonomy represents simultaneously some of the best and worst in the organization of information*”. In order to design an augmented kind of organization system, a promising direction would hence be to keep the adaptive aspects of Social Annotations, while applying methods to enhance precision and resolve ambiguity in order to “harvest” more formalized knowledge structures.

5.2. Ontology Learning to capture Emergent Semantics

The idea of extracting meaningful concepts and relations from massive corpora of semantically informal content stems traditionally from the discipline of *ontology learning*. According to (Cimiano et al., 2009), the latter defines a family of “*data-driven techniques supporting the task of engineering ontologies*”. Typically, such approaches are applied to structured (e. g., databases), semi-structured (e. g., HTML or XML documents) or unstructured (e. g., textual) resources. Based on data mining and machine learning principles, the goal hereby is to detect structures within the data which correspond to meaningful relations. These are then extracted, intended to support an ontology engineer in the task of modeling a particular domain of interest. Obviously, a prerequisite for such a methodology is the *existence* of semantic structures within the data.

In this section, we will first report on evidences of *emergent semantics* within Social Annotation Systems, which makes them an appropriate data source for ontology learning. In the following, we will detail on the different tasks involved, and explain which ones are relevant and feasible for the case of Social Annotations. Furthermore, the characteristics of Social Annotation data compared to more “traditional” ontology learning input like natural language text are worked out. The section closes with a precise description and definition of the ontology learning approach pursued within this dissertation.

5.2.1. Emergent Semantics

According to the definition given by Philippe Cudré-Mauroux in the Encyclopedia of Database Systems (Cudré-Mauroux, 2009), emergent semantics “*refers to a set of principles and techniques analyzing the evolution of decentralized semantic structures in large scale distributed information systems*”. In this article, Cudré-Mauroux also mentions collaborative tagging as a key application to be analyzed using an emergent semantic paradigm. In fact one of the main reasons of the growing interest of different research communities in social bookmarking data were early evidences for the formation of stable semantic patterns within the large bodies of human-annotated content.

An early systematic analysis was performed by (Golder and Huberman, 2006). One core finding was that the openness and uncontrolled nature of these systems did not give rise to a “tag chaos”, but led on the contrary to the development of stable patterns in tag proportions assigned to a given resource. (Cattuto,

2006) reported similar results and denoted the emerging patterns as “*semantic fingerprints*” of resources. Kome (2005) provided further empirical evidence for the existence of hidden hierarchical relationships among tags. Cattuto et al. (2007) analyzed statistical properties of tag co-occurrence networks; by using a shuffling approach, they discovered local patterns of co-occurrence indicating a possible underlying semantic hierarchical organization. Al-Khalifa and Davis (2007) compared folksonomy tags with (i) keywords from trained human indexers and (ii) content-extracted keywords and judged folksonomies as a rich source for semantic metadata. Hotho et al. (2006a) showed that existing knowledge discovery techniques like association rule mining provide meaningful results when applied to tagging data. From a slightly different perspective, (Kennedy et al., 2007) showed that tagging patterns of pictures upload to Flickr together with geographical labels can be useful to generate summaries of important locations and events.

Though coming from different disciplines and analyzing different aspects, the consensus of these works is that the large bodies of human-annotated content resulting from collaborative tagging systems contain evidences for emergent semantics.

5.2.2. Ontology Learning Tasks

Similar to the differentiation of ontologies by their semantic spectrum (cf. Section 4.1.2), methods of ontology learning can be distinguished by the specific ontology components they are targeting. Because these are inherently of different complexity, the involved tasks can be arranged in a “layer cake” hierarchy (see Figure 5.2) as proposed by (Cimiano, 2006).

Starting from the bottom, the least complex task is to discover **terms**. This is more relevant for textual resources; for the case of Social Annotations, existing approaches typically regard the keywords used for annotation as terms. In this way, Social Annotation data greatly simplifies the process of term extraction. As a next goal, the extracted terms need to be grouped into **synonym** sets, i. e., clusters of terms with similar meanings. According to (Cimiano, 2006), apart from its lexical representation $Ref_C(c)$, learning a **concept** c also requires an intensional definition $i(c)$ (i. e., a specification of its meaning, given, e. g., by a natural language description or typical attributes) and an extension $||c||$ (i. e., the set of its instances). As a next step, it is desirable to arrange the learned concepts into a **concept hierarchy**, whose relation semantics may

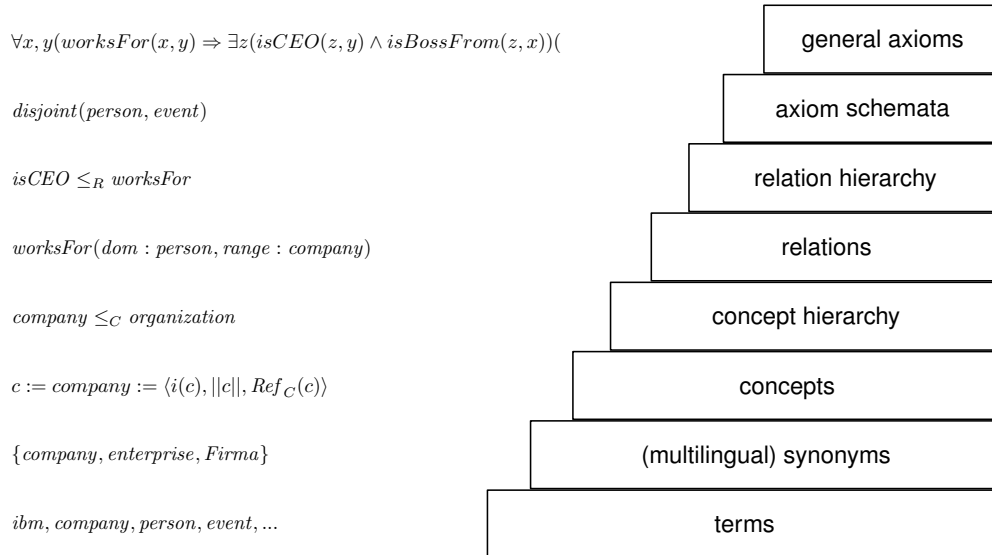


Figure 5.2.: Summary of the different tasks involved in ontology learning, arranged in a hierarchical “layer cake” as proposed by (Cimiano, 2006).

be strict or non-strict. Besides this taxonomic relation, the next ontology learning task addresses general **relations** among concepts. Due to the great variety of possible relations, the complexity of this task is considerably higher compared to the lower layers – and is becoming even more complex when trying to arrange the relations into a **relation hierarchy**. Finally, the top two layers are concerned with **axiom schemata** and **general axioms**, which are the most formal representation variants. However, due to the very high complexity of deriving rigorous logical axioms by a data-driven methodology, it is a less tackled problem (Terrientes et al., 2010).

5.2.3. Comparison to Ontology Learning from other input

In the context of ontology learning, the question is now to which extent the data resulting from Social Annotation Systems differs from “traditional” input data of ontology learning algorithms, like database schemata, dictionaries or plain text documents. A fundamental difference lies hereby within the way *how*

and *by whom* the data is created. This difference can be best explained among the following comparison dimensions:

- Motivation of contributors
- Communication among contributors
- Requirements for contribution

The following sections explain each of these in detail.

Motivation of contributors. The first aspect which can help to characterize Social Annotation data is the motivation of the people involved in its creation. Having a look at, e. g., approaches of ontology learning from text, one can observe that these are often performed on specialist literature of a given domain (e. g., tourism websites in (Maedche and Staab, 2005) or abstracts from medical journals in (Cimiano, 2006)). This implies that the underlying text corpora have often been produced *for a specific purpose* and often *targeted towards a specific audience*. Especially for datasets stemming from scientific literature, newspapers, periodicals or journals, it is justified to presume that an important motivation of the authors is to *capture and convey information such that it can be used by others*. This is not necessarily the case for Social Annotations; for social bookmarking systems, (Golder and Huberman, 2006) presented evidence that “*users bookmark primarily for their own benefit, not for the collective good*”. While this distinction is less sharp for, e. g., weblogs, Twitter posts or wikis, the aspect that an intrinsically private information management task (like the maintenance of a personal bookmark collection) is performed in a public space is a novel characteristic of the resulting data. To use a more pointed formulation, the main motivation for users to “produce” data in a Web 2.0 system is not necessarily to make the data available to the public (as it is when writing articles or books), but rather to solve a specific personal task.

Communication among contributors. Apart from the motivation of the users, the way how the contributors are interacting differs fundamentally compared to more traditional data used for ontology learning. Taking a look at, e. g., a group of people writing a book, creating a dictionary, or designing a database schema to describe a given domain, one can usually assume that each participant is aware of all other involved individuals (e. g., group members or authors of cited

papers). It is furthermore probable that the contents of the book, dictionary or schema were discussed and negotiated via (direct or computer-mediated) exchange of messages. This allows also to trace back particular contributions on an individual level. On the contrary, the connection among all users involved, e. g., in the creation of a tag cloud describing a certain resource is much less explicit. (Xia et al., 2009) refers to these communication features as *ballot box communication*. Compared to “traditional” computer-mediated communication, one of its distinguishing characteristics is the many-to-one nature of information flows. It describes the paradigm that the many individual contributions are aggregated and presented to each user as a single “voice of the crowd”. This exposure leads to an implicit influence on individuals by local or global trends within a certain community - without necessarily being able to pin down this influence to one or more particular users.

Requirements for contribution. As a third and last comparison dimension, we’ll have a look at what is required from each contributor, and which cognitive processes are involved in contribution. Sticking to the example of ontology learning approaches based on specialist literature of a given domain, it is clear that authors are typically required to have a demonstrated expert knowledge in the area of interest. The process of structuring and capturing this knowledge in a written form is very time-consuming and requires further communication skills as well as an elaborate audience design. These high entry barriers are a natural cause for the relatively small number of contributors involved in the creation of “traditional” ontology learning input. Among the characteristics of social annotation systems, here we can find probably the most distinguishing feature: A central aspect of them is their very low entry barrier and immediate usability for a large number of users. (Xia et al., 2009) noted that this is partially due to limited interaction options which lower the participation costs compared to, e. g., reading and writing messages. A cognitive analysis by Sinha (2005) attributed the simplicity of free annotation to the observation that after the mental activation of related categories, no choice between them is necessary. In addition, most systems impose no special requirements of domain knowledge, which leads to the situation that the knowledge is not concentrated within a small circle of experts, but more fragmented and distributed among a broad community. Table 5.3 summarizes the main dimensions of comparison discussed before.

Table 5.3.: Dimensions of comparison between “traditional” input of ontology learning algorithms and input from social tagging systems. The different kinds of “traditional” input are explained at the beginning of Section 5.2.

		Social Annotation data		“traditional” OL input
<i>Motivation</i>	user goal	solution of specific (personal) task, primarily for own benefit	↔	capturing and conveying information for a specific audience
	community goal	aggregate user preferences	↔	provide more content in higher quality
	publicity of medium	side-effect	↔	core aspect
	user types	producers, consumers	↔	contributors, lurkers
<i>Communication</i>	communication richness	low	↔	high
	type of communication	many-to-many, many-to-one	↔	one-to-many
	communication cost	low	↔	high
	influence on users	by actions, implicit	↔	by messages, explicit
<i>Requirements</i>	amount of domain knowledge required	small	↔	large
	cognitive processes involved	related category activation, lightweight conceptualization, comparatively effortless	↔	thorough structuring & preparation of content, audience design, communication skills
	number of contributors	high	↔	low

In the following section, we broaden again our perspective from ontology learning to the general research direction of “bridging the gap” between Social Annotations and the Semantic Web. Specifically, the State of the Art in this area will be elaborated, based on a comparison framework which summarizes the most important distinguishing dimensions.

5.3. State of the Art

The early evidences of emergent semantics in Social Annotation Systems and the availability of large test data sets quickly motivated a large number of approaches coming from different disciplines targeted towards making the implicit semantic structures explicit. In the sequel, we will first propose a set of comparison dimensions and values in Section 5.3.1, which are intended to cover as exhaustively as possible all approaches. Then, the related work is explained along the core dimension of “Learning Tasks” (see Table 5.4), which comprise the ones mentioned in the ontology learning layer cake (see Section 5.2). The selection of the presented approaches was intended to be as complete as possible. If a selection was still necessary (e.g., for space reasons), it was guided by the goals of (i) preferring early works on a specific direction (ii) covering a possibly broad range of relevant authors and (iii) providing of at least one example for each of the values defined in our comparison model.

5.3.1. Comparison dimensions

The main dimensions and values used for comparison are summarized in Table 5.4. Garcia-Silva et al. (2011) also provided an extensive review of approaches to discover tag semantics, including a suggestion for a unified process model. Complementing their work, our main focus is not to compare existing approaches based on the different process steps, but rather using dimensions similar to prior comparisons of approaches of general ontology learning like (Omelayenko, 2001; Shamsfard and Barforoush, 2003; Biemann, 2005). As we consider the “Learning Tasks” to be the core dimension of comparison, we will cover each of its values in a subsequent section; the other dimensions will now be briefly explained.

First of all, a core question is of course which **Data Source** is exactly used to derive semantics from. A large number of approaches (e.g., (Heymann and Garcia-Molina, 2006; Schmitz, 2006)) is solely based on the *folksonomy structure*.

Table 5.4.: Comparison dimensions and possible values to compare methods of making semantics in folksonomies explicit.

<i>Dimension</i>	<i>Values</i>
Data Sources	folksonomy structure, resource content, tag content, user-defined tag relations, additional metadata, external sources
Data Filtering	by tag / resource / user properties, by external source, manual
Learning Technique	statistical data mining & machine learning (clustering, association rules, generative models, latent semantic analysis), SNA measures, NLP techniques, custom algorithms
Learning Task	ontology construction (terms, synonyms, concepts, concept hierarchy, relations & axioms), semantic measures (relatedness, generality), tag sense disambiguation, ontology maintenance, ontology population
Evaluation	human assessment, gold-standard based, application-centered (folksonomy), application-centered (external application)

By this we mean the tripartite structure itself as defined in Section 3.1.2, as well as derived structures like tag-resource (Mika, 2005) or tag co-occurrence networks. The advantage of this kind of approaches is their independence from tag language and content type of the shared resources (e. g., bookmarks, videos or pictures). When taking into the *tag content* (i. e., the lexical representation of a tag itself (Tatu and Moldovan, 2010)) or the *resource content* (Brooks and Montanez, 2006), the advantage of more information is to be traded off against a restricted applicability to different languages and content types. If available, *additional metadata* like time and location information can also be exploited (Kennedy et al., 2007). Some systems also allow their users to define *tag relations*, which can also be used for taxonomy induction (Plangprasopchok and Lerman, 2009). Finally, approaches which consider *external sources* like existing ontologies or thesauri (Marinho et al., 2008; Angeletou, 2008) benefit from rich prior knowledge, but of course naturally depend on the availability of such.

Having chosen a specific data source, most approaches perform an a-priori **Data Filtering** in order to exclude inappropriate content or optimize the input for a specific learning procedure. Typically, data items are hereby disregarded based on certain *tag*, *user* or *resource properties*. Very common are minimum frequency thresholds (Wu et al., 2006b), or top-k selections which restrict the analysis to popular folksonomy partitions. But also more sophisticated strategies like including only resources annotated with a least one verb (Maala et al., 2007) can be found. Spam removal is another motivation to this end. Another possible data restriction is to only keep items which are present in external repositories like tags in Tagpedia⁵ (Tesconi et al., 2008) or resources in Wikipedia⁶ (Meder, 2010). In addition, the fine-tuning of the resulting dataset is also often done manually, e. g., by removing system tags like *system:unfiled*.

The variety of disciplines interested in learning tag semantics also led to a variety of applied **Learning Techniques**. Clustering is an obvious candidate of *statistical data mining and machine learning* techniques to form groups of (semantically) related tags (Begelman et al., 2006; Giannakidou et al., 2008; Gemmell et al., 2008). Closely related are association rule mining methods (Hotho et al., 2006a; Lin et al., 2009) which are able to detect semantic relations among items. A more specialized example from this area are statistical models of subsumption (Schmitz, 2006), targeting the discovery of is-a relations among tags. From a different perspective, generative approaches like the separable mixture model (SMM, (Zhang et al., 2006)) are modeling the users' behavior in assigning tags to resources. Starting from an observed tag co-occurrence distribution, a conditional distribution of tags over a fixed number of topics is computed. Another theoretically well-founded approach is to apply dimensionality reduction techniques like latent semantic analysis (LSA, (Eda et al., 2009; Levy and Sandler, 2008)) to the high-dimensional tag vector space, resulting in a mapping of tags to "topics" or "concepts". Because some folksonomy-induced networks exhibit suitable properties, also measures stemming from *social network analysis* (SNA) like centrality or clustering coefficient were applied (Mika, 2005; Heymann and Garcia-Molina, 2006), mostly in order to distinguish between general and specific tags. Despite the fact that the assignment of tags to resources does not follow any kind of syntactical pattern, researchers from the natural language processing community (NLP) have used part-of-speech

⁵<http://www.tagpedia.org>

⁶<http://www.wikipedia.org>

taggers to gain a deeper syntactic understanding of a tag and finally to discover equality and synonymy relations among tags (Tatu and Moldovan, 2010). A last family of approaches are custom algorithms specifically tailored for the task of capturing tag semantics, like the incremental tag taxonomy induction algorithm proposed by (Heymann and Garcia-Molina, 2006).

Besides the specific learning tasks (which will be covered in the subsequent subsections), the last comparison dimension is which **Evaluation** paradigm was used to assess the quality of the learned semantics. Because this dimension differs least from general ontology learning, we just recapitulate briefly the three main classes mentioned by (Dellschaft and Staab, 2006) of (i) human assessment (e. g., (Schmitz, 2006)), (ii) gold-standard based and (iii) application-centered approaches. The latter is often performed within the folksonomy system itself, e. g., by using the learned semantics to improve tag recommendations (Wetzker et al., 2010) or information retrieval (Marinho et al., 2008). Examples of integration into external systems are less frequent, but found for example in the context of e-learning applications (Doush and Pontelli, 2010).

5.3.2. Capturing Semantic Relatedness

Motivated by the existence of ontology-based measures of *Semantic Relatedness* (see Section 4.2.1), an early research question was to which extent such measures could be derived from a *folksonomy*. While a number of approaches successfully applied several kinds of such measures (e. g., an adapted version of Jaccard similarity coefficient in (Meo et al., 2009)), a systematic analysis has largely been missing. This gap is addressed by this dissertation, which examines several variants of folksonomy-based measures of semantic relatedness in Section 7.1.

5.3.3. Learning Concepts

Although, e. g., measures of semantic tag relatedness do have a value on their own, the goal of formalizing the implicit knowledge in folksonomies by constructing ontologies remains desirable. For the coverage of such approaches, we will stick to the ontology learning layer cake (see Figure 5.2) and start with term extraction methods. Because the processes of synonym discovery and concept formation is sometimes hard to disentangle, both will be presented in a common paragraph. Approaches of tag sense disambiguation as an important aspect will then be explained in a separate paragraph.

Terms: A main advantage of folksonomies compared to text documents as input for ontology learning algorithms is that the process of term extraction is much simpler – in fact, many approaches skip this step completely and regard tags directly as terms. However, observing a great variety of spelling and abbreviation variants among tags, some works perform *tag normalization* at different levels of complexity. This reaches from the simple removal or replacement of special characters (Cantador et al., 2008) (e. g., ü → u) to the unification of morphological variants by string distance measures (Specia and Motta, 2007) or information from external lexical resources (Tesconi et al., 2008).

Synonyms & Concepts: Among the most often mentioned problems of Social Annotation Systems is that synonymy within the uncontrolled tag vocabulary hampers information retrieval tasks. As an example, in order to collect references for an ontology learning book, some people will use the tag *ol*, others *ont-learn* or *ontology_learning*. Searching by only one of these will consequently fail to retrieve the relevant resources tagged with the others. This obvious deficiency has motivated researchers to come up with various methods to form groups of tags with a similar meaning, often referred to as *concepts*. In a strict sense, the latter does not conform to the definition given by (Cimiano, 2006), according to which the process of concept formation should also provide an intensional definition (e. g., a natural language description or a set of typical attributes) of each concept. Despite that, we will treat for simplicity reasons both variants (i. e., semantic groupings of tags with and without⁷ an additional intensional definition) uniformly as “concepts”. Begelman et al. (2006) proved the applicability of clustering to discover concepts by using an algorithm based on spectral bisection. Their approach as well as the ones from (Grahl et al., 2007; Giannakidou et al., 2008) requires a predefinition of the number of clusters – a parameter which is usually hard to come up with in advance. Examples of approaches which produce a variable number of clusters are (Specia and Motta, 2007; Zhou et al., 2008; Gemmell et al., 2008; Radelaar et al., 2011). Jung (2010) explores the possibility to match tags with the same meaning across different languages. Jäschke et al. (2008a) applied FCA techniques to discover so-called frequent tri-concepts (i. e., sets of users, tags and resources belonging

⁷Of course one could also argue that the set of tags themselves can always be interpreted as a lightweight intensional description.

to an implicit concept) within folksonomies. Wetzker et al. (2010) apply a translation approach to compute mappings between tags used by different users to describe similar concepts.

Another class of approaches is to take into account external sources containing prior knowledge about semantic relations. Angeletou (2008) proposed a method to enrich folksonomies by mapping tags to concepts defined in WordNet; (Cantador et al., 2008; Tesconi et al., 2008; Garcia-Silva et al., 2009) described similar approaches using categories and concepts derived from Wikipedia. Grineva et al. (2008) exploited a Wikipedia-derived measure of semantic relatedness for tag clustering. Though coming from a different direction, approaches like (Abbasi and Staab, 2009; Lee and Yong, 2007) are exploiting WordNet for query expansion, geared towards enhancing retrieval quality by querying with “concepts on the fly”.

Tag Sense Disambiguation Besides the aforementioned problem of synonymy, another major and often mentioned weakness of tagging systems is ambiguity of tags. Of course this is not an intrinsic problem of social tagging, but in a way “inherited” from the fact that tags can mostly be considered as natural language entities. However, the openness and uncontrolled nature of these systems makes this issue more visible.

In principle, the task of disambiguating tag meanings belongs to the process of concept identification (see Figure 5.2; but in order to clarify the different aspects and approaches, it is treated separately in this chapter. Statistical natural language processing distinguishes between supervised, dictionary-based and unsupervised disambiguation (Manning and Schütze, 1999). In all cases, information taken from the *context* of a term forms the basis for its assignment to a certain sense. In the process model of discovering tag semantics by (Garcia-Silva et al., 2011), “context identification” is also included as a major step.

Supervised approaches are based on labelled training data, and learn usually a classifier based on context features of a given word. Such approaches have rarely been applied to social tagging systems. Dictionary-based approaches rely on sense definitions defined in dictionaries or thesauri. Angeletou et al. (2008) first identifies a set of candidate senses for a given tag within WordNet, interprets co-occurring tags as context and uses a measure of semantic relatedness to choose the most appropriate sense. In a similar manner, (Garcia-Silva et al., 2009) uses cosine similarity between tag co-occurrence vectors and a bag-of-words

representation of Wikipedia pages to identify the most suitable sense definition within DBPedia.⁸ Lee et al. (2009) also computes a relevance score between tags and Wikipedia articles for the same purpose.

Unsupervised approaches are trying to partition the context of a given term into clusters corresponding to its different senses. Au Yeung et al. (2007, 2009a) analyzed several folksonomy-derived networks with regard their suitability to derive senses by graph clustering algorithms. Zhang et al. (2006) proposed an entropy-based metric to capture the level of ambiguity of a given tag. Si and Sun (2009) take into account a web-based measure of semantic relatedness as well as textual article content to disambiguate tags in weblogs by spectral clustering.

As a last class of approaches, methods like the one proposed by (Passant and Laublet, 2008; Passant, 2007) require the user to define the intended meaning during the tagging process by choosing among a set of possible senses.

5.3.4. Capturing Semantic Generality

Apart from semantic relatedness, some knowledge extraction algorithms are also based on a notion of *Semantic Generality* of tags, which is used, e. g., to distinguish between broader and narrower terms. Again, several measures were used in the literature (e. g., network centrality by (Heymann and Garcia-Molina, 2006), or a statistical model of subsumption in (Schmitz, 2006)). However, a systematic comparison has largely been missing. This gap is addressed within this thesis, more precisely in Section 7.3.

5.3.5. Learning Concept Hierarchies

While synonym resolution is mainly targeted towards improving retrieval tasks, the reconstruction of hierarchical relationships among tags (or learned concepts) is often mentioned in the context of enhanced browsing facilities. Because the maintenance of larger hierarchies is a difficult task and hence the idea of self-organizing structures is very appealing, many researchers have focused on this aspect. Mika (2005) pioneered in deriving broader / narrower tag relations from a user-tag graph (called “actor-concept network”), which are effectively based on subcommunity relationships. Hamasaki et al. (2007) extended his work by taking into account information from the users’ neighbourhood in the

⁸<http://www.dbpedia.org>

folksonomy graph. Heymann and Garcia-Molina (2006) suggested a custom algorithm to induce a “tree of tags”, based on a measure of tag generality and a measure of tag similarity. Schmitz (2006) applied a statistical model of subsumption (originally stemming from work on deriving concept hierarchies from text) for a similar purpose. Eda et al. (2009) used probabilistic latent semantic indexing (PLSI) to induce a taxonomy of tags. Zhou et al. (2008) used a divisive clustering technique based on deterministic annealing to iteratively split the set of tags into semantically coherent subsets. Lin et al. (2009) uses association rule mining together with hypernym relations from WordNet to derive a hierarchical tag structure. The approach of (Meo et al., 2009) consists in building first a directed weighted tag graph using a notion of generality, and then removing edges until a maximum spanning tree is found.

Based on a different data source, (Plangprasopchok and Lerman, 2009) suggested to integrate user-specified tag relations⁹ into a global consensus structure.

5.3.6. Learning Attributes, Relations and Axioms

Apart from learning taxonomic relations among concepts as described in the previous paragraph, literature on learning other kinds of relations from social tagging data is still sparse. A possible reason for this is that a large portion of relation learning techniques based on text (see (Cimiano, 2006) for an overview) comprise the exploitation of syntactic dependencies or lexico-syntactic patterns, which do not exist in folksonomies. However, when reviewing the results of taxonomy learning techniques, it turns out that in some cases the learned taxonomic relations do not always convey a sharp and precise “is-a” semantics. In the examples given by the authors, one can also find occurrences of e. g., “part-of”-relationships or purpose-related connections. But the “disambiguation” of these relations remains so far an open and interesting research problem, as well as the extraction of more complex constructs like axioms.

⁹Recall that some social tagging systems like BibSonomy or Flickr allow users to create explicit directed tag relationships.

5.4. Evaluation Paradigms

Because the methods and techniques used to harvest emergent semantics build mostly upon data mining and machine learning models, evaluation is a crucial aspect. However, compared to other disciplines in this area like information retrieval or speech recognition, standardized benchmark datasets and measures are largely missing (Dellschaft and Staab, 2006). Among the possible reasons for this, one factor can surely be seen in the nature of semantic modeling itself: The establishment of valid semantic structures which represent precisely the conceptual elements and relations of a certain domain is an inherently challenging task. Especially when learning ontologies, their inherent complexity makes a holistic evaluation approach hardly feasible:

“An ontology is a fairly complex structure and it is often more practical to focus on the evaluation of different levels of the ontology separately rather than trying to directly evaluate the ontology as a whole.”
(Brank et al., 2005)

In addition, a global quality criterion may not even be desirable for the following reason: Ontologies are usually constructed not only for the mere purpose of representing knowledge, but are often targeted towards a particular application. Depending on the application type and the anticipated user population, different aspects of the ontology will be more or less important, which should be reflected in any evaluation approach. If, e. g., there exists a direct interface where humans work with the ontology, the quality of the lexical labels to describe concepts will be more important than in a case where the ontologically captured knowledge serves only as an input for an automatic process. Generally speaking, the different aspects or levels correspond to the ontology learning layer cake (see Figure 5.2). So in general, the design of a meaningful evaluation needs to consider at least two aspects: (i) which ontology learning tasks are to be evaluated and (ii) which task is most relevant for the targeted application, task or audience.

Because a core question when capturing emergent structures within social annotation data is to which extent they represent “correct” relations, most works in this direction have applied *semantic evaluation* approaches. Hereby, one can broadly distinguish between three evaluation paradigms (cf. Dellschaft and Staab (2006); Brank et al. (2006)):

- *Application-centered*: Especially within the Semantic Web, there exist several applications (like semantic search engines or recommender systems) which are based on background knowledge in form of explicitly represented semantics. In such cases, a natural measure of semantic quality would be the performance improvement achieved by using different semantic structures as input. A requirement hereby is the existence of measures to compare the achieved results. Though this paradigm reflects clearly the actual “utility” of an ontology, a problematic issue is how to disentangle the influence of the semantic input from other application parameters.
- *Human Assessment*: This paradigm relies on the assessments of human experts how well an formal semantic representation meets a set of predefined criteria. Hereby it is obviously an important question on which criteria to agree. This paradigm can be expected to provide valuable assessments of semantic quality at a high cost due to the heavy involvement of human interaction.
- *Reference-based*: The prerequisite of this methodology is the existence of a semantic “gold-standard”, to which the learned semantic structures can be compared. The gold standard can be an ontology itself, but also, e. g., a set of documents covering the domain in question. The key issues hereby are how to assess the quality of the gold-standard itself, and the establishment of valid comparison measures.

Comparing the paradigms, (Dellschaft and Staab, 2006) concludes that application-centered evaluation is suitable for ontology engineering scenarios; for large-scale and frequent evaluations of ontology learning algorithms themselves, only reference-based methods are practically feasible. In the context of this dissertation, the main evaluation goal is to assess the degree to which the captured semantic structures exhibit desirable properties (e. g., precision, consistency, unambiguity) of more formally engineered structures. Hence, we will mostly adopt a reference-based evaluation paradigm, including manually built semantic resources. Because our focus hereby lies in the formation of concepts and the induction of concept hierarchies, the two most relevant aspects for evaluation are the *lexical* layer (i. e., which terms are used to denote the concepts) and the *structural* layer (i. e., how the concepts are arranged hierarchically). In section 5.4.1 and 5.4.2, existing evaluation measures for both are

presented. For a deeper discussion of evaluation issues, refer to (Dellschaft and Staab, 2006; Bade and Benz, 2008; Brank et al., 2005).

More recently, a fundamentally different paradigm of evaluation was proposed in the context of inducing hierarchies from social tagging systems (Helic et al., 2011). It is essentially concerned with an assessment how useful a learned structure is to fulfil a particular task like e. g., navigation within a folksonomy. The difference to application-centered evaluation is that this task is not necessarily implemented within a running system, but can be e. g., simulated within a model. Compared to semantic evaluation, this is an orthogonal dimension of comparison, because a learned hierarchy does not necessarily be semantically precise in order to serve as a useful navigational aid. However, the application of both paradigms to a set of candidate hierarchies actually returned consistent quality assessments (Strohmaier et al., 2011). For this reason and because the main focus of this dissertation lies in the assessment of *semantic* properties of the emergent structures, we will adhere to the semantic evaluation paradigm as is customary in the literature.

5.4.1. Lexical Layer

Because an important issue in the field of knowledge organization is the assignment of meaning to terms, a first criterion when comparing two organization structures is to which extent they are using the same vocabulary to denote concepts. Several measures have been proposed in the literature (see (Dellschaft and Staab, 2006; Bade and Benz, 2008) for an overview). Most of them are assessing the size of the intersection between the term sets used in the learned and reference structure; (Sabou et al., 2005) proposed to adapt the precision / recall measures known from information retrieval for this purpose. Dellschaft and Staab (2006) picked up this idea and defined *lexical precision*, *lexical recall* and *lexical F1-measure* as :

Definition 5.1 *Lexical precision, recall, F1-measure*

Given a learned ontology

$$\mathcal{O}^* := (C^*, \leq_C^*, \text{root}_C^*, R^*, \sigma_R^*, \leq_R^*)$$

and a reference ontology

$$\mathcal{O}^\Delta := (C^\Delta, \leq_C^\Delta, \text{root}_C^\Delta, R^\Delta, \sigma_R^\Delta, \leq_R^\Delta)$$

with associated lexicons

$$\mathcal{L}^* := (L_C^*, L_R^*, Ref_C^*, Ref_R^*)$$

and

$$\mathcal{L}^\Delta := (L_C^\Delta, L_R^\Delta, Ref_C^\Delta, Ref_R^\Delta)$$

then the lexical precision LP and lexical recall LR between \mathcal{O}_L and \mathcal{O}_R are given by

$$LP(\mathcal{O}^*, \mathcal{O}^\Delta) = \frac{|L_C^* \cap L_C^\Delta|}{|L_C^*|} \quad LR(\mathcal{O}^*, \mathcal{O}^\Delta) = \frac{|L_C^* \cap L_C^\Delta|}{|L_C^\Delta|}$$

The lexical F1-measure LF combines LP and LR according to

$$LF(\mathcal{O}^*, \mathcal{O}^\Delta) = \frac{2 \cdot LP(\mathcal{O}^*, \mathcal{O}^\Delta) \cdot LR(\mathcal{O}^*, \mathcal{O}^\Delta)}{LP(\mathcal{O}^*, \mathcal{O}^\Delta) + LR(\mathcal{O}^*, \mathcal{O}^\Delta)}$$

Though other measure exist, which take, e. g., into account the edit distance between lexical labels (Maedche and Staab, 2002), the above measures will be used to compute similarity on the lexical level. Besides their clarity and simplicity, another reason for this choice was to avoid potential mismatches induced by matching semantically unrelated terms with a small edit distance (e. g., punk – funk, bank – bunk).

5.4.2. Structural Layer

Even if two conceptualizations are described by exactly the same terms (which would be reflected in lexical precision and recall values of 1), their *structure* can still be completely different. However, on this structural layer, it is a non-trivial task to judge the similarity between a learned concept hierarchy and a reference hierarchy, especially regarding the absence of well-established and universally accepted evaluation measures. While measures for the similarity of trees, concept lattices and graphs exist, their implication when applied to concept hierarchies (Cimiano, 2006). Yet, a number of useful measures have been proposed by past research. A number of them is based on the comparison of so-called *characteristic excerpts* (Maedche and Staab, 2002; Cimiano, 2006;

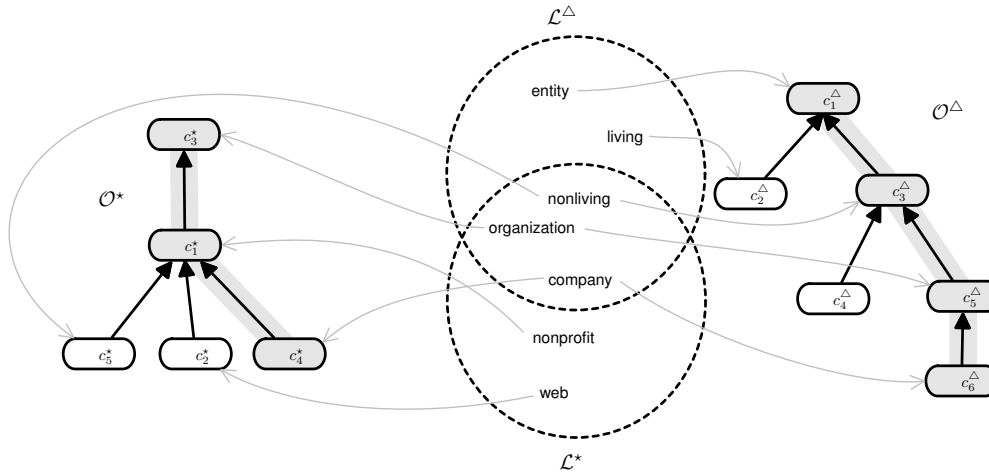


Figure 5.3.: Visualization of taxonomic similarity measures between a learned ontology \mathcal{O}^* and a reference ontology \mathcal{O}^Δ . The grey shade corresponds to a possible characteristic excerpt of the concept denoted as “company” in both structures.

(Dellschaft and Staab, 2006): The idea hereby is to represent a concept c^* within a hierarchy by some kind of a “representative neighbourhood” $ce(c^*)$, which reflects its position within the hierarchical structure. Given that a corresponding concept c^Δ exists in another hierarchy, its excerpt $ce(c^\Delta)$ can be compared to $ce(c^*)$ – if both are very similar, then it is assumed that both are found at similar positions within both hierarchies. Such a single computation corresponds to the *local part* of these measures. Figure 5.3 is adapted from (Maedche, 2002) and depicts this principle. The highlighted parts within both example hierarchies correspond to the two characteristic excerpts of the concept denoted as “company”. As a next step, the local values are aggregated into a *global part*, which finally indicates the overall similarity of both structures.

In all proposed variants, these taxonomic similarity measures are strongly influenced by (i) the specific composition of the characteristic excerpt and (ii) the schema by which the local parts are aggregated into a global value. Dellschaft

and Staab (2006) formulated three criteria which should be met when making these choices, namely (i) *multi dimensionality* (i. e., that different aspects can be separately evaluated without interference effects), (ii) *proportional error effect* (i. e., “more severe” differences like those close to the hierarchy root should be reflected in stronger dissimilarities) and (iii) *usage of interval* (i. e., the range of similarity values should be fully used, and gradual increases / decreases be reflected accordingly). Analyzing existing approaches, Dellschaft argued that the *common semantic cotopy* as a characteristic excerpt fulfils these criteria.

It is defined as:

$$csc(c, \mathcal{O}^*, \mathcal{O}^\Delta) := \{c' : c' \in Ref_C^*(L^* \cap L^\Delta) \wedge (c' \leq_C^* c \vee c \leq_C^* c') \wedge c \neq c'\}$$

Simply spoken, a concept is described by its sub- and superconcepts whose lexical representations are present in both hierarchies. Based on this excerpt, Dellschaft defines:

Definition 5.2 *Taxonomic precision, recall, F1-measure* Given a learned and a reference ontology and the corresponding lexicons as defined in Definition 5.1, the local taxonomic precision and recall values of two concepts $c^* \in C^*$ and $c^\Delta \in C^\Delta$ are defined according to:

$$tp_{csc}(c^*, c^\Delta, \mathcal{O}^*, \mathcal{O}^\Delta) = \frac{|Ref_C^*(Ref_C^{*-1}(csc(c^*, \mathcal{O}^*, \mathcal{O}^\Delta)) \cap Ref_C^{*-1}(csc(c^\Delta, \mathcal{O}^\Delta, \mathcal{O}^*)))|}{|csc(c^*, \mathcal{O}^*, \mathcal{O}^\Delta)|}$$

$$tr_{csc}(c^*, c^\Delta, \mathcal{O}^*, \mathcal{O}^\Delta) = \frac{|Ref_C^*(Ref_C^{*-1}(csc(c^*, \mathcal{O}^*, \mathcal{O}^\Delta)) \cap Ref_C^{*-1}(csc(c^\Delta, \mathcal{O}^\Delta, \mathcal{O}^*)))|}{|csc(c^\Delta, \mathcal{O}^\Delta, \mathcal{O}^*)|}$$

The global taxonomic precision TP is computed by averaging over the concept overlap between both ontologies according to:

$$TP(\mathcal{O}^*, \mathcal{O}^\Delta) = \frac{1}{|Ref_C^*(L^* \cap L^\Delta)|} \sum_{c \in Ref_C^*(L^* \cap L^\Delta)} tp_{csc}(c, c, \mathcal{O}^*, \mathcal{O}^\Delta)$$

The global taxonomic recall TR is computed analogously. Finally the taxonomic F-measure is computed as the harmonic mean of taxonomic precision and recall

according to

$$TF(\mathcal{O}^*, \mathcal{O}^\Delta) = \frac{2 \cdot TP(\mathcal{O}^*, \mathcal{O}^\Delta) \cdot TR(\mathcal{O}^*, \mathcal{O}^\Delta)}{TP(\mathcal{O}^*, \mathcal{O}^\Delta) + TR(\mathcal{O}^*, \mathcal{O}^\Delta)}$$

The same idea underlies the measure of *taxonomic overlap* proposed by Maedche (Maedche, 2002), which is defined as follows:

Definition 5.3 *Taxonomic Overlap* Given a learned and a reference ontology and the corresponding lexicons as defined in Definiton 5.1, the local taxonomic overlap values of two concepts $c^* \in C^*$ and $c^\Delta \in C^\Delta$ are defined according to:

$$to_{csc}(c^*, c^\Delta, \mathcal{O}^*, \mathcal{O}^\Delta) = \frac{|Ref_C^*(Ref_C^{*-1}(csc(c^*, \mathcal{O}^*, \mathcal{O}^\Delta)) \cap Ref_C^{*-1}(csc(c^\Delta, \mathcal{O}^\Delta, \mathcal{O}^*)))|}{|Ref_C^*(Ref_C^{*-1}(csc(c^*, \mathcal{O}^*, \mathcal{O}^\Delta)) \cup Ref_C^{*-1}(csc(c^\Delta, \mathcal{O}^\Delta, \mathcal{O}^*)))|}$$

The global taxonomic overlap TP is computed by averaging over the concept overlap between both ontologies according to:

$$TO(\mathcal{O}^*, \mathcal{O}^\Delta) = \frac{1}{|Ref_C^*(L^* \cap L^\Delta)|} \sum_{c \in Ref_C^*(L^* \cap L^\Delta)} to_{csc}(c, c, \mathcal{O}^*, \mathcal{O}^\Delta)$$

While these measures have not been applied widely, they are theoretically sound and interesting. This was the reason to choose them for the evaluations on the structural layer within this dissertation.

5.5. Approach of this dissertation

Having laid the groundwork by (i) describing ontology learning as a methodology to combine the advantages of Social Annotations and ontologies, (ii) introducing a comparison framework of related work in this direction and (iii) giving an overview of evaluation approaches, this section is intended to specify which tasks will be addressed exactly, what kind of data they will be based on and which evaluation paradigms will be chosen.

Learning Tasks: As was pointed out in Figure 5.1, the implicitness of concepts and relations within Social Annotation Systems is a problematic issue regarding their reusability. Hence, one core focus of the approaches presented within this dissertation is to make these concepts and relations explicit. In terms of the ontology learning layer cake (see Figure 5.2), we adopt hereby the notion from the literature that the set of *terms* is defined by the set of *keywords* (e. g., the set of tags T within a folksonomy $\mathbb{F} = (U, T, R, Y)$). Starting from those terms, the tasks of synonym discovery and concept formation will be pursued (see Sections 7.1 and 7.2). As a next goal, the arrangement of the learned classes into a hierarchical structure is approached (see Sections 7.3 and 7.4). Hereby, the goal to derive a *formal* taxonomic structure (as described in Section 4.1.3) is not regarded as mandatory. In other words, the learned hierarchical structures are not restricted to strong generalization/specialization relationships only. Tasks which are found in layers above (i. e., relations, relation hierarchy and axioms) are regarded as beyond the scope of this dissertation and left for future investigation. The rationale for this decision as well as preliminary ideas will be presented in Section 7.5.

Data Foundation: As pointed out in Section 5.3.1, an important question is which data source is exploited to capture emergent semantic structures. In order to remain independent from different languages and different resource types, the approaches presented in this dissertation are solely based on the *structure* of Social Annotation data. This implies that when performing the ontology learning tasks, no prior knowledge from external sources (e. g., existing ontologies) is taken into account. The main intention behind this is to observe the conceptual equivalent of “desire lines” (Mathes, 2004), i. e., semantic structures which are formed by *actual needs and actions* of users, in an unbiased manner.

Evaluation: In order to evaluate the quality of the learned semantic structures, our primary paradigm will be the comparison against gold-standard references. We stick hereby to the common notion in the literature that gold-standard based methods are the only feasible tools for frequent and large-scale evaluations (Dellschaft and Staab, 2006). However, when possible, these evaluations will be completed by user studies.

5.6. Summary

The main goal of this chapter was to concretize which motivations, aspects and methods are relevant for the goal of exploiting synergies between Social Annotations and ontologies (as core part of the Semantic Web). In order to highlight contrasting characteristics, the paradigms of *bottom-up* and *top-down* knowledge organization were discussed, and the need to make implicit concepts within Social Annotations explicit was identified. This process was then related to the transition from the collaborative *categorization* to a more formal *classification*, based on capturing “matured” structures. Further motivation was given by a summary of the inverse relation between the respective *strengths and weaknesses* of both approaches. As a concrete family of techniques to capture emergent semantics, *ontology learning* was introduced, along with the ontology learning layer cake which summarizes the different tasks involved like term extraction or concept hierarchy induction. Three main differences between *ontology learning from more “traditional” input* were identified, namely (i) the motivation of contributors, (ii) the communication among contributors and (iii) the requirements for participation. In the sequel, related work in the field of making implicit semantics within Social Annotations explicit was described extensively in order to convey a clear picture to the reader of the *state of the art* in this field. The description was structured by a comparison framework, whose dimensions were (i) data sources, (ii) data filtering, (iii) learning technique, (iv) learning task and (v) evaluation. Furthermore, the last dimension was elaborated more deeply by presenting approaches of assessing the quality of learned semantics, both on the lexical and the structural layer. The chapter closed with a precise specification which tasks are being addressed within the scope of this dissertation.

Part II.

Capturing Emergent Semantics: Data, Methods and Influencing Factors

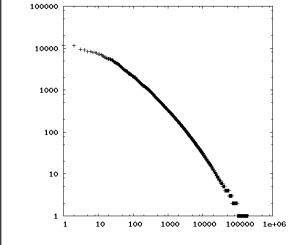
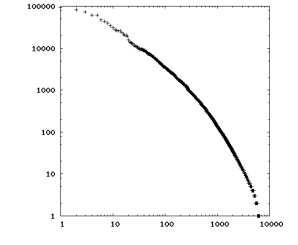
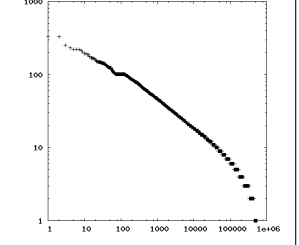
Chapter 6.

Data

Because the collaborative emergence of semantics within Social Annotation Systems is finally a data-driven process, one can assume that different dataset properties have an influence on the result of methods and algorithms which capture the emergent semantic structures. For this reason, we did not restrict ourselves to a single system, but selected several datasets intended to provide a broad coverage of system characteristics. These differ in various dimensions, like the size, the type of resources which are annotated, or the permission model (see also the distinction between broad and narrow folksonomies from Section 3.1.4). Furthermore, we also included data which does not stem from “pure” social resource sharing systems, but also from related applications like search engines (see Section 3.2.4 on logsonomies) or Question-Answering platforms (Section 3.2.3). This choice allows us to assess the usefulness of the methods presented later in Chapter 7 on a variety of Social Annotation data. Each dataset will be briefly introduced in Section 6.1, along with a summary of its respective statistical properties.

Apart from the question of *input* data, another important aspect considering the empirical evaluation of the *output* of the presented algorithms is the choice of reference datasets. Because we will stick in most cases to a gold-standard evaluation paradigm, we hence require a “ground truth”, to which the learned semantics can be compared. Hereby we are naturally facing the problem that one can hardly imagine the existence of a single “universal” semantic gold standard, which is globally accepted. Instead, we will also focus on a set of manually and semi-automatically built ontologies, characterized by different sizes and scopes. Those will be introduced in Section 6.2.

Table 6.1.: Statistics about the BibSonomy dataset.

<i>popular tags</i>	<i>Statistics</i>				
software	<i>Complete dataset</i>				
zzz.to_sort	Tags	Users	Resources	TAS	Posts
deutschland	192 445	6 463	551 540	2 434 387	636 479
web2.0	<i>Restricted to 10 000 most popular tags</i>				
nn	Tags	Users	Resources	TAS	Posts
programming	10 000	5 777	504 709	1 829 211	580 763
theorie	<i>WordNet overlap of k most popular tags</i>				
web	100	1 000	10 000	100 000	all
university	60	544	4 027	18 521	25 182
<i>Frequency Distributions</i>					
Tags		Users		Resources	
					

6.1. Systems and Datasets

As stated above, we will now briefly introduce a set of Social Annotation datasets, which will serve as a basis for the methods of capturing emergent semantics presented in Chapter 7. More precisely, we will introduce three broad folksonomies (BibSonomy, CiteULike and Delicious), a narrow one (Flickr), and two related datasets, namely a search engine clicklog (AOL logsonomy) and a snapshot from a Question-Answering platform (Stackoverflow).

6.1.1. BibSonomy

BibSonomy¹ is a social bookmark and publication sharing system developed at our group. Established in 2004, its user base consists currently mainly of students, scientists and knowledge workers. Being a broad folksonomy, it allows to share two kinds of resource types, namely URLs to web pages (i. e., bookmarks) and references to scientific papers. The latter is based on

¹<http://www.bibsonomy.org>

the BibTeX-format (Patashnik, 1988), commonly used especially in natural scientific areas. Resource identity is established for web pages by their URL; for publications, a *bibliographic hashkey*² is used, which is computed over a set of normalized metadata fields (namely title, author, editor and year). The BibSonomy team provides dumps for research purposes on a regular basis³. For the analyses presented in this thesis, a snapshot from September 2010 was used. In order to allow for an aggregated view on the data, bookmarks and publications were merged into a single dataset. Table 6.1 gives an overview about the statistical properties of the resulting folksonomy. Compared to the other datasets (which will be presented thereafter), the user base of BibSonomy is comparatively small. However, an important issue hereby is that BibSonomy performs an extensive semi-automatic filtering of inappropriate system usage by spammers (which is the topic of Section 8.4). The current cleaned dataset consists solely of non-spam contributions. The relatively small number of users makes the popular vocabulary sensitive towards very active users; as an example, the tag `zzz_to_sort` was used by just a single user, but is still the 2nd most popular tag within the whole system. We included this dataset to see whether this sensitivity has implications for the emergent semantics within the system. The last table row shows the frequency distributions of tags, users and resources. In all cases, the objects are found on the x -axis, ordered in descending order by their frequency; the y -axis depicts then the frequency of each object. Both axes are log-scaled. The shape of all curves exhibits the folksonomy-typical characteristics of long tailed distributions – i. e., there are few very frequent objects, and a large number of infrequent ones. The latter are denoted as the “long tail”.

6.1.2. CiteULike

CiteULike⁴ has a similar focus like the BibSonomy system, because it is also a broad folksonomy which facilitates the sharing of bibliographic references. Initiated in 2004 by Richard Cameron⁵, its user base now has a noticeable topical focus in the field of biology and genetics. This becomes visible in the list of most popular tags (see Table 6.2), where several keywords denoting a worm

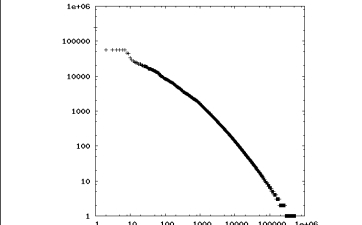
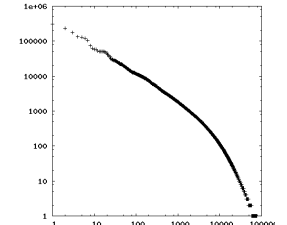
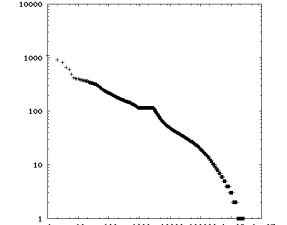
²http://www.gbv.de/wikis/cls/Bibliographic_Hash_Key

³<http://www.kde.cs.uni-kassel.de/bibsonomy/dumps>

⁴<http://www.citeulike.org>

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

Table 6.2.: Statistics about the CiteULike dataset.

<i>popular tags</i>	<i>Statistics</i>				
no-tag	<i>Complete dataset</i>				
bibtex-import	Tags	Users	Resources	TAS	Posts
elegans	549 145	77 846	2 656 969	12 014 185	3 480 953
celegans	<i>Restricted to 10 000 most popular tags</i>				
c_elegans	Tags	Users	Resources	TAS	Posts
nematode	10 000	72 249	2 373 352	8 856 879	3 073 457
caenorhabditis.elegans	<i>WordNet overlap of k most popular tags</i>				
wormbase	100	1 000	10 000	100 000	all
humans	82	808	5 593	23 008	45 764
<i>Frequency Distributions</i>					
<i>Tags</i>		<i>Users</i>		<i>Resources</i>	
					

organism called “caenorhabditis elegans” are found. CiteULike also provides regular data dumps⁶. For our study, we used a snapshot from September 2010. Its user base is roughly an order of magnitude larger than BibSonomy, and its vocabulary contains a larger portion of English words. It is an empirical question if these properties lead also to a larger degree of emergent semantics. The operators of CiteULike do not expose which kinds of preventions exist against system abuse and spam.

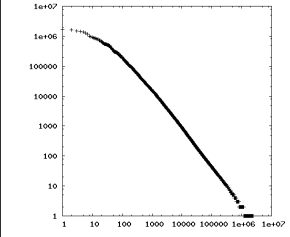
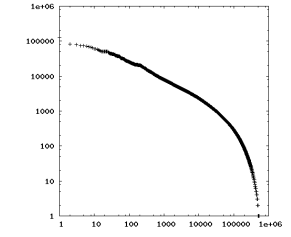
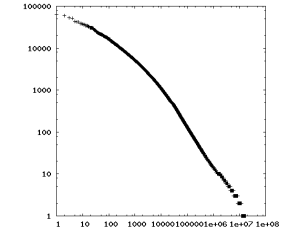
6.1.3. Delicious

Being a pioneer of the social bookmarking movement, Delicious⁷ is probably among the most well-known systems of its kind. Founded in 2003 by Joshua Schachter, it grew quickly and claimed to serve 5.3 million users sharing roughly

⁶<http://www.citeulike.org/faq/data.adp>

⁷<http://www.delicious.com>

Table 6.3.: Statistics about the Delicious dataset.

popular tags	Statistics				
design software blog web programming reference tools music css	<i>Complete dataset</i>				
	Tags	Users	Resources	TAS	Posts
	2 454 546	532 938	17 296 850	140 333 714	47 342 391
	<i>Restricted to 10 000 most popular tags</i>				
	Tags	Users	Resources	TAS	Posts
	10 000	511 348	14 567 465	117 319 016	42 202 093
	<i>WordNet overlap of k most popular tags</i>				
	100	1 000	10 000	100 000	all
	83	797	6 117	25 367	79 528
	<i>Frequency Distributions</i>				
Tags		Users	Resources		
					

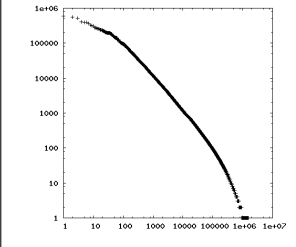
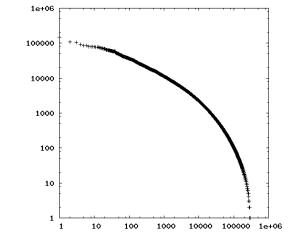
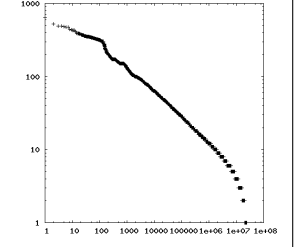
180 million unique URLs after 5 years of existence⁸. The Delicious dataset⁹ used within this thesis stems from a crawl performed in the context of the Tagora project¹⁰ in November 2006. Table 6.3 summarizes its statistical properties. Similar to BibSonomy and CiteULike, it is also a broad folksonomy; however, due to the much larger user base, the percentage of shared resources (i. e., those annotated by more than one user) is higher. Its popular vocabulary does have a bias towards technophilic areas (like web design and programming) (Michlmayr, 2005), but contains less idiosyncratic terms compared to, e. g., BibSonomy. This is also reflected in the comparatively high overlap of its vocabulary overlap with WordNet: Among the 100 most popular tags, 83 % are proper English words. Even when taking into account the top 10 000 tags, there is still a significant overlap of ≈ 61 %. Although not all tags can be mapped, one can see clearly here the existence of a vocabulary of “common” words. We included

⁸<http://blog.delicious.com/blog/2008/11/delicious-is-5.html>

⁹<https://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Research/DataSets/PINTSEperimentsDataSets/>

¹⁰<http://www.tagora-project.eu>

Table 6.4.: Statistics about the Flickr dataset.

<i>popular tags</i>	<i>Statistics</i>				
2005	<i>Complete dataset</i>				
wedding	Tags	Users	Resources	TAS	Posts
2004	1 547 678	298 954	24 599 875	110 345 103	24 599 875
party	<i>Restricted to 10 000 most popular tags</i>				
japan	Tags	Users	Resources	TAS	Posts
travel	10 000	271 359	21 633 082	72 002 331	21 633 082
family	<i>WordNet overlap of k most popular tags</i>				
friends	100	1 000	10 000	100 000	all
vacation	89	835	5 568	20 387	61 370
<i>Frequency Distributions</i>					
<i>Tags</i>		<i>Users</i>		<i>Resources</i>	
					

the Delicious dataset in our studies, because the system offers a large amount of data and has been analyzed by several other researchers. Furthermore, its high WordNet overlap allows a reliable evaluation of learned keyword structures.

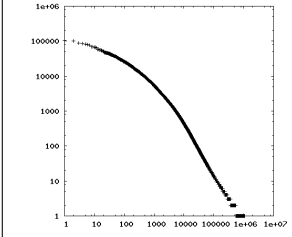
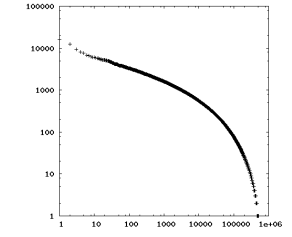
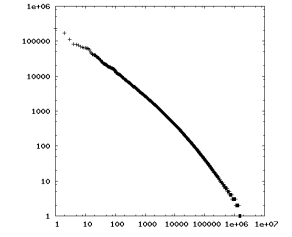
6.1.4. Flickr

As a representative example of a narrow folksonomy, Flickr¹¹ is a social photo sharing site. Launched in 2004, it was also among the “stars” of the Web 2.0 movement, it claimed to host 6 billion images in 2011¹². We use a dataset created within the Tagora project, more precisely a snapshot from the system containing roughly 25 million pictures, uploaded in 2004 and 2005 (see Table 6.4). Please note that within a narrow folksonomy, the number of posts is always equal to the number of resources, because each resource can only be annotated by its owner. Despite that, the frequency distribution of tags, users and resources show a similar behavior compared to broad folksonomies. Unsurprisingly, the

¹¹<http://www.flickr.com>

¹²<http://news.softpedia.com/news/Flickr-Boasts-6-Billion-Photo-Uploads-215380.shtml>

Table 6.5.: Statistics about the AOL logsonomy dataset.

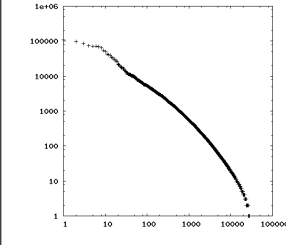
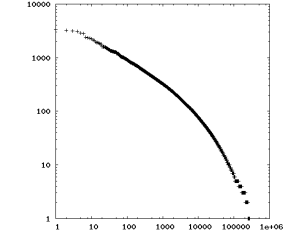
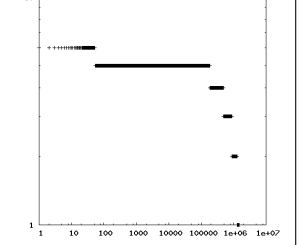
popular tags	Statistics				
free	<i>Complete dataset</i>				
	Tags	Users	Resources	TAS	Posts
county	1 074 640	519 203	1 619 871	34 500 590	12 758 653
pictures	<i>Restricted to 10 000 most popular tags</i>				
school	Tags	Users	Resources	TAS	Posts
lyrics	10 000	463 380	1 284 724	26 227 550	10 513 533
florida	<i>WordNet overlap of k most popular tags</i>				
sale	100	1 000	10 000	100 000	all
sex	95	944	8 166	36 154	65 235
google					
<i>Frequency Distributions</i>					
Tags		Users	Resources		
					

popular vocabulary comprises keywords denoting common situations for taking pictures, like *party* or *vacation*. Despite its obvious different nature, we included the Flickr dataset to assess to which extent methods from analyzing broad folksonomies are applicable to narrow folksonomies as well.

6.1.5. AOL Logsonomy

Having observed similar structural properties within search engine clicklogs (i. e., logsonomies as introduced in Section 3.2.4) and folksonomies, the question arises to which extent emergent semantics can be captured from this kind of data as well. For this purpose, we used a click dataset from the AOL search engine, collected from March, 1st to May, 31st 2006. Based on that, we constructed the logsonomy as presented in Definition 3.3. We hereby interpreted each individual query term (using the space character as separator) as a keyword. Since the AOL data was only available with truncated URLs, we reduced the URLs to host-only URLs, i. e., we removed the path of each URL leaving only the host name. Table 6.5 gives an overview of the result. A first interesting observation is that the popular vocabulary consists to a large extent from proper English

Table 6.6.: Statistics about the Stackoverflow dataset.

<i>popular tags</i>	<i>Statistics</i>				
<i>c#</i>	<i>Complete dataset</i>				
java	Tags	Users	Resources	TAS	Posts
php	28 221	272 882	1 468 486	4 307 918	1 468 486
javascript	<i>Restricted to 10 000 most popular tags</i>				
jquery	Tags	Users	Resources	TAS	Posts
.net	10 000	272 313	1 466 062	4 191 551	1 466 062
iphone	<i>WordNet overlap of k most popular tags</i>				
asp.net	100	1 000	10 000	100 000	all
c++	41	438	2 939	2 939	5 231
<i>Frequency Distributions</i>					
<i>Tags</i>		<i>Users</i>		<i>Resources</i>	
					

terms. Apart from that, the frequency distributions come close to the ones observed in the “real” annotation systems. We included this dataset as a representative of implicit annotations.

6.1.6. Stackoverflow

As a more controlled and hence more widely related example for social annotations, we used a dataset from the Question-and-Answering platform Stackoverflow¹³. Being part of the Stack Exchange network¹⁴, its focus lies within topics around computer programming. When posing a question, users are forced to use between one and five tags to categorize their inquiry. While these can be in principle freely chosen, the operators impose directly visible guidelines (like “favor existing popular tags, avoid creating new tags”, “don’t include synonyms” or “combine multiple words with dashes”). Furthermore, annotation is supported by an elaborated suggestion mechanism, which also includes a snippet explaining

¹³<http://www.stackoverflow.com>

¹⁴<http://stackexchange.com/>

the meaning of a particular keyword. Popular keywords also have an “info page”¹⁵ containing a detailed explanation and manually identified synonyms. We used a data dump from April 2011¹⁶ and converted the information of users annotating questions with tags into a (narrow) folksonomy-like structure, described in Table 6.6. Naturally, the specialized technical terminology has a small overlap with WordNet, and the resource frequency distribution mirrors directly the tagging limitations mentioned above. Despite these differences, the reason to include this dataset was to broaden the spectrum of types of analyzed data and to assess if a more controlled and supported way of annotation leads to a clearer kind of emergent semantics.

As a followup of the presentation of the Social Annotation datasets, the upcoming section deals with semantic resources used in the remainder of this dissertation.

6.2. Gold-standard Ontologies

As already mentioned in Section 5.4 in the context of evaluating ontology learning methods, an inherent problem of the automatic discovery and capturing of semantic structures is how to judge their quality. Because gold-standard based paradigms do have strong benefits (Dellschaft and Staab, 2006) for frequent and systematic evaluations (as required within the context of this thesis), we will make use of a number of “established” semantic resources. Hereby we are aware that the term “gold-standard” needs to be interpreted with caution, as the process of establishing and maintaining a “correct” knowledge repository is afflicted with difficulties and potential sources of error. Nevertheless we have chosen a set of resources which facilitate a process of *semantic grounding* of derived relations. Because the grounding process becomes more meaningful the more complete a mapping between learned and reference resources is, there is a tradeoff between coverage and semantic precision: While carefully crafted ontologies made by experts are usually smaller, but semantically precise, semi-automatically enriched resources or collaboratively created hierarchies are more fuzzy, but cover a greater amount of domain aspects and terms. The following choice of ontologies was guided by the motivation to select exemplars equally distributed between both ends of this scale: While *WordNet* is a prototype

¹⁵see <http://stackoverflow.com/tags/java/info> as an example for *java*

¹⁶<http://blog.stackoverflow.com/category/cc-wiki-dump/>

Table 6.7.: Statistical properties of the gold-standard datasets. C denotes the set of concepts, L_C the set of lexical items which is part of the associated lexicon, Ref_C is the lexical reference relation and \geq_C the taxonomic relation (see the ontology model defined in Section 4.1.1).

<i>Dataset</i>	$ C $	$ \geq_C $	$ L^C $	$ Ref_C $
<i>WORDNET</i>	79 690	81 866	141 391	141 692
<i>YAGO</i>	244 553	249 465	206 418	244 553
<i>WIKI</i>	2 445 974	4 447 010	2 445 974	2 445 974
<i>DMOZ</i>	767 019	767 019	241 910	767 019

of a comparatively small and precise resource, its semi-automatic extension in *YAGO* exhibits a much larger coverage. Finally, the category hierarchies derived from the *Open Directory project* and from *Wikipedia* focus even less on strict semantics, but cover the largest amount of concepts and terms.

6.2.1. WordNet

WordNet (Fellbaum, 1998) is a semantic lexicon of the English language. In WordNet, words are grouped into *synsets*, sets of synonyms that represent one concept. Synsets are nodes in a network and links between synsets represent semantic relations. The “meaning” of most synsets is described by means of a *gloss*, which is a short textual description (somewhat similar to dictionary definitions). WordNet provides a distinct network structure for each syntactic category (nouns, verbs, adjectives and adverbs). For nouns and verbs it is possible to restrict the links in the network to (directed) *is-a* relationships only, therefore a subsumption hierarchy can be defined. The *is-a* relation connects a *hyponym* (more specific synset) to a *hypernym* (more general synset). A synset can have multiple hypernyms, so that the graph is not a tree, but a directed acyclic graph. Further relations which are defined among synsets are, e. g., holonymy, meronymy or antonymy.

Compared to the remaining datasets, WordNet has the smallest coverage, but is probably also the semantically most “precise” one. This means that in general, one can assume, e. g., the taxonomic *is-a* relation to exhibit the strong semantics described in Section 4.1.3. Apart from that, WordNet is known for a comparatively fine-grained distinction among meanings of a given term.

6.2.2. YAGO

A popular dataset which can be seen as an “extension” of WordNet is YAGO, a large ontology which was derived automatically from Wikipedia and WordNet (Suchanek et al., 2007). Manual evaluation studies have shown that its precision (i. e., the percentage of correct facts) lies around 95 %. It has a much higher coverage than WordNet (see Table 6.7), because it also contains named entities like people, books or products. The complete ontology contains 1.7 million entities and 15 million relations; as our main interest lies in the taxonomy hierarchy, we restricted ourselves to the contained *is-a* relation¹⁷ among concepts.

YAGO is – as said – much larger than WordNet, but this comes at the cost of having not exclusively manually-defined relations. This means despite the precision is high, one cannot be sure that the taxonomic relation does in all cases exhibit strong semantics. However, even in such a case the hierarchical information can still be of great use for our evaluation purposes.

6.2.3. DMOZ

DMOZ¹⁸ (also known as the open directory project or ODP) is an open content directory for links of the World Wide Web. Although it is hierarchically structured, it differs from the above-mentioned datasets insofar as its internal link structure does not always reflect a sub-concept / super-concept relationship. Taking a cursory look at its category hierarchy reveals a rather “mixed” semantics of the contained links, which resembles much more the way how, e. g., people organize their bookmarks hierarchically – which is in a certain sense exactly what DMOZ was built for. Hence one has to be careful when talking about this dataset as an “ontology”. However, we included this dataset as a reference because it was built for a similar purpose than many social bookmarking systems, namely to organize references to web pages. Because it uses hereby the contrary approach of *hierarchical* structuring (in contrast to the flat paradigm of social bookmarking), it might be the case that the emergent semantic structures from Social Annotation data resemble to those found within DMOZ. Apart from that, especially for our analysis of term generality in Section 7.3, its category hierarchy is also a valuable resource, as its top level

¹⁷<http://www.mpi-inf.mpg.de/yago-naga/yago/subclassof.zip> (v2008-w40-2)

¹⁸<http://www.dmoz.org/>

categories (like “arts” or “business”) are described by rather abstract terms, becoming more specific towards the leaf categories.

6.2.4. Wikipedia Category Hierarchy

The last dataset used for evaluation is a “Wikitaxonomy”, which was derived from the Wikipedia category hierarchy (Ponzetto and Strube, 2007). This large scale domain independent taxonomy¹⁹ was derived by evaluating the semantic network between Wikipedia concepts and labeling the relations as *is-a* and *not-is-a*, using methods based on the connectivity of the network and on lexico-syntactic patterns. It contains by far the largest number of lexical items (see Table 6.7), but this comes at the cost of a much lower level of manual control. Despite this, the interesting point about this reference dataset is that the assignment of Wikipedia pages to user-defined categories has also been interpreted as a kind of “social annotation” (Voss, 2007) – however, in contrast to, e. g., typical social bookmarking systems, including a structure among the categories used for annotations. Hence we expect this dataset to serve as the most “closely related” gold-standard, and it will be especially interesting to see if the learned semantic relations resemble those stemming from Wikipedia annotators.

After this presentation of the social and semantic datasets which are used for learning and evaluation purposes, the following main chapter of this dissertation presents methods to capture emergent semantic structures within Social Annotation data.

¹⁹<http://www.h-its.org/english/research/nlp/download/wikitaxonomy.php>

Chapter 7.

Methods

Within the previous chapter of this dissertation, social and semantic approaches to Knowledge Organization on the Web were introduced, as well as the promising idea to use ontology learning approaches to “bridge the gap” between both worlds. Using the datasets described lastly, the current chapter can be seen as a core part of this dissertation, which introduces several concrete methods and algorithms to capture emergent semantics in various forms. Similar to the description of the state of the art in this direction (see Section 5.3), we will broadly stick hereby to the structuring given by the different tasks contained in the ontology learning layer cake (see Section 5.2.2). More precisely, given the observed differences between “traditional” ontology learning, we will propose an adapted layer model for capturing emergent semantics from Social Annotation data; it is displayed in Figure 7.1.

It comprises basically some levels of the ontology learning layer cake (like, e. g., learning concepts or concept hierarchies), but adds two layers (mainly the measures of semantic relatedness / generality) which have shown to be useful for our purpose at hand. Furthermore, the “keywords” layer (which corresponds to the term layer in the original version) is of lesser importance, because a positive aspect about most Social Annotation Systems is that one can directly interpret the contained keywords as terms. Finally, higher levels which seem to be hardly reachable from the current point of view (like learning relations or axioms) are represented in a condensed way.

Based on this model, we will now present in a bottom-up manner methods which tackle the respective task of each layer. Because a prerequisite for higher levels is a precise understanding of different notions of keyword relatedness, we will start with a systematic analysis to this end. Based on that, we will continue by presenting methods of concept learning, mainly concerned with tackling two problems of Social Annotation Systems, namely synonymy and polysemy. As a further prerequisite to induce structure into the initially flat tag space, we will

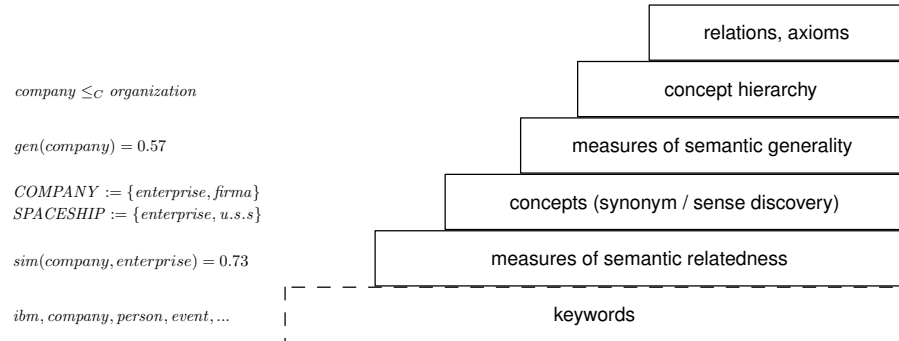


Figure 7.1.: Layer cake model of capturing emergent semantics from Social Annotation data.

then have a look at different folksonomy-derived notions of tag generality. The latter are an integral input for the last class of methods, which targets towards learning concept hierarchies from keywords.

7.1. Capturing Semantic Relatedness

As pointed out in Section 4.2.1, the *explicit* representation of knowledge within ontologies allows to derive measures of *semantic relatedness*, which encode the degree of likeness between concepts. Although concepts are not explicitly present within Social Annotation data, the reported evidences for emergent semantics (see Section 5.2.1) suggest that such measures could also be derived from these emergent *implicit* structures. Because concepts inherently do not exist, the goal hereby is to compute relatedness among the keywords used for annotation. In analogy to Definition 4.5, we will denote such a measure based on a folksonomy structure $\mathbb{F} = (U, T, R, Y)$ as $\rho_T : T \times T \rightarrow \mathbb{R}^+$. Because these keywords can be considered as natural language entities, one way of defining such measures is to map the keywords to explicitly defined concepts within, e. g., a thesaurus or lexicon like the ones presented in Section 6.2, and to compute keyword relatedness based on the well-known metrics among the mapped concepts. However, this approach has the following drawbacks:

- When observing the dynamic vocabulary of Social Annotations, it turns out that it contains many community-specific terms, neologisms and other

lexical items which are with high probability not (yet) present in any external lexical resource. Hence a restriction to well-established lexicon vocabulary would leave out a significant portion of metadata. Furthermore, this would narrow the possibility to include “matured” Social Annotation vocabulary back into the thesauri or lexicons, which could potentially address the inherent knowledge acquisition bottleneck problem of these systems.

- The process of mapping keywords to concepts is inherently afflicted with problems like polysemy or homonymy. Hence it is a non-trivial task to identify the “correct” concept for a given keyword. Non-optimal choices would lead here to inappropriate relatedness assessments.
- Even when a lexical item is present within an external resource, one can not be sure to which extent its usage and meaning in the Social Annotation System corresponds to the meaning captured in the external resource. As an example, the technical meaning of the keyword `ajax`¹ is not necessarily present within a lexicon, where it might be associated to a Greek mythology figure.

For the above reasons, it is highly desirable to define measures of relatedness directly on the network structure of the Social Annotation System. From a linguistic point of view, this corresponds to a structuralist approach of lexical semantics (de Saussure, 1916). Because relatedness is – from a semantic point of view – a relatively unspecific relation, a deeper understanding in the characteristics of different relatedness measures is an indispensable prerequisite for capturing more precise semantic relations. For this reason, we will present in this chapter a detailed analysis of several relatedness measures, partially inspired by corpus-based approaches (Lin, 1998; Cimiano, 2006), information theory and information retrieval. We will start with a presentation of three main classes of measures (namely co-occurrence based, distributional and graph-based), followed by a qualitative evaluation of their respective properties. We then go one step further and present an in-depth analysis of the semantic characteristics of each measure based on a grounding approach against an external lexical resource (namely WordNet). After that, we present alternative aggregation and

¹AJAX (Asynchronous JavaScript and XML) is a programming technique used in many web applications.

weighting approaches, whose intention is to reduce complexity and enhance the quality of the captured relations. In summary, we are laying the groundwork in this section for choosing an appropriate measure of relatedness for the different tasks of concept learning presented in Section 7.2.

7.1.1. Relatedness Measures

Measures of keyword relatedness in a Social Annotation System can be defined in several ways. Most of these definitions use statistical information derived directly from the tripartite structure or from induced networks (see Section 3.1.3 for an overview). Hereby especially the different types of *co-occurrence* networks are playing an important role. Please recall that a co-occurrence event between two keywords corresponds to their common usage either within (i) the same resource, (ii) the same post or (iii) the same user. This relationship of “direct contact” is referred to as *first-order co-occurrence* (Rapp, 2002) in corpus-based approaches. Other approaches adopt the *distributional hypothesis* (Firth, 1957; Harris, 1968), which states that words found in similar contexts tend to be semantically similar. Of course a crucial question hereby is what kind of “contextual” information is taken into account. A typical approach having its roots in the field of information retrieval and automatic text processing is to project a word into a suitable vector space, whose dimensions correspond to contextual semantic features. As a last class, *graph-based* approaches are operating directly on the tripartite annotation graph, computing relative relevancy among the contained items.

From a linguistic point of view, the first two families of measures focus on orthogonal aspects of structural semiotics (de Saussure, 1916; Chandler, 2007). The first-order co-occurrence measures address the so-called *syntagmatic* relation, where words are considered related if they occur in the same part of text. The contextual measures address the *paradigmatic* relation (originally called associative relation by Saussure), where words are considered related if they can replace one another without affecting the structure of the sentence.

Apart from the aforementioned three broad classes of approaches, the following issues need to be considered when designing a measure of semantic relatedness based on Social Annotation Systems:

- *Co-occurrence*: Which kind of co-occurrence (post-based, resource-based, user-based) is used?

- *Aggregation*: When projecting the tripartite network structure into two-dimensional vectors, how are the values aggregated?
- *Weighting Scheme*: Which weighting scheme is used to post-process the co-occurrence counts or the vector entries?
- *Vector Similarity*: Which metric is chosen to measure similarity among the projected vectors?

For each individual design choice there exist several candidate methods stemming from different fields. Furthermore, dataset characteristics might also have an influence. In order to disentangle the effects of the resulting large number of possible combinations, our approach is to start with our largest annotation dataset (namely Delicious as introduced in Section 6.1.3) and a systematic analysis of “standard representatives” of each dimension. More precisely, these are five measures: First, the post-based *co-occurrence count*; then *three distributional measures* which use the cosine similarity (Salton, 1989) in the vector spaces spanned by users, tags, and resources, respectively; and finally *FolkRank* (see Section 3.1.5), a graph-based measure that is an adaptation of PageRank (Brin and Page, 1998) to folksonomies. For those, we will present a in-depth analysis in a qualitative (Section 7.1.2) and semantically grounded (Section 7.1.3) manner. In order to avoid noise caused by the inherent sparseness of the Delicious folksonomy, we restricted our dataset to the 10 000 most frequent tags of Delicious, and to the resources/users that have been associated with at least one of those tags. One could argue that tags with low frequency have a higher information content in principle – but despite this fact they are less useful for the study of both co-occurrence and distributional measures. In the sequel, we will focus on alternative aggregation and weighting schemes and vector similarity measures (Section 7.1.4).

Co-Occurrence The most “direct” way to assess the relatedness between two keywords is to count how often they were used together within the same post. This corresponds to computation of the edge weights within the *post-based* tag co-occurrence graph described in Section 3.1.3. Please recall that the set of nodes of this graph is the set of tags T of the folksonomy (U, T, R, Y) , and that the co-occurrence count for a pair of tags (t_1, t_2) is incremented each time t_1 and t_2 were used to annotate the same resource by the same user. Based on this graph, the co-occurrence relatedness between tags is given directly by the

edge weights. For a given tag $t \in T$, the tags that are most related to it are thus all the tags $t' \in T$ with $t' \neq t$ such that $w(t, t')$ is maximal. We will denote this co-occurrence relatedness by ρ_T^{co-occ} or simply *co-occ*. For its computation, we first create a sorted list of all tag pairs which occur together in a post. The complexity of this can be estimated as $O(\frac{|Y|^2}{2|P|} \log(\frac{|Y|^2}{2|P|}))$. Then, we group this list by each tag and sort by count, which corresponds to an additional complexity of $O(|T|^2 \log(|T|^2))$. Y, P, T denote the set of tag assignments, posts and tags, respectively (see Section 3.1.2).

Distributional measures For capturing distributional relatedness, we adopt the standard approach of vector space representations. A core aspect hereby – and thus a core aspect of the measures – is the feature space used to describe the keywords. Having users, tags and resources as possible dimensions of the folksonomy, we vary over these and introduce three representations. Specifically, for $X \in \{U, T, R\}$ we consider the vector space \mathbb{R}^X , where each tag t is represented by a vector $v_t \in \mathbb{R}^X$, as described below.

- *Tag Context Similarity.* The Tag Context Similarity ($\rho_T^{TagCont}$ or *TagCont*) is computed in the vector space \mathbb{R}^T , where, for tag t , the entries of the vector $v_t \in \mathbb{R}^T$ are defined by $v_{tt'} := w(t, t')$ for $t \neq t' \in T$, where w is the co-occurrence weight defined above, and $v_{tt} = 0$. The reason for giving weight zero between a node and itself is that we want two tags to be considered related when they occur in a similar context, and not when they occur together. The complexity of this measure comprises the cost of computing co-occurrence (see above), i. e., $O(\frac{|Y|^2}{2|P|} \log(\frac{|Y|^2}{2|P|}) + |T|^2 \log(|T|^2))$, plus the cost of comparing each tag pair, which is $O(|T|^2 2|X|)$, $X \subseteq T$.
- *Resource Context Similarity.* The Resource Context Similarity ($\rho_T^{ResCont}$ or *ResCont*) is computed in the vector space \mathbb{R}^R . For a tag t , the vector $v_t \in \mathbb{R}^R$ is constructed by counting how often a tag t is used to annotate a certain resource $r \in R$: $v_{tr} := |\{u \in U \mid (u, t, r) \in Y\}|$. In terms of complexity, the tag-resource counts amount for $O(|Y| \log(|Y|))$, plus the pairwise comparison cost of $O(|T|^2 2|R|)$.
- *User Context Similarity.* The User Context Similarity ($\rho_T^{UserCont}$ or *UserCont*) is built similarly to ResCont, by swapping the roles of the sets R and U : For a tag t , the vector $v_t \in \mathbb{R}^U$ is defined as $v_{tu} := |\{r \in R \mid (u, t, r) \in Y\}|$. In this case, the complexity is $O(|Y| \log(|Y|) + |T|^2 2|U|)$.

In all three representations, we adopt a standard method from information retrieval (Salton, 1989) and measure vector similarity by using the cosine measure: If two tags t_1 and t_2 are represented by $v_1, v_2 \in \mathbb{R}^X$, their cosine similarity is defined as:

$$\text{cossim}(t_1, t_2) := \cos \angle(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\|_2 \cdot \|v_2\|_2}.$$

The cosine similarity is thus independent of the length of the vectors. Its value ranges from 0 (for totally orthogonal vectors) to 1 (for vectors pointing into the same direction).

FolkRank As described in Section 3.1.5, FolkRank employs the principle of the PageRank algorithm (Brin and Page, 1998) to folksonomies: A resource which is tagged with important tags by important users becomes important itself. The same holds, symmetrically, for tags and users. By modifying the weights for a given tag t in the random surfer vector, FolkRank can compute a ranked list of relevant tags for t . To establish the connection to searching the web, this corresponds to “querying” the FolkRank search engine with t , getting relevant other tags as results. We then interpret relevancy as an indicator of relatedness, and assume the most relevant tags to be semantically related.

More specifically, we have set the weights in the random surfer vector as follows: Initially, each tag is assigned weight 1. Then, the weight of the given tag t is increased according to $w(t) = w(t) + |T|$. Afterwards, the vector is normalized. The random surfer has an influence of 15% in each iteration. The tags that, for a given tag t , obtain the highest FolkRank score are considered to be the most relevant in relation to t . This measure will be denoted as ρ_T^{FolkRank} or simply *FolkRank*. Its complexity can be estimated as $O(i|Y|)$, where i is the number of iterations (the typical values used in this study were 30-35).

7.1.2. Qualitative Evaluation

Using each of the measures introduced above, we computed, for each of the 10 000 most frequent tags of Delicious, its most closely related tags. As we used different (partially existing) implementations for the measures we investigate, runtimes do not provide meaningful information on the computational cost of the different measures. We refer the reader to the prior discussion on computational complexity.

Table 7.1.: Examples of most related tags for each of the presented measures co-occurrence (CO), FolkRank (FR), tag context (TC), resource context (RC) and user context (UC) relatedness.

tag	meas.	1	2	3	4	5
web2.0	CO	ajax	web	tools	blog	webdesign
	FR	web	ajax	tools	design	blog
	TC	web2	web-2.0	webapp	“web	web_2.0
	RC	web2	web20	2.0	web_2.0	web-2.0
	UC	ajax	aggregator	rss	google	collaboration
howto	CO	tutorial	reference	tips	linux	programming
	FR	reference	linux	tutorial	programming	software
	TC	how-to	guide	tutorials	help	how_to
	RC	how-to	tutorial	tutorials	tips	diy
	UC	reference	tutorial	tips	hacks	tools
games	CO	fun	flash	game	free	software
	FR	game	fun	flash	software	programming
	TC	game	timewaster	spiel	jeu	bored
	RC	game	gaming	juegos	videogames	fun
	UC	video	reference	fun	books	science
java	CO	programming	development	opensource	software	web
	FR	programming	development	software	ajax	web
	TC	python	perl	code	c++	delphi
	RC	j2ee	j2se	javadoc	development	programming
	UC	eclipse	j2ee	junit	spring	xml
opensource	CO	software	linux	programming	tools	free
	FR	software	linux	programming	tools	web
	TC	open_source	open-source	open.source	oss	foss
	RC	open-source	open	open_source	oss	software
	UC	programming	linux	framework	ajax	windows
tobuy	CO	shopping	books	book	design	toread
	FR	toread	shopping	design	books	music
	TC	wishlist	to_buy	buyme	wish-list	iwant
	RC	wishlist	shopping	clothing	tshirts	t-shirts
	UC	toread	cdm	todownload	todo	magnet

Table 7.1 provides a few examples of the related tags returned by the measures under study. A first observation is that in many cases the tag and resource context similarity provide more synonyms than the other measures. For instance, for the tag *web2.0* they return some of its alternative spellings.² For the tag *games*, the tag and resource similarity also provide tags that could be regarded as semantically *similar*. For instance, the morphological variations *game* and

²The tag “*web*” at the fourth position (tag context) is likely to stem from users who typed ‘*web 2.0*’, which the early Delicious interpreted as two separate tags, ‘*web* and *2.0*’.

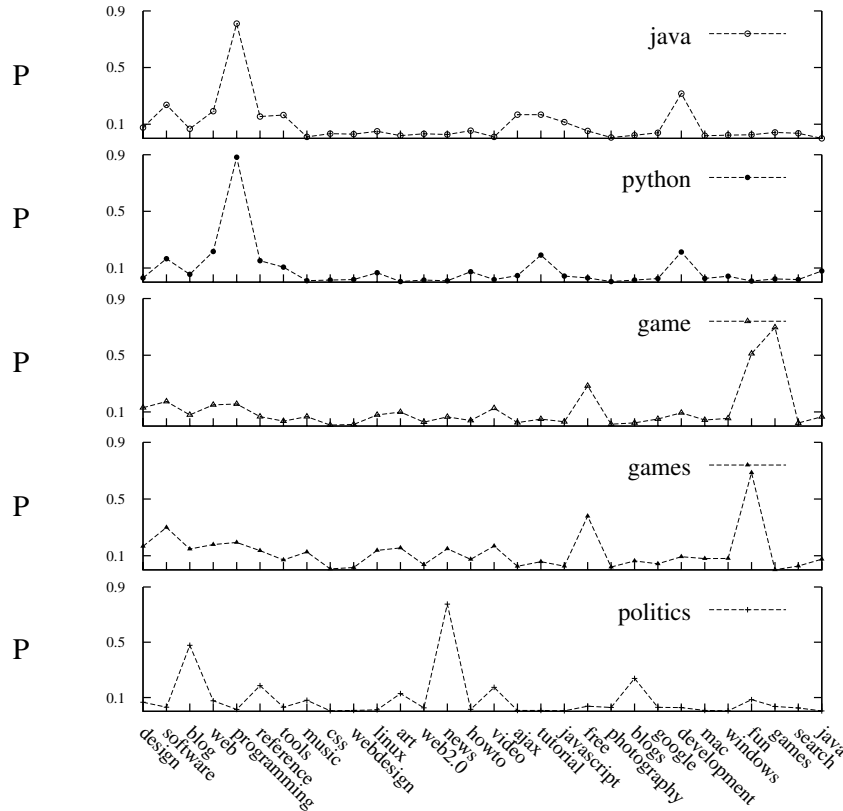


Figure 7.2.: Tag co-occurrence fingerprint of five selected tags in the first 30 dimensions of the tag vector space.

gaming, or corresponding words in other languages, like *spiel* (German), *jeu* (French) and *juegos* (Spanish). This effect is not obvious for the other measures, which tend to provide rather *related* tags instead (*video*, *software*). The same observation holds for the “functional” tag *tobuy* (see (Golder and Huberman, 2006)), for which the tag context similarity provides tags with equivalent functional value (*to_buy*, *buyme*), whereas the FolkRank and co-occurrence measures provide categories of items one could buy. The user context similarity also yields a remarkable amount of functional tags, but with different

target actions (*toread*, *todownload*, *todo*). The latter might be due to the behavior that if a user utilizes functional tags, he will use them for more than one action. The aggregation across all “functional tag users” would then yield high correlations among these tags.

An interesting observation about the tag *java* is that *python*, *perl* and *c++* (provided by tag context similarity) could all be considered as siblings in some suitable concept hierarchy, presumably under a common parent concept like **PROGRAMMING LANGUAGES**. An approach to explain this behavior is that the tag context is measuring the frequency of co-occurrence with other tags *in the global context* of the folksonomy, whereas the co-occurrence measure and – to a lesser extent – FolkRank measure the frequency of co-occurrence with other tags *in the same posts*.

Another insight offered by this first visual inspection is that context similarities for tags and resources seem to yield equivalent results, especially in terms of synonym identification. The tag context measure, however, seems to be the only one capable of identifying sibling tags, as it is visible for the case of *java* in Table 7.1. This is also visible in Figure 7.2, which displays the tag co-occurrence vectors of 5 selected tags. The vectors are restricted to co-occurrence with the 30 most frequent tags of the folksonomy, i. e., to only 30 dimensions of the vector space \mathbb{R}^T introduced earlier in this section.³ The figures shows that both *java* and *python* appear frequently together with *programming*, and (to a lesser degree) with *development*. These two common peaks alone contribute approximately 0.68 to the total cosine similarity of the two tags *java* and *python* of 0.85.

A similar behavior can be seen for *game* and *games* both displaying peaks at *fun* and (to a lesser degree) *free*. Here we also see the effect of imposing $v_{tt} = 0$ in the definition of the cosine measure: while the tag *game* has a very high peak at *games*, the tag *games* has by definition a zero component there. The high value for tag *game* in the dimension *games* shows that these two tags are frequently assigned together to resources (probably because users anticipate that they will not remember a specific form at the time of retrieval).

In the case of *python*, on the other hand, we observe that it seldom co-occurs with *java* in the same posts (probably because few web pages deal with both java and python). Hence – even though *python* and *java* are “most related” according to the tag context similarity – they are less so according to the other

³The length of all the vectors was normalized to 1 in the L_2 -norm.

Table 7.2.: Overlap between the 10 most closely related tags according to the measures under consideration.

	<i>co-occurrence</i>	<i>FolkRank</i>	<i>tag context</i>	<i>resource context</i>
<i>user context</i>	1.77	1.81	1.35	1.55
<i>resource context</i>	3.35	2.65	2.66	
<i>tag context</i>	1.69	1.28		
<i>FolkRank</i>	6.81			

measures. In fact, in the lists of tags most closely related to *java*, *python* is at position 21 according to FolkRank, 34 according to co-occurrence, 97 according to user context similarity, and 476 according to resource context similarity.

Our next step is to substantiate these first insights with a more systematic analysis. We start by using simple observables that provide qualitative insights into their behavior.

The first natural aspect to investigate is whether the most closely related tags are shared across measures of relatedness. We consider the 10 000 most popular tags in Delicious, and for each of them we compute the 10 most related tags according to each of the relatedness measures. Table 7.2 reports the average number of shared tags for the relatedness measures we investigate. We first observe that the user context measure does not exhibit a strong similarity to any of the other measures. The same holds for the tag context measure, with a slightly higher overlap of 2.65 tags with the resource context measure. Based on the visual inspection above, this can be attributed to shared synonym tags. A comparable overlap also exists between resource context and FolkRank / co-occurrence similarity, respectively. Based on the current analysis, it is hard to learn much on the nature of these overlapping tags. A remarkable fact, however, is that relatedness by co-occurrence and by FolkRank share a large fraction (6.81) of the 10 most closely related tags. That is, given a tag t , its related tags according to FolkRank are – to a large extent – tags with a high frequency of co-occurrence with t . In the case of the context relatedness measures, instead, the suggested tags seem to bear no special bias towards high-frequency tags. This is due to the normalization of the vectors that is implicit in the cosine similarity, which disregards information about global tag frequency.

To better investigate this point, for each of the 10 000 most frequent tags in Delicious we computed the average rank (according to global frequency) of its

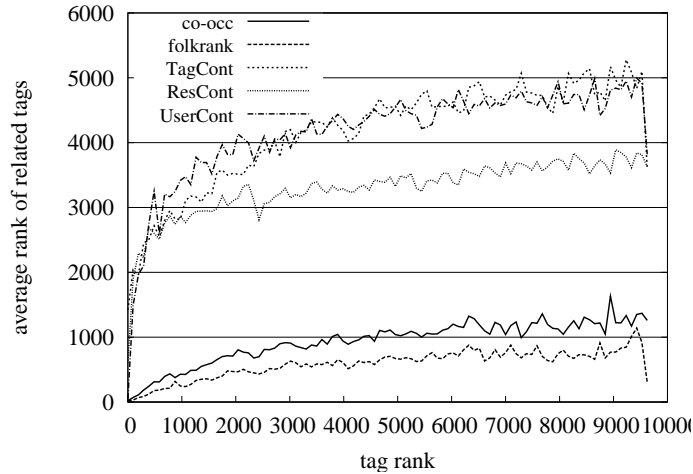


Figure 7.3.: Average rank of the related tags as a function of the rank of the original tag.

10 most closely related tags, according to each of the relatedness measures under study. Figure 7.3 shows the average rank of the related tags as a function of the original tag's rank. The average rank of the tags obtained by co-occurrence relatedness and by FolkRank is low and increases slowly with the rank of the original tag: this points out that most of the related tags are high-frequency tags, independently of the original tag. On the contrary, the context (distributional) measures display a different behavior: the rank of related tags increases much faster with that of the original tag. That is, the tags obtained from context relatedness span a broader range of ranks.⁴

7.1.3. Evaluation by Semantic Grounding

In this section we shift perspective and move from the qualitative discussion of Section 7.1.2 to a more formal validation. Our strategy is to ground the relations between the original and the related tags by looking up the tags in a formal representation of word meanings. As structured representations of

⁴Notice that the curves for the tag and user context relatedness approach a value of ≈ 5000 for high ranks: this is the value one would expect if the rank of the related tags was independent from the rank of the original tags.

knowledge afford the definition of well-defined metrics of semantic similarity (see Section 4.2.1, one can investigate the type of *semantic* relations that hold between the original tags and their related tags, defined according to any of the relatedness measures under study).

In the following we ground our measures of tag relatedness by using WordNet (see Section 6.2.1, a semantic lexicon of the English language. This choice was guided by its large coverage of English terms (i. e., a language also frequently used on the Web), and its careful engineering by language experts. As a consequence, the semantic relations among its comprised terms can be regarded as precise and well-defined. Furthermore, there exist a number of well-established measures of semantic similarity based on WordNet – which is exactly what is required for our presented methodology of semantic grounding. Most WordNet-based measures of semantic similarity take into account its taxonomic *is-a* relation. Since the *is-a* WordNet network for nouns and verbs consists of several disconnected hierarchies, it is useful to add a fake top-level node subsuming all the roots of those hierarchies, making the graph fully connected and allowing the definition of several graph-based similarity metrics between pairs of nouns and pairs of verbs. We will use such metrics to ground and characterize our measures of tag relatedness in folksonomies. In WordNet, we will measure the semantic similarity by using both the taxonomic shortest-path length and a measure of semantic distance introduced by Jiang and Conrath (Jiang and Conrath, 1997) that combines the taxonomic path length with an information-theoretic similarity measure by Resnik (Resnik, 1995).

For our studies, we used the implementation of those measures available in the `WordNet::Similarity` library⁵, using WordNet 2.1 as basis. It is important to remark that (Budanitsky and Hirst, 2006) provides a pragmatic grounding of the Jiang-Conrath measure by means of user studies and by its superior performance in the context of a spell-checking application. Thus, our semantic grounding in WordNet of the similarity measures is extended to the pragmatic grounding in the experiments of (Budanitsky and Hirst, 2006).

The program outlined above is only viable if a significant fraction of the popular tags in Delicious is also present in WordNet. Several factors limit the WordNet coverage of Delicious tags: WordNet only covers the English language and contains a static body of words, while Delicious contains tags from different languages, tags that are not words at all, and is an open-ended system. Another

⁵<http://search.cpan.org/dist/WordNet-Similarity/>

Table 7.3.: WordNet coverage of Delicious tags.

# top-frequency tags	100	500	1 000	5 000	10 000
fraction in WordNet	82 %	80 %	79 %	69 %	61 %

limiting factor is the structure of WordNet itself, where the measures described above can only be implemented for nouns and verbs, separately. Many tags are actually adjectives (Golder and Huberman, 2006) and although their grounding is possible, no distance based on the subsumption hierarchy can be computed in the adjective partition of WordNet. Nevertheless, the nominal form of the adjective is often covered by the noun partition. Despite this, if we consider the popular tags in Delicious, a significant fraction of them is actually covered by WordNet: as shown in Table 7.3, roughly 61 % of the 10 000 most frequent tags in Delicious can be found in WordNet. In the following, to make contact with the previous sections, we will focus on these tags only.

A first assessment of the measures of relatedness can be carried out by measuring – in WordNet – the average semantic distance between a tag and the corresponding most closely related tag according to each one of the relatedness measures we consider. Given a measure of relatedness, we loop over the tags that are both in Delicious and WordNet, and for each of those tags we use the chosen measure to find the corresponding most related tag. If the most related tag is also in WordNet, we measure the semantic distance between the synset that contains the original tag and the synset that contains the most closely related tag. When measuring the shortest-path distance, if either of the two tags occurs in more than one synset, we use the pair of synsets which minimizes the path length.

Figure 7.4 reports the average semantic distance between the original tag and the most related one, computed in WordNet by using both the (edge) shortest-path length and the Jiang-Conrath distance. The tag and resource context relatedness point to tags that are semantically closer according to both measures. We remark once more that the Jiang-Conrath measure has been validated in user studies (Budanitsky and Hirst, 2006), and because of this the semantic distances reported in Figure 7.4 correspond to distances cognitively perceived by human subjects.

The best performance is achieved by similarity according to resource context. This is not surprising as this measure makes use of a large amount of contextual

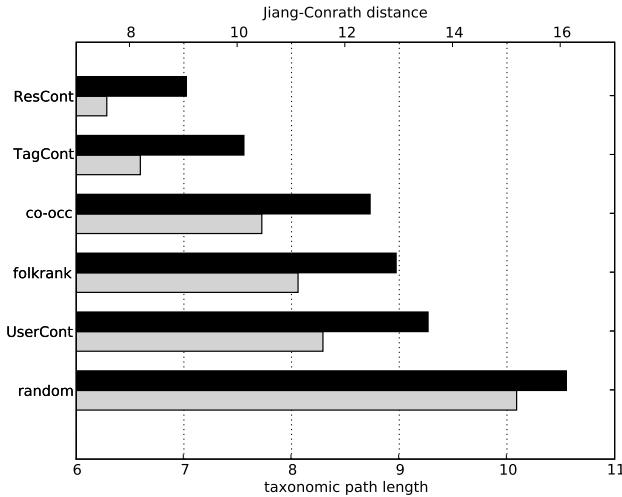


Figure 7.4.: Average semantic distance, measured in WordNet, from the original tag to the most closely related one. The distance is reported for each of the measures of tag similarity discussed in the main text (labels on the left). Grey bars (bottom) show the taxonomic path length in WordNet. Black bars (top) show the Jiang-Conrath measure of semantic distance.

information (the large vectors of resources associated with tags). While similarity by resource context is computationally very expensive to compute, it can be used as a reference for comparing the performance of other measures. To this end, we also computed the distances for the worst case scenario of a measure (marked as *random* in Figure 7.4) that associates every tag with a randomly chosen one. All the other measures of relatedness fall between the above extreme cases. Overall, the taxonomic path length and the Jiang-Conrath distance appear strongly correlated, and they induce the same ranking by performance of the similarity measures. Remarkably, the notion of similarity by tag context (*TagCont*) has an almost optimal performance. This is interesting because it is computationally lighter than the similarity by resource context, as it involves tag co-occurrence with a fixed number (10 000) of popular tags, only. The closer semantic proximity of tags obtained by tag and resource context relatedness was intuitively apparent from direct inspection of Table 4.1, but now we are able to

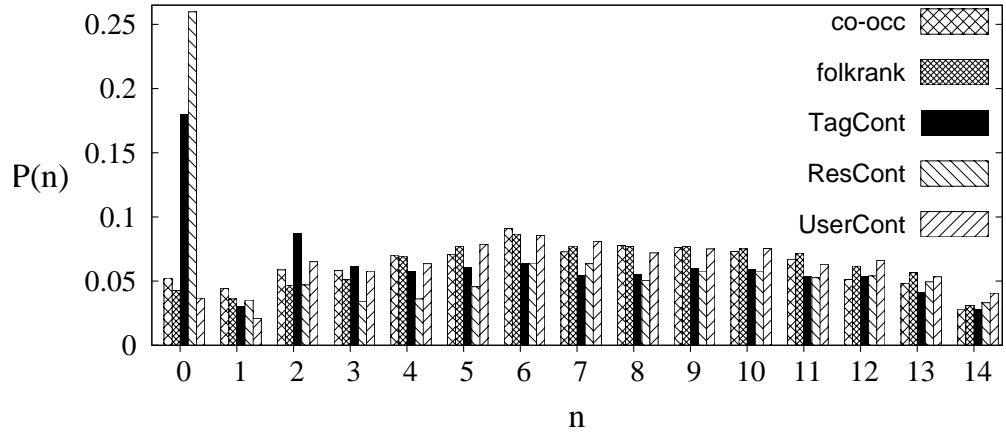


Figure 7.5.: Probability distribution for the lengths of the shortest path leading from the original tag to the most closely related one. Path lengths are computed using the subsumption hierarchy in WordNet.

ground this statement through user-validated measures of semantic similarity based on the subsumption hierarchy of WordNet.

As already noted in Section 7.1.2, the related tags obtained via tag context or resource context appear to be “synonyms” or “siblings” of the original tag, while other measures of relatedness (co-occurrence and FolkRank) seem to provide “more general” tags. The possibility of looking up tags in the WordNet hierarchy allows us to be more precise about the nature of these relations. In the rest of this section we will focus on the shortest paths in WordNet that lead from an initial tag to its most closely related tag (according to the different measures of relatedness), and characterize the length and edge composition (hypernym/hyponym) of such paths.

Figure 7.5 displays the normalized distribution $P(n)$ of shortest-path lengths n (number of edges) connecting a tag to its closest related tag in WordNet. All similarity measures share the same overall behavior for $n > 3$, with a broad maximum around $n \simeq 6$, while significant differences are visible for small values of n . Specifically, similarity by tag context and resource context display a strong peak at $n = 0$. Tag context similarity also displays a weaker peak at $n = 2$ and a comparatively depleted number of paths with $n = 1$. For higher values of n , the histogram for resource context and tag context has the same shape as the

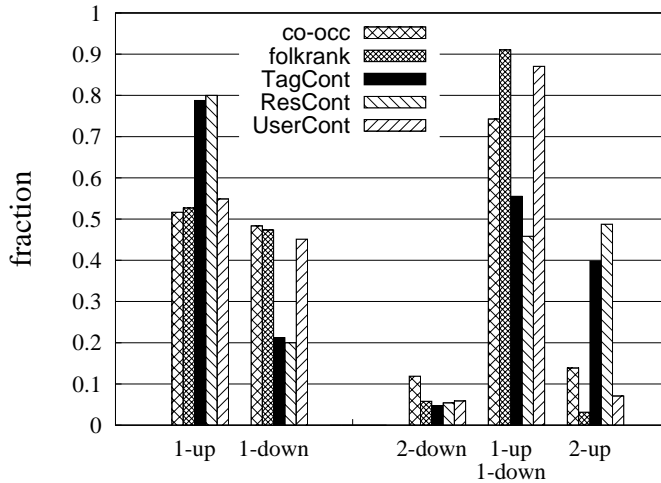


Figure 7.6.: Edge composition of the shortest paths of length 1 (left) and 2 (right). An “up” edge leads to a hypernym, while a “down” edge leads to a hyponym.

others, but is systematically lower due to the abundance of very short paths and the normalization of $P(n)$. The peak at $n = 0$ is due to the detection of actual synonyms in WordNet. As nodes in WordNet are synsets, a path to a synonym appears as an edge connecting a node to itself (i. e., a path of length 0). Similarity by tag context points to a synonym in about 18% of the cases, while using resource context this figure raises to about 25%. In the above cases, the most related tag is a tag belonging to the same synset of the original tag. In the case of tag context, the smaller number of paths with $n = 1$ (compared with $n = 0$ and $n = 2$) is consistent with the idea that the similarity of tag context favors siblings/synonymous tags: moving by a single edge, instead, leads to either a hypernym or a hyponym in the WordNet hierarchy, never to a sibling. The higher value at $n = 2$ (paths with two edges in WordNet) for tag context may be compatible with the sibling relation, but in order to ascertain this we have to characterize the typical edge composition of these paths.

Figure 7.6 displays the average edge type composition (hypernym/hyponym edges) for paths of length 1 and 2. The paths analyzed here correspond to $n = 1$ and $n = 2$ in Figure 7.5. For tag context, resource context and user context, we

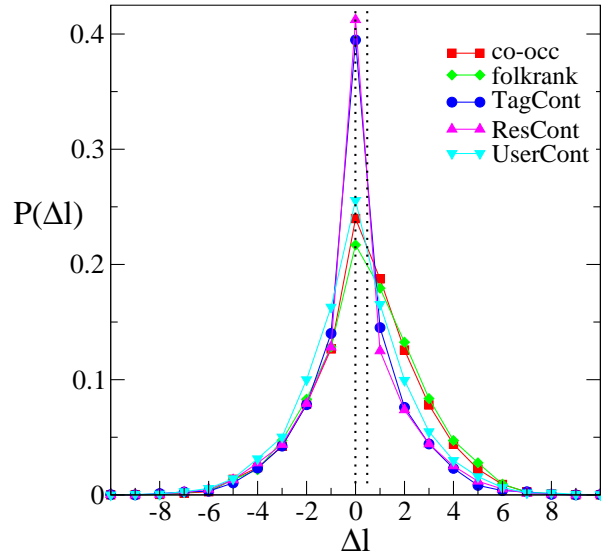


Figure 7.7.: Probability distribution of the level displacement Δl in the WordNet hierarchy.

observe that the paths with $n = 2$ (right-hand side of Figure 7.6) consist almost entirely of one hypernym edge (up) and one hyponym edge (down), i. e., these paths do lead to siblings. This is especially marked for the notion of similarity based on tag context, where the fraction of paths leading to a sibling is about 90 % of the total. Notice how the path composition is very different for the other non-contextual measures of relatedness (co-occurrence and FolkRank): in these cases roughly half of the paths consist of two hypernym edges in the WordNet hierarchy, and the other half consists mostly of paths to siblings. We observe a similar behavior for paths with $n = 1$, where the contextual notions of similarity have no statistically preferred direction, while the other measures point preferentially to hypernyms (i. e., 1-up in the WordNet taxonomy). As mentioned above, for $n > 2$ the distribution $P(n)$ of paths lengths for tag context and resource context (Figure 7.5) is similar to the ones for the other distributions. We can say that if synonyms or siblings are present, then these measures are able to find them. If not, they behaves similarly to the other measures, i. e., they points to related tags that preferentially lie higher in the WordNet hierarchy.

We now generalize the analysis of Figure 7.6 to paths of arbitrary length. Specifically, we measure for every path the *hierarchical displacement* Δl in WordNet, i. e., the difference in hierarchical depth between the synset where the path ends and the synset where the path begins. Δl is the difference between the number of edges towards a hypernym (up) and the number of edges towards a hyponym (down). Figure 7.7 displays the probability distribution $P(\Delta l)$ measured over all tags under study, for the five measures of relatedness. We observe that the distribution for the tag context and resource context is strongly peaked at $\Delta l = 0$ and highly symmetric around it. The fraction of paths with $\Delta l = 0$ is about 40 %. The average value of Δl for all the contextual measures is $\overline{\Delta l} \simeq 0$ (dotted line at $\Delta l = 0$). This reinforces, in a more general fashion, the conclusion that the contextual measures of similarity involve no hierarchical biases and the related tags obtained by them lie at the same level of the original one, in the WordNet hierarchy. Tag context and resource context are more peaked, while the distribution for user context, which is still highly symmetric around $\Delta l = 0$, is broader. Conversely, the probability distributions $P(\Delta l)$ for the non-contextual measures (co-occurrence and FolkRank), look asymmetric and both have averages $\overline{\Delta l} \simeq 0.5$ (right-hand dotted line). This means that those measures – as we have already observed – point to related tags that preferentially lie higher in the WordNet hierarchy.

Generalization to other datasets As stated at the beginning of Section 7.1.1, the presented analysis of relatedness measures has been done so far based solely on the Delicious dataset. The main motivations behind this choice were:

- Because the emergence of semantics is finally a user-driven process, we expect more reliable results from systems with (i) larger user populations and (ii) more active users (i. e., a higher total number of annotations produced by them). For both criteria, the Delicious dataset is most suitable among the ones presented in Section 3.1.
- Golder and Huberman (2006) suggested that *imitation* is a crucial aspect of the evolution of stable tagging patterns. This influence can be expected to be stronger within *broad* folksonomies (see Section 3.1.4), in which a user is typically exposed to other users' keyword choices while annotating a resource. Delicious also exhibits this property.

- As Delicious was among the first Social Annotation Systems which became very popular early in the Web 2.0 movement, its presence within the scientific literature is also comparatively strong. In this way, the presented analysis becomes comparable with other approaches.
- Despite a clearly visible bias towards technical topics, the sheer amount of contributing users introduced still a relatively broad coverage across several domains of interest. This is also reflected in a comparatively “general” vocabulary (i. e., not restricted to highly specialized terminology), consisting to a substantial amount (namely around 60 %) of proper English words. This makes our proposed grounding methodology against WordNet more reliable.

To summarize, we expected Delicious to be the most “representative” Social Annotation dataset. Despite this fact, the question remains to which extent the presented results can be generalized to other systems and datasets. Because an in-depth analysis on both a qualitative and semantically grounded level is beyond scope, we focus on on a single core aspect. Specifically, we think that the comparison of measures based on how well they correspond to WordNet-based metrics (as done in Figure 7.4) gives a good overall impression. For this reason, we applied the same procedure to datasets from BibSonomy, CiteULike, Flickr, AOL logsonomy and Stackoverflow. In the same manner as earlier in Section 7.1.1, we restricted all datasets to the 10 000 most popular tags in order to make the vector-based similarity assessment more meaningful. Figure 7.8 depicts the results. For comparison results, the prior results of Delicious are also included.

To start with the most “similar” systems, BibSonomy (Figure 7.8b) and CiteULike (Figure 7.8c) are also collaborative tagging systems which allow resources to be annotated by more than one user (i. e., broad folksonomies). A first commonality to Delicious is that both FolkRank and the user context relatedness (*UserCont*) lead to semantically more distant tags. The other measures (i. e., co-occurrence, tag and resource context relatedness) show a very similar performance (± 0.3 for BibSonomy, ± 0.5 for CiteULike). This differs from Delicious insofar as the resource context measure is not consistently best. A possible explanation lies in the high dimensionality of the resource context vector space: Despite the fact (as argued before) that the latter encodes potentially a large amount of contextual information, it requires a comparatively

7.1. Capturing Semantic Relatedness

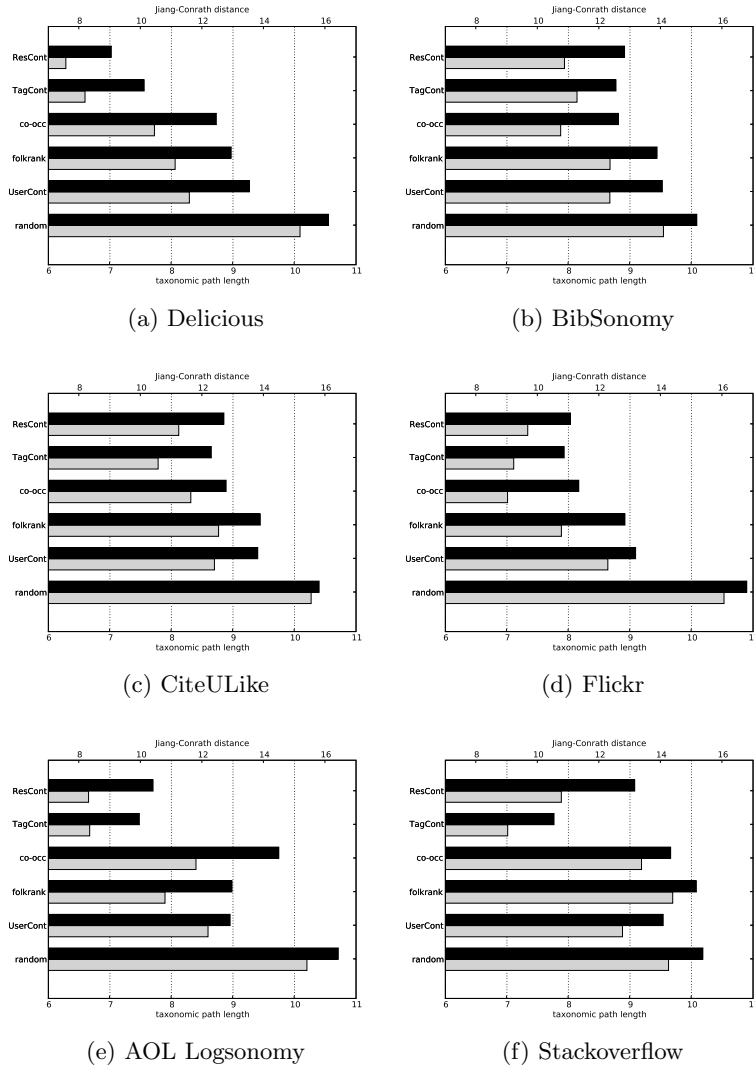


Figure 7.8.: Semantic grounding of tag relatedness measures for different datasets. Grey bars (bottom) show the taxonomic path length in WordNet, black bars (top) show the Jiang-Conrath measure of semantic distance (see Section 7.1.3).

high amount of annotations to form meaningful “resource context fingerprints” (similar to the ones exemplified in Figure 7.2 for the tag context). Because the total number of tag assignments is much smaller within BibSonomy (≈ 1.8 million) and CiteULike (≈ 8.8 million) compared to Delicious (≈ 117.3 million), one could hypothesize that the information does not suffice to establish precise semantic fingerprints in the resource context. Another difference is that the tag context relatedness differs not strongly from plain co-occurrence. Again, this can be probably attributed to the fact that precise tag context vectors also need a sufficient number of tag assignments – which is lower than the one needed for the resource context, but still higher than when just counting plain co-occurrences (as done by the co-occurrence relatedness).

Although tagging also plays a crucial role within Flickr, an important difference is its permission model which allows only owners to annotate their pictures – leading to a *narrow* folksonomy. This naturally makes aggregation in the resource context less meaningful, which is reflected in a non-optimal performance of the resource context relatedness in Figure 7.8d. While FolkRank and user context relatedness are consistently less precise, co-occurrence relatedness performs better compared to Delicious. This might be explained by the different purposes of these systems: While users of Delicious primarily use keywords to structure their personal collection of bookmarks, the exchange of pictures among Flickr users has a much more extrovert nature. In order to attract a broad audience, picture publishers might use more keywords, including semantically closely related ones as well as synonyms. In fact, the average number of keywords per resource is slightly higher (≈ 3.01) compared to Delicious (≈ 2.62).

Lastly, the AOL logsonomy and Stackoverflow datasets do not stem from genuine social tagging systems. For the logsonomy case, the annotation process is much more *implicit* (namely by clicking on search engine results), while the keyword annotation is not the main focus of the question/answering platform Stackoverflow. For both reasons, we observe the strongest differences here. Within the AOL logsonomy dataset, a remarkable difference compared to the Delicious is that the co-occurrence relatedness yields tags whose meanings are comparatively distant from the one of the original tag. A cursory manual analysis revealed that co-occurrence often “reconstructs” compound expressions; e. g., the most related tag to *power* according to co-occurrence relatedness is *point*. This is a natural consequence of splitting queries and consequently splitting compound expressions as we did; so our results confirm the intuitive assumption

that the semantics of isolated parts of a compound expression usually are semantically complementary. For the Stackoverflow case, all measures except the tag context relatedness show a comparatively weak performance. We attribute this to the strong topical focus of its computer programming community, which introduces highly specialized relations within its technical terminology not necessarily present within a general resource like WordNet.

Despite the mentioned individual differences, a common observation across all datasets is that the tag context relatedness shows an optimal or nearly optimal performance – independent of the size or type of the underlying dataset. This is especially interesting because its computation is computationally much less expensive compared to the high-dimensional resource context relatedness. For both reasons, we will stick in the further analysis to the tag context relatedness as a measure of semantic similarity among keywords in Social Annotation Systems.

7.1.4. Alternative Aggregation, Weighting and Similarity Approaches

As stated at the beginning of Section 7.1.1, the variety of options within the different steps of computing semantic relatedness leads to a large number of possible combinations. In the previous analysis, standard choices for each phase (i. e., co-occurrence computation, aggregation, weighting and similarity) were selected in order to assess the core properties of co-occurrence, distributional and graph-based measures. An important observation hereby was that the tag context relatedness seems to be a suitable proxy for semantic similarity. In this section, we aim to complete this picture by presenting alternative choices for each step of its computation.

Because the entries at each dimension of a tag context vector are based on *co-occurrence* events, the first question is which scheme is used for their computation. Coming back to the explanations of Section 3.1.3, we will first analyze the semantic implications of post-based, resource-based and user-based co-occurrence. Furthermore, for the distributional measures presented in the previous section, a crucial aspect is how the dimensionality of the tripartite hypergraph is reduced into the two-dimensional vector space used for similarity computation. Because this process is inherently afflicted with the loss of correlation information, an inappropriate aggregation choice may have a detrimental effect on the quality of the captured semantic relations. The intuitive scheme

which was used in the previous chapter can be regarded as “micro-aggregation”, in analogy to micro-averaging in text mining and when one thinks of users as classes. In this section, we will present two further aggregation methods (namely macro and collaborative aggregation) and compare their characteristics. Another important issue stemming from the field of information retrieval is the scheme by which the vector entries are weighted after aggregation. The goal hereby is usually to assign lower weights to less informative dimensions. In order to complement the approach of uniform weighting (i. e., all dimensions were regarded as equally informative) from the previous section, we will apply to two standard weighing methods from the field of information retrieval, namely term frequency \times inverted document frequency (TFIDF) and positive pointwise mutual information (PPMI) (Turney and Pantel, 2010). A last and important choice is which method is used to compute the similarity (or dissimilarity) among the context vectors obtained from the previous steps. In the field of natural language processing, several measures of distributional similarity were proposed (Lee, 1999). These include geometrically-motivated functions (as the cosine similarity used in the previous chapter), as well as metrics of “distance” between probability distributions. We will detail here on the semantic implications of the α skew divergence and the $L1$ norm, which showed superior performance in the context of estimating the probability of unseen co-occurrences (Lee, 1999). For judging the effect of the presented alternative choices, we will stick to the methodology of semantic grounding presented in the previous section. Specifically, we will use the average Jiang-Conrath distance (measured in WordNet) between a tag t and its most related tag t_{sim} according the alternative metrics (compare Figure 7.4). For the observed distances, we will use the term *semantic precision*: Hereby we consider a measure as semantically more precise if it leads to smaller average distances while grounding its top relatedness pairs in WordNet.

Co-occurrence computation

As explained in Section 3.1.3, there exist three basic notions of a co-occurrence event between tags. These differ in the definition of the context in which the two tags co-occur: The *post-based* variant counts a co-occurrence when two tags were used together *within the same post* – i. e., by the same user for the same resource. On the other hand, *resource-based* counting disregards the users and only records co-occurrences *within the same resource*. The *user-based*

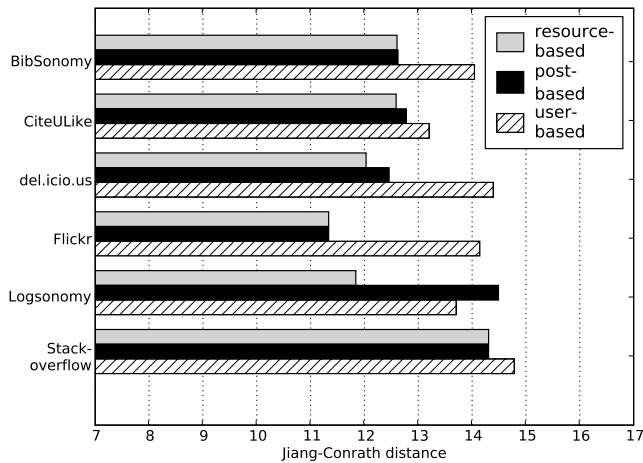


Figure 7.9.: Semantic grounding of different types of co-occurrence.

approach finally captures co-occurrences only within individual vocabularies of users, independent of resources. Figure 7.9 shows the semantic precision of each variant. The black bar corresponds to the post-based variant, which has been used in the previous in-depth analysis (Section 7.3.2). On the y -axis, the different Social Annotation datasets are found. First of all, one can observe that within all *narrow* folksonomies (i. e., Flickr and Stackoverflow), post-based and resource-based are identical – which is obvious: As each resource is present in exactly a single post in these systems, both variants are effectively identical. In the other systems, resource-based co-occurrence is on the same or a slightly better level than post-based co-occurrence. This is also not too surprising, because a single user is less likely to assign, e. g., two synonym tags to a particular resource. Conversely, because different users might use different keywords for the same thing, co-occurrence on the resource levels is more likely to establish connections among semantically more closely related keywords. This is most clearly visible for the logsonomy, where resource-based co-occurrence seems to partially compensate the aforementioned effects of reconstructing compound expressions (see Section 7.1.3). Throughout all conditions, user-based co-occurrence performs worst: A possible explanation is that users will probably be interested in several independent (and semantically diverse) topics.

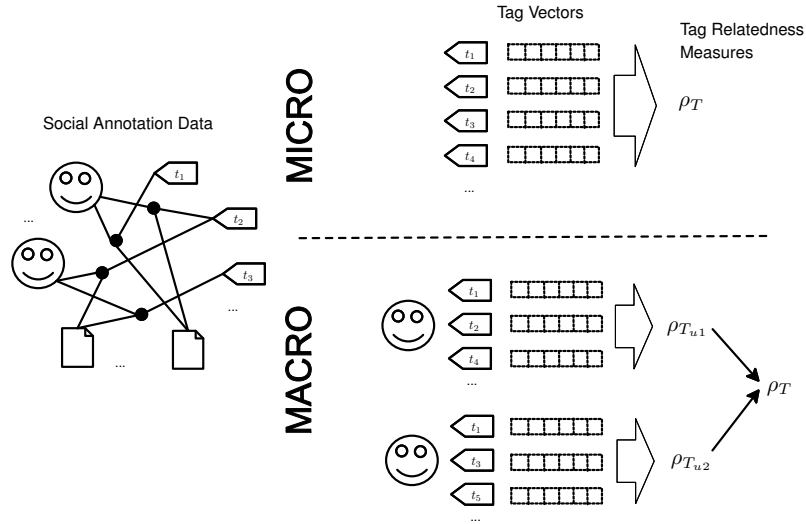


Figure 7.10.: Schematic overview of micro and macro aggregation.

Hence, co-occurrence in the user context tends to introduce relations among unrelated keywords.

In summary, the resource-based variant of co-occurrence computation seems to be the best choice for the task of capturing first-order tag-tag correlation information from Social Annotation data. However, one has also to take into account complexity aspects: Co-occurrence always comprises the pairwise processing of all keywords found in a given context. This quadratic complexity can prove problematic especially in large folksonomies, where the number of keywords attached to a resource is much larger than the number of keywords found in a post. In such a situation, post-based co-occurrence can be seen as an almost optimal replacement in most cases.

Aggregation Schemes

For the process of aggregating the tripartite Social Annotation graph into a two-dimensional representation for similarity computation, a “global” scheme was used in the previous analysis – i. e., all co-occurrences produced by all users were aggregated into a single “system-wide” vector for each keyword. This method can be seen as an analogy to micro-averaging in the field of text

mining (Manning et al., 2008). Hereby each annotation is given the same weight, so that a more active user would have a larger impact on the weights and consequently on any derived similarity measure. We will now present two alternative aggregation schemes, which first treat each user’s annotation set independently, and then aggregate across users.

Macro-aggregation Instead of “system-wide” aggregation based on a complete folksonomy $\mathbb{F} = (U, T, R, Y)$, the idea behind the macro approach is to work on personomies P_u (see Definition 3.1), i. e., “user-wide” folksonomy partitions. Figure 7.10 gives a schematic overview about both variants. One can see that in the macro aggregation case, each personomy is processed in the same manner as the complete folksonomy in the micro case, i. e., co-occurrence is computed, context vectors are built, and the similarity among these vectors is computed. The resulting per-user relatedness measures $\rho_{T_u}, u \in U$ are then combined into a global measure via a voting, i. e., summing across all users according to

$$\forall t_1, t_2 \in T : \rho_T(t_1, t_2) = \sum_{u \in U} \rho_{T_u}(t_1, t_2)$$

Macro-aggregation does not have a bias toward users with many annotations. However, in giving the same importance to each user, the derived similarity measures amplify the relative impact of annotations by less active users. It is an empirical question which of these biases is more effective.

Collaborative aggregation The approach of macro aggregation comes close to the idea of collaborative filtering: Global relatedness is constructed by many votes similar to “other users think that these keywords are related”. However, because all context relatedness measures are essentially based on feature-based representations, they will yield zero similarities if two keywords do not share a feature. This effect is less severe for the case of micro aggregation, but may prove problematic within the necessarily smaller per-user vocabularies used for macro aggregation. In order to tackle this issue, we interpret the common usage of two keywords by a user for annotation as an implicit evidence for their relatedness. Technically, we achieve this by adding a “special tag” t_u^* to each post of a given user u . In this way, we “connect” all keywords used by u with a common feature. In order to avoid the thematic mixing effects observed for the user-based co-occurrence, the motivation behind these special annotations is to

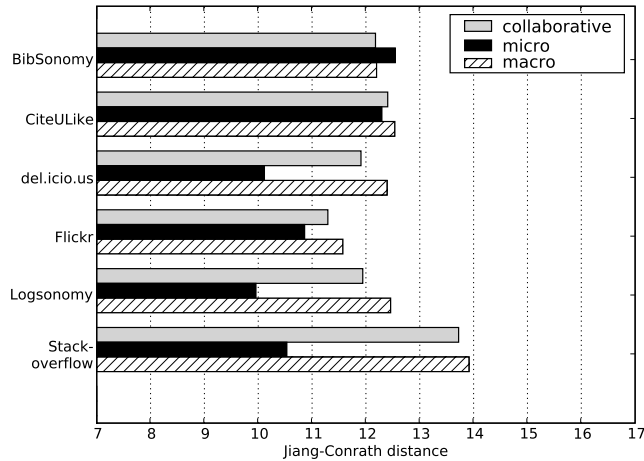


Figure 7.11.: Semantic grounding of different types of aggregation.

yield a small but non-zero contribution to the per-user relatedness measures. The remaining procedure is identical to macro aggregation as described above.

Figure 7.11 displays the semantic implications of macro (top bar) and collaborative (lower bar) aggregation, compared to micro aggregation (middle bar). A first consistent observation is that collaborative and macro aggregation perform similar, with a slight advantage of the collaborative variant throughout. Hence, the small contribution of the special annotation seems to lead to a slight positive effect. For BibSonomy and CiteULike, both alternative aggregation methods perform comparable with the micro approach; for all other datasets, they do not reach its semantic precision. In fact, this is especially visible for datasets with (i) larger user populations and (ii) where the original tag context relatedness performed particularly well. Hereby the fact that each user is given the same weight seems to be problematic: Because within the latter systems, there exists a large number of less active users, their aggregated contribution has a strong influence – probably towards more popular tags. Table 7.4 corroborates this assumption: It lists the overlap of the 10 most related tags according to each aggregation method with the 10 most related tags according to (post-based) co-occurrence, averaged across all tags. In all cases, one can see that macro and collaborative aggregation introduce a higher overlap (second and third row).

Table 7.4.: Average overlap of the 10 most related tags according to different aggregation schemes with the 10 most frequently co-occurring tags (post-based co-occurrence).

	<i>Bib-Sonomy</i>	<i>CiteULike</i>	<i>Flickr</i>	<i>Stackoverflow</i>	<i>AOL Logsonomy</i>	<i>Delicious</i>
micro	2.77	2.20	2.90	1.25	0.88	1.69
macro	3.40	3.79	3.01	5.97	4.82	5.59
collaborative	3.20	3.64	3.10	6.48	5.96	6.20
# of users	5 777	72 249	271 359	272 313	463 380	511 348

However, this effect is much stronger for systems with larger user populations (fourth row). This means that while alternative aggregation schemes seem to be an option for smaller systems, the long tail of less active users in larger systems introduces a strong bias towards popular keywords.

Weighting Schemes

When representing keywords as feature vectors, another crucial aspect is the weighting scheme applied to capture the relative importance of each dimension in a given vector. The underlying idea is that “*surprising events, if shared by two vectors, are more discriminative of the similarity between vectors than less surprising events*” (Turney and Pantel, 2010). A typical approach from the field of information retrieval are *tf-idf* (term frequency \times inverted document frequency) weighting functions. According to those, a term (i. e., a dimension) is relevant for a document (i. e., a vector) if its frequency *tf* within the document is high, and if it occurs seldom within other documents (i. e., its inverted document frequency *idf* is high). Translated to the tag context relatedness, a co-occurrence event of a tag *t* with another tag *t'* should be given higher weight the more often it occurs, and if *t'* co-occurs with only a small number of other tags besides *t*. More formally, if we denote the co-occurrence weight with $w(t, t')$ as defined in Section 7.1.1, and the set of all tags co-occurring with *t* as $cooc(t) := \{t' \in T : w(t, t') > 0\}$, the adaptation of *tf-idf* weighting to tag context vectors v_t can be written according to:

$$v_{tt'} = w(t, t') \log \frac{|T|}{|cooc(t')|}$$

An alternative weighting scheme which has shown good performances for

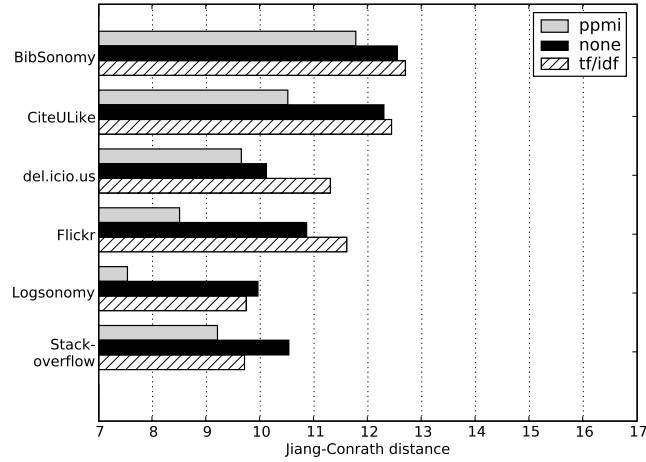


Figure 7.12.: Semantic grounding of different types of weighting schemes (tag context relatedness). The values used for Laplace smoothing were $k = 0$ (BibSonomy, CiteULike, Flickr), $k = 18$ (Delicious, AOL logsonomy) and $k = 6$ (Stackoverflow).

measuring semantic similarity with word-context matrices is Pointwise Mutual Information (PMI) (Turney and Pantel, 2010), and especially its strictly positive variant (PPMI) (Bullinaria and Levy, 2007). Its basic idea is to relate the probability that a term t occurs in a context c with the “marginal” probabilities of t and c , respectively. Applied to our problem at hand, we define:

$$p_{tt'} = \frac{w(t, t')}{\sum_{t \in T} \sum_{t' \in T} w(t, t')}$$

$$p_{t*} = \frac{\sum_{t' \in T} w(t, t')}{\sum_{t \in T} \sum_{t' \in T} w(t, t')}$$

$$p_{*t'} = \frac{\sum_{t \in T} w(t, t')}{\sum_{t \in T} \sum_{t' \in T} w(t, t')}$$

$$pmi_{tt'} = \log \left(\frac{p_{tt'}}{p_{t*} p_{*t'}} \right)$$

$$ppmi_{tt'} = \begin{cases} pm_{i_{tt'}} & \text{if } pm_{i_{tt'}} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Consequently we set $v_{tt'} = ppmi_{tt'}$ in the PPMI-weighted vector. As one can see, the PMI weighting increases when there is a statistical dependency between t and t' such that $p_{tt'} > p_{t*}p_{*t'}$. Similar to the *tf-idf* scheme, this introduces a bias towards frequent co-occurrences among infrequent tags. A way to weaken this bias is to apply Laplace smoothing to the probability estimates $p_{tt'}$, p_{t*} and $p_{*t'}$ by adding a constant $k > 0$ to the raw frequencies (Turney and Pantel, 2010), i. e., $w(t, t') + k$. We did a simple grid search and varied $k = 0, 2, 4, 6, \dots, 30$ in our experiments in order to find the optimal choice regarding the minimization of the average Jiang-Conrath distance. Figure 7.12 depicts the effects of both weighting schemes on the results of the tag context relatedness. Compared to using no weighting scheme (black bar), a first common observation is that PPMI weighting leads to an improvement in semantic precision in all cases (the respective values for the Laplace smoothing are given in the caption). The gain is partially impressive: The ratio of synonyms among the most similar tags improves from 5.4 to 11.6% for CiteULike, from 11.0 to 20.0% for Flickr and from 20.3 to 30.9% for the AOL logsonomy. Even for Delicious, the improvement from 17.3 to 21.1% indicates that PPMI weighting seems to be able to compensate partially for the loss of context information compared to the resource context relatedness (whose synonym ratio is 25.6%, see Figure 7.5).

On the contrary, for all social tagging systems (i. e., BibSonomy, CiteULike, Delicious and Flickr), the *tf-idf* scheme has a detrimental effect. For the AOL logsonomy and Stackoverflow, its gain is noticeably smaller compared to the PPMI weighting. A main difference between both weighting schemes is that for a given dimension t' , the inverse document frequency takes into account only the *number* of other vectors in which t' has a non-zero value, while the marginal probabilities of PPMI are based on the *sum* of all vector entries at dimension t' . The inherent bias towards less frequently used dimensions (i. e., keywords in our case) seems to have beneficial effects against the domination of the similarity by overly popular dimensions. However, especially for larger and broad folksonomies (like Delicious or AOL logsonomy), a slight dampening of this effect by Laplace smoothing can further improve the results. A possible explanation for this is that while within smaller datasets, popular terms might be strongly influenced by individual users, which might increase the percentage of popular idiosyncratic terms. These are probably no good “partners” in terms

of semantic similarity to other tags. In larger datasets, this effect is probably less severe because the influence of individual users is smaller, and hence the popular tags might tend to resemble more closely an “agreed” terminology. Furthermore it can be hypothesized that the latter is even more visible in broad folksonomies, because of imitation effects when users “talk” about the same resources – which would explain why not smoothing yields optimal values for the large, but narrow Flickr folksonomy.

Similarity Measures

The distributional representation of tags based on a set of context dimensions can either be interpreted as a vector or as a probability distribution. Both perceptions afford a variety of possibilities to assess tag similarity: Especially in the field of information retrieval, geometrically-motivated metrics like the cosine similarity (which was used in the previous analysis), the Euclidean distance or the $L1$ norm are commonly applied (Turney and Pantel, 2010). Other measures like the *Kullback-Leibler*, Jensen-Shannon or skew divergence stem from information theory and focus on quantifying the “distance” between probability distributions (Lee, 1999). For choosing representative measures to compare with, we rely on prior work by Weeds et al. (2004), who found that different measures tend to select neighbors of particular frequencies. Specifically, the authors proposed a division into three classes, namely measures that favor (i) high frequency terms, (ii) low frequency terms or (iii) terms with a similar frequency compared to the target term. Our choice of measures was guided by the motivation to select a measure for each class which has shown good performances in the literature.

For the high-frequency class, we selected the α -skew divergence. It is based on the Kullback-Leibler (KL) divergence measure (also known as relative entropy), which quantifies the distance between two discrete probability distributions p and q according to:

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Tag co-occurrence vectors can be turned into probability distributions by normalization according to $v_{tt'} = \frac{v_{tt'}}{\sum_{t'' \in \text{cooc}(t)} v_{tt''}}$. Based on this, we would rewrite the KL divergence as $D(v_t||v_{t'}) = \sum_{t'' \in \text{cooc}(t)} v_{tt''} \log \frac{v_{tt''}}{v_{t't''}}$. Obviously, an infinite

distance occurs when $v_{tt''} = 0$, i. e., if there exists a single dimension of v_t where v_t' has a zero entry – which is very likely to happen. Different kinds of smoothing techniques were proposed to alleviate this problem; among them, the α -skew divergence showed the best performance for the task of predicting unseen co-occurrences (Lee, 2001). It is defined as:

$$s_\alpha(p, q) = D(r || \alpha q + (1 - \alpha)r)$$

Hereby the distribution q is smoothed by p ; the parameter α controls the degree to which the original KL distance is approximated. We stick to a choice of $\alpha = 0.99$, because it has given good results in previous studies (Lee, 2001).

Another measure which belongs to the low-frequency biased class is *precision* according to the additive MI-based co-occurrence retrieval model (AM-CRM) (Weeds et al., 2004). Given two tags t and t' , it is defined as:

$$\text{prec}(t, t') = \frac{\sum_{t'' \in \text{cooc}(t) \cap \text{cooc}(t')} \text{ppmi}_{tt''}}{\sum_{t'' \in \text{cooc}(t)} \text{ppmi}_{tt''}}$$

ppmi denotes the positive pointwise mutual information, as introduced above. Hereby, tags t' are considered as more similar when they have co-occurred with fewer tags that t did *not* co-occur with. This leads to a preference towards less frequent and consequently distributionally more specific words.

A similar measure which tends to select terms with the same frequency as the target term is *Lin's similarity* (Lin, 1998), which is defined as:

$$\text{lin}(t, t') = \frac{\sum_{t'' \in \text{cooc}(t) \cap \text{cooc}(t')} \text{ppmi}_{tt''} + \text{ppmi}_{t't''}}{\sum_{t'' \in \text{cooc}(t)} \text{ppmi}_{tt''} + \sum_{t'' \in \text{cooc}(t')} \text{ppmi}_{t't''}}$$

It is based on information theory and relates “*the amount of information needed to state the commonality of A and B*” to “*the information needed to fully describe what A and B are*” (Weeds et al., 2004).

Including in total four measures (i. e., the cosine, α -skew divergence, precision and Lin's), we have a representative selection for both geometrically-motivated and information-theoretic approaches, covering all frequency-bias classes mentioned by (Weeds et al., 2004). Figure 7.13 displays the semantic precision for each measure. In general, the picture looks less consistent: While precision performs worst across almost all datasets, it seems to have an advantage for the Flickr case. On the other hand, the skew divergence has benefits in most

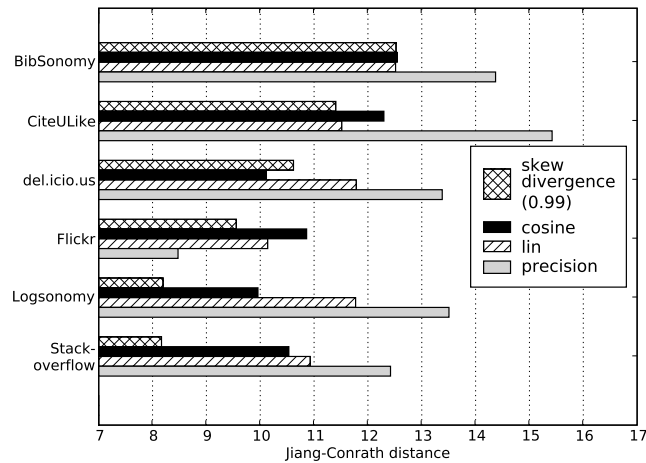


Figure 7.13.: Semantic grounding of different types of similarity measures.

situations, but is slightly worse compared to cosine for Delicious. First of all, manual inspection of the results turned out that the bias of precision towards infrequent terms led in fact to a large number of idiosyncratic terms being judged as most related. As most of these are not present in WordNet, the number of pairs which afforded a semantic grounding by the Jiang-Conrath measure was significantly lower compared to the other measures – e. g., 229 versus 2002 for the case of Flickr. The fewer the number of mappable pairs is, the less reliable becomes the evidence provided by semantic grounding. Hence, the only valid conclusion which can be drawn is that the precision measure seems to avoid “agreed” vocabulary (when we consider the existence of a term in a dictionary as a sign of agreement).

Generally speaking, the α -skew divergence seems to be a good choice for a broad range of Social Annotation datasets, especially if they are not too large: A further experiment based on roughly half of the total annotations within the Delicious dataset led to superior results of the skew divergence compared to cosine similarity.

7.1.5. Summary

In this section, we have presented an in-depth analysis of the semantic characteristics of several measures of keyword relatedness based purely on the structure of Social Annotation Systems. Using a methodology based on semantic grounding within WordNet, the first observation was that the notions of FolkRank and co-occurrence relatedness exhibit a tendency towards *more general* keywords – i. e., they seem to have a bias towards hypernyms. Distributional approaches based on resources or tags as context showed a different behavior, and turned out to favor synonym or “sibling” keywords. Especially for the tag context relatedness, this property was found to be consistent across different kinds of Social Annotation Systems. Hence, the context of co-occurring keywords can be seen as a valuable source of information for capturing the semantics of a given keyword.

In the following, this finding was further elaborated by presenting various alternatives for the different steps involved in measuring distributional similarity based on the tag context – namely co-occurrence computation, aggregation, weighting and similarity assessment. While resource-based co-occurrence was found to be optimal in most cases, it is approximated well by the less complex post-based co-occurrence. Aggregation schemes based on the idea of collaborative voting seemed to be problematic for larger datasets, because they introduce a strong bias towards popular (and not necessarily semantically close) keywords. Hence micro-aggregation (i. e., computing similarity on the system instead of the user level) seems to be the most suitable scheme. Because of the domination effects by popular tags, weighting approaches proved to be crucial: Using positive pointwise mutual information (PPMI) to weigh the context vector entries led to a substantial improvement throughout all conditions, partially in conjunction with Laplace smoothing on larger datasets. Finally the choice of the vector similarity measure showed a less consistent picture, pointing to skew divergence and cosine similarity as generally robust metrics for social bookmarking systems.

When taking into account all analyzed factors, a common conclusion is that appropriately weighted keyword co-occurrence seems to encode a large amount of the implicit semantics within Social Annotation Systems. In the next section, these insights will be exploited for the purpose of learning concepts.

7.2. Learning Concepts

Having identified measures of keyword relatedness which are able to discover semantically close keywords, the next step according to the ontology learning layer cake (see Section 5.2.2) is to use these measures to identify *concepts*. To put it more formally, one can interpret the keyword vocabulary T of a Social Annotation System as a set L_C of lexical labels for an unknown set C of concepts. Hence, the process of concept formation has two aspects, namely (i) identifying this set C of underlying concepts and (ii) learning the relation Ref_C among these concepts and the keyword vocabulary L_C . Hereby, two problems which are inherent to the flexibility of Social Annotations need to be solved, namely synonymy (i. e., the existence of different keywords for the same concept) and polysemy (i. e., a keyword denoting more than one concept). In other words, this corresponds to “cleaning up” the collaboratively created vocabulary.

As explained earlier when detailing on the state of the art in this direction (see Section 5.3.3), (Cimiano, 2006) mentions that the process of concept discovery “*should ideally provide (i) an intensional definition of concepts, (ii) their extension and (iii) the lexical signs which are used to refer to them*”. In order to assure the applicability of our proposed methods to a variety of systems (and not only to those dealing with textual resources), we interpret concepts in a less strict manner as “semantic groupings” of keywords. In this way, we restrict ourselves to the keywords themselves as the only intensional and lexical vocabulary. More precisely, this means that we will use the keywords for the intensional definition and as lexical signs. Our primary focus hereby is furthermore the *identification* of the set of underlying concepts; this means that the task of instance assignment (i. e., creating the extension of the concepts) is secondary. The main reason for this decision is twofold: First, the discovery of concepts *without* their extension does have a value on its own, e. g., for automatically extending structured lexical resources like WordNet. Second, we think that a prerequisite for instance assignment is naturally the existence of meaningful concepts.

In this section, we will first analyze different methods to resolve the usage of synonymous keywords based on the measures of semantic relatedness presented in the previous section. In a next step, we discuss approaches to discover different meanings of individual keywords. In summary, both corresponds to the discovery of concepts underlying the collaboratively created vocabulary of Social Annotation Systems.

7.2.1. Synonym Resolution

As mentioned in Section 3.1.1, the retrieval of relevant resources within Social Annotation Systems is negatively affected when different keywords are being used to denote identical or very similar concepts. Due to the open nature of these systems, this can happen in a variety of ways: First of all, users have different habits of separating multiple words within a single keyword, like *ontology_learning*, *ontology-learning* or *ontologyLearning*. The usage of singular and plural is also subject to the same kind of fluctuations (e.g., *ontology* and *ontologies*). Apart from that, various acronyms and abbreviations like *ol* or *ontolearn* are also typically found. The international population of many systems introduces also cross-language phenomena like *Ontologie* (German) or *ontologi* (Swedish). Lastly, there exist potentially also very similar keywords like *ontologyDefinition* for which it may be desirable to be subsumed under the same concept. As one can see, we are interpreting “synonymy” hereby in a comparatively broad manner. So instead of detecting linguistically precise synonyms, our goal is to “shrink” the vocabulary T of a Social Annotation System by grouping all keywords with a very similar meaning.

In order to be able to use the existing formalisms for this newly created structure, we will refer to it as a *synsetized folksonomy* defined as follows:

Definition 7.1 *A synsetized folksonomy is a tuple $\mathbb{F}^s := (U, S, R, Y^S)$ where U and R are the sets of users and resources present in a folksonomy \mathbb{F} . S is the synsetized vocabulary of F , whereby each original tag $t \in T$ has been replaced by a synset $s \subset S$. The synsets $s \in S$ may overlap and contain exhaustively all tags $t \in T$, i. e., $\bigcup_{s \in S} s = T$.*

The synsetized tag assignments Y^S are created by replacing each original triple $(t, u, r) \in Y$ with its corresponding synsetized variants $(s_1, u, r), \dots, (s_n, u, r)$. Hereby s_1, \dots, s_n are the synsets associated to t .

In most cases, the synsetized vocabulary is at most as big as the original vocabulary, i. e., $|S| \leq |T|$ and $|Y^S| \leq |Y|$. Of course, the crucial point here is how to map a tag to its synset. Two classes of approaches are predestined for this kind of task, namely dimensionality reduction and clustering techniques. A common approach in the first direction is *latent semantic indexing* (Deerwester, 1988), which maps from a high-dimensional space (i. e., the ones spanned by all keywords) into a latent *latent semantic space*, whose dimensions correspond

to concepts (Eda et al., 2009). While such approaches are found in the literature (Eda et al., 2009; Zhang et al., 2006), most of them were performed on comparatively small datasets due to the involved complexity. For this reason and in order to complement these works, we are focusing on clustering techniques in the scope of this dissertation. Another reason for this decision is that we can build hereby on elaborate measures of tag relatedness, which we presented earlier.

The general idea of clustering is to form groups of objects (so-called clusters) which exhibit a greater similarity to each other than to objects within other clusters. A great number of algorithms and variants has been developed within different research areas; cf. (Jain et al., 1999) for an overview. For some algorithms (like, e. g., k-means) the number of clusters needs to be fixed in advance. Because the number of underlying concepts within a Social Annotation System is unknown, such approaches are not very well suited. A popular brand of algorithms which do not require a predefined number of clusters is *hierarchical agglomerative clustering* (HAC). While there exist several other approaches, we have chosen HAC because it is the most “direct” way to derive synonym classes from the relatedness measures presented in Section 7.1. Our goal hereby is not primarily to perform an exhaustive study of the suitability of different clustering algorithms for the purpose of identifying synonyms, but more to ensure that our studied measures do provide valuable input for clustering. We will now first detail on the variants of hierarchical clustering we analyzed, and explain how clusters can be derived; additionally, we will introduce a simple baseline to compare against.

Hierarchical Agglomerative Clustering

Given a set of objects O , the basic idea of hierarchical agglomerative clustering is to build a “linkage” of clusters, starting from $|O|$ individual clusters and merging these in a bottom-up manner until all objects are in a single cluster. Hereby often the notion of *distance* instead of similarity is used; however, given a properly normalized similarity measure, one can easily transform a similarity value s into a distance value d according to $d = 1 - s$. The general hierarchical clustering algorithm is explained by (Jain et al., 1999) as follows:

1. Compute the distance matrix containing the distance between each pair of objects. Treat each object as a cluster.

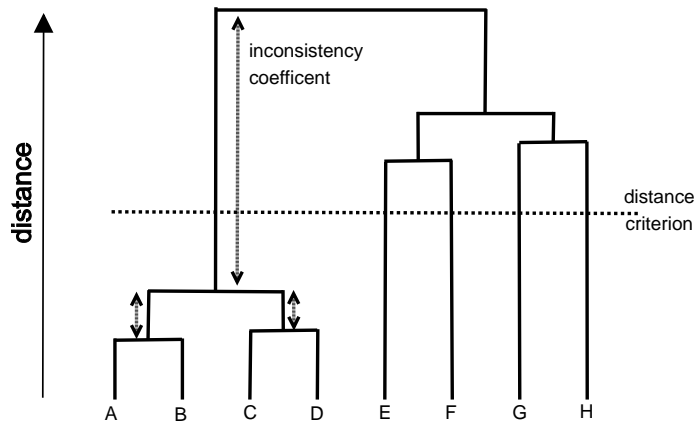


Figure 7.14.: Example dendrogram obtained from hierarchical agglomerative clustering.

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 results in a so-called *dendrogram*, which depicts graphically the level of distance at which each merging step took place (See Figure 7.14). Hierarchical agglomerative clustering algorithms can mainly be differentiated by the scheme how the distance matrix in step 2 is updated. The core question hereby is how the distance between merged clusters is computed. Let $U = \{u_1, \dots, u_i\} \subseteq O$ be a the 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. Then the following standard schemes for updating the matrix with distances between U and each V exist:

- *Single Link:*

$$\text{dist}(U, V) = \min_{u \in U, v \in V} \text{dist}(u, v)$$

- *Complete Link:*

$$\text{dist}(U, V) = \max_{u \in U, v \in V} \text{dist}(u, v)$$

- *Average Link:*

$$\text{dist}(U, V) = \sum_{u \in U, v \in V} \frac{\text{dist}(u, v)}{|U| \cdot |V|}$$

- *Weighted Average Link:*

$$\text{dist}(U, V) = \frac{\text{dist}(S, V) + \text{dist}(T, V)}{2}$$

- *Centroid Link:*

$$\text{dist}(U, V) = \text{dist}(c_U, c_V)$$

Hereby c_U and c_V are the centroids of clusters U and V , respectively. The centroid of a cluster is that object which minimizes the distance to all other contained objects.

- *Ward's variance minimization Link:*

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

Hereby $x = |V| + |S| + |T|$

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 our case), this dendrogram needs to be further analyzed. Hereby there exist two standard ways to “cut” the latter into a set of flat object clusters by using a threshold t :

- *Distance criterion:* The most intuitive approach is to fix a threshold τ_{dist} and to keep only those clusters in which all objects have the maximum distance of τ_{dist} . This corresponds to “cutting” the dendrogram by a horizontal straight line, and keeping those clusters underneath the cut (this is denoted as “cut criterion” in Figure 7.14).
- *Inconsistency coefficient:* Another way to identify “natural” divisions within the set of objects is to compare the distance at which a given cluster c was formed with the distances at which its components were created. If this distance is very low, then c can be expected to contain a coherent set of highly similar objects (see Figure 7.14). Leaf nodes in the cluster

tree and clusters merged from those are hereby assigned an inconsistency coefficient of 0. In our experiments, we used the implementation of inconsistency computation present in SciPy’s corresponding module⁶. By setting a threshold τ_{inc} , one can keep only those clusters with a maximum inconsistency of τ_{inc} .

In the following, we will explain a simple baseline against which the performance of the aforementioned algorithm variants can be compared.

Baseline Approach

Apart from clustering, another very direct way to group keywords based on their similarity is to use a direct similarity threshold τ_{sim} . More precisely, we defined the baseline synonym mapping function *syn* as follows:

$$syn(t) = \{t' \in T : sim(t, t') > \tau_{sim}\}$$

Because $sim(t, t) = 1$ a tag t is always contained in its “own” synset, i. e., $\forall t \in T : t \in syn(t)$. It is clear that when setting τ_{sim} to a very high value, then the size of synonym sets will be very small, while setting it very low will yield large sets of unrelated tags.

In order to assess the suitability of the different variants of hierarchical agglomerative clustering to build meaningful synonym sets, we computed clusterings for all datasets mentioned in Section 6.1. As similarity measure among keywords, we used the tag context relatedness, along with a PMI weighting including appropriate Laplace smoothing as shown in Figure 7.12. Because data sparsity can be problematic for its computation, we restricted ourselves again to the sub-folksonomies induced by the 10 000 most frequently used tags. For identifying concrete clusterings, we performed a stepwise incrementation of the distance criterion and inconsistency coefficient thresholds starting from 0, and ending at the value where only a single cluster was formed. The same was done for the baseline approach, varying the similarity threshold τ_{sim} between 0 and 1.

A first cursory analysis of the computed synsets showed promising results; especially using Ward’s method for distance computation together with an

⁶<http://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.inconsistent.html>

inconsistency coefficient thresholding seemed to produce consistent keyword clusters. In order to allow a more systematic assessment of the influence of the different algorithm variants, the following section performs a semantically grounded evaluation against WordNet synonyms.

7.2.2. Evaluation by Semantic Grounding

Similar to the procedure of grounding measures of keyword relatedness against reference measures, we also apply a gold-standard based evaluation paradigm to the task of synonym resolution. Because within WordNet, synonym words are grouped into synsets, we can exploit those as ground truth to compare the clustered synsets against. More precisely, we define the set of keywords which is present in WordNet as $T^{WordNet}$, and based on that a reference synonym function:

$$syn^{\Delta} : T^{WordNet} \rightarrow \mathbb{P}(T^{WordNet})$$

This function maps each keyword t to all other keywords contained in all synsets where t is included. We are aware that we are “mixing” hereby different senses of a word within WordNet; however, the same mixing effects can be expected to take place on the folksonomy side. Analogously, we define the synonym functions induced by the aforementioned algorithms as:

$$syn^{\star} : T^{WordNet} \rightarrow \mathbb{P}(T^{WordNet})$$

Based on these functions, we can now measure to which extent both functions retrieve the same synonyms. We will employ a variant of the well-known IR measures precision and recall for this purpose. More precisely, we define local precision and recall functions according to:

$$p(t) = \begin{cases} \frac{syn'^{\star}(t) \cap syn'^{\Delta}(t)}{syn'^{\star}(t)} & \text{if } |syn^{\star}(t)| > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$r(t) = \begin{cases} \frac{syn'^{\star}(t) \cap syn'^{\Delta}(t)}{syn'^{\Delta}(t)} & \text{if } |syn^{\Delta}(t)| > 0 \\ 0 & \text{otherwise} \end{cases}$$

Hereby syn'^{\star} and syn'^{Δ} exclude the term itself from its synset, i. e., $syn'^{\star}(t) = syn^{\star}(t) \setminus t$. The global precision and recall values, as well as their combination

in the F-measure are then computed according to:

$$P = \frac{1}{|T^{WordNet}|} \sum_{t \in T^{WordNet}} p(t)$$

$$R = \frac{1}{|T^{WordNet}|} \sum_{t \in T^{WordNet}} r(t)$$

$$F = \frac{2 * P * R}{P + R}$$

Because our focus lies more on the question whether the learned synsets are correct rather than to reach a complete coverage of all WordNet senses, we will focus in the evaluation on precision and F-measure.

Figure 7.15 shows the results of the semantic evaluation for the Delicious dataset⁷. The subfigures (a) to (d) depict the hierarchical clustering approaches, using the inconsistency coefficient (subfigures (a) and (c)) and the distance criterion (subfigures (b) and (d)). Subfigure (e) contains the baseline approach based on a simple similarity threshold. Higher values on the y -axis correspond in each figure to a higher agreement with the reference synsets, i. e., a better performance. When reading the figures, one has to keep in mind that lower thresholds imply a *higher* number of smaller clusters, while higher thresholds produce *fewer* clusters of larger size.

A first observation is that almost all hierarchical clustering conditions yield a higher agreement to WordNet than the baseline approach. This is not too surprising, as setting a single global threshold can be seen as a relatively coarse instrument, which does not fit well the potentially very different levels of similarity produced by the tag context relatedness.

Among the clustering variants, single and centroid linkage show a weaker performance throughout all conditions. This can be probably attributed to the “chaining effect” of single linkage, which tends to create large clusters. Additionally, taking into account only a centroid as a representative for a cluster seems to disregard a crucial amount of information encoded in the remaining cluster members.

When comparing the inconsistency coefficient and the distance criterion, it turns out that the latter does not reach the performance of the first: Its optimal

⁷For space reasons, we omit detailed results for the other datasets; however, they show a very similar behaviour, whose optimal values are depicted in Table 7.5.

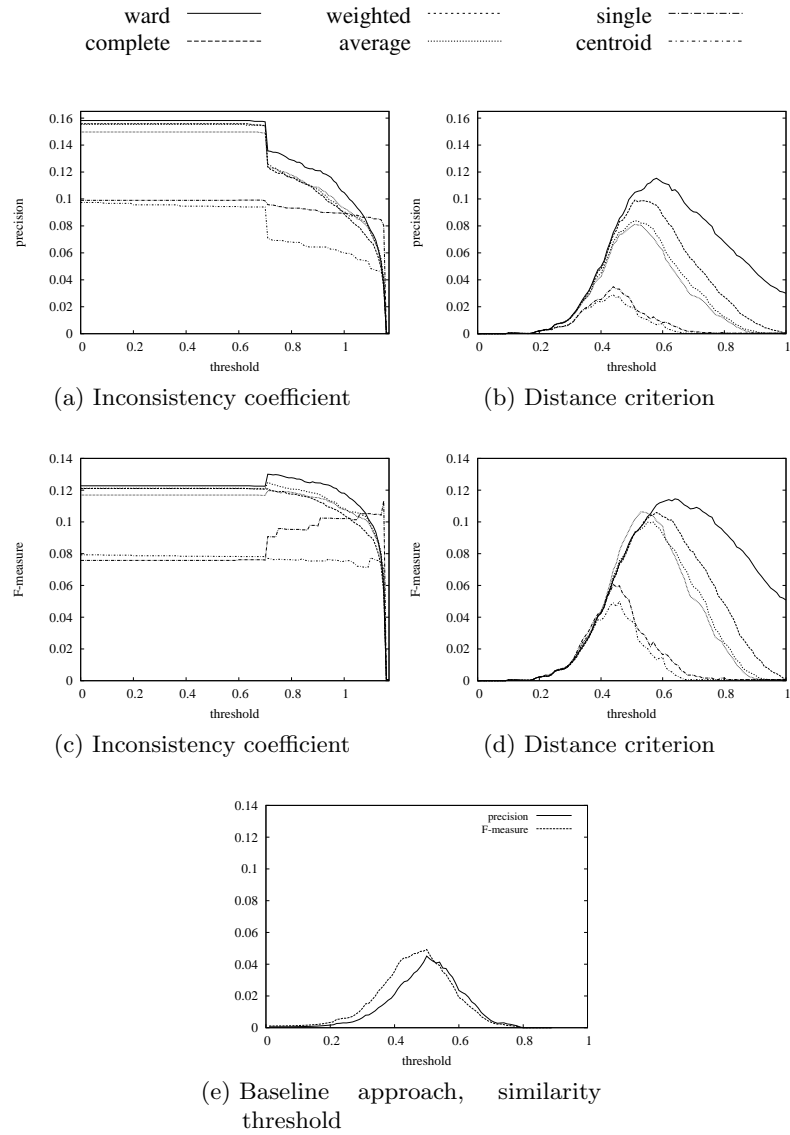


Figure 7.15.: Performance of different clustering methods for the task of synonym resolution.

values for both precision and F-measure are considerably lower. Hence the inconsistency coefficient seems to be the more appropriate method to “flatten” the hierarchical keyword clustering.

Though the remaining clustering variants show a similar performance, Ward’s method almost consistently has a slight advantage. Taking a look at the precision values obtained by varying the inconsistency threshold, it turns out that the clustering result is basically identical between a threshold range from 0.0 to roughly 0.6, yielding a set of 6 424 clusters. Recall that in contrast to the distance criterion, setting the inconsistency threshold to 0.0 already provides non-singleton clusters, because the inconsistency of leaf nodes and clusters composed of leaf nodes in the cluster tree is defined as zero. Interestingly, the resulting synset / keyword ratio of ≈ 1.6 comes close to the one of WordNet (≈ 1.73). So when the goal is to optimize synset precision, then choosing Ward’s method along with an inconsistency coefficient threshold of 0.0 seems to be a good choice.

In order to exclude the possibility that this is a dataset-specific phenomenon, Table 7.5 depicts the thresholds for the optimal precision for all other datasets (upper table). Apart from CiteULike and Flickr, the same optimal threshold is obtained. However, a closer look reveals that $\tau_{inc} = 0.0$ yields almost the same precision in these cases (0.1255 and 0.23, respectively). The same holds for the BibSonomy dataset, where Ward’s method yields a precision of 0.0679 for a threshold of 0.0.

When shifting perspective to the F-measure, we find a similar common peak around $\tau_{inc} = 0.71$. Taking a look at the number of clusters produced, one can observe a decrease of around $\approx 2\,000$ clusters in this region. Again, this effect is visible across all datasets (see Table 7.5). Increasing the threshold further has a detrimental effect in all cases. Taking a cursory look at the resulting synsets at this level reveals naturally slightly larger clusters, conveying the impression of “topics”, which embrace closely related, but not necessarily synonym keywords.

In summary, one can condense the results of the semantic evaluation to two statements: First, Ward’s method together with an inconsistency coefficient threshold seems to be most well-suited among all analyzed variants for the purpose of detecting synonym clusters. And second, when using the latter, there seem to be two natural “steps” of cluster formation, taking place when (i) highly related “leaf” tags are combined into smaller clusters (threshold of 0.0) and when (ii) those clusters are combined into topic-like larger clusters (threshold of 0.71). It depends on the requirements of a knowledge engineer

Table 7.5.: Optimal parameters and cluster sizes for synonym detection on different datasets. The upper table depicts depicts the optimal thresholds and cluster sizes for precision, the lower one for the F-measure (see also Figure 7.15)

	<i>threshold</i>	<i>precision</i>	<i>method</i>	<i># clusters</i>
BibSonomy	0.0	0.0695	complete	6 387
CiteULike	0.59	0.1257	ward	6 407
Delicious	0.0	0.1582	ward	6 426
Flickr	0.65	0.2302	ward	6 260
AOL logsonomy	0.0	0.1334	ward	6 341
Stackoverflow	0.0	0.1655	ward	5 994

	<i>threshold</i>	<i>F-measure</i>	<i>method</i>	<i># clusters</i>
BibSonomy	0.71	0.0706	ward	3 808
CiteULike	0.71	0.1174	ward	3 838
Delicious	0.71	0.1301	ward	3 777
Flickr	0.77	0.2254	ward	3 676
AOL logsonomy	0.71	0.1192	ward	3 845
Stackoverflow	0.84	0.1527	ward	3 870

which step is most suited for a particular application at hand. For the purpose of building a synsetized folksonomy, the first step is probably the preferred one, because it corresponds to a more “cautious” consolidation of highly related keywords.

In order to convey an impression of the resulting keyword clusters obtained on both step levels, Table 7.6 depicts some examples. The synonyms of “step 1” (i. e., those obtained by a threshold of 0.0) seem to come in fact closer to the notion of “identical meaning” (or to meaningful concepts, like exemplified by “desperate housewives”⁸ or “evolutionary selection”), while those obtained in the second step (i. e., by a threshold of 0.71) add slightly more widely related keywords.

In order to tackle the next problematic issue of the open vocabulary of Social Annotation Systems, the following chapter will explain methods to resolve

⁸“Desperate Housewives” is the title of an American television series.

Table 7.6.: Example synonym sets created by hierarchical clustering, using Ward’s method together with an inconsistency threshold. The two steps correspond to setting its value to $\tau_{inc} = 0.0$ and $\tau_{inc} = 0.71$, respectively.

<i>Dataset</i>	<i>keyword</i>	<i>synonyms step 1</i>	<i>synonyms step 2</i>
BibSonomy	football	soccer	soccer, basketball
	genderstudies	gender	geschlechterforschung, gender, feminismus, queer, frauen, gendercrawl
	broadcast	broadcasting	broadcasting, audience
CiteULike	amplifier	amplifiers	amplifiers, gain
	selection	evolutionary	adaptation, drift, evolutionary
	kernel	kernels	kernels, factorization
Delicious	parser	parsing	parsing, compiler, syntax
	remove	removal	removal, hijacker, nuker, remover
	soviet	ussr	ussr, chernobyl
Flickr	smiling	laughing	laughing, posing, smiles
	jewelry	jewellery	jewellery, beads, bracelet, necklace
	nyc	newyorkcity	newyorkcity, newyork, ny, manhattan
AOL logsonomy	housewives	desperate	desperate, elliot, wright, regina
	conroller	controls	controls, active, features
	fence	fences	fences, privacy
Stackoverflow	digraphs	directed-graphs	directed-graphs, directed-acyclic-graphs, dag, igraph, networkx
	empty	blank	blank, none, check
	pixel	pixels	pixels, processing, raw

polysemy by discovering multiple senses of a single keyword.

7.2.3. Tag Sense Disambiguation

As mentioned in Section 3.1.1, another weakness of Social Annotation Systems is related to polysemy, i. e., when a single keyword may have more than one meaning. This clearly hampers retrieval as well as browsing facilities: Because the different senses of a keyword may be semantically unrelated (e. g., **apple** as a fruit and a company), the user is presented with irrelevant content. Naturally this problem is not restricted to Social Annotations, but is present basically within all systems dealing with natural language; however, the open vocabulary

as well as the lack of structure (compared to, e. g., the syntax of a written text) makes this issue more visible.

Despite that, the problem of word sense disambiguation has been addressed in a large number of studies in the field of Natural Language Processing (Manning and Schütze, 1999, p. 229ff). The applied methods can be broadly distinguished in supervised and unsupervised approaches. Supervised methods see disambiguation mainly as a classification task and require a defined set of senses for each word, as well as a correctly disambiguated training set. Both requirements make their application to the dynamic vocabularies of Social Annotation Systems difficult, because the number of senses of a given keyword is generally unknown and subject to change. Unsupervised methods can be distinguished according to the necessity to specify the number of senses in advance. Clearly, for our problem at hand only those approaches are viable which do not require an a-priori determination of this end.

Accordingly, the first desirable goal in our case is best described by *tag sense discovery*. NLP approaches in this field like (Dorow and Widdows, 2003; Pantel and Lin, 2002) 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 two crucial issues hereby are (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 prior work, (Au Yeung et al., 2009a) performed extensive studies on the characteristics of different context definitions for the task of tag sense discovery. More precisely, the authors examined tag- and user-based document networks, as well as tag co-occurrence and similarity networks. Among those, it was found that tag context similarity networks provided *the most clear-cut results among all the network types..* Although hereby some more specific senses were missed in some cases, we will adopt tag similarity networks for building contexts because we are primarily interested in “generally agreed” emergent senses – which we assume to be more general ones. Hence, we will use the tag context relatedness to depict the relations among the items present in the context of a given keyword.

The next question is which keywords to include in the context of a given keyword t . The goal hereby is to choose a sample of context keywords which are representative for t 's main senses. Hereby we follow the procedure described by (Rapp, 2003), who found that the “*20 strongest first-order associations [...] are [...] a good mix of the two main senses for each word*”. First-order

associations correspond to tag-tag co-occurrence in our case. Although we do not necessarily target to discover *two* main senses, we follow these steps to construct a context for a given keyword t :

1. Let $t \in T$ be a keyword 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 $w : V_t \rightarrow \mathbb{R}$. We call this graph 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 maximize the co-occurrence relatedness to t .
4. The edges are constructed by computing the pairwise tag context relatedness as introduced in Section 7.1.2 among all $t \in V_t$; we add an edge 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.

Given this graph representation of the context, the next step is how it can be divided into partitions which denote different meanings. As stated above, clustering techniques have been used to this end, e. g., Clustering By Committee (Pantel and Lin, 2002), Markov Clustering (Dorow and Widdows, 2003) or graph clustering (Au Yeung et al., 2009a). From a different direction, this problem is similar to community detection, for which modularity-based clustering techniques showed promising results (Leskovec et al., 2010), and were also applied to sense detection by (Au Yeung et al., 2009a).

For the scope of this study, we will stick to hierarchical agglomerative clustering as described in the previous chapter as a representative of a standard algorithm; furthermore we include a graph partitioning method inspired by (Widdows and Dorow, 2002), as well as a modularity-based clustering used by (Au Yeung et al., 2009a). Each approach will be briefly introduced in the following subsections.

Hierarchical Agglomerative Clustering

Based on the similarities among the context keywords which form the edges of the sense context graph SC_t , the hierarchical clustering procedure described in the previous Section 7.2.1 can be directly applied to form “sense clusters”. Because it showed optimal behavior for the task of synonym clustering, we will

stick to Ward’s method for updating the distance matrix. For obtaining distinct clusters, we will vary the distance criterion threshold for simplicity reasons, and because the inconsistency coefficient showed suboptimal behavior during initial studies.

Graph Clustering

Another observation made by (Widdows and Dorow, 2002) is that the different senses of a given word are often semantically unrelated. Based on a similarity-based graph representation of a word’s context as in our case, the authors proposed to identify different senses as different *connected components* within the context graph. This follows the intuition that context words belonging to a particular sense will be highly similar (i. e., connected) to each other, while less similar (i. e., unconnected) to the other senses. In other words, with an appropriate definition of similarity, the context graph should “automatically” decompose itself into sense components. Of course a crucial question hereby is how to fix an appropriate similarity threshold when creating the edges within the context graph: A too low value will yield a single connected component (i. e., a single sense), while a too high value will result in a set of singleton components – both of which is not desired. In order to check if a suitable threshold setting produces meaningful senses, we varied the thresholds between 0.0 and 1.0.

Modularity-based Clustering

An inherent property of the two aforementioned algorithms is that the number of resulting clusters depends strongly on the specified threshold. Another approach to detect a “natural” number of clusters which does not require parameterization is *modularity-based* clustering. It has been applied to community detection in social networks (Leskovec et al., 2010), and also to tag sense discovery (Au Yeung et al., 2009a). Modularity is hereby a measure of quality of a given graph partitioning; see (Leskovec et al., 2010) for a detailed explanation. As a concrete implementation, we used the algorithm by (Blondel et al., 2008). It basically takes the context graph as an input, and returns a “best partition” of its nodes according to a modularity optimization strategy.

Having introduced three clustering variants to discover tag senses, it is important to notice that our intention is not a complete coverage of all relevant

clustering techniques in order to find the optimal method. Moreover, we want to assess whether our proposed measures of tag relatedness serve as a useful input for an exemplary set of applicable clustering approaches, coming from different fields. Furthermore, we want to assess whether the relatively small set of 20 context keywords suffices to identify different senses.

In all cases, a further question is how the computed senses should be labeled. In the literature, typically the most popular tags within the resulting clusters are used (Au Yeung et al., 2009a). Instead, we propose to choose a single label by choosing the keyword t_i within the sense cluster which maximizes the tag context relatedness to the keyword 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 the tag context relatedness. Then we choose the label t_S for S as follows:

$$t_S = \operatorname{argmax}_{t' \in S} sim(t, t')$$

While we do not propose to disregard the remaining keywords within the sense cluster, we hypothesize that t_S is a concise description of what the cluster is about. We will refer to the remaining keywords as *preference tags*, because they serve as an additional description of what the sense is about. In order to assess the performance of all methods under consideration, the following section presents a gold-standard based evaluation approach.

7.2.4. Evaluation by Semantic Grounding

As mentioned by (Pantel and Lin, 2002), evaluating the performance of a sense discovery algorithm is difficult, because there is no completely reliable way to determine the “correct” set of senses within a given corpus. Existing sense definitions like those within WordNet may be too fine-grained or incomplete. For this reason, researchers have used manual case studies instead (Au Yeung et al., 2009a). However, those do not scale for the purpose of frequent evaluation across several datasets, as is required in our case. Because of that, we stick to a WordNet-based evaluation approach, because we think that despite its shortcomings, it can still give us some insights on the quality of the results produced by the different algorithms. Within WordNet, a polysemous word is being assigned to several synsets, each of which is being described by a *gloss*, a short textual description what the synset is about. For the computed sense clusters, one could also interpret the contained keywords within each cluster

as a lightweight form of “gloss”. Hence, we propose to match the computed senses against those defined within WordNet based on overlapping words in both “glosses”. Because WordNet is reported on the one to be too fine-grained, and on the other hand to miss other senses (Au Yeung et al., 2009a), precision and recall are not very well suited. Instead, we suggest to take the absolute number of matching senses as an indicator of quality. The intuition behind this decision is that if an algorithm produces a large number of senses which can be mapped correctly to WordNet senses, then its performance is better.

More precisely, the sense mapping process between learned sense clusters and WordNet glosses involves the following steps:

1. Let t be a keyword whose sense context graph $SC_t = (V_t, E_t)$ has been disambiguated by an algorithm into k sense clusters $C_t = \{S_1, \dots, S_k\}$, $S_1 \dot{\cup}_{i=1, \dots, k} S_i = V_t$.
2. First we remove all singleton clusters from C_t (i. e., those with $|S_i| = 1$), because we want focus more on “generally agreed” senses.
3. After that, we retrieve the set of glosses $G_t = \{g_1, \dots, g_j\}$ for each noun sense of t within WordNet.
4. We consider a sense cluster S_i to match a gloss g_j if at least two keywords present in S_i occur in the gloss g_j . We have chosen two keywords in order to exclude matches by chance of only a single keyword, and in order to still be able to match also smaller sense clusters (consisting of just a few keywords). Based on this definition, we compute all matching pairs of sense clusters and glosses.
5. From the resulting set of matches, we remove all those which do not correspond to a one-to-one mapping, i. e., we remove all matches where a sense cluster matches more than one gloss, and all matches where one gloss matches more than one sense cluster.

Based on the resulting set of matches for each keyword t , we compute the global match count by summing over those keywords for which one of the following conditions holds:

- $|C_t| = 1 \wedge |G_t| = 1$

- $|C_t| > 1 \wedge |G_t| > 1$

The first condition corresponds to the case in which an algorithm has correctly predicted that a keyword has exactly one sense; the second one to the case where polysemy has been correctly identified. This comparatively elaborate kind of aggregation is necessary because otherwise algorithms which produce exactly a single sense for each keyword would be strongly biased.

Figure 7.16 shows the results for varying the threshold for hierarchical clustering and Widdow’s graph clustering algorithm, exemplified for the Delicious dataset. Please note the different scale of the two x-axes; in order to ease comparison, the x -axis for the distance criterion threshold is inverted. In this way, the starting point on the left hand side corresponds for both algorithms to the state where each keyword is assigned exactly one sense (i. e., all context keywords lie within a single cluster). When moving from left to right (i. e., lowering the distance criterion threshold and increasing the similarity threshold), the number of sense clusters is growing. The interesting question hereby is if we can obtain a higher number of matches by splitting up the “correct” keywords (namely those which are polysemous within WordNet). It becomes apparent that this is the case for both algorithms; however, the number of matches produced by the hierarchical clustering algorithm is consistently higher compared to the graph clustering approach. So a first impression is that according to the chosen evaluation criterion, hierarchical clustering seems to be slightly better suited than the graph clustering approach by Widdows to discover keyword senses which exist also in an external lexical resource like WordNet. Furthermore, based on the number of matches it is now also possible to deduce an “optimal” parameter for each case, i. e., a setting which maximizes the number of matches. For the dataset at hand, those are 0.35 for the similarity threshold and 1.38 for the distance criterion threshold.

In order to give a more detailed insight, we performed the same parameter optimization for all other datasets under consideration. Table 7.7 summarizes the properties of the results. Hereby we are also including the modularity-based clustering approach for comparison. In terms of the absolute number of matches, hierarchical clustering reaches the best performance in all cases except for the Flickr dataset, where it scores slightly lower, but still better than the graph clustering approach. This points further in the direction that hierarchical clustering seems to be the most adequate choice among all examined methods to discover tag senses. An additional insight is that while modularity-based

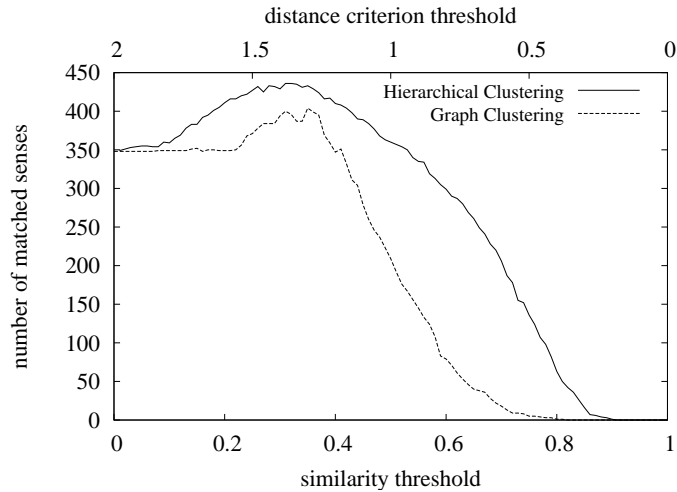


Figure 7.16.: Parameter variation for two clustering algorithms for tag sense discovery. The performance is measured against matched senses within WordNet (y -axis). The graph clustering algorithm is inspired by Widdows (see Section 7.2.3).

clustering also performs quite well in some conditions, it has a strong tendency to judge *all* keywords to be ambiguous (as can be seen for the very high *perc_ambig* values in Table 7.7). While this fine-grained distinction of senses might be desired in some cases, the possibility to obtain only “generally agreed” senses by an appropriate threshold might be more suitable in others.

Despite that, the results so far have to be seen in light of the shortcomings of the chosen evaluation paradigm; because, e. g., the matching between WordNet glosses and sense clusters might be afflicted with noisy matches, the dominance of hierarchical clustering is best regarded as a tendency. In order to establish a connection between our chosen evaluation method and a user study, we reuse a manually compiled set of disambiguated keywords from Delicious introduced by (Au Yeung et al., 2009a). It consists of 10 ambiguous tags, which were disambiguated to in total 22 senses by human judges. Though the concrete dataset used in this study is not identical to ours, we can still identify each tag and compare the manually created senses with the ones produced by our algorithms, using the obtained optimal thresholds. In this setting, we can use precision, recall and F-measure as valid performance metrics, because both

Table 7.7.: Optimal sense clusters according to different clustering algorithms. *num_matches* denotes the number of matches, *opt_thresh* the optimal threshold obtained by parameter variation, *perc_ambig* the percentage of ambiguous keywords and *avg_senses* stands for the average number of senses, computed over all ambiguous keywords.

	<i>Hierarchical clustering</i>	<i>Widdow's graph clustering</i>	<i>Modularity-based clustering</i>
BibSonomy			
<i>num_matches</i>	167	159	144
<i>opt_thresh</i>	1.26	0.15	n.a.
<i>perc_ambig</i>	79.7 %	31.4 %	98.3 %
<i>avg_senses</i>	2.21	2.31	2.67
CiteULike			
<i>num_matches</i>	406	384	275
<i>opt_thresh</i>	1.62	0.15	n.a.
<i>perc_ambig</i>	20.9 %	1.1 %	99.6 %
<i>avg_senses</i>	2.01	2.02	2.47
Delicious			
<i>num_matches</i>	436	404	387
<i>opt_thresh</i>	1.38	0.35	n.a.
<i>perc_ambig</i>	78.8 %	57.6 %	99.7 %
<i>avg_senses</i>	2.07	2.67	2.55
Flickr			
<i>num_matches</i>	435	375	453
<i>opt_thresh</i>	1.24	0.22	n.a.
<i>perc_ambig</i>	95 %	71.3 %	99.9 %
<i>avg_senses</i>	2.49	3.02	2.87
AOL logsonomy			
<i>num_matches</i>	602	437	598
<i>opt_thresh</i>	1.04	0.41	n.a.
<i>perc_ambig</i>	81.2 %	48.3 %	92.2 %
<i>avg_senses</i>	2.19	2.51	2.14
Stackoverflow			
<i>num_matches</i>	63	38	62
<i>opt_thresh</i>	1.22	0.11	n.a.
<i>perc_ambig</i>	77.7 %	6.6 %	99.7 %
<i>avg_senses</i>	2.16	2.04	2.74

Table 7.8.: Manual evaluation of tag sense discovery for the Delicious dataset. Precision, recall and F-measure are computed relative to a manually disambiguated set of 10 keywords compiled by Au Yeung et al. (2009a).

	<i>Hierarchical clustering</i>	<i>Widdow's graph clustering</i>	<i>Modularity-based clustering</i>
<i>precision</i>	0.78	0.63	0.56
<i>recall</i>	0.78	0.46	0.68
<i>F-measure</i>	0.78	0.53	0.62

sense sets are now defined over the same dataset. The matching between the senses was done manually.

Table 7.8 depicts the results. Because the average number of senses is higher, the modularity-based clustering scores comparatively low precision values. Apart from that, the hierarchical clustering approach outperforms the other examined methods. Although the manual evaluation scores will change with different threshold settings, we think it is a fair comparison after having selected the “optimal” parameters by the aforementioned procedure. This provides evidence that our WordNet-based evaluation approach is valid, and corroborates further the impression that hierarchical clustering is the best choice among all analyzed methods to discover tag senses. Of course the setting of the distance criterion threshold is a crucial issue; however, its variation also gives the knowledge engineer the possibility to adjust the granularity of discovered senses. For particular applications, also an optimization strategy like the one presented in this analysis (i. e., including external lexical repositories) may be the method of choice. In order to provide the reader with an impression of the disambiguation quality, Table 7.9 depicts some exemplary discovered senses within all datasets under consideration. Hereby the usefulness of our proposed method to denote the sense label of a cluster becomes visible.

7.2.5. Summary

This chapter on concept learning was mainly intended to examine techniques addressing a major shortcoming of Social Annotation Systems, namely that retrieval of relevant resources is negatively affected by the usage of different keywords for the same concept, and by keyword ambiguity. The first problem

Table 7.9.: Examples of discovered senses of selected keywords produced by hierarchical clustering along with the optimal distance criterion threshold (see Table 7.7. Samples are taken from all datasets under consideration.

<i>keyword</i>	<i>sense label</i>	<i>preference keywords (excerpt)</i>
lemon ^a	recipe	dessert food recipes chicken baking recipe cake cooking
	attorney	consumer auto lawyer car lawyers automobile law usa
paper	craft	diy art craft origami cool crafts design fun blog shopping
	research	software reference toread tools article science web
desert	photography	art photography blog design photos
	chocolate	dessert food recipes cooking chocolate recipe
	arizona	california high travel home to plants in arizona palm usa
bug	insect	fly macro bugs closeup insects nature flower yellow animal
	beetle	2005 vw volkswagen beetle
game	games	videogame screenshot games video
	football	highschoolfootball show play v blue high hockey texas night
ant ^b	colony	optimization colony algorithms genetic systems artificial
	build-tools	tool tools apache build-tools java develop development build
cloud	data	grid computing storage data amazon analysis
	tagcloud	tag tagcloud tagging web 2.0 daily tags google blog web2.0
library	applications	07_system soft applications bar
	opensource	tools web javascript java opensource programming software
	books	blog information search bibliothek science research web2.0
taxonomy	phylogeny	vertebrates primates evolution genetics morphology animals
	web	social folksonomy software ontology web classification
screen ^c	yeast	genetic genomics drug yeast bioinformatics protein rnai
	printing	technology interaction methods touch surface multi
flea	pet	pictures black red free dogs white dog pet
	market	market city markets house home florida water center sale day
disc ^d	system	pictures dvd cd video free high make home http car music
	pain	problems disk body pain back
index	mysql	oracle sql sql-server postgresql query mysql database
	indexing	indexing search lucene
	arrays	c# python performance javascript arrays optimization php
pattern	matching	matching regex string perl javascript matcher php match
	c#	c# .net asp.net
	design-patterns	mvc architecture factory oop design design-patterns

^a“Lemon laws” is a nickname of a special kind of American state laws concerning the rights of automobile purchasers, see http://en.wikipedia.org/wiki/Lemon_law

^bApache Ant is a build-tool for developing Java applications, see <http://ant.apache.org/>

^c“Yeast screening” is a technique used in molecular biology, see http://en.wikipedia.org/wiki/Two-hybrid_screening

^dA common cause for backaches are problems with the intervertebral discs.

was tackled by methods of *synonym resolution*, whose goal is to form groups of highly related keywords, yielding a so-called *synsetized folksonomy*. Based on an evaluation against reference synonym sets from WordNet, hierarchical clustering using Ward’s method together with an inconsistency coefficient thresholding seemed to produce meaningful clusters of higher quality than a simple similarity-based thresholding approach. For the second problem, the core idea was to perform *sense discovery* for a given keyword by partitioning its context into *sense clusters*. For this purpose, again hierarchical clustering together with a distance criterion threshold exhibited advantages compared to graph clustering and modularity-based clustering approaches. This result was based on matching disambiguated senses to WordNet senses, and a subsequent small-scale manual evaluation against a manually disambiguated set of keywords.

Despite hierarchical clustering proved to be useful for both purposes, it is not the intended core contribution of this section to have identified the “optimal” clustering technique. We see the main contribution in (i) the used methodology, showing up a reference-based paradigm of evaluating keyword-based concept learning, and (ii) the confirmation that the measures of keyword relatedness introduced in Section 7.1.1 are a valuable input to standard clustering algorithms with the goal to synthesize and disambiguate keywords.

In summary, our results point at the direction that our applied methods are able to capture “emergent concepts” within the vocabularies of our analyzed Social Annotation Systems. Although these can not directly be expected to exhibit the same degree of semantic precision than those defined by experts, the learned structures can be of great help for (i) enhancing retrieval within Social Annotation Systems, and (ii) tackling the knowledge acquisition bottleneck of semantic resources by discovering new meanings of existing terms.

Having identified concepts, a crucial aspect when trying to assemble those into a hierarchical structure is how to capture the different levels of generality. The following section presents a thorough analysis of generality measures derived from Social Annotations.

7.3. Capturing Semantic Generality

As introduced in Section 4.2.2, a semantic representation of concepts like a taxonomy allows to differentiate the contained concepts based on their level of “generality”: Because the hierarchical structure is typically obtained by a top-

down approach of subsequently subdividing a domain, more general concepts are typically found closer to the taxonomy root. Despite those hierarchical structures are not explicitly present within Social Annotation Systems, the existence of emergent semantics suggests that it might still be possible to assess the “level of generality” of a given keyword. Within the literature, several folksonomy-derived notions of “tag generality” have been used for this purpose. In a similar fashion to the systematic analysis of relatedness measures presented in Section 7.1, the objective of this chapter is to study the characteristics of the different approaches. Our chosen methodology is hereby to examine how close different measures come to the notion of generality which is encoded in a set of reference taxonomies.

Building on the formalization of semantic generality measures presented in Section 4.2.2, we will primarily make use of *ranking functions* γ , which induce the generality measures. Because these functions are not defined based on ontologies, but rather on folksonomies or derived networks, we generalize the definition 4.6 of term generality measures to allow a common terminology as follows: Let \mathbb{S} be either a Folksonomy $\mathbb{F} = (U, T, R, Y)$ or a derived term graph $\mathbb{T} = (T, V)$ (as defined in Sections 3.1.2 and 3.1.3, respectively). Then we refer to a *term generality measure* based on \mathbb{S} as a partial ordering among the contained vocabulary:

$$\sqsubseteq^{\mathbb{S}} \subseteq T \times T$$

Accordingly, the corresponding *generality ranking functions* are defined as:

$$\gamma_T : T \rightarrow \mathbb{R}^+$$

Apart from that, these definitions inherit all remaining properties from those based on ontologies (see Section 4.2.2). We will now continue by presenting generality measures coming from different fields; this will be followed by a semantic grounding of each measure against “ground-truth” measures, which gives us insights into the characteristics of each measure.

7.3.1. Generality Measures

Based on the problem formalization from the previous chapter, we will now introduce a set of ranking functions γ_T which are supposed to order lexical items within a folksonomy \mathbb{F} by their degree of generality, inducing a partial order

$\sqsubseteq_{\gamma^T}^{\mathbb{F}}$ among the set of tags. The measures are partially based on prior work in related areas, and build on different intuitions. One commonality they all share is that none of them considers the textual content of a tag itself (e.g., with linguistic methods). All measures operate solely on the folksonomy structure itself or on a derived term network, making them language-independent.

Frequency A first natural intuition is that more abstract tags are simply used more often, because there exist more resources which they describe - as an example, the number of *computers* in the world is much larger than the number of *notebooks*. Hence one might assume that within a folksonomy, the tag *computer* is used more often than the tag *notebook*. We capture this intuition in the abstractness measure $\sqsubseteq_{freq}^{\mathbb{F}}$ induced by the ranking function $freq(t)$ which counts the number of tag assignments according to

$$freq(t) = |\{(u, t', r) \in \mathcal{Y} : t = t'\}| \quad (7.1)$$

SNA measures In network theory the *centrality* of a node in a network is usually an indication of how important the vertex is (Wasserman and Faust, 1994). Such metrics typically capture the connectedness and the position of a node; in the field of social network analysis, this allows, e.g., to identify members which are close to the network “core”, and could hence potentially be more influential. Applied to our problem at hand, centrality can also be interpreted as a measure of abstractness or generality, following the intuition that more abstract terms are also more “important”. The same idea underlies the an approach by (Heymann and Garcia-Molina, 2006) to infer hierarchical tag relationships. We adopted three standard centralities (degree, closeness, betweenness). All of them can be applied to a term graph \mathbb{T} , leaving us with three measures $\sqsubseteq_{dc}^{\mathbb{T}}$, $\sqsubseteq_{bc}^{\mathbb{T}}$ and $\sqsubseteq_{cc}^{\mathbb{T}}$ as follows:

Degree centrality simply counts the number of direct neighbors $d(v)$ of a vertex v in a graph $G = (V, E)$:

$$dc(v) = \frac{d(v)}{|V| - 1} \quad (7.2)$$

According to *betweenness centrality* a vertex has a high centrality if it can be found on many shortest paths between other vertex pairs:

$$bc(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (7.3)$$

Hereby σ_{st} denotes the number of shortest paths between s and t and $\sigma_{st}(v)$ is the number of shortest paths between s and t passing through v . As its computation is obviously very expensive, its application to larger networks is typically only feasible by using approximations (Brandes, 2001) like calculating the shortest paths only between a fraction of points.

Finally, a vertex ranks higher according to *closeness centrality* the shorter its shortest path length to all other reachable nodes is:

$$cc(v) = \frac{1}{\sum_{t \in V \setminus v} d_G(v, t)} \quad (7.4)$$

$d_G(v, t)$ denotes hereby the geodesic distance (shortest path) between the vertices v and t .

Entropy measures Another intuition stems from information theory: *Entropy* measures the degree of uncertainty associated with a random variable. Considering the application of tags as a random process, one can expect that more general tags show a more even distribution, because they are probably used at a relatively constant level to annotate a broad spectrum of resources. Hence, more abstract terms will have a higher entropy. This approach was also used by Heymann et al. (2008) to capture the “generality” of tags in the context of tag recommendation. We adapt the notion from there and define:

$$entr(t) = - \sum_{t' \in cooc(t)} p(t'|t) \log p(t'|t) \quad (7.5)$$

whereby $cooc(t)$ is the set of tags which co-occur with t , and the conditional probabilities are computed based on the co-occurrence counts according to $p(t'|t) = \frac{w(t',t)}{\sum_{t'' \in cooc(t)} w(t'',t)}$. $w(t',t)$ is hereby the co-occurrence weight defined in section 3.1.3. $entr(t)$ induces the term abstractness measure $\sqsubseteq_{entr}^{\mathbb{F}}$.

Statistical Models Schmitz et.al. (Schmitz, 2006) applied a statistical *model of subsumption* between tags when trying to infer hierarchical relationships. It is based on the assumption that a tag t subsumes another tag t' if $P(t|t') > k$

Table 7.10.: Statistical properties of the term graphs derived from the Delicious dataset.

<i>Term Graph</i>	$ T $	$ E $
<i>COOC</i>	892 749	38 210 913
<i>SIM</i>	10 000	405 706

and $P(t'|t) < k$ for a suitable threshold k . This approach has its roots in a model proposed by (Sanderson and Croft, 1999) to derive concept hierarchies from text. For measuring generality, the number of subsumed tags according to this model can be seen as an indicator of abstractness – the more tags a tag subsumes the more general it is:

$$subs(t) = |\{t' \in T : p(t|t') > k \wedge p(t'|t) < k\}| \quad (7.6)$$

The resulting abstractness measure will be denoted as $\sqsubseteq_{subs}^{\mathbb{F}}$.

While some of these measures were used in different contexts of analyzing Social Annotation data, their choice has been in most cases rather ad-hoc. In the following section, we present a comparative study of the introduced measures in order to allow a more systematic selection of a suitable measure for a given purpose.

7.3.2. Evaluation by Semantic Grounding

In order to assess the quality of the tag abstractness measures introduced in the previous section, a natural approach is to compare them against a ground truth. A suitable grounding should yield reliable assessments about the “true” abstractness of a given lexical item. Of special interest are hereby taxonomies and concept hierarchies, whose hierarchical structure typically contains more abstract concepts like `ENTITY` or `THING` close to the taxonomy root, whereby more concrete ones are found deeper in the hierarchy. Hence, we have chosen a set of established ontologies and taxonomies, which cover each a rather broad spectrum of topics. They vary in their degree of control – WordNet (Section 6.2.1) on the one hand being manually crafted by language experts, while the Wikipedia category hierarchy (Section 6.2.4) and DMOZ (Section 6.2.3) on the other hand are built in a much less controlled manner by motivated web users. For an introduction of each used reference dataset, refer to Section 6.2; an overview about their statistical properties can be found in Table 6.7.

As an experimental testbed for the proposed term abstractness measures, we used data from the social bookmarking system Delicious as described in Section 6.1.3. For space reasons, we will perform a detailed analysis solely for this dataset. Furthermore, because Delicious is the largest corpus within our collection which has been analyzed in other studies as well, we think that this choice makes the study most representative. From the raw data, we first derived the post-based *tag-tag co-occurrence graph* $COOC = (T', E_{cooc}, w_{cooc})$ as described in Section 3.1.3. Recall that two tags t_1 and t_2 are connected by an edge, if there is at least one post (u, T_{ur}, r) with $t_1, t_2 \in T_{ur}$. The edge weight is given by $w_{cooc}(t_1, t_2) := |\{(u, r) \in U \times R \mid t_1, t_2 \in T_{ur}\}|$. In order to exclude co-occurrences introduced by chance and to enable an efficient computation of the centrality measures, we removed all tags from the resulting graph with a degree of less than 2.

In a similar way to (Heymann and Garcia-Molina, 2006), we also derived a *tag-tag similarity graph* $SIM = (T'', E_{sim}, w_{sim})$ based on the resource context relatedness described in Section 7.3.1. We have chosen this measure because it was also used by (Heymann and Garcia-Molina, 2006), and showed furthermore good results for capturing semantic similarity (see Section 7.1.3). However, because rarely used tags have very sparse vector representations, we restricted ourselves to the 10 000 most frequently used tags. Based on the resulting pairwise similarity values, we added an edge (t_1, t_2) to the edge list E_{sim} when the similarity was above a given threshold $min_sim = 0.04$. This threshold was determined by inspecting the distribution of all similarity values. Table 7.10 summarizes the statistics of all obtained term graphs.

Subsequently, we computed all term abstractness measures introduced in the previous chapter based on the Delicious folksonomy DEL and the derived term graphs $COOC$ and SIM , i. e., \sqsubseteq_{freq}^{DEL} , \sqsubseteq_{entr}^{DEL} , \sqsubseteq_{dc}^{COOC} , \sqsubseteq_{bc}^{COOC} , \sqsubseteq_{cc}^{COOC} , \sqsubseteq_{bc}^{SIM} , \sqsubseteq_{cc}^{SIM} and \sqsubseteq_{subs}^{DEL} .

Grounding by reference taxonomies

As stated above, our grounding datasets (namely the reference taxonomies) contain information about concept subsumptions. If a concept c_1 subsumes concept c_2 (i. e., $(c_1, c_2) \in \geq_C$), we assume c_1 to be more abstract than c_2 ; as the taxonomic relation is transitive, we can infer $(c_1, c_2), (c_2, c_3) \in \geq_C \Rightarrow (c_1, c_3) \in \geq_C$ and hence that c_1 is also more abstract than c_3 . In other words, thinking of the taxonomic relation as a directed graph, a given concept c is more abstract than

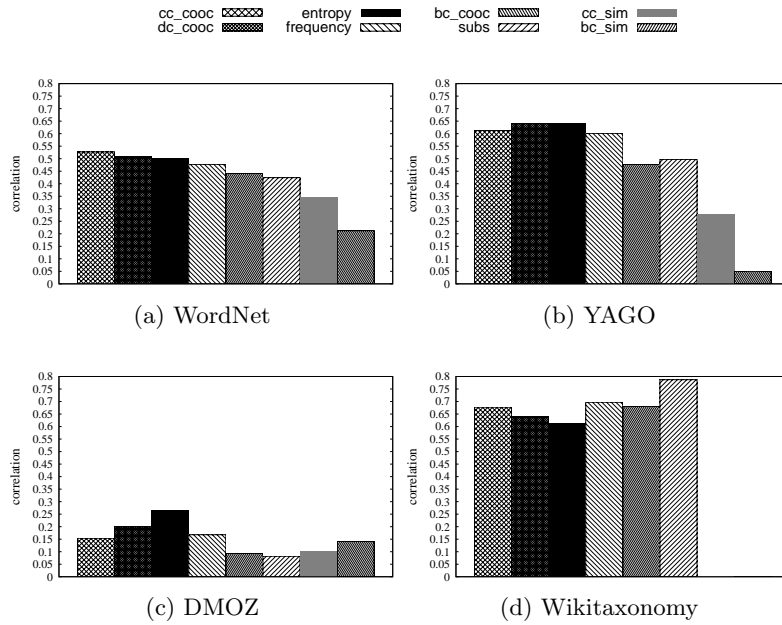


Figure 7.17.: Grounding of each introduced term abstractness measure \sqsubseteq^S against four ground-truth taxonomies. Each bar corresponds to a term abstractness measure; the y -axis depicts the gamma correlation as defined in Equation 7.8. .

all other concepts contained in the subgraph rooted at c .

As we are interested in the abstractness of lexical items, we can consequently infer that concept labels for more abstract concepts are more abstract themselves. However, hereby we are facing the problem of polysemy: A given lexical item l can be used as a label for several concepts of different abstractness levels. Consequently, l has “several” abstractness levels, depending in which context it is used. As a most simple approach, which removes possible effects of word sense disambiguation techniques, we “resolve” ambiguity in the following way: Let $\sqsubseteq^{\mathbb{O}} \subseteq C(\mathbb{O}) \times C(\mathbb{O}) =_{\geq_C}$ be the abstractness relation on the concepts of \mathbb{O} given by the taxonomic relation. Then, we construct a corresponding abstractness relation \sqsubseteq^{L_C} on the *lexical labels* L_C contained in the associated lexicon (L_C, L_R, Ref_C, Ref_R) according to:

$$(l_1, l_2) \in \sqsubseteq^{LC} \Leftrightarrow \{(c_1, l_1), (c_2, l_2)\} \in Ref_C \wedge (c_1, c_2) \in \geq_C \quad (7.7)$$

Due to the polysemy effect described above, \sqsubseteq^{LC} is not necessarily a partial order, as it may contain cycles. But despite this fact, \sqsubseteq^{LC} contains the complete information which terms $l_i \in L^C$ are more abstract than other terms $l_j \in L^C$ according to the taxonomy of \mathbb{O} . Hence we can use it as a “ground truth” to judge the quality of a given term abstractness measure \sqsubseteq^S . We are interested how well \sqsubseteq^{LC} correlates to \sqsubseteq^S ; picking up the idea of the *gamma rank correlation* (Cheng et al., 2010), we define *concordant* and *discordant* pairs between two partial orderings $\sqsubseteq, \sqsubseteq_*$ as follows: A pair of terms l and k is called concordant when both partial orderings $\sqsubseteq, \sqsubseteq_*$ agree on it, i. e., $(l \sqsubseteq k \wedge l \sqsubseteq_* k) \vee (k \sqsubseteq l \wedge k \sqsubseteq_* l)$. It is called discordant if they disagree, i. e., $(l \sqsubseteq k \wedge k \sqsubseteq_* l) \vee (k \sqsubseteq l \wedge l \sqsubseteq_* k)$. Based on these notions, the gamma rank correlation is defined as:

$$CR(\sqsubseteq, \sqsubseteq_*) = \frac{|C| - |D|}{|C| + |D|} \quad (7.8)$$

whereby C and D denote the set of concordant and discordant pairs, respectively. In our case, \sqsubseteq_* is not a partial ordering, but only a relation – which means that in the worst case, a pair l, k can be concordant and discordant at the same time. As is obvious from the definition of the gamma correlation (see Eq. 7.8), such inconsistencies lead to a lower correlation. Hence, our proposed method of “resolving” term ambiguity by constructing \sqsubseteq^{LC} according to Eq. 7.7 leads to a lower bound of correlation.

Figure 7.17 summarizes the correlation of each of our analyzed measures, grounded against each of our ground truth taxonomies. First of all, one can observe that the correlation values between the different grounding datasets differ significantly. This is most obvious for the DMOZ hierarchy, where almost all measures perform only slightly better than random guessing. A slight exception is the entropy-based abstractness measure \sqsubseteq_{entr}^F , which in general gives greater than 0.25 across all datasets. Another relatively constant impression is that the centrality measures based on the tag similarity graph (*cc_sim* and *bc_sim*) show a smaller correlation than the other measures. The globally best correlations are found for the Wikitaxonomy dataset, namely by the subsumption-model-based measure *subs*. Apart from that, the centrality measures based on the tag co-occurrence graph and the frequency-based measure show a similar behavior.

Hence the first impression is that there is no “clear winner” regarding the

ability of the measures to capture the notion of generality which is defined in external resources. However, another interesting first insight is that computationally rather “cheap” measures (like the frequency count or the degree centrality) perform equally or in some cases even better than more complex ones like closeness or betweenness centrality. In order to further investigate this point, we will now take into account information which is not explicitly encoded in the taxonomy, as well as a case study with human subjects.

Grounding by taxonomy-derived measures

The grounding approach of the previous section gave a first impression of the ability of each measure to predict term abstractness assessments which are *explicitly* present in a given taxonomy. This methodology allowed only for an evaluation based on term pairs between which a connection exists based on their associated concepts c_1 and c_2 in the taxonomy, i. e., pairs where c_1 is either a predecessor or a successor of c_2 in the subsumption hierarchy. However, our proposed measures make further distinctions among terms between which no connection exists within a taxonomy (e. g., the *freq* states that the most frequent term t is more abstract than *all* other terms). This phenomenon can probably also be found when asking humans – e. g., if one would ask which of the concepts **DOG** or **INANIMATE ENTITY** is more abstract, most people will probably choose **INANIMATE ENTITY**, even though both words are not connected by the is-a relation in (at least most) general-purpose taxonomies. This issue has also already been discussed in Section 4.2.2; see especially Figure 4.7.

In order to extend our evaluation to these cases, we derived two straightforward measures from a taxonomy which allow for a comparison of the abstractness level between terms occurring in disconnected parts of the taxonomy graph. Because this approach goes beyond the explicitly encoded abstractness information, the question is justified to which extent it makes sense to compare the generality of completely unrelated concepts, e. g., between **WATERFALL** and **CHAIR**. Besides our own intuition, we are not aware of any reliable method to determine when humans perceive the abstractness of two terms as *comparable* or not. For this reason, we validated the derived measures – namely (i) the shortest path to the taxonomy root and (ii) the number of subordinate terms – by an experiment with human subjects.

Table 7.11.: Results from the user study on judging semantic generality.

<i>Category</i>	<i>Number of classifications</i>
One tag more general	41
Same level	11
Not comparable	154
Do not know one or two tags	3

Shortest path to taxonomy root As stated above, most taxonomies are built in a top-down fashion, whereby more abstract terms are more likely to occur closer to the taxonomy root. Hence, a natural candidate for judging the abstractness of a term is to measure its distance to the root node. This corresponds to a ranking function $sp_root(c)$, which ranks the concepts c contained in a taxonomy in ascending order by the length of the shortest path between $root$ and c . Simply spoken, different shades of generality are hereby translated to different depth levels in the taxonomic hierarchy.

Number of subordinate terms Another measure is inspired by Kammann and Streeter (1971), who stated that “*the abstractness of a word or a concept is determined by the number of subordinate words it embraces[...]*”. Given a taxonomy \mathbb{O} and its comprised taxonomic relation \leq_C , we can easily determine the number of “sub-concepts” by $subgraph_size(c) = |\{(c, c') \in \leq_C\}|$. We are aware that this measure is strongly influenced, e. g., by fast-evolving domains like, e. g., **MOBILE COMPUTING**, whose rapid growth along with a strong expansion of the included vocabulary might lead to an overestimation of its abstractness level. This is another motivating reason for the user study presented in the next paragraph.

Validation by user study In order to check whether $sp_root(c)$ and $subgraph_size(c)$ correspond to human perception of abstractness, we performed an exemplary user study with 12 participants⁹. As a test set, we drew a random sample of 100 popular terms occurring in each of our datasets; for each term, we selected 3 candidate terms, taking into account co-occurrence information from the folksonomy *DEL*. The resulting 300 term pairs were shown to the each subject

⁹students and staff from two IT departments

Table 7.12.: Accuracy of the taxonomy-derived abstractness measures.

	WordNet	YAGO	DMOZ	Wikitaxonomy
<i>sp_root</i>	0.94	0.42	0.88	0.45
<i>subgraph_size</i>	0.94	0.96	0.8	0.87

via a web interface¹⁰, asking them to label the pair by one of 5 options (see Table 7.11). Hereby, the option “One tag more general” was split into two options, indicating which term is more general than the other one.

We calculated Fleiss’ κ (Fleiss, 1971) to take a closer look at the agreement of the study participants. In the case of this experiment $\kappa = 0.2836$ indicating fair agreement. Table 7.11 shows the results of the number of classifications given that an agreement of 6 or more participants signalizes significant agreement. The relatively high number of “not comparable” choices show that even with our elaborate filtering, the task of differentiating abstractness levels is quite difficult.

Despite this fact, our user study provided us with a well-agreed set of 41 term pairs, for which we got reliable abstractness statements. Denoting these pairs as $\sqsubseteq_{\text{manual}}$, we can now check the accuracy of the term abstractness measures introduced by *sp_root* and *subgraph_size*, i. e., the percentage of correctly predicted pairs. Table 7.12 contains the resulting accuracy values. From our sample data, it seems that the subgraph size (i. e., the number of subordinate terms) is a more reliable predictor of human abstractness perception. Hence, we will use it for a more detailed grounding of our folksonomy-based abstractness measures.

The ranking function *subgraph_size* naturally induces a partial order $\sqsubseteq_{\text{subgraph_size}}^{\mathbb{O}}$ among the set of concepts present in an ontology \mathbb{O} . In order to check how close each of our introduced term abstractness measures correlate, we computed – as already done within the first evaluation – again the *gamma correlation coefficient* (Cheng et al., 2010) between the two partial orders (see Equation 7.8). The transformation from a generality measure based on concepts to one which is based on lexical labels of concepts was also done analogously to Eq. 7.7 according to:

¹⁰http://www.kde.cs.uni-kassel.de/benz/generality_game.html

7.3. Capturing Semantic Generality

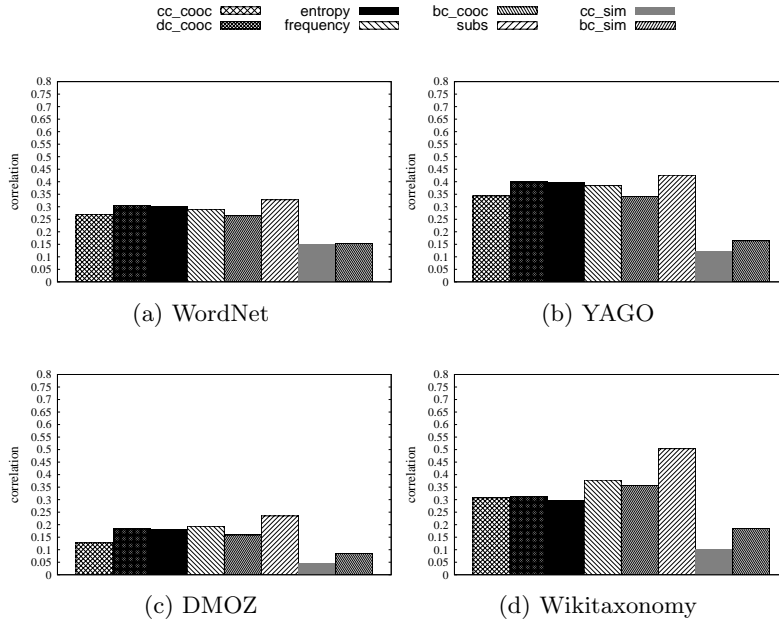


Figure 7.18.: Grounding of the term abstractness measure \sqsubseteq^S against $\sqsubseteq_{subgraph_size}^{\mathbb{O}}$ derived from four ground-truth taxonomies. Each bar corresponds to a term abstractness measure; the y -axis depicts the gamma correlation as defined in Equation 7.8.

$$(l_1, l_2) \in \sqsubseteq_{subgraph_size}^{L_C} \Leftrightarrow \{(c_1, l_1), (c_2, l_2)\} \in Ref_C \wedge (c_1, c_2) \in \sqsubseteq_{subgraph_size}^{\mathbb{O}} \quad (7.9)$$

Figure 7.18 shows the resulting correlations. Again, the correlation level between the datasets differs, with DMOZ having the lowest values. This is consistent with the first evaluation based solely on the taxonomic relations (see Figure 7.17). Another consistent observation is that the measure based on the tag similarity network (bc_sim and cc_sim) show the weakest performance. The globally best value is found for the subsumption model, compared to the Wikitaxonomy (0.5); for the remaining conditions, almost all correlation values lie in the range between 0.25 and 0.4, and correlate hence weakly.

Discussion Our primary goal during the evaluation was to check if folksonomy-based term abstractness measures are able to make reliable decisions about the relative abstractness level of terms. A first consistent observation is that measures based on frequency, entropy or centrality in the tag co-occurrence graph do exhibit a correlation to the abstractness information encoded in gold-standard-taxonomies. One exception is DMOZ, for which almost all measures exhibit only very weak correlation values. We attribute this to the less precise semantics of the DMOZ topic hierarchy, compared to the other grounding datasets. As an example, the category

`TOP/COMPUTERS/MULTIMEDIA/MUSIC_AND_AUDIO/SOFTWARE/JAVA`

does hardly imply that `SOFTWARE` “is a kind of” `MUSIC_AND_AUDIO`. WordNet on the contrary subsumes the term *Java* (among others) under taxonomically much more precise parents:

[...] > `ARTIFICIAL LANGUAGE` > `PROGRAMMING LANGUAGE` > `JAVA`

The same holds for YAGO, and the Wikitaxonomy was also built with a strong focus on *is-a* relations (Ponzetto and Strube, 2007). This is actually an interesting observation: Although both DMOZ and Delicious were built for similar purposes (namely organizing WWW references), the implicit semantics within Delicious resembles more closely to well-established semantic repositories than to the bookmark-folder-inspired hierarchical organization scheme of DMOZ.

Another consistent observation is that abstractness measures based on tag similarity graphs (as used, e. g., by (Heymann and Garcia-Molina, 2006)) perform worst through all experimental conditions. This is consistent with observations in Section 7.1.3, where we showed that distributional similarity measures (like resource context relatedness used by us and by (Heymann and Garcia-Molina, 2006)) induce connections preferably among tags having the same generality level. On the contrary, applying, e. g., centrality measures to the “plain” tag co-occurrence graph yield better results. Hence, a justifiable conclusion is that tag-tag co-occurrence encodes a considerable amount of “taxonomic” information.

But this information is not solely present in the co-occurrence graph - also a probabilistic model of subsumption (Schmitz, 2006) yields good results in some conditions, especially when grounding against the taxonomy-derived *subgraph_size* ranking. We attribute this to the fact that both measures (the subsumption model and the subgraph size) are based on the same principle, namely that a term is more general the more other terms it subsumes.

Apart from that, even the simplest approach of measuring term abstractness by the mere frequency (i. e., the number of times a tag has been used) already

exhibits a considerable correlation to our gold-standard taxonomies. This has an interesting application to the *popularity / generality problem*: Our results point in the direction that popular tags are on average more abstract (or more general) than less frequently used ones.

7.3.3. Summary

Analogous to the prior systematic characterization of tag relatedness measures, this section presented a comparative study of different families of measures which capture the degree of “semantic generality” of keywords within Social Annotation Systems. These were partially used in the literature, and comprise different levels of computational complexity: While frequency-based metrics as well as degree centrality and entropy exhibit a comparatively lightweight nature, the shortest path computations involved in closeness or betweenness centrality are much more demanding. For the evaluation based on the comparison against generality information encoded in reference taxonomy, all network-based measures were computed on two kinds of keyword graphs, namely a co-occurrence and a similarity graph. While the results exhibited a mixed picture in general, similarity-based tag networks seemed to be less well suited as an information source for capturing semantic generality. In summary, the interpretation of the results can be condensed in two statements: First, folksonomy-based measures of term abstractness do exhibit a partially strong correlation to well-defined semantic repositories; and second, the abstractness level of a given tag can be approximated well by simple measures.

To conclude, these results can be interpreted as further empirical evidence for the feasibility of making emergent semantics within Social Annotation Systems explicit. The presented characterization of generality measures is intended to inform the decision process of knowledge engineers concerned with deriving, e. g., hierarchical keyword relationships for a given purpose. In the same spirit, the following chapter builds on these insights, and presents methods to learn concept hierarchies based on Social Annotation data.

7.4. Learning Concept Hierarchies

Apart from the synonymy and polysemy problem, which was tackled by methods of concept learning in Section 7.2, another main shortcoming of Social Annotation Systems is their lack of structure. As explained in Section 3.1.1,

the inherently “flat” nature of the keyword space does not allow for tasks like narrowing or broadening a search, or browsing the system content from more general to more specific topics. Simply spoken, in this aspect Social Annotation Systems are missing exactly the advantages of the hierarchical arrangement of concepts within the taxonomic relation defined in an ontology. However, although explicit hierarchies are not present, researchers like (Cattuto et al., 2007) analyzed the network properties of folksonomies and found characteristics which “*could be related to an underlying hierarchical organization of tags*”. This very appealing vision of “emerging taxonomies” has motivated a large number of works targeted towards making the latent hierarchical structures explicit (see Section 5.3.5).

While some methods (like lexico-syntactic patterns or dictionary-based approaches, see (Cimiano, 2006)) of learning concepts from text are not applicable to Social Annotation data due to the lack of syntactical structure, others can be adapted. Among the latter, *clustering* techniques, especially based on distributional similarities like those introduced in Section 7.1 can be quite analogously applied. Apart from that, one can find in the literature some specific algorithms tailored towards Social Annotation Systems; an example is (Heymann and Garcia-Molina, 2006). These are best described by *generality-based* methods, because they are based on measures of keyword generality as those introduced in the previous Section 7.3.

In this section, we will examine approaches from both fields regarding their ability to induce a meaningful hierarchical structure among the initially flat keyword space. In a similar way to the approaches presented in the previous chapters, it is hereby not our intention to perform an exhaustive study on, e. g., a large set of clustering techniques in order to identify the optimal one. On the contrary, our goal is to pick representative approaches from different fields (which were partially already used in the literature for the same purpose), and assess which ones perform better in exploiting keyword *similarity* and / or keyword *generality* measures to build taxonomy-like keyword structures.

As an evaluation criterion, we will again primarily stick to a reference-based paradigm and compare the learned structures with reference taxonomies taken from several gold-standard datasets (see Section 6.2). In order to disentangle the effects of concept learning, we will first apply clustering and generality-based methods to the “plain” folksonomies (Sections 7.4.1 and 7.4.2). The quality of the results will then be judged first by a reference-based evaluation (Section 7.4.3), and then be backed up against a study involving human assessment

(Section 7.4.4). In a last step, we will present an enhanced version of the best-performing algorithm, which involves the techniques of synonym resolution and concept disambiguation described earlier.

7.4.1. Clustering Approaches

While the typical goal of clustering is to create partitions of similar objects, there exist several approaches in which not a flat set of clusters is produced, but rather a hierarchical arrangement of those (Bade, 2009). As input, those algorithms usually require information about the pairwise similarity among all objects to be clustered. Based on the analysis of semantic relatedness from Section 7.1, we can now build on its results and choose the *tag context relatedness* as a distributional measure which captures *semantic* relations among keywords.

Having fixed the similarity measure, the next important choice is which kind of clustering algorithm to use. Analogously to flat clustering methods, one possible criterion to differentiate the available approaches is the need to specify a “preferred” branching degree within the resulting taxonomy tree. As an example of a method which requires a fixed specification of how many “sub-clusters” to create in each splitting step, we will use a hierarchical variant of the well-known *k-means* algorithm (Jain et al., 1999). Though the a-priori fixation of a branching factor (by choosing a concrete parameter *k*) is surely a rather rigid control of the result, it can be motivated by cognitive limitations of humans interacting with the hierarchy: Hereby, structures with a very large number of sub-nodes are probably, e. g., quite difficult to navigate. In other words, the k-means approach corresponds to the effort to “force” the emergent hierarchies into a predefined structural schema.

As an example of a technique which leaves a greater degree of freedom to the properties of the resulting hierarchy, we will use *affinity propagation* as a new clustering method which has been successfully applied to derive hierarchies from Social Annotation data (Plangprasopchok et al., 2010). We prefer this method to the standard hierarchical agglomerative one (which showed good results for the tasks of synonym detection and sense discovery before) in order to enable a better comparison with state-of-the-art clusterings for our purpose at hand. The following two paragraphs will describe the concrete implementation¹¹ of

¹¹Please note that because the current study was done in collaboration with the Knowledge Management Institute of the Graz University of Technology, the implementation of the algorithms as well as the creation of the keyword hierarchies were performed by our

each clustering method.

Hierarchical K-Means The basic idea behind the k-means clustering algorithm is to produce a partition by iterative relocation of k centroids. The number of resulting clusters is hereby controlled by the parameter k . As a similarity metric among the data points (i. e., the keywords), we will use the tag context relatedness introduced in Section 7.1.1, using a “spherical” k-means variant introduced by (Dhillon et al., 2001). The latter differs from the original algorithm mainly in a heuristic which speeds up the convergence of the algorithm towards a (local) maximum of the cluster coherence criterion. Furthermore, in order to allow efficient computation, we used an optimized variant proposed by (Zhong, 2005). In order to derive a hierarchical structure among keywords, we utilize k-means iteratively in a top-down manner. Basically, in the first step, the whole input data set is used for clustering the data into 10 clusters. While the choice of $k = 10$ seems arbitrary, it was motivated by the above-mentioned cognitive limitations of humans when interacting with the resulting structure. Clusters containing more than 10 connected samples are further partitioned while ones with less than 10 samples are considered as leaf clusters. However, since a cluster containing 11 samples would also be partitioned into 10 clusters, we introduced a special case to give some freedom to the clustering process for these border cases by setting the cluster number to the maximum of 10 or number of data samples divided by 3 what would result in 3 clusters in case of 11 samples. The tag representing a node is selected by taking the nearest tag to the centroid. Furthermore, this tag is removed from the actual tags contained in a cluster and which are further clustered in the next step, if there are more than 10 samples left.

Affinity Propagation Frey and Dueck introduced Affinity Propagation (AP) as a new clustering method in (Frey and Dueck, 2007). A set of similarities between data samples provided in a matrix represents the input for this method. The diagonal entries (self-similarities) of the similarity matrix are called preferences and are set according to the suitability of the corresponding data sample to serve as a cluster center (called “exemplar” in (Frey and Dueck, 2007)). Although it is not required to set a cluster number explicitly, the preference values correlate

collaborators.

with the number of resulting clusters (lower preference values result in fewer clusters and vice versa).

In several iterations, AP exchanges messages between data samples to update their “responsibility” and “availability” values. Responsibility values reflect how well data samples serve as exemplars for other data, and the availability values show the suitability of other data samples to be the exemplars for specific data samples. Responsibility and availability are refined iteratively with a parameter λ as an update factor. A full description of AP is beyond the scope of this section, we point the interested reader to (Frey and Dueck, 2007) for further information.

Based on (Frey and Dueck, 2007), the authors of (Plangprasopchok et al., 2010) have introduced an adaption of affinity propagation to infer hierarchies from social tagging data. The authors incorporated structural constraints directly into the global objective function of affinity propagation, so that a tree evolves naturally from execution. Their work is based on incorporating user-defined keyword relations. Because in this dissertation, we solely rely on the network structure of Social Annotation Systems (as some systems do not allow the specification of keyword relations), we follow a simpler approach by applying the original AP recursively in a bottom-up manner. In a first step, the top 10 Cosine similarities (pruned for memory reasons) between the tags in a given dataset serve as the input matrix, and the minimum of those serves as preference for all data samples. Then, AP produces clusters by selecting examples with associated data samples. If the ratio between the number of clusters and the data samples is between 3 and 15 (which we use as an adjustable parameter), then the result will be retained, otherwise another run with lower (too many clusters have been selected) or higher preference values (too few clusters have been selected) will be executed. Finally, the centroids of the clusters are calculated by using the sum of the connected data samples normalized to unit length. Now the Cosine similarities between the centroids serve as the input matrix for the next run of affinity propagation. This approach is executed until the top-level is reached.

Since our objective is to construct a tag hierarchy where each node represents a unique tag, a tag in each cluster is used as a label. The label is selected by choosing the nearest tag to the centroid. Furthermore, this tag is removed from the actual tags contained in the leaf cluster and is not used as a representative in lower hierarchy levels. We set the AP parameter λ_0 to 0.6 with increasing values depending on the iteration count (i) ($\lambda_i = \lambda_{i-1} + (1.0 - \lambda_0) * i/i_{max}$).

AP will terminate after either a maximum of 5 000 iterations (i_{max}) or if the exemplars of clusters are stable for at least 10 iterations.

Following the description of the two clustering approaches, the next section presents generality-based approaches developed specifically for Social Annotation data.

7.4.2. Generality-based Methods

In (Heymann and Garcia-Molina, 2006), the authors introduce an incremental algorithm to assemble a tag hierarchy from social tagging data. It is comprised mainly of two building blocks, namely (i) a measure of keyword generality, and (ii) a measure of keyword similarity. As introduced in Section 7.3.1, in the original version generality is computed based on betweenness centrality in a tag similarity network. As a similarity measure, the authors stick to the resource context relatedness introduced in Section 7.1.1. Based on these two components, the scheme to induce a keyword hierarchy is as follows:

1. Create a list of all keywords, ordered in descending order by their degree of generality (according to the chosen generality measure).
2. Create an empty tree, and add the most general (i. e., topmost) keyword from this list as root node.
3. Iterate through the remaining keywords in order of descending generality and add each keyword t to the tree according to:
 - a) If a sufficiently similar keyword t' is present within the tree, add t and a child node to t' . The degree of sufficient similarity is controlled by a threshold τ_{sim} .
 - b) Otherwise, add t as child node to the root node.

Though its simplicity, the authors found that their proposed algorithm produces consistent hierarchical structures based on manual inspection. Furthermore, the algorithm is described as extensible due to the possibility to apply different similarity and centrality measures. As one can see, this property fits nicely into our general approach, and allows us to “plug in” different notions of generality and similarity, which were analyzed in depth in the previous sections. For the current study, we have chosen two settings: The first one is intended

Table 7.13.: Statistical properties of the tag-tag-networks derived from four social tagging systems.

	<i>BibSonomy</i>	<i>CiteULike</i>	<i>Delicious</i>	<i>Flickr</i>
<i>Tags</i>	56 424	347 835	380 979	395 329
<i>Links</i>	2 003 986	27 536 381	39 808 439	17 524 927

to come close to the setup of the original algorithm. Hence, as a generality measure, we have chosen *closeness centrality* within a tag similarity network. The latter is based on cosine similarity within the tag context, which is also used as similarity measure for the algorithm itself. We will denote this setup as *Clo/Cos*.

The second condition is driven by the idea of “simplicity“: Instead of using computationally more expensive measures like closeness centrality and cosine similarity, we were interested if “cheaper” measures like simple co-occurrence and degree centrality perform equally. For this reason, we have chosen tag-tag co-occurrence as similarity measure, and degree centrality within the co-occurrence network as generality measure. This condition will be referred to as *Deg/Cooc*.

Subsequently, we computed the keyword hierarchies by all aforementioned algorithms based on the BibSonomy, CiteULike, Delicious and Flickr dataset. The AOL logsonomy and Stackoverflow data is excluded, because social bookmarking systems were of primary interest, and in addition some data was compiled after the collaboration had taken place. Furthermore, in order to facilitate the computation of the clustering algorithms, we extracted the data of a single month from the Delicious and the Flickr dataset (November 2006 and December 2005, respectively). The BibSonomy dataset differs insofar as only the publications (not the bookmarks) were taken into account. For each dataset, we first computed the tag-tag co-occurrence network; Table 7.13 summarizes the size of the resulting network. This network was the direct input to the *Deg/Cooc* condition of the generality-based method introduced above. Then, we created a tag-tag similarity network by re-weighting the edges of the co-occurrence network with the tag context similarity. As a baseline, we also created a random hierarchy by repeatedly adding 10 randomly chosen tags as child nodes to each tag.

Table 7.14 summarizes some statistical properties of the produced hierarchies. The first impression is that on the structural level, there are some substantial

Table 7.14.: Statistical properties of the induced hierarchies by all proposed methods. *br* depicts the average branching factor, computed over all non-leaf nodes (For comparison, the branching factors of the reference taxonomies are: WordNet 4.86, YAGO 14.32, Wikitaxonomy 48.82). *dia* depicts the full network diameter, based on 500 randomly selected nodes (For comparison, the diameters of the reference taxonomies are: WordNet 7, YAGO 7, Wikitaxonomy 2).

	<i>BibSonomy</i>		<i>CiteULike</i>		<i>Delicious</i>		<i>Flickr</i>	
	<i>br</i>	<i>dia</i>	<i>br</i>	<i>dia</i>	<i>br</i>	<i>dia</i>	<i>br</i>	<i>dia</i>
<i>Affprop</i>	3.36	6	3.42	4	3.61	5	3.23	4
<i>Clo/Cos</i>	2.21	13	2.1	16	2.46	12	2.25	13
<i>Deg/Cooc</i>	8.04	8	6.82	9	8.14	9	9.17	7
<i>KMeans</i>	3.78	6	3.81	10	3.71	36	3.81	5
<i>Random</i>	10	2	10	3	10	2	10	4

Table 7.15.: Lexical overlap among concepts present in the learned and reference taxonomies. The values are approximated, as some induction algorithms led to slight variations of the overlap, but to a negligible amount (+/- 100 concepts).

	<i>BibSonomy</i>	<i>CiteULike</i>	<i>Delicious</i>	<i>Flickr</i>
<i>WordNet</i>	8 680	22 380	21 830	23 480
<i>YAGO</i>	5 970	14 180	13 620	13 770
<i>Wikitaxonomy</i>	11 280	33 430	40 270	37 950

differences: While the *Clo/Cos* approach seems to produce relatively “deep” hierarchies with a small branching factor, the *Deg/Cooc* variant has a bias towards more shallow ones. Both clustering variants lie mostly in between both extremes. Apart from these first insights, the following section compares each hierarchy with several gold-standard ones, targeting a deeper comparison on the *structural* level.

7.4.3. Gold-standard based Evaluation

A prerequisite for the application of hierarchy comparison metrics like those introduced in Section 5.4.2 (namely taxonomic overlap, precision and recall) is the existence of a sufficient overlap between the vocabulary between a learned and a reference taxonomy. Table 7.15 summarizes the overlap among all Social Annotation datasets and the reference hierarchies. The significant overlap in all cases makes the comparison on the structural level by the aforementioned metrics possible.

Figure 7.19 displays the results of the reference-based semantic evaluation. On the y -axis of each figure, the similarity between each learned hierarchy and a reference gold-standard taxonomy is depicted. We measure similarity using different measures, including taxonomic precision (TP), taxonomic recall (TR), taxonomic F1-measure (TF) and taxonomic overlap (TO), each based on the common semantic cotopy (*csc*) as characteristic excerpt. As explained in Section 5.4.2, all these measures have a local part based on the direct comparison of excerpts, while the global value is obtained by averaging the local ones.

At a first glance, the results from our experiments convey a consistent picture: Taking the taxonomic F1-measure (black bars) as an example, one can observe that across almost all experimental conditions the hierarchies induced by generality-based methods (*Clo/Cos* and *Deg/Cooc* in the figures) outperform the clustering-based ones (*Affprop* and *Kmeans*). A similar distribution is found for the other measures (TP, TR and TO). In all cases, the folksonomy induced by the random algorithm performs worst and yields a similarity score of close to zero.

So one justified conclusion which can be drawn from these empirical results is that the clustering techniques we investigated seem to produce hierarchies which exhibit a smaller degree of similarity to gold-standard taxonomies than techniques based on term generality. Especially those produced by degree centrality as generality measure and co-occurrence as similarity measure seem to resemble most closely to the reference taxonomies. This is an interesting observation, especially regarding that these measures are computationally much more lightweight compared to, e. g., closeness centrality, cosine similarity or elaborate clustering mechanisms.

When comparing the clustering techniques among each other, it seems that affinity propagation has a slight advantage compared to k-means, however to a much lesser extent than the difference to the generality-based methods. An open

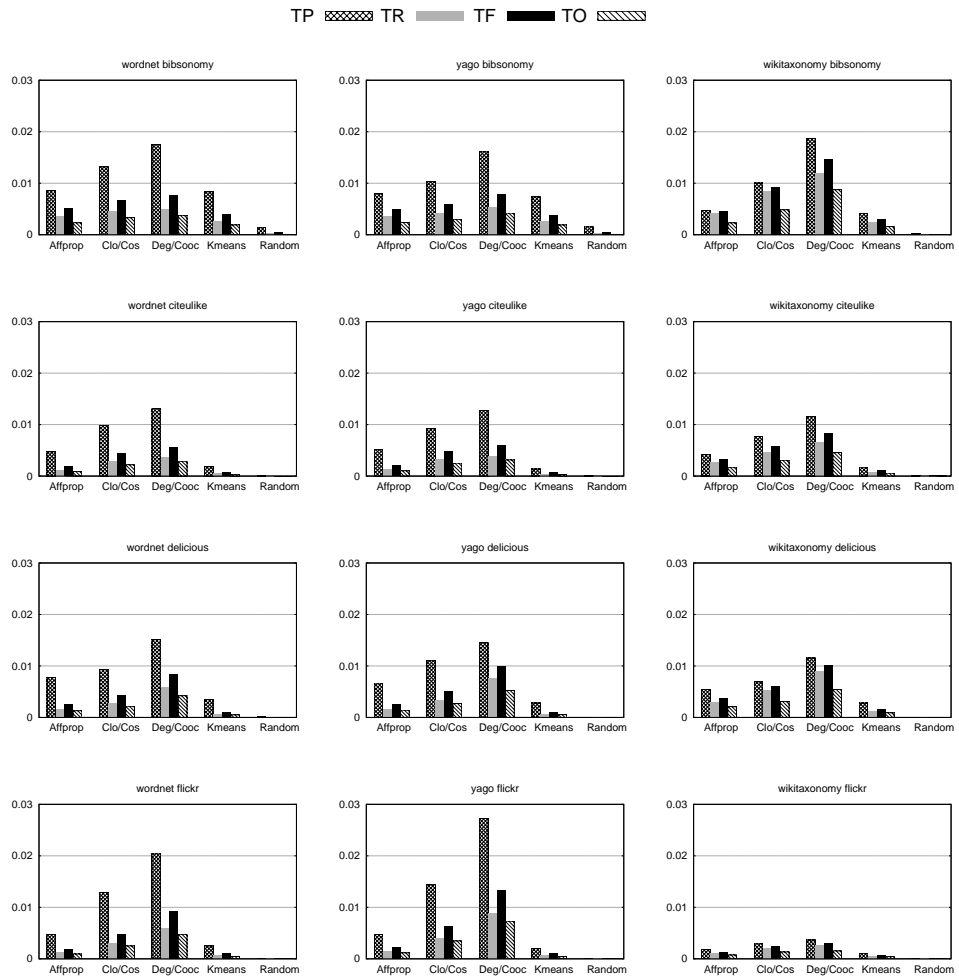


Figure 7.19.: Results of the reference-based semantic evaluation. Higher values on the y -axis depict stronger similarity between a learned hierarchy and the gold-standard, and hence a better performance.

question which remains is how to interpret the absolute similarity values, or in other words: Is, e. g., a score of 0.02 captured by the taxonomic F1-measure an indication of a “strong” similarity between the learned and the reference taxonomy? Due to the complexity and the size of the involved structures, it is

difficult to make a clear decision to this end. Because the values are averaged over the complete concept overlap, it is possible that some branches are very similar, while others are not. In order to facilitate a better understanding of the “true” quality of the learned hierarchies, we also performed a small-scale user study, which will be presented in the subsequent section.

7.4.4. Evaluation by Human Assessment

Although the human ability to interpret and integrate information in a meaningful way can surely be seen as superior to current automatic approaches, the task of evaluating the “quality” of a learned hierarchical structure remains challenging even for skilled subjects. Especially the manual comparison of two (potentially very large and complex) taxonomies will probably not lead to consistent and reproducible evaluation results. For this reason, we have chosen a simpler approach targeted towards the assessment of the consistency of each learned taxonomy. Our basic idea hereby was to sample a subset of all direct taxonomic subsumption pairs from a learned hierarchy, and then to let humans judge if (and if yes – how) the two contained terms are related. We used a web interface to present each human subject one term pair (A, B) at a time, asking “*What’s the relation between the two terms A and B?*”. As an answer, the participant could choose between selecting one of the following options:

- *A is the same as B.*
- *A is a kind of B.*
- *A is a part of B.*
- *A is somehow related to B.*
- *A is not related to B.*
- *I don’t know the meaning of A or B.*

In order to allow as many meaningful answers as possible from a broad audience, we performed an a-priori filtering of the term pairs by a list of “common” words, namely the 5 000 nouns which were used most often in the

Brown corpus¹². We only kept those pairs (A, B) as candidates for the study where both terms A and B were present in this list of popular nouns.

The intuition behind this approach is that a “better” taxonomy will yield a lower percentage of pairs being judged as unrelated. The reason why we allowed for a further distinction of relations (i. e., “same as”, “kind of”, “part of” and “somehow related”) is that we do not expect our analyzed algorithms to produce exclusively semantically sharp taxonomic (i. e., “kind of”) relations.

In order to come up with a concrete set of samples for our study, we first extracted all subsumption pairs containing “common” terms (as described before) present in each hierarchy induced from the Flickr dataset. We focused on this dataset because its scores in the reference-based evaluation were comparatively high. Furthermore, data from this system was used in related work on folksonomy induction before (Plangprasopchok et al., 2010), which allows for comparison with their results. From the resulting sets of candidate pairs, we randomly selected 25 pairs for each hierarchy induction algorithm under consideration, leading to 125 term pairs. As a control condition, we also added 25 term pairs randomly sampled from one of our reference hierarchies (namely the WordNet noun taxonomy), leading to a total number of 150 term pairs to be judged for each of our subjects. We then sent a link¹³ pointing to the online study to students and staff from two IT departments. In summary, 27 persons took part in the evaluation. Because some of them did not completely finish the rating of all pairs, we received 3 381 votes, including 249 “don’t know” choices – leading to a total of 3 132 useful answers for our study. In order to consider only pairs for which we have reliable assessments, we removed 22 pairs with very sparse voting data. This left us with a final set of 128 term pairs. For each term pair, we computed the fraction of each possible answer, and averaged these values subsequently over each hierarchy induction algorithm. Figure 7.20 shows the results.

The topmost five rows correspond to the algorithms used, while the lowermost row depicts the control condition based on the WordNet noun taxonomy. The values on the y -axis depict the average fraction of choices for each possible answer – as an example, among all assessments on subsumption pairs produced by affinity propagation, the average fraction of “part of” answers was roughly

¹²This corpus was compiled in 1960 and contains roughly 2 million words from a general set of English texts (see <http://khnt.aksis.uib.no/icame/manuals/brown/>)

¹³http://www.kde.cs.uni-kassel.de/benz/relations_and_cartoons.html

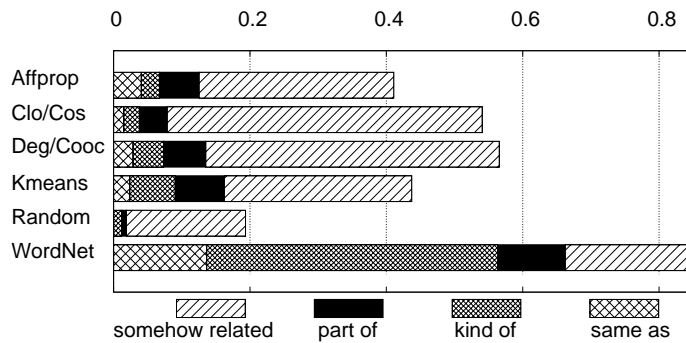


Figure 7.20.: Results of the semantic evaluation performed by a user study. The upper five horizontal bars correspond each to a folksonomy induced on the Flickr dataset by each algorithm under consideration; the lowest bar depicts a control condition based upon the WordNet noun taxonomy. The different patterns correspond to the average fraction of choices the human subjects have made when presented with a sample of subsumption pairs from each hierarchy.

5,8% (0.058, black part of the uppermost bar). Please note that only “positive” answers are included in this plot (i. e., answers stating that there is a meaningful relation among two terms). However, the percentage of “negative” answers (i. e., explicit statements by the users that two terms are not related) can be deduced from the figure by subtracting the sum of positive votes from 1. As an example, we received for affinity propagation in average a fraction of roughly 59% (0.59, topmost row) of “not related” answers for each pair. So as a shortened statement, one can say that the “longer” the bars are, the higher is the quality of the corresponding folksonomy.

To start with the lower and upper bounds, the folksonomy produced by the random algorithm performs worst - all “positive” relation assessments are adding up to roughly 0.2. On the contrary, the control assessments on the WordNet noun taxonomy sum up to ≈ 0.82 , including a large portion (≈ 0.42) of “kind of” answers. So as a first observation, we can say that the random folksonomy was judged to be the worst and the WordNet noun taxonomy was judged to be the best hierarchy – which confirms our intuition and validates our experimental methodology. In between these bounds, the sum of positive votes seems to confirm the impression from the reference-based evaluation: Again,

the two generality-based methods yield a higher percentage of positive votes compared to the two clustering approaches. Despite this fact, taking a more detailed look one can also see that the percentage of “kind of” and “part of” votes (which are semantically more sharp compared to “somehow related”) is highest for the k-means clustering algorithm. This could of course be an artifact of sampling, but could also point towards a greater semantic precision of the folksonomy induced by k-means clustering. However, taking a closer look at the “somehow related” pairs, it turns out that despite their lesser degree of semantic sharpness, the obtained relations can still be useful especially for organizational purposes of a category hierarchy (e. g., “pot / stove”). In light of this viewpoint, the results of the user study can be seen as a confirmation of the validity of the measures we used in our reference-based evaluation setting.

7.4.5. Enhancement by Synonym Resolution and Disambiguation

The alert reader will have noticed that the “concepts” within the concept hierarchies learned within the previously described methods corresponded directly to keywords used within the Social Annotation Systems. While this was done (as explained earlier) intentionally in order to allow an unbiased comparison of the hierarchy induction algorithms, the question remains if the methods of synonym resolution and sense disambiguation (i. e., concept learning) can further enhance the quality of the learned taxonomies. Because the generality-based approaches led to the best results, we will focus on those, and present an enhanced variant of the algorithm proposed by (Heymann and Garcia-Molina, 2006). For the sake of brevity, we omit a complete coverage of all datasets used within this dissertation, but focus again on Delicious as a representative for a large and broad folksonomy. However, we will not use the exact dataset as described in Section 6.1.3, but a slightly smaller crawl performed in July 2005. Originally, it contained $|U| = 75\,260$ users, $|T| = 533\,191$ tags, and $|R| = 3\,151\,353$ resources, related by $|Y| = 17\,364\,552$ triples. Because “singletons” (i. e., nodes without any connection) in the tag-tag *co-occurrence* graph are not relevant for our purpose, we removed all tags used only once by a single user; this left us with a dataset of $|U| = 74\,680$ users, $|T| = 373\,690$ tags, and $|R| = 2\,972\,695$ resources, related by $|Y| = 17\,181\,896$ triples.

Based on this data, we performed a very careful synonym resolution by the baseline approach described in Section 7.2.1. Hereby we applied a minimum similarity threshold of $\tau_{sim} = 0.96$. The main reason for choosing the baseline

approach instead of the synonym clustering methods is that the latter were developed after the collaboration, during which the study presented in this section was performed. Apart from that, the choice of a rather high threshold can be expected to lead to the inclusion of only few, but rather “correct” synonym sets. Based on this setting, the resulting synsetized folksonomy (recall the Definition 7.1) contained $|S| = 373\,572$ synsets and $|Y^s| = 17\,154\,948$ synsetized tag assignments.

In a next step, we applied tag sense discovery by the hierarchical clustering method described in Section 7.2.3, using a context size of $k = 10$ tags. As a distance criterion threshold, a value of $\tau_{dist} = 0.55$ led to good results. In order to exploit the information about ambiguity, we adapted the original generality-based algorithm by (Heymann and Garcia-Molina, 2006) in such a way that it considers each sense of a keyword separately. More precisely, this adaptation led to the following algorithm (differences to the original version are highlighted in italics):

1. Create a list of all keywords, ordered in descending order by their degree of generality (according to the chosen generality measure).
2. Create an empty tree, and add the most general (i. e., topmost) keyword from this list as root node.
3. Iterate through the remaining keywords in order of descending generality and add each keyword t to the tree according to:
 - a) Identify the most similar existing tag within the hierarchy t_{sim} to t .
 - b) *If the generality score of t is above a given threshold τ_{gen} , or if the t_{sim} is not sufficiently similar according to a threshold τ_{sim} , append t as a child node to the root node.*
 - c) *If t_{sim} is an ambiguous tag, identify the correct sense of t_{sim} in the context of the t by taking into account the preference tags of t_{sim} and t*
 - d) Append t_i as a less general term underneath *the correct sense of t_{sim} .*
 - e) *If t is an ambiguous tag, repeat steps 3.a - 3.e for each of its senses.*

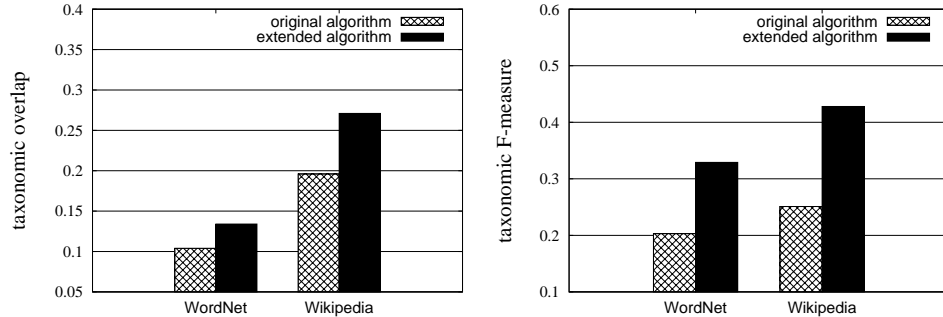


Figure 7.21.: Results of the reference-based evaluation of enhanced concept hierarchy learning based on Delicious data. The learned structures are compared against two gold-standard hierarchies from WordNet and Wikipedia.

4. Apply a post-processing to the resulting tree by re-inserting orphaned tags underneath the root node in order to create a balanced representation. The re-insertion is done based on steps 3.a-3.c.

So in summary, the main enhancements are the handling of ambiguous tags and the post-processing step. In order to assess whether these modifications lead to a “better” learned taxonomy, we computed a hierarchy based on the original and the extended version of the algorithm. To ensure comparability, we used in both cases co-occurrence as tag similarity measure and degree-centrality in the co-occurrence graph as generality measures. This corresponds to the *Deg/Cooc* from the previous chapter, which seemed to be the best performing variant. For the adjustment of the algorithm parameters τ_{sim} and τ_{gen} , we performed a variation procedure and kept those settings which maximized the similarity to the reference taxonomies, namely those from WordNet (see Section 6.2.1) and the Wikipedia category hierarchy (Section 6.2.4). Figure 7.21 summarizes the results obtained for the best parameter settings.

For both reference ontologies, our proposed extended algorithm leads to taxonomies which resemble more closely to the gold-standard. While the increase is smaller when using taxonomic overlap, it is still clearly visible. To come back to our initial question, the results point in the direction that synonym resolution and tag sense discovery are in fact suitable to further

enhance the quality of concept hierarchy learning. While this is intuitively clear, we provided in this section empirical evidence for this assumption by reaching a higher structural agreement with gold-standard taxonomies.

7.4.6. Summary

After laying the groundwork by analyzing measures of semantic relatedness and generality, as well as concept learning methods based on Social Annotation, this chapter was intended to examine algorithms which induce hierarchical relationships among the initially flat keyword space. The main motivation hereby was to analyze standard clustering methods as well as custom generality-based procedures which exploit similarity and generality information for establishing a concept hierarchy. Based on a gold-standard based evaluation setting, it turned out that *k-means* and *affinity propagation* clustering approaches produced hierarchies which were less similar to several reference taxonomies compared to *generality-based* algorithms specifically tailored for Social Annotation data. This assumption was confirmed by a user study, which involved human assessments about the consistency of the learned subsumption relations. Beside that, further results seemed to be that tag co-occurrence as a similarity measure together with degree centrality based on the co-occurrence graph have advantages compared to closeness centrality and distributional similarity measures. These findings are in line with the prior analysis in Section 7.1.3, where co-occurrence relatedness exhibited a bias towards semantically more general tags. As a last step, the beneficial influence of “preprocessing” the raw Social Annotation data by synonym grouping and keyword disambiguation was shown in an exemplary case study.

Despite these results are promising, the same limitations for the concept learning methods in general apply here as well: It can be probably not be expected that the resulting taxonomies exhibit in their current state the same semantic precision as manually built ones. However, they can surely be helpful to address the problems related to the lack of structure within Social Annotation Systems by providing additional navigation facilities, or query refinement / broadening possibilities. Furthermore, it can be expected that they can serve as useful input for further refinement steps, targeted towards, e. g., suggesting potential new relations for more controlled semantic repositories.

7.5. Learning Attributes, Relations and Axioms

As already mentioned in the earlier literature review (Section 5.3.6), approaches which address higher levels in the ontology layer learning cake are still scarcely found. Hence most researchers “stop” at the level of concept hierarchies as the most complex structure to be learned. Though there exists rich networks among users, tags and resources within Social Annotation Systems, the application of methods stemming from, e. g., learning attributes or relations from text often require further syntactical or lexical structure. Among the common approaches in this field mentioned by (Cimiano, 2006), in fact only *collocations* have a counterpart in Social Annotation data, namely in the co-occurrence of keywords. However, it is further mentioned that “*A collocation [...] typically reveals a strong but unknown relation between words.*”. For this reason, we will also refrain from tackling further layers of the ontology learning layer cake within the scope of this dissertation.

In order to provide some potential methodological starting points for further research in this direction, the interested reader is referred to Section 7.1. By using an approach of semantic grounding, a deeper insight into the *semantic* properties of different folksonomy-based relations was obtained. The “target” used for grounding was hereby the taxonomic relation of WordNet. Despite – as said before – the definition of measures will be more difficult, this procedure is in principle viable for other kinds of externally defined relations as well. As an example, the same setup could be used to assess, e. g., if there exists a folksonomy-based relation which captures information similar to meronymy (i. e., part-of relationship) within WordNet. A cursory analysis of the learned hierarchies from the previous chapter also revealed that such relations are partially already present, giving the learned taxonomies a somewhat “mixed” semantics. It may be another good starting point to try to “disambiguate” these different kinds of relations, eventually by taking into account background knowledge from an external source.

With these pointers, the methodological chapter of making emergent semantics within Social Annotation Systems is closing. While within this chapter, several “instruments” to capture emergent structures were introduced, the next chapter changes perspective and focuses on analyzing factors which play a crucial role in the process of emergence itself.

Chapter 8.

Influencing Factors

When trying to make the implicit semantics within Social Annotation Systems explicit, all methods and approaches presented in the previous chapter were mainly addressing the “final” state of these systems, up to the time when the snapshot for the dataset was taken. In addition, besides some a-priori filtering based on keyword frequency (especially in order to counteract sparse vectors when computing context relatedness), all available data was taken into account in an unmodified manner. This was done in order to enable a “clear” view on the kind of structures which these systems produce, and in order to check how well this “pure” data serves as an input to algorithms which discover semantics.

In this chapter, we will shift perspective and focus on the processes and aspects which influence the emergence of implicit semantic structures. A deeper understanding of the factors which have a (desirably positive) effect hereby would be highly desirable for at least two purposes: (i) Operators of Social Annotation Systems interested in fostering emergent semantics would be enabled to adapt the design of their systems (e. g., by optimized interfaces) in order to stimulate the beneficial processes. Furthermore, (ii) analysts of Social Annotation data would be able to perform intelligent data filtering, retaining, e. g., only a subset which exhibits potentially “more precise” semantics.

In order to address this issue, we will first describe which methodology we applied in order to “measure” the effects of different factors regarding the evolution of emergent semantics. Then, we will focus on three major aspects – namely keyword properties, tagging pragmatics and spam – and analyze their influence on the emergence process. In summary, these studies are intended to complement the methods presented in the previous chapter by getting a deeper understanding of when they produce especially good results.

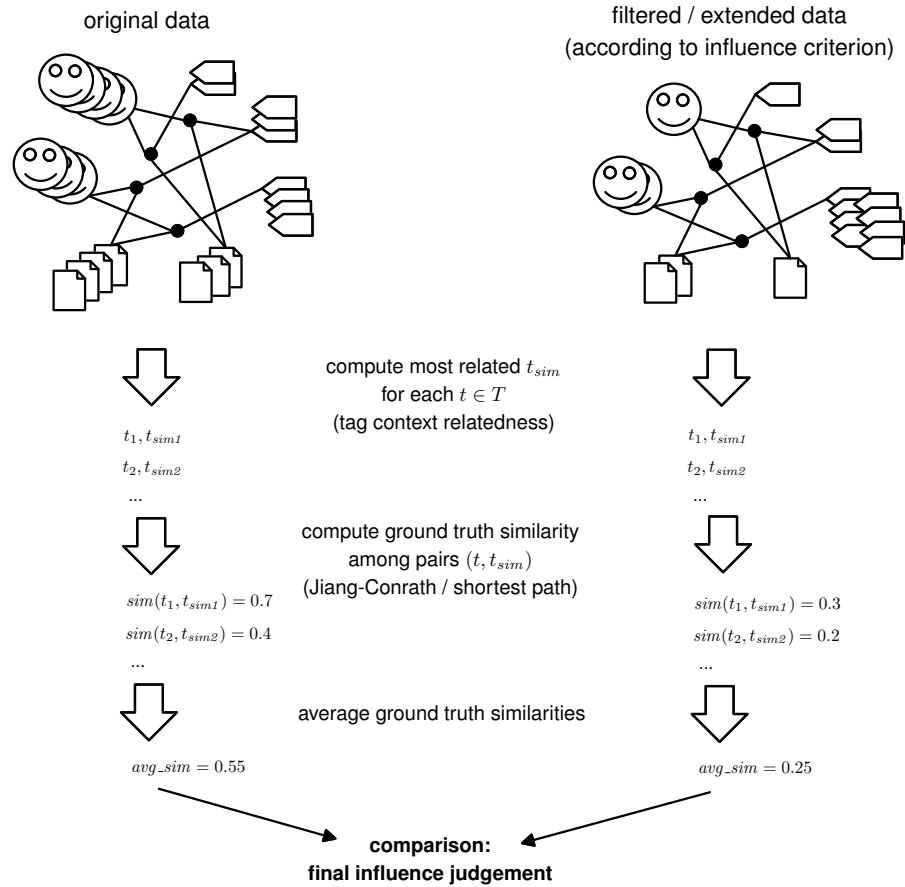


Figure 8.1.: Methodology of influence assessment concerning factors of emergent semantics. The similarity values are for exemplary purposes only. See Section 8.1 for a detailed explanation.

8.1. Methodology of Influence Assessment

The first crucial question when interested in the factors which influence the emergence of semantics is how to quantify their impact. The goal hereby is to assess that certain influences lead to “better” or “worse” semantics. Because a manual assessment is hereby hardly feasible, we stick to the gold-standard based evaluation paradigm used so far in this dissertation and validate the

different outcomes against “reference semantics”, defined in (semi-)manually built ontologies. More precisely, because many of the methods presented in Chapter 7 (like, e. g., synonym resolution or tag sense discovery) benefit from a precise notion of semantic relatedness among keywords, we will use the *tag context relatedness* (*TagCont*, see Section 7.1.1) as our main indicator dimension. The intuition hereby is as follows: If a particular factor (e. g., the keyword frequency, as will be analyzed in Section 8.2.1) has a *beneficial* effect on the *semantic* quality of the relations obtained by *TagCont*, then this has a positive influence on the overall quality of semantic structures produced by our methods.

In order to operationalize this idea, we will focus on the *most related* keyword, according to the tag context relatedness. We will then perform its semantic grounding (as already done in Section 7.1.3), and measure the “true” semantic relatedness among the current keyword pair, using established WordNet-based metrics. Depending on the context, we will employ the *Jiang-conrath distance* or the *shorttext taxonomic path length* (as introduced also in Section 7.1.3) for this purpose. This semantically grounded relatedness will then be used as our final indicator of the “quality” of the emergent implicit structures. While we consider these values individually in some cases, in others we will use their average as the final criterion. Because many factors can be analyzed best when performing a *filtering* or an *extension*, we will often perform the final comparison between the “original” data, and another “filtered” (or extended) dataset according to a the given criterion. Figure 8.1 depicts graphically this chosen approach.

8.2. Keyword Properties

Because within the approach of this dissertation, keywords and their interconnections within Social Annotation Systems are regarded as the main “medium” of semantics, it is natural to start with the question if there exist certain keyword properties for which the process of harvesting semantics works especially well. Especially having observed that the tag context relatedness yields semantically very close keywords in some cases (e. g., synonyms), while more widely related ones in others, it is worth investigating why this could be the case. Hereby we will focus on two aspects: First, because we get intuitively the more information about a particular keyword the more often it has been used, we will check if the usage *frequency* has an impact. An obvious hypothesis hereby is that a higher frequency will yield to more precise semantics, simply because we have a more

“informed” relatedness measure in such case.

Second, because the choice of keywords is also often strongly influenced by technical or visual properties of the system at hand, we will analyze *interface characteristics* and their influence on the resulting vocabulary. Specifically, we will check if the removal of certain artifacts (e.g., system tags or erroneous delimiters) has a beneficial effect. In all cases, we will stick to the methodology of influence assessment described in the previous section.

8.2.1. Frequency

An implicit assumption when talking about “emergent semantics” is that the semantic structures emerge over time, as more and more information is added to the system. Because the methods to capture these structures are hence mostly data-driven, an justified hypothesis is that the semantics of a particular keyword can be captured better when there is more data about it, i.e., when it has been used more often. The alert reader will have noticed that this hypothesis was also the basis for restricting the datasets in the previous chapter (especially Section 7.1) to popular (mostly the top 10 000) keywords: While this decision was motivated by the idea to avoid sparse context vector representations, we will now assess the effect of frequency in a more systematic manner.

Because we can quantify for each keyword (i) its frequency and (ii) the “true” semantic similarity to its most related “partner” according to the tag context relatedness (see Figure 8.1), a first natural thing is to check for a correlation among both. Figure 8.2 depicts a plot, where the frequency is found on the y -axis by means of the “tag rank” (the most frequent tag is assigned rank 1, the second one rank 2, and so on), and the x -axis shows the semantic quality of the most related keyword. Because one can obviously expect a strong variation of the values, the distributions are smoothed by Bézier curves. Apart from the BibSonomy dataset, there is an evident tendency that more often used keywords (i.e., low-rank ones, found on the left) are assigned to a semantically more similar partner than less often used ones (right side of the figure). This is especially visible for the Delicious dataset, which exhibits an additional peak roughly among the 200 most popular keywords. This points in the direction that our initial hypothesis was correct, i.e., that a more frequent usage provides substantially more information which can be exploited by the tag context relatedness. The impression that the BibSonomy dataset seems to follow a slightly different pattern could also be explained herewith, because

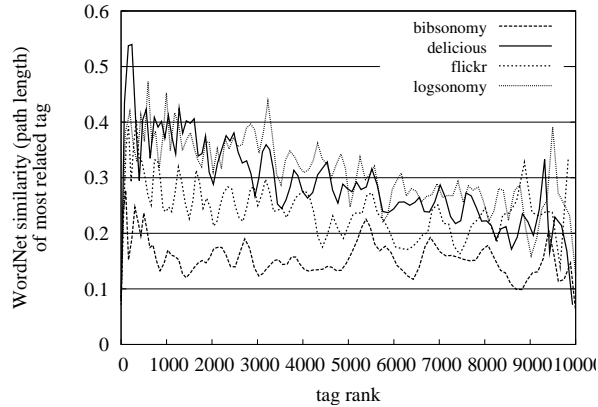


Figure 8.2.: Influence of tag frequency on emergent semantics. The x -axis depicts the tag rank (whereby 1 corresponds to the most frequently used tag), and the y -axis shows the path length similarity (measured in WordNet) between each tag and its most related one according to the tag context relatedness. The results are smoothed using a Bézier curve.

due to its comparatively small size (see Section 6.1.1), there is still quite little data for the more popular keywords. The remaining datasets (i. e., CiteULike and Stackoverflow) are left out for readability reasons, but exhibited a similar behavior.

In order to confirm these visual impressions, we computed Kendall's τ as a rank correlation coefficient; the two rankings of keywords were induced (i) by the tag rank and (ii) the semantic similarity to its most closely related keyword. Based on that, Kendall's τ is defined as:

$$\tau = \frac{|C| - |D|}{\frac{1}{2}n(n-1)}$$

Hereby C is the set of *concordant* pairs (i. e., those where both rankings agree), and D the set of *discordant* pairs (i. e., those where they disagree). Its value ranges from 1 (perfect positive correlation) to -1 (perfect negative correlation). We found for BibSonomy and CiteULike a value of ≈ -0.05 , and for Delicious, Flickr, AOL logsonomy and Stackoverflow a value of ≈ -0.1 . This confirms the

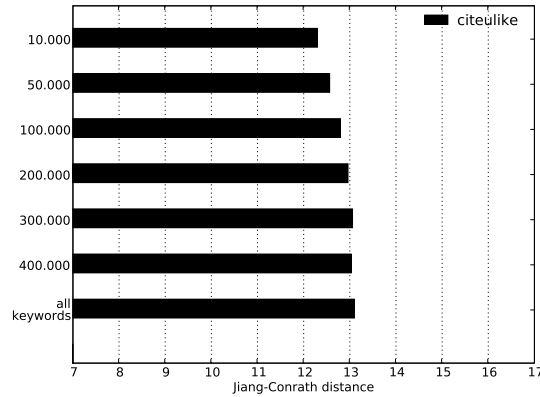


Figure 8.3.: Influence of restrictions to different fractions of popular keywords (CiteULike dataset). The x -axis measures semantic quality by means of the Jiang-Conrath distance, the y -axis depicts the number of popular keywords each sub-dataset was derived from.

slight, but recognizable negative correlation, which seems to be stronger for larger datasets.

In order to better investigate this point (and in order to justify the restriction to the 10 000 most popular tags often applied in Chapter 7), we extended the analysis presented in Section 7.1.3 on the CiteULike dataset to less frequently used tags. While we originally used the sub-folksonomy induced by the 10 000 most frequent keywords, we varied the threshold and used the ones induced by the 50 000, 100 000, 200 000, 300 000, 400 000 most frequent keywords, as well as the complete dataset (549 145 keywords). Based on each condition, we computed the average semantic similarity to the most closely related tags, as explained in the previous section. Figure 8.3 shows the results. The y -axis depicts the number of popular keywords, which induced each dataset; the x -axis shows in the customary manner the semantic quality (shorter bars indicate higher quality, as we are using the Jiang-Conrath distance). In line with the prior observations, the overall quality of semantics decreases when taking into account less frequent keywords. Though being intuitive, these results confirm in an empirical way our assumption that it is difficult to assess the semantics of the “long tail” of seldom used keywords, because their inherent sparsity

hampers a reliable assessment.

To summarize, our results suggest that frequent usage facilitates the harvesting of keyword relationships with a higher degree of semantic “precision”. This implies that, e. g., that operators of Social Annotation Systems should focus on enhancing user activity, and analysts of semantics within tagging data could benefit from restricting their studies to the denser parts of the folksonomy. As another keyword-related aspect, we will now turn our attention to the characteristics of the tagging interfaces.

8.2.2. Interface characteristics

Because they provide the direct interaction points with the users, the design and characteristics of the *interfaces* by which the users submit their annotations can also be expected to be relevant. Typically there exists several of those, e. g., a “common” website where users fill out forms, or a programming interface like a REST-API, which allows local client applications to interact directly with the system. In addition, many systems allow a batch import of resources in various formats. When compiling the datasets for this dissertation, especially two artifacts introduced by these variants became visible. Among the popular keywords of the CiteULike dataset, many “import-related” tags were found, like *file-import*, *file-import-10-02-08*, *jabref-import* and many others. Apart from their minor importance for, e. g., learning concepts from tags, the question arises if the batch processing (probably widely without human interaction) could even have *detrimental* effects on the global semantics. As an example, the tag context relatedness is based on the 10 000 most popular keywords as vector dimensions, which are intended to capture the semantic context of a keyword. If among those, there are many “import-related” tags without any meaningful relevance to the other keywords, it could happen that the vector representations get blurred.

The next phenomenon is related to our own BibSonomy system. Because we are using the whitespace character as keyword delimiter, it turned out that quite an number of tags could be found which ended in a comma – pointing towards users erroneously thinking that the latter can be used as a delimiter. Because tags have been treated in an unfiltered way so far, this would lead to, e. g., *java* and *java,* being handled as separate terms. Apart from that, we could also find the typical variations of using multi-word tags, like e. g., *ontology-learning*, *ontology_learning* or *ontology.learning*. In all those

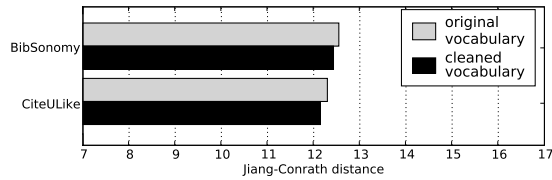


Figure 8.4.: Influence of vocabulary cleaning on emergent semantics. For the BibSonomy dataset, all non-alphanumeric characters were removed from each keyword; for the CiteULike case, all “system tags” related to imports from different systems were removed.

cases, it can be assumed that a careful “normalization” of tags might have beneficial effects.

In order to assess both issues, we cleaned the vocabulary of both the CiteULike and the BibSonomy dataset. For the CiteULike case, we manually removed all “import-related” keywords, leaving us with 9 431 instead of 10 000 tags. Within the BibSonomy case, we applied a comparatively thorough tag normalization: First, we replace all dashes (–) by underscores (_), and removed subsequently all non-alphanumeric characters (except underscore). This reduced the vocabulary size from 10 000 to 9 311 keywords. Based on both cleaned datasets, we computed similar tags as done before.

Figure 8.4 compares the semantic precision of the original datasets to their cleaned alternatives. Evidently, both led to a very small improvement. Despite the effect is not strong, these results suggest that within Social Annotation data, there may exist fractions of interface-related artifacts ($\approx 6.5\%$ of keywords in our case) which should be considered and can – at least – be safely disregarded from a further semantic analysis.

8.2.3. Summary

In this first step of assessing influences of emergent semantics, we focused on keyword-related aspects. By checking the correlation between the usage frequency and the semantic precision of related keywords, a first observation was that a high usage frequency has in general beneficial effects on the resulting semantics. This was attributed mainly to the corresponding higher density of the

co-occurrence network, which encodes a larger amount of semantic information. Subsequently, properties of the tagging interfaces themselves were analyzed, namely artifacts introduced by batch-processing as well as particular system design choices like the keyword delimiter. The analysis here suggested that such artifacts can safely be removed, without negatively affecting the inherent semantic structures.

As a next step, we will move away from the keywords themselves as objects of investigation, but focus rather on users and *pragmatic* aspects, i. e., how and why Social Annotation System are used.

8.3. Tagging Pragmatics

As introduced in Section 3.1.6, the analysis of *pragmatic* aspects of tagging (i. e., how and why users tag) has gained considerable attention in the research community. Especially the distinction proposed by (Körner et al., 2010) of users into *categorizers* (who follow an “ontology-like” style of tagging) and *describers* (who are characterized by a more verbose and descriptive behavior) intuitively suggests an influence on the quality of the resulting semantic structures. In other words, the question arises if those different tagging practices and motivations affect the processes that yield emergent semantics. This would mean that in order to assess the usefulness of methods for harvesting semantics from folksonomies, we would need to know whether these methods produce similar results across different user populations characterized by different tagging practices and driven by different motivations for tagging. Given these implications, it is interesting to explore *whether and how emergent semantics of tags are influenced by the pragmatics of tagging*.

For a general introduction into the characteristics of categorizers and describers, refer to Section 3.1.6. A prerequisite to study their influence is the ability to assign users to either of the two types. Because a “direct” assessment would require an interaction with the users and is hence difficult, we will first present a set of quantitative measures which indicate – for each user – the degree of membership in both classes. Based on these metrics, our approach is then to systematically induce “sub-folksonomies”, which are comprised of varying proportions of categorizers and describers. As an indicator of the quality of the emergent semantic structures, we will stick to the approach described in Section 8.1 and compute tag relatedness on each of these subsets. Based on

the results, our last step is to discuss implications for methods which target towards harvesting emergent semantics from Social Annotation Systems.

8.3.1. Measures of Tagging Pragmatics

As stated above, because the motivation behind tagging is difficult to measure without direct interaction with users, we are defining in the following surrogate measures which capture different pragmatic aspects of tagging. Since these measures correspond to different intuitions how the different behavior of categorizers and describers could become apparent, we will use them later to approximate a user's membership in either of the two classes.

Vocabulary size: The first measure is based on the intuition that *describing* a resource in a detailed manner will typically require a larger amount of keywords than categorizing it according to its membership in a few classes. Hence, the *vocabulary size*, i. e., the number of distinct keywords used by a user, is our first proposed measure:

$$vocab(u) = |T_u| \tag{8.1}$$

Describers would likely produce an open set of tags with an unlimited and dynamic tag vocabulary while categorizers would try to keep their vocabulary limited and would need far fewer tags.

Tag/resource ratio (trr): While the plain size of the vocabulary provides a first coarse estimation, an inherent limitation is that it does not consider the number of resources annotated by the vocabulary. If a user has used, e. g., a large number of keywords to annotate just a few resources, but uses very few keywords to annotate the majority of them, then having a look at the vocabulary size alone might be misleading. Because we expect describers to introduce more and more keywords as they annotate further resources, a natural adaptation is to relate the vocabulary size with the total number of annotated resources according to:

$$trr(u) = \frac{|T_u|}{|R_u|} \tag{8.2}$$

Taggers who use lots of different tags for their resources would score higher

values for this measure than users that use fewer tags. Due to the limited vocabulary, a categorizer would likely achieve a lower score on this measure than a describer who employs a theoretically unlimited vocabulary. Equation 8.2 shows the formula used for this calculation where R_u represents the resources which were annotated by user u .

Average tags per post (tpp): Although the tag/resource ratio encompasses further information, its scores do not necessarily reflect that a user uses *consistently* more keywords for annotation. In order to better capture this aspect, one would have to take a look at each individual post, and average the number of applied keywords. This is exactly what our next proposed measure is doing:

$$tpp(u) = \frac{\sum_{r \in R_u} |T_{ur}|}{|R_u|} \quad (8.3)$$

Taggers who usually apply lots of tags to their resources get higher scores by this measure than users who use few tags during the annotation process. Describers would score high values for this measure because of their need for detailed and verbose tagging. In contrast, categorizers would score lower values because they try to annotate their resources in an efficient way.

Orphan ratio: Our final measure builds upon a different intuition. Hereby we have a look at “orphaned” tags, i. e., tags which are assigned to very few resources. Because of the expected “verbose” annotation style of describers, we hypothesize that we will find a higher percentage of such orphaned tags within their vocabulary, because they might – to use a pointed formulation – annotate in a somewhat “tag-and-forget” manner. This is not a problem, as their goal is not to establish a small and consistent vocabulary; however, because the latter is the explicit target of categorizers, we expect to find fewer orphaned tags within their vocabularies. The last measure is defined as:

$$orphan(u) = \frac{|T_u^o|}{|T_u|}, T_u^o = \{t : |R_{ut}| \leq n\}, n = \left\lceil \frac{|R_{ut_u^{max}}|}{100} \right\rceil \quad (8.4)$$

The *orphan ratio* thus captures the percentage of items in a user’s vocabulary that represent orphaned tags. T_u^o denotes the set of orphaned tags in a user’s tag vocabulary T_u (based on a threshold n). The threshold n is derived from each user’s individual tagging style in which t_u^{max} denotes the tag that was used

the most. R_{ut} denotes the set of resources which are tagged with tag t by user u . The orphan ratio values range from 0 to 1 where a value of 1 identifies users with lots of orphaned tags and 0 identifies users who maintain a more consistent vocabulary. Considering the categorizer / describer paradigm this would mean that categorizers tend more towards values of 0 because orphaned tags would introduce noise to their personal taxonomy. For a describer's tag vocabulary, this measure would produce values closer to 1 because describers tag resources in a verbose and descriptive way, and do not mind the introduction of orphaned tags to their vocabulary.

Properties of the measures While these measures of tagging pragmatics were inspired by the dichotomy between categorizers and describers, we do not require them to accurately capture this distinction. Another aspect is that these measures might not only capture intrinsic user characteristics, but can also be influenced by, e. g., elements of user interfaces (such as recommenders). What is important in the light of our hypothesis is that all of *the above measures are independent of semantics* – they capture *usage patterns* of tagging (the pragmatics of tagging) only. This allows us to explore a potential link between tagging pragmatics and the emergent semantics of tags.

8.3.2. Influence Assessment

In the analysis of tag relatedness measures in Section 7.1, an important finding was that the definition of an adequate *context* plays a crucial role for the task of capturing emergent (tag) semantics. However, given the massive amounts of data available in social tagging systems, the question is not only to identify a *valid* context, but also to identify the *minimal* context which retains the relevant structures while allowing for efficient computation. As human annotators are the creators of implicit semantic structures, an important aspect hereby is which users should be included in an optimal context composition. Following our discussion in the previous section, our hypothesis is that individual tagging pragmatics can play an important role for selecting “productive” users. The question is whether the categorizers – who follow the ontology engineering principle of a clean vocabulary – or the describers – who provide more descriptions to their resources – are the more “productive” ones.

In order to answer this question, our strategy is to analyze the suitability of each of our previously introduced pragmatic measures to assemble a (pref-

Table 8.1.: Statistic for the Delicious dataset variants used for the influence assessment of tagging pragmatics.

<i>dataset</i>	$ T $	$ U $	$ R $	$ Y $
full	10 000	511 348	14 567 465	117 319 016
min100res	9 944	100 363	12 125 476	96 298 409

entially small) subset of users which provides a sufficient context to harvest emergent tag semantics. The general idea hereby is to start at both ends of the scale with the “extreme” categorizers and describers, and then to subsequently add more users (in the order given by the respective measure). In each step, we check how well the folksonomy partition defined by the current user subset serves as a basis to compute semantically related tags. For the latter, we revert to the tag context relatedness measure (TagCont) that has shown to produce valid results (cf. Sec. 7.1.1). The assumption hereby is that the TagCont measure will yield more closely related tags when better implicit semantic structures are present. Hence, the whole procedure allows us to assess the quality of the emergent semantics and finally the degree to which it was influenced by tagging pragmatics.

Experiments

The intention of our experiments is to quantify the influence of individual tagging practices on emergent tag semantics in a folksonomy. We will first provide details on data preprocessing and then explain each experimentation step before discussing the results.

Data preprocessing In order to validate our hypothesis on real-world data, we used the Delicious dataset described in Section 6.1.3. More precisely, because we are relying heavily on the computation of meaningful distributional similarity measures, we stick to our dataset induced by the 10 000 most frequent tags. We will refer to the resulting folksonomy as the *full* dataset (see Table 8.1). In order to eliminate noise introduced by our measures misjudging new users, we furthermore removed all users having less than 100 resources in their collection. The reason behind this is that, e. g., the tag/resource ratio is not very informative in the case of a new user with very few resources. Interestingly, our results show that removing this “long tail” of new (or inactive) users already increases

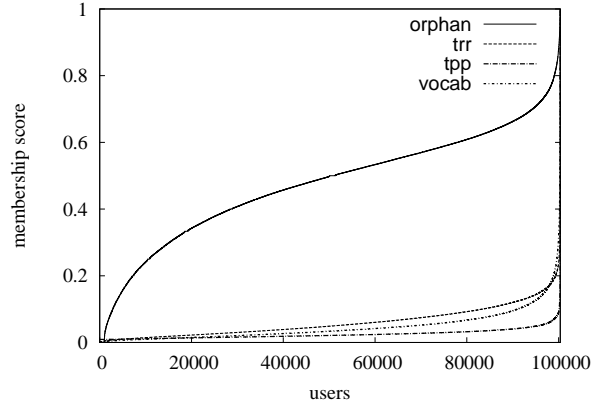


Figure 8.5.: Distribution of the membership scores for each introduced measure of tagging motivation (orphan ratio, tag/resource ratio, tags per post and vocabulary size), computed for the 100 393 users present in our Delicious dataset (x -axis). Values close to 0 on the y -axis indicate strong categorizers, while values close to one 1 point to describers. All measures were normalized to the interval $[0, 1]$.

the quality of the learned semantic relations. Details of this observation will be discussed in Section 8.3.3. We will denote the resulting dataset as *min100res* (see Table 8.1).

Experimental setup In order to assess the capability of each of our measures to predict “productive” users, we followed an incremental approach: For each of our measures $m \in \{orphan, vocab, trr, tpp\}$, we first created 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, while giving high scores to describers. This means that, e. g., the first user in the orphan ratio list (denoted as $L_{orphan}[1]$) is assumed to be the most extreme categorizer, while the last one ($L_{orphan}[k], k = |U|$) is assumed to be the most extreme describer. Figure 8.5 depicts the obtained distribution of membership scores for each ordered list L_{tpp} , L_{trr} , L_{orphan} and L_{vocab} . An observation which can be made in this figure is that the distribution of the *orphan* measure differs clearly from the other three measures. This implies that the orphan ratio seems to be able to make a more fine-grained distinction between users. However, our results did not exhibit a positive impact

on the resulting semantics; rather contrary, the orphan ratio performs often worse than the other measures (see below for details).

Because we are interested in the minimum amount of users needed to provide a valid context, we start at both ends of L and extract two folksonomy partitions CF_1^m and DF_1^m based on 1% of the “strongest” categorizers ($Cat_1^m = \{L_m[i] \mid i \leq 0.01 \cdot |U|\}$) and describers ($Desc_1^m = \{L_m[i] \mid i \geq 0.99 \cdot |U|\}$). $CF_1^m = (CU_1^m, CT_1^m, CR_1^m, CY_1^m)$ is then the sub-folksonomy of F induced by Cat_1^m , i. e., it is obtained by $CU_1^m := Cat_1^m$, $CY_1^m := \{(u, t, r) \in Y \mid u \in Cat_1^m\}$, $CT_1^m := \pi_2(CY_1^m)$, and $CR_1^m := \pi_3(CY_1^m)$. The sub-folksonomy DF_1^m is determined analogously.

As a next step, we took the first extracted partition CF_1^m as input to extract semantic tag relations, in the way described in Section 7.1. We check whether the data produced by a very small subset of “extreme” categorizers already suffices to compute meaningful semantic relations. More specifically, for each tag $t \in CT_1^m$, we computed its most similar tag t_{sim} according to the tag context relatedness defined in Section 7.1.1. We then looked up each resulting pair (t, t_{sim}) in WordNet and measured – whenever both t and t_{sim} were present – the Jiang-Conrath distance $JCN(t, t_{sim})$ between both words (see Section 7.1.3). After that we took the average JCN distance of all mapped tag pairs as an indicator of the quality of emergent semantic structures contained in CF_1^m :

$$JCN_{avg}(CF_1^m) = \frac{\sum_{t \in CT_1^m} JCN(t, t_{sim})}{wn_pairs(CT_1^m)}$$

Here, $wn_pairs(DT_1^m)$ denotes the number of tag pairs (t, t_{sim}) (i. e., a tag and its most similar tag) for which both t and t_{sim} are present in WordNet. The corresponding describer partition DF_1^m was processed in the same manner.

As discussed in Section 7.1.3, we use the Jiang-Conrath distance as an indicator of the “true” semantic relatedness between tags. However, in order to avoid the dependency of our results on a single measure of semantic similarity, we also measured the *taxonomic path length* for each mapped tag pair (t, t_{sim}) between the two synsets s_1 and s_2 containing t and t_{sim} , respectively.¹ This measure counts the number of nodes in the WordNet subsumption hierarchy along the shortest path between s_1 and s_2 . We noticed that the outcomes of both measures (JCN and taxonomic path length) was almost perfectly correlated

¹If t and t_{sim} were present in more than one synset, we took the shortest possible path.

throughout our experimentation; for this reason, we will stick to the JCN distance in the remainder of this paper, because it has been shown to be a better surrogate for the human perception.

For or each of our measures $m \in \{orphan, vocab, trr, tpp\}$, we repeated this overall procedure, using the following percentages i :

$$i \in \{1, 2, 3, \dots, 24, 25, 30, 40, 50, 60, 70, 80, 90\}$$

As we keep adding users while incrementing i , it is important to notice that the size of the resulting “sub-folksonomy” is growing towards the size of the full dataset, i. e., $DF_{100}^m = CF_{100}^m = F$. Another important aspect is the fact that users are added in descending order of their membership degree in the respective user class: This means that CF_1^m contains users u who score high on measure m , while, e. g., CF_{50}^m contains a more mixed population. “Mixed” in this context means that there exist users in CF_{50}^m which are to a certain degree assumed to exhibit describer characteristics as measured by m . This implies that the distinction between both user groups is blurred while incrementing i . In other words, one can also read these partitions from the other side, namely that CF_{90}^m contains all users *except* 10% of the most extreme describers.

In summary, we created 64 partitions for each of our 4 measures (32 categorizer + 32 describer), summing up to a total of 256 sub-folksonomies, each being extracted by a different composition of users according to their tagging characteristics. Before presenting our results on the most suitable partitions for extracting semantic tag relations, we discuss upper and lower bounds. As we measured the quality of an extracted relation between two tags t and t_{sim} by its Jiang-Conrath distance within WordNet, a lower bound can be identified by computing the pairwise JCN distance between all tags $t \in T$ and averaging over the minimum distance found for each tag:

$$JCN_{lower}(F) = \frac{\sum_{t \in T} \min_{t_{sim} \in T} JCN(t, t_{sim})}{wn_pairs(T)}$$

As an upper bound we assume that the respective folksonomy subset does not contain any inherent semantics and hence only randomly related tags are returned by our measure. We simulate this by defining a random relatedness function $rand(t)$, which returns a randomly selected tag $t_{sim} \in T, t_{sim} \neq t$. The upper bound is then:

$$JCN_{upper}(F) = \frac{\sum_{t \in T} JCN(t, rand(t))}{wn_pairs(T)}$$

For the Delicious dataset it turned out that $JCN_{upper} \approx 15.8$ and $JCN_{lower} \approx 0.7$. Please recall that JCN is a semantic *distance* measure – which means a low JCN distance corresponds to a high degree of semantic relatedness.

As seen later (cf. Figure 8.6), none of our experimental conditions (including the full dataset) came close to the lower bound. There are (at least) two explanations for this. First, the lower bound was determined independent of a sub-folksonomy of the full dataset. It would be interesting to determine the sub-folksonomy that provides the optimal average Jiang-Conrath distance. Then one could check how far it is away from this optimum, and one could try to learn a classifier for this target dataset. Unfortunately, the computation of this sub-folksonomy requires the consideration of all subsets of the user set U and is thus computationally unfeasible.

Second, WordNet is built by language experts with the goal to capture *all* existing senses of a given word. Given two tags t_1 and t_2 , our JCN implementation searched for the smallest possible distance between *any* two senses of each tag. By doing so for all possible pairs of tags $t \in T$, the probability is quite high to find two closely related (or even equal) senses. Contrary to that, the technophilic bias of the user population of Delicious leads to some usage-induced relations which are not reflected well within WordNet; as an example, the most related tag to `doom` in a folksonomy subset was `quake`², leading to a large JCN distance of ≈ 18.08 , while the optimal distance was found between `doom` and `will` with ≈ 1.88 . This observation does not invalidate the procedure of semantic grounding as a whole, because we *do* find matching semantics in both systems. The same approach has also been taken in Section 7.1.3.

Results In Figures 8.6a and 8.6b we present the results of our analysis of the different sub-folksonomies which were created in each of our 256 experimental conditions.

The horizontal axis displays the percentage of included users; the vertical axis displays the average JCN distance obtained from computing semantically related tags based on the respective partition. The dashed line at the bottom of each figure represents the level of semantic precision obtained from the full

²Doom and Quake are popular videogames.

dataset. A first impression is – in all diagrams, independently of the selection strategy – that mass matters: the average JCN distance decreases and hence the results get better while more users are included. This equally holds for the random selection strategy (solid line, +). In other words, the more people contribute to a collaborative tagging system, the higher is the quality of the semantic tag relations which can be obtained from the folksonomy structure they produce. This matches the intuition that a sufficient “crowd” is necessary to facilitate the emergence of the “wisdom of the crowds”.

However, the obvious differences between the two Figures 8.6a and 8.6b suggest that the *composition* of the crowd also seems to make a difference: When incrementally adding users ordered from categorizers to describers (starting from the left of Figure 8.6a), all resulting folksonomy partitions yield systematically weaker semantic precisions compared to adding users in random order (solid line, +). This effect can be observed most clearly for the vocabulary size measure *vocab* (dotted line, ▲), which judges users as categorizers when the size of their tag vocabulary is small (see Eq. 8.1). Only after the addition of 90 % of all users in this order, the quality of the inherent semantics is on the same level of randomly selected 90 %. The other measures – with an exception of the tags per post ratio (dotted line, ●) which will be discussed later – show a very similar behavior, namely the tag/resource ratio (dotted line, ■) and the orphan ratio (dotted line, *).

When incrementally building sub-folksonomies starting from describer users (Figure 8.6b), we see a completely different picture: most measures start on the same or even on a slightly higher level of contained semantics compared to adding users in a random order. Beginning from roughly 10 % included users, all sub-folksonomies yield better results than the random case. In addition, after having added 40 % of the users in the order of the tag/resource ratio (dotted line, □), we can even observe a first improvement of the results compared to the full dataset. This implies that a bit less than the “better half” of the complete folksonomy population produces equally precise semantic structures compared to the whole unfiltered “crowd”. This improvement increases and reaches its maximum after adding 70 % of all users, before it decreases again to the global level.

Especially for very small partitions (roughly $\leq 20\%$), users selected in descending order by their vocabulary size yield the best results (dotted line, △). Interestingly, this effect is inverse when adding users the other way round (dotted line, ▲, in Fig. 8.6a): Even quite a large number of users with small

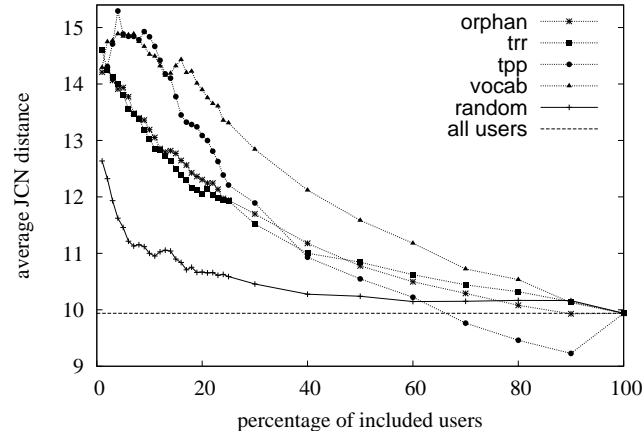
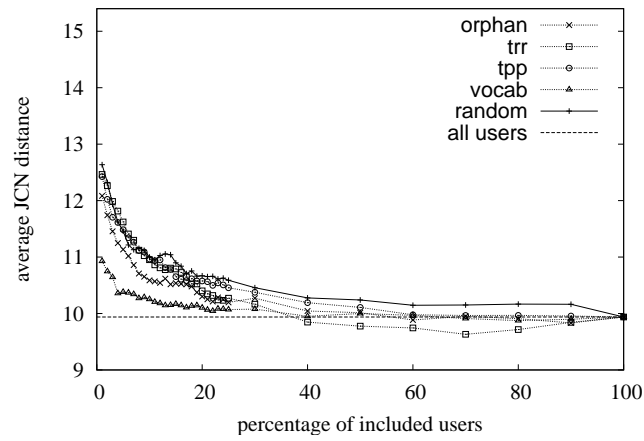
(a) CF_i^m (b) DF_i^m

Figure 8.6.: Results of the influence assessment of tagging pragmatics. Each datapoint corresponds to a “sub-folksonomy” CF_i^m (a) / DF_i^m (b) induced by different pragmatic measures. The x -axis denotes the percentage of all folksonomy users included in the subset, and the y -axis depicts the quality of the semantic tag relations obtained from the respective partition (lower values are better).

Table 8.2.: Statistical properties of selected folksonomy partitions. %t denotes the fraction of the tags from the complete dataset included in the respective partition; %w denotes the number of similar tag pairs (t, t_{sim}) found in WordNet for the respective partition divided by the number of mapped pairs from the whole dataset. For the entire dataset, $|T| = 9944$ and $wn_pairs(T) = 4335$.

i	DF_i^{trr}		DF_i^{tpp}		DF_i^{orphan}		DF_i^{vocab}	
	%t	%w	%t	%w	%t	%w	%t	%w
1	0.93	1.03	0.96	1.01	0.97	1.02	0.98	1.04
3	0.96	1.02	0.98	1.02	0.99	1.01	0.99	1.03
5	0.97	1.02	0.99	1.02	0.99	1.02	0.99	1.03
10	0.97	1.03	0.99	1.02	1.00	1.02	0.99	1.01
20	0.98	1.02	0.99	1.00	1.00	1.03	0.99	1.01
50	0.98	1.02	1.00	1.00	1.00	1.00	1.00	1.01
70	0.99	1.01	1.00	1.00	1.00	1.00	1.00	1.00

i	CF_i^{trr}		CF_i^{tpp}		CF_i^{orphan}		CF_i^{vocab}	
	%t	%w	%t	%w	%t	%w	%t	%w
1	0.56	0.48	0.44	0.00	0.48	0.59	0.27	0.18
3	0.86	0.77	0.74	0.23	0.78	0.77	0.59	0.44
5	0.94	0.83	0.87	0.49	0.89	0.88	0.76	0.59
10	0.97	0.90	0.95	0.80	0.95	0.95	0.91	0.78
20	0.99	0.95	0.97	0.88	0.97	0.98	0.97	0.88
50	1.00	1.00	0.98	0.96	0.98	1.01	0.98	0.95
70	1.00	1.00	0.98	0.98	0.99	0.99	0.98	0.98

vocabularies performs considerably worse than most other folksonomy partitions. This means that scale still matters, as the quality almost constantly increases while adding users; but the “collaborative verbosity” of a small subset of users with large vocabularies seems to lead to much richer inherent semantics than the contributions of a larger set of more “tight-lipped” users.

One could suspect now that this comparison is not completely fair: Especially when selecting users with small vocabularies, the question is to which extent semantic relations *can* be present at all in the data. In other words: If the

aggregated small vocabularies of a subset of categorizers result in a considerably smaller global vocabulary compared to aggregating more verbose users, then the probability to find semantically close tags would consequently be much lower. In the worst case, the vocabulary would be so small that the “right partner” for a given tag *does not exist*.

In order to eliminate this concern, we counted the size of the collective tag vocabulary for each sub-folksonomy. In addition, we measured how many tag pairs (t, t_{sim}) could be mapped to WordNet during the computation of the JCN distance. By doing this we want to make sure that the average semantic distance is computed roughly over the same number of tag pairs. Table 8.2 summarizes some selected statistics relative to the complete dataset.³

The first observation is that in all partitions based on describers (upper half of the table) the global vocabulary is almost completely contained ($\geq 93\%$). For partitions larger than 20%, this value raises to 98%. The same holds for the fraction of tag pairs mapped to WordNet. On the first sight, values > 1 might appear counter-intuitive here. The explanation is the following: It can happen that for a given tag t , its most similar tag t_{sim} based on the complete dataset is not present in WordNet, but its most similar tag t'_{sim} based on a particular partition is contained. A high percentage of mapped tags does not imply better semantics per se (as the two mapped tags can still be semantically distant); but the comparison of different sub-folksonomies is more meaningful when they both allow for a roughly equal number of mapped pairs. As expected, the coverage observed for the describer-based case is not as complete as for the categorizer-based excerpt: For very small samples, the collective tag pool is in fact small. However, this effect is mitigated already for samples of 3%; and starting from roughly 10-20% sample size, a sufficient global vocabulary exists ($\approx 97\%$). This means that the comparison in general is performed on a fair basis, because the underlying vocabulary sizes are comparable.

Our results suggest that sub-folksonomies based on describers contain more precise inherent semantic structures than partitions based on categorizers. However, there seems to be a limitation in this observation: Inspecting the curve for the tpp measure on the right side of Figure 8.6a, one can observe that the most precise semantic relations among all experimental conditions are found after the addition of 90% of the categorizers according to this measure. As

³We did not include the statistics for every partition for space reasons; missing values can be interpolated from the given examples.

stated above, this partition can also be read from the other side and corresponds to a removal of 10 % of the most extreme describers. As the *tpp* measure captures the average numbers of tags per post, there seems to be a number of “ultra-taggers” who use a large number of tags per post (many spammers, typically more than 9 tags per post in our case) and have detrimental effects on the global tag semantics. In other words, removing these users seems to eliminate “semantic noise”, leading to more precise tag semantics.

8.3.3. Implications

A core topic of this dissertation as well as of other works in the literature (e.g., (Wu et al., 2006a; Schmitz, 2006)) are methods to harvest emergent semantic structures from Social Annotation Systems. Our results show that the effectiveness of current semantic measures for tag relatedness are influenced by factors originating outside of the semantic realm. On small data samples (up to 40 % of users in our dataset), we have singled out a group of users (categorizers) that has particularly detrimental effects on the performance of current semantic measures compared to random sampling. At the same time, describers (based on the tags-per-resource measure) consistently outperform random sampling, and can level and even outperform the results achieved on the entire dataset with as little as 40 % of the users. This suggests that methods for harvesting *semantics* from samples of tagging systems can be made more effective when utilizing knowledge about the *pragmatics* of tagging, considering individual user behavior. For analysts of small data samples who wish to improve semantic relatedness measures, this would mean focusing on those users that use tagging systems in a verbose ‘Stop Thinking, Start Tagging’ fashion. With increasing sample sizes (>50 % of users), we can observe that adding more categorizers does not produce significantly better results. However, when adding more describers, we see significant improvements in performance until we hit an accuracy limit at approximately 90 % of users. This suggests that rewarding verbose taggers comes with limitations itself: The most verbose taggers (in our case: mostly spammers) negatively influence the results as well.

The practical implications of our results concern mainly two questions: (i) What is the minimum amount of users needed to produce meaningful tag semantics in collaborative tagging systems and how can these users be selected? (ii) Does the quality of emerging tag semantics increase with the available amount of data, or can it be improved by eliminating “semantic noise”?

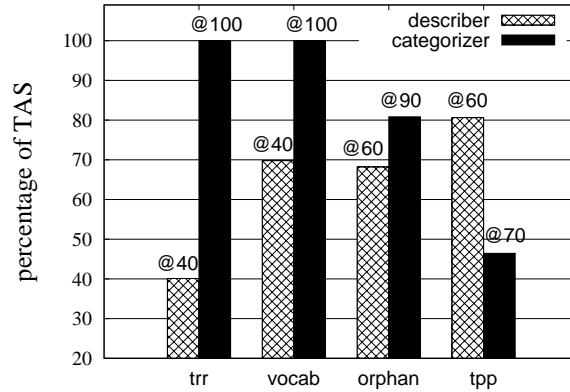


Figure 8.7.: Minimum size of the folksonomy partitions created by each measure sufficient to reach the semantic precision of the complete dataset. The y -axis denotes the percentage of tag assignments contained in the smallest folksonomy partition which reached the global semantic precision; the labels above the bars depict the percentage of users the respective sub-folksonomies are based on.

A main contribution of our analysis lies in the observation that tagging pragmatics, i. e., individual tagging characteristics, play an important role in both cases. The experiments described above reveal that not all users contribute equally to emerging semantics; we could show that a relatively small subset of describers yields significantly better results than a group of categorizers. Figure 8.7 summarizes the minimum sizes of the folksonomy partitions identified by each of our introduced measures necessary to reach the level of semantic precision for the entire dataset. The white bars correspond to sampling users ordered from describers to categorizers (Fig. 8.6b) while the black bars correspond to sampling users ordered in the opposite direction (Fig. 8.6a). The number on top of each bar displays the user fraction needed to reach the global semantic precision; the y -axis depicts the size of the respective sub-folksonomy relative to the complete one.

In general, most describer-based selection strategies create smaller folksonomies which produce meaningful semantics. The “smallest” one consists of 40 % describers according to the *trr* measure, responsible for roughly 40 % of all tag assignments. However, the observation that uncontrolled verbosity is not a good

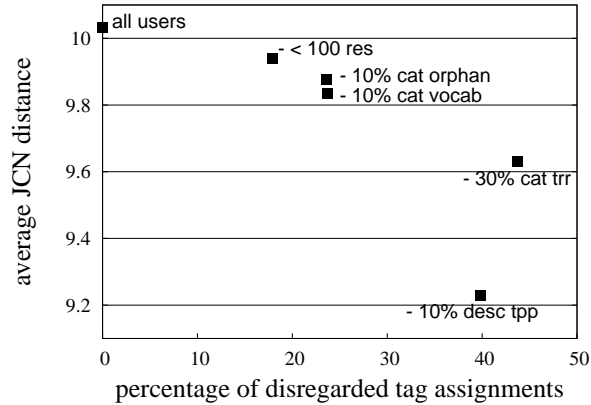


Figure 8.8.: Improvement of semantic precision by removing users from the complete dataset. The y -axis depicts the semantic precision of the (sub-)folksonomies, while the x -axis denotes the percentage of tag assignments which were disregarded by removing certain users. The labels at each data point describe which users were removed.

thing is confirmed by the observation that removing 30% of the most extreme describers according to the tpp measure (rightmost black bar) also creates a comparatively small and semantically precise partition. According to Figure 8.7, two adequate strategies for creating the smallest possible scaffolding for global tag semantics can be identified: (1) include roughly half of the users with a high tag/resource ratio, and (2) remove roughly one third of “ultra-taggers” identified by a large average number of tags per post.

The next interesting question to ask is whether, and to which extent we can even infer *more precise* semantics when removing users. Figure 8.8 displays the obtained semantic precision (y -axis) plotted against the amount of tag assignments removed when removing users according to different selection strategies. The first and most simple strategy is to remove the “long tail” of users with less than 100 resources in their collection. This already eliminates roughly 18% of the data, while interestingly slightly improving the semantic precision. One cannot conclude from that that the long tail of users does not contain valuable information at all. But with regard to *popular* tags (recall that we restricted our dataset to the top 10 000 tags), a valid first insight is that the long tail of inactive users can be discarded during the computation of

semantic tag relations.

As discussed before, our results indicate that categorizers also have a detrimental effect on the quality of the emerging structures. Removing 30% of them as determined by the tag/resource ratio leads to a further improvement in semantic precision. The best result in all of our experimental conditions however was reached by eliminating 10% of the extreme describers according to the tags-per-post measure. Those “hyper-active” users (in our case mostly spammers as confirmed by manual inspection) generate roughly 40% of the global amount of tag assignments. Spammers typically use a large number of semantically disjoint tags to attract other users and to bias search engines towards their posted URLs. Unsurprisingly, they are not very helpful for creating meaningful tag relations. Rather the contrary is the case: we can see in our results that spammers introduce significant semantic noise – a removal of them leads to an overall improvement in accuracy of the resulting semantic structures. Turning the tables around, this insight can of course also be useful for spammer detection itself – but because our dataset does not contain explicit spammer labels for each user, determining the exact ratio of spammers detected by each of our pragmatic measures is subject to future work.

Generalization to other datasets In order to exclude the possibility that the implications mentioned above are influenced by characteristics of the Delicious dataset, we repeated the experimental procedure described in Section 8.3.2 on a dataset from January 2010 of our own social bookmarking system BibSonomy⁴. It differs from the dataset mentioned in Section 6.1.1 insofar as it also contains spam posts. In summary, it contained 17 777 users, 10 000 tags and 4 520 212 resources connected by 34 505 061 TAS. We omit a detailed explanation for the sake of brevity; but in general, all measures exhibited a very similar behavior as observed for the Delicious dataset in Figures 8.6a and 8.6b. Especially the practical implications discussed before were valid in a nearly identical way for the BibSonomy data: 30% of describers according to the *trr* measure were sufficient to reach the semantic precision of the whole dataset, and removing 20% of describers according to the *tpp* measure led to the best overall semantics.

⁴<http://www.bibsonomy.org>

8.3.4. Summary

In this section, we analyzed the influence of individual tagging practices in collaborative tagging systems on the emergence of global tag semantics. After proposing a number of statistical measures to assign users to two broad classes of categorizers and describers, we systematically built folksonomy partitions by incrementally adding users from each class. We then judged the quality of the emergent semantics contained in each of these “sub-folksonomies” by means of semantically grounded tag relatedness measures. Apart from the observation that adding more users is beneficial in many – but not all – cases, our results reveal a dependence of the obtained semantic structures on the different partitions. In general, the collaborative verbosity of describers provides a better basis for harvesting meaningful tag semantics. However, this observation comes with a limitation: The most verbose taggers (in our case mostly spammers) negatively influenced semantic accuracy. From a practical perspective, the pragmatic measures can be used to select a comparatively small subset of users which produce tag relations of equal or better quality than the entire set of users. In addition, the measures can facilitate improvement of the global semantic precision by eliminating users that introduce “semantic noise”.

A main implication hereby is the presentation of first empirical evidence for a causal link between the pragmatics of tagging (individual tagging practices) and the emergent semantics of tags. This link is *not* dependent on our choice of a particular semantic relatedness measure, because 1) the chosen Jiang-Conrath distance has been shown to best reflect the human perception of semantic relatedness in previous validation studies (Budanitsky and Hirst, 2006) and 2) our experiments with alternative measures for semantic relatedness have produced similar results.

This finding has interesting implications for the research questions addressed within this dissertation: First, while our results focus on semantic relatedness, it appears plausible that other semantic tasks, such as hypo/hyponym detection, exhibit similar effects. We argue that a general link between tagging pragmatics and tag semantics could yield new ways of thinking and new algorithm designs for learning ontologies from folksonomies. Second, a further promising idea is to “stimulate” the emergence of semantics by utilizing tag recommenders to influence tagging behavior and to “steer” evolution of folksonomies into semantically richer directions.

8.4. System Abuse and Spam

As with many other services on the Web, the popularity of Social Annotation Systems may be seen as a mixed blessing: While they are on the one hand of great help for a large number of users, they also attract malicious activities. Among the latter, especially the system abuse in form of *spamming* (i. e., the annotation of inappropriate content) is relevant for studies of emergent semantics. Despite the discovery of spammers is not a core topic of this dissertation, we will start by giving a brief introduction of our understanding of “spam” in the context of Social Annotation Systems, and describe the countermeasures we undertake within BibSonomy to prevent system abuse. Because we are in the unique position to have a dataset with explicitly marked up users, we will present in the following a comparison between the kind of semantics introduced by spammers and non-spammers.

8.4.1. Spam Definition and Detection

While the existence of, e. g., email spam is widely known, its counterpart in Social Annotation Systems is sometimes less visible for the average user. A potential explanation for this is that the system operators are forced to undertake preventive measures, which filter out large parts of the inappropriate content. In order to convey our understanding of spam for the context of this dissertation to the reader, we stick to the following definition:

“[...]we consider spam in folksonomies as (1) content which legitimate users do not wish to share and (2) content which is tagged in a way to mislead other users. The first part refers to web spam: For commercial or political interests, to simply distract the system, or to run down other companies, spammers try to score high with their web sites by posting their content in the system. The second part considers the tagging behavior: spammers add keywords that do not match the content of the bookmarks. Again the motivation may be self-promotion (users looking for a specific tag will receive advertisements) or to distract and destroy the serendipitous browsing facilities that make folksonomies special.” (*Krause et al., 2008b*)

So briefly spoken, spam corresponds to unintended and disturbing usage of a Social Annotation System. Especially the last mentioned aspect can be expected

to have an especially detrimental effect: Because emergent semantic structures are mostly based on *meaningful* co-occurrences, the intentional introduction of semantically irrelevant correlations presents a potential threat. This has been visible, e. g., in the previous Section 8.3.2, where a comparatively small number of excessively “spamming describers” significantly affected the global folksonomy semantics. On the other hand, in case that spammers are using “correct” annotations (for whatever purpose), it is not impossible that they may be also useful: As shown in Section 8.2.1, more data has usually a beneficial effect.

In any case, from our own experience of running BibSonomy, we know that spam is an issue that needs definitely to be considered by each Social Annotation System operator. While it is (for good reasons) typically hard to find out which countermeasures are taken by other systems, our position of running an own platform enables us to take a look behind the scenes. Due to a growing popularity since its establishment in 2004, the high search engine ranking of BibSonomy has been an attractive target for a large number of users trying to increase their own ranking by submitting large amounts of content at our site. In the beginning, those were filtered out manually by a small group of evaluators. This process is described best as follows:

“The flagging of spammers by different evaluators is a very subjective process. There were no official guidelines, but a common sense of what distinguishes users from spammers, based on the content of their posts. To narrow down the set of potential spammers, the evaluators normally looked at a user’s profile (e. g., name, e-mail address), the composition of posts (e. g., the semantics of tags, the number of tags) before assessing the content of the bookmarked web sites. Borderline cases were handled from a practical point of view. BibSonomy intends to attract users from research, library and scholarly institutions. Therefore, entries referring to commercial advertisements, Google Ad clusters, or the introduction of specific companies are considered as spam.” *(Krause et al., 2008b)*

Because this approach naturally did not scale with further growth of the system, the manually labeled training data was used to optimize machine learning algorithms, which make automatic predictions if a user is a spammer or not. Those decisions are based on a varying set of features from users’ profiles, locations or activities (see (Krause et al., 2008b) for detailed explanations). In

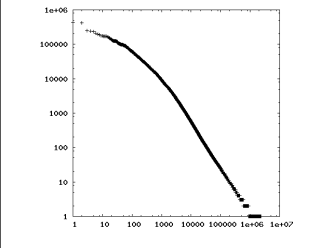
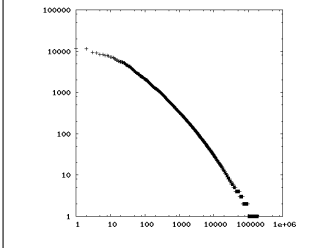
order to highlight the “dimension” of the spammer problem, Table 8.3 depicts some statistics of the “good” and the “bad” part of BibSonomy. As one can see, the amount of data produced by spammers is immense, and outnumbers, e. g., the number of tag assignments produced by non-spammers by the factor of ≈ 25 . While those large amounts of actually unwanted contents pose some technical challenges, they might give rise to the hope that at least some parts of it could still be useful for complementing the comparatively small size of the non-spammer data.

In order to investigate the possibility on a qualitative level, Table 8.3 compares a number of statistical properties of the folksonomies produced by both groups. We start by contrasting the two vocabularies in order to check whether spammers and non-spammers “use the same language” or “talk about the same things”, at least within certain domains. From inspecting the 10 most popular tags, this seems not to be case: While we find partially offensive and business-related terms for spammers (as well as artifacts from automatic content submission systems⁵), the non-spammer vocabulary is mostly free from the latter. Also when exploring this issue in a more general fashion by computing the overlap between both vocabularies for the top 10 000, 100 000 and all keywords, it becomes clear that only roughly every 4th or 5th keyword is used by both parties. This confirms the assumption that the fundamentally different motivations for using the system are reflected in differing vocabularies. The same holds for the resource overlap: Here we counted how many resources (bookmarks or publications) were annotated by both spammers and non-spammers. The comparatively low value of 30 951 (which corresponds to roughly 5.6 % of all non-spammer resources) indicates that spam content is of limited interest for “good” system users.

Despite these indications for a limited usability of spam data, the comparison of the tag frequency distribution shows a similar pattern. In addition, the WordNet overlap is even higher for the spammer case. Although a cursory manual analysis exhibited that the higher overlap seems to lie mostly within typical “spammer topics”, it is still worth exploring if some parts of this data can be useful to counteract the sparsity problem when capturing emergent semantics. The following section analyzes this issue by comparing the different *semantic* qualities of both datasets.

⁵Onlywire <http://onlywire.com> is a platform which offers automated submission of content to several social sharing platforms.

Table 8.3.: Comparison of spammer and non-spammer data within BibSonomy.

<i>spammers</i>		<i>non-spammers</i>
<i>statistics</i>		
2 283 703	T	192 445
297 174	U	6 463
6 701 874	R	551 540
62 128 872	Y	2 434 387
<i>popular tags</i>		
onlywire, free, online, to, for, home, business, sex, and, video	top 10	zzz.to.sort, deutschland, web2.0, nn, programming, theorie, web, university, media
<i>Keyword overlap Spammers / Nonspammers</i>		
2 751	top 10 000	2 751
21 853	top 100 000	21 853
74 450	all	74 450
<i>tag frequency distribution</i>		
		
<i>Resource overlap Spammers / Nonspammers</i>		
30 951	all	30 951
<i>WordNet Overlap</i>		
771	top 1 000	544
6 098	top 10 000	4 027
22 154	top 100 000	18 521
62 366	all	25 182

8.4.2. Influence Assessment

As stated above, the ultimate question is whether or not spammers are contributing the emergence of implicit semantic structures within Social Annotation Systems. Following our method influence assessment, we computed tag similarities on three different datasets, comprising (i) only non-spammer data, (ii) only spammer data and (iii) spammer and non-spammer data. The goal of the last “mixing” condition was to check if spammers would eventually even destroy the

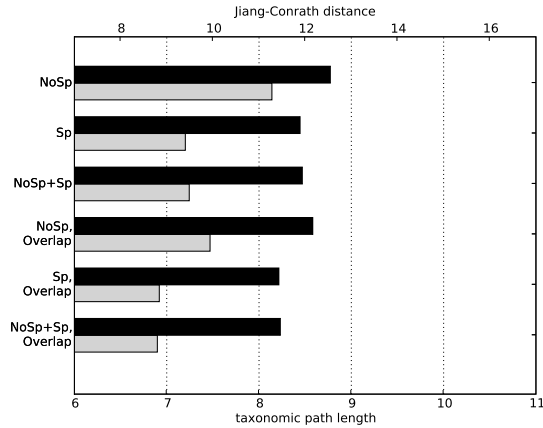


Figure 8.9.: Influence assessment of spammers on emergent semantics. “Sp” stands for Spammers, “NoSp” for non-Spammer users. “Overlap” symbolized the keyword overlap among both ($\approx 2\,700$ keywords).

“good” kind of semantics produced by non-spammer users.

The top 3 rows of Figure 8.9 depict the results, measured by both the Jiang-Conrath distance and the taxonomic path length. Somewhat contrary to our expectations, all conditions including spammers lead to better results, i. e., to semantically more closely related tags. In order to exclude the possibility that this effect is due to “good” semantic relations within “unwanted” domains (like, e. g., explicit terms), we repeated the experiment based only on the vocabulary overlap. Because the keywords in the overlap were also used by trustworthy users, we expect the latter to be a filtering to appropriate terms only. The lower three rows of Figure 8.9 depict the result – which shows a similar pattern like the unfiltered condition. So an impression obtained from these results is that spammers – especially after an elaborate filtering of content has been undertaken – could in fact be useful to counteract data sparsity issues. Table 8.4 shows some exemplary keyword pairs for which the inclusion of spam data had a positive effect.

However, these results need to be interpreted with caution: First of all, within the scope of this dissertation there is evidence for both beneficial and detrimental effects of spam. The latter was especially visible in the analysis of tagging pragmatics (see Section 8.3.2), where excessive “describers” negatively

Table 8.4.: Examples of better semantic relations obtained from spam data. For the given pairs, the semantic quality of the most related tag was higher when taking into account spammer data.

<i>keyword</i>	<i>spammers + non-spammers</i>	<i>non-spammers</i>
education	teaching	english_linguistics
bike	motorcycle	earth
physicians	physicians	medical
plans	program	participatory
pipes	pipe	bloglines
tax	taxes	financial_performance
bags	bag	boomerang
exercises	exercise	imaging
e-mail	email	privacy
converter	converters	editing

affected the global semantics. Second, one has to take into account the strong difference in the size of the datasets: While the non-spammer BibSonomy dataset is the smallest one we used (roughly 2.5 million tag assignments, see Table 8.3, its spammer counterpart is an order of magnitude larger (roughly 60 million tag assignments). This obvious discrepancy may also have biased the results. And finally, as explained in the previous section, the labelling of spammers has been done both manually and by automatic algorithms – both of which can not be expected to make solely perfect decisions. Hence one can not exclude the possibility that there are “normal” users among the “spammers”, and vice versa.

Despite these considerations, a contribution of this analysis is that spam data should not be disregarded as a matter of principle when analyzing emergent semantics. Though a lot of care has to be taken, it seems to be worth to explore those cases in which the intrinsically *malicious* spammer activities can be exploited in a *beneficial* way.

8.4.3. Summary

In this section, the influence of inappropriate system usage or *spam* on the emergence of implicit semantic structures was analyzed. After introducing a definition of two main spammer groups, namely those who (i) annotate inap-

propriate content and (ii) target to mislead other users, manual and automatic spam classification as countermeasures undertaken within the BibSonomy platform were explained. Based on the resulting dataset, the quality of inherent semantics within both was measured. Despite the vocabulary of spammers and non-spammers significantly differs, there seemed to be cases in which the large amounts of spam data had actually a beneficial effect of tackling data sparseness. As a final recommendation, the outcome of the experiments was summarized by stating that spam data should not be excluded from semantic analyses as a matter of principle.

With the presented spam analysis, the chapter on factors of emergent semantics is closing, as well as the second part of this dissertation, which was concerned with data, methods and influencing factors. The following chapter is now concerned with concrete applications which can benefit from the insights of the studies presented so far in this dissertation. As a final step, the conclusions which can be drawn from the latter are presented.

Part III.

Applications and Conclusions

Chapter 9.

Applications

While the previous and main part of this dissertation was concerned with a thorough theoretical and empirical analysis of methods to harvest emergent semantics from Social Annotation Systems, our next goal within the current Chapter is to discuss how to make best use of the insights gained hereby. More precisely, we will present a set of applications which implement or operationalize the most important findings. Based on the two broad classes of the “Semantic Web” and “Social Annotations” introduced in the first part of this dissertation, a first question hereby is which class is targeted: We will start by describing applications which focus on exploiting the gained insights for enhancing Social Annotation Systems themselves. As mentioned in Section 3.1.1, those do exhibit a number of weaknesses, which can be nicely addressed by, e. g., the outcomes of concept learning approaches.

As a next step, we will discuss how some of the shortcomings of typical Semantic Web applications (see Section 4.1.4) can benefit from the captured emergent semantics from Social Annotation Systems. In general, all mentioned applications in this chapter are in different “development states”: While some of them (e. g., those mentioned in Section 9.1.1) are already being used for productive purposes, others represent ideas or prototypical implementations. The reason to present all of them in a unified manner is to present to the reader an overview of existing and potential future tools in the field of “bridging the gap” between social and semantic resource annotation on the Web.

9.1. Enhancing Social Annotation Systems

For the purpose of providing an enhanced user experience of Social Annotation Systems based on the previous methods and results, there exist two natural integration points: First, based on the insights on influencing factors of emergent semantics, it is a desirable goal to develop tools and adaptations which have

a positive effect to this end. This ideas of “steering” the evolution of the folksonomy into a semantically richer direction would lead to better results of tools which capture these semantics, and finally to an enhanced overall experience. Such tools which foster the emergence process will be presented in Section 9.1.1

The second natural idea is to use the previously described methods to tackle *directly* the inherent shortcomings of Social Annotation Systems, like synonymy, polysemy or the lack of structure. Because such ideas are basically making the implicit semantics available to the user in an explicit form, this corresponds to “feeding back” the learned semantics into the original system. This idea is especially appealing for two reasons: First, because no external semantic repository is involved, the kind of semantics should – in the ideal cases – match exactly the users’ needs, because it essentially does not produce “new” things, but rather makes existing latent structures visible. Second, because of imitation processes it might be possible that the exposure to meaningful semantic relations influences the users towards a “better” kind of annotation, ultimately fostering the convergence to a globally accepted and universal vocabulary. Such applications will be discussed in Section 9.1.2.

9.1.1. Fostering the Emergence of Semantics

The main results which are relevant for applications targeted towards stimulating the emergence of semantic structures within Social Annotation Systems are the ones from Chapter 8 on influencing factors. An important insight hereby was that having more data is (with certain limitations) desirable in order to allow a more precise assessment of keyword semantics. Consequently, one could argue that all tools which enhance user interaction and enlarge the resulting set of annotations help to foster the emergence of semantics. Based on the experiences of developing and maintaining the BibSonomy system, a crucial point when trying to leverage user contributions is to enhance the general usefulness of a system for a broad variety of tasks. Because a large part of BibSonomy’s audience stems from an academic environment, an especially important factor is hereby that the system integrates as seamlessly as possible into the working routine of doing research and writing papers. The maintenance of publications lists is hereby a critical use case, as well as the integration of BibSonomy into existing applications and tools. Because the latter was a core responsibility of the author as a member of the BibSonomy development team, we will briefly

describe a set of integration tools. After this, we will discuss how the insights from analyzing *pragmatic* aspects of tagging can be operationalized by means of tag recommendation engines.

Integration tools

When trying to attract users and their contributions to BibSonomy, it is an important question which specific *task* they are typically using the system for. From our own experience, there are “power users” which use BibSonomy as their personal resource repository, making use of all of its facets. Besides those, especially the maintenance of publication lists (e. g., on institution or group websites) seemed to be a use case for which a support by BibSonomy would be highly desirable. This means if we can provide an added value by offering a comfortable and automatic maintenance of publication lists, then our user base might significantly grow. Furthermore, despite BibSonomy offers a convenient web interface, some users are already accustomed to local solutions for maintaining their bibliographies. In order to make BibSonomy more attractive, and ultimately to enable an enhanced user experience by a semantically richer folksonomy, the author has been the leading developer of the following tools and applications:

- **JabRef Integration:** JabRef¹ is a popular open source application for client-side management of bibliography files based on the BibTeX-format². It offers various convenient facilities which ease the process of creating, maintaining and using a repository of bibliographic references. As an example, it allows to directly search and retrieve entries from various online libraries, without the need of manually typing each publication detail. Furthermore, it offers a built-in grouping mechanism, which allows among others to organize a collection based on keywords. These features have led to its popularity, reflected in roughly between 800 and 1 000 downloads each day³. In order to tap into this large user base and its potentially high number of keyword-annotated bibliographic records, we developed a *plugin* which connects a local repository in a convenient way

¹<http://jabref.sourceforge.net>

²<http://www.bibtex.org>

³According to <http://sourceforge.net/projects/jabref/files/jabref/stats/timeline>, retrieved on 2011/10/17.

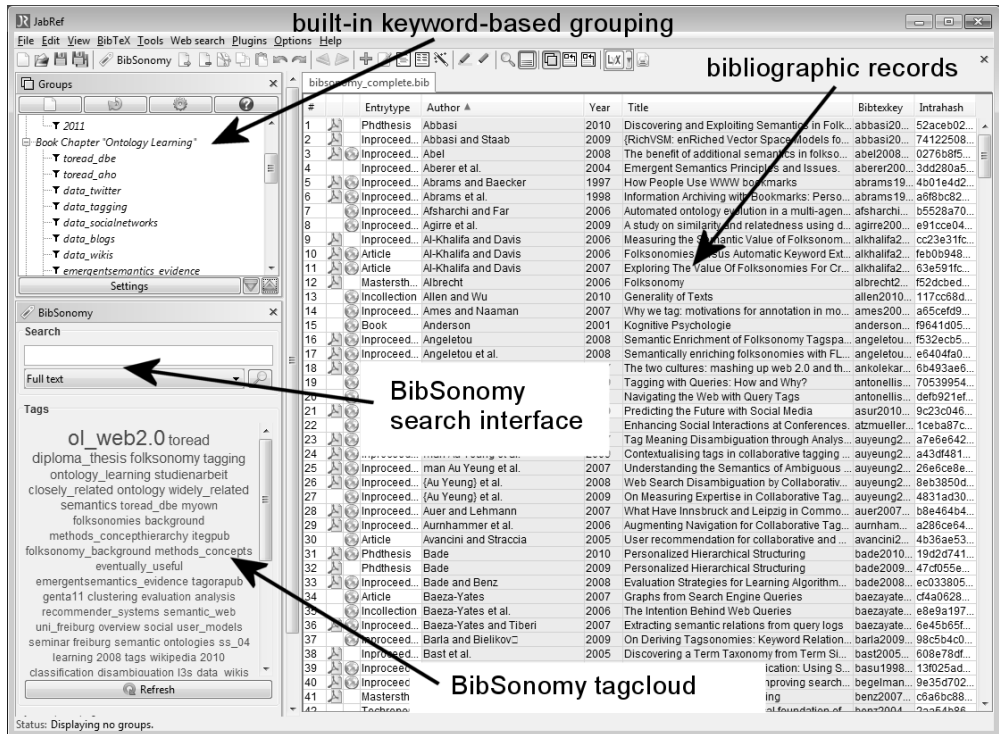


Figure 9.1.: Screenshot of JabRef, including the BibSonomy plugin.

to an online variant within BibSonomy. Figure 9.1 shows a screenshot highlighting some of its functionalities. On the top left corner, one can see JabRef’s built-in grouping facility; below that, there is a BibSonomy tag cloud, which allows browsing and direct retrieval from entries present in BibSonomy. The latter is complemented by a full-text search facility.

In order to push the entries from client side to our servers, there exist two mechanisms: First, there is a *synchronization* option, which guides the user through a dialog targeted towards aligning his local and remote bibliography. Furthermore, one or more entries can be individually selected and uploaded via a menu option. In this way, our goal was to combine the advantages of a comfortable local bibliography client with those from a central repository like BibSonomy. From a viewpoint of emergent semantics, our goal was to enlarge our corpus of annotated content by

greatly simplifying the process of uploading, directly from within the JabRef application. Furthermore, because the local grouping facility follows a hierarchical scheme, this opens up future possibilities to, e. g., feed back automatically induced keyword hierarchies from BibSonomy back into JabRef. The usage of the plugin is also extensively documented⁴.

- **Typo3 Integration:** While the JabRef targets to stimulate user contributions mainly by supporting the creation and maintenance phase of bibliographies, the next tool is focused towards enhancing BibSonomy’s usefulness for the purpose of making the latter available to the public. Typically, on personal or group websites, one can find a *publication list* for this purpose. Because an up-to-date list of published papers can be seen as a promotionally effective “asset”, great efforts are sometimes undertaken to this end. Instead of manually created static web pages, many researchers are using content management platforms like Typo3⁵ for this purpose. However, instead of potentially maintaining two reference collections (one for the publication list and one as a personal library), our idea was to build an extension which allows to display entries from BibSonomy directly within a Typo3-based website, formatted in an appropriate manner. The resulting extension can be found in the Typo3 Extension Repository⁶. Figure 9.2 shows a screenshot of its configuration interface, while Figure 9.3 displays a publication list as an exemplary output. The offered features mainly correspond to controlling the options of BibSonomy’s internal layout rendering engine, i. e., selection of content via a URL pattern, and choosing a layout template to render the selected entries. Apart from that, tag clouds can also be displayed. While these functionalities are not *directly* related to adding content to BibSonomy, we think that providing a solution to an often requested task (namely the comfortable and centralized management of publication lists) provides a further incentive to use BibSonomy and hence fosters *indirectly* the annotation of content.
- **REST-API:** While the above tools were targeted towards integration with two specific existing applications, the establishment of an interface which

⁴<http://www.bibsonomy.org/help/doc/jabref-plugin/index.html>

⁵<http://typo3.org>

⁶http://typo3.org/extensions/repository/view/ext_bibsonomy/current/

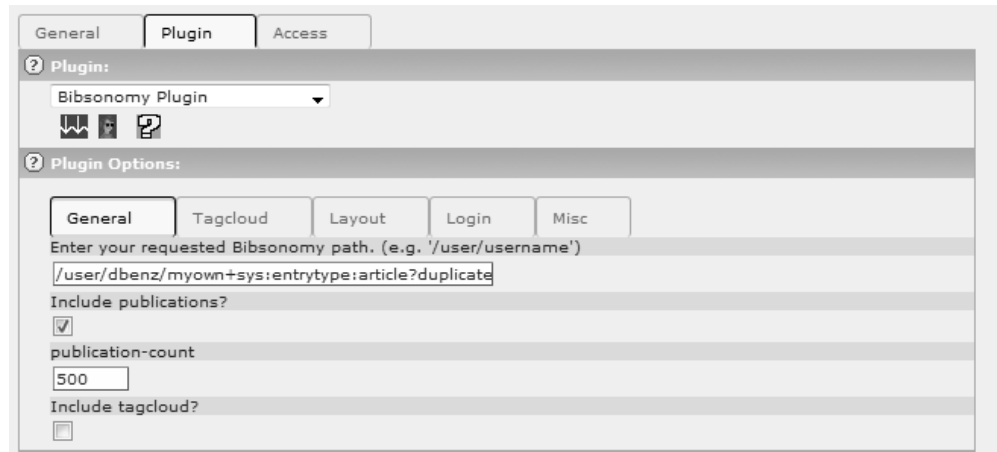


Figure 9.2.: Screenshot of the Typo3 plugin configuration.

allows for integration with a variety of potential future applications was the driving force behind developing a REST-based programming interface (i. e., a *REST-API*). Initially built within the scope of a student project (Bork, 2006), its further development and maintenance was primarily lead by the author. The basic functionality of the API is to allow the interaction with the system using a standardized XML data format, making it possible to create, read, update and delete (CRUD) publication and bookmark entries. For this purpose we are offering a dedicated URL syntax⁷, against which different HTTP requests can be sent in order to access the corresponding functionalities. Again, the REST-API on its own does not lead directly to new content being added to BibSonomy; however, the existence of a programmatic access to BibSonomy data is intended to catalyze the development of third-party applications based on the latter, which finally can be expected to have positive effects on the total amount of Social Annotation data ending up in our system.

⁷<http://www.bibsonomy.org/help/doc/api.html>



Figure 9.3.: Screenshot of a publication list created by the Typo3 plugin.

Tag Recommendation

Besides the goal to “collect” more data and hence to improve potentially emerging semantics, another insight from analyzing pragmatic aspects of tagging (Section 8.3) was that a more “verbose” style of tagging is desirable. In other words, apart from attracting new active users, an additional goal is to provide incentives for a richer annotation, especially one which includes more keywords. While the goal of changing user habits is hereby probably hard to reach (and may also not be desired), another viable strategy is to make richer annotations *easier* by providing support during the annotation process. A typical tool in this field are *tag recommenders*, which are suggesting a set of keywords. Because the process of choosing among the latter is cognitively easier than coming up with own keywords, the hope is that this alleviation is reflected in a more verbose annotation. A large number of approaches exists in this direction; see (Hotho et al., 2008) for an overview. Some of the mentioned works are also available as implemented recommenders within BibSonomy; it will be an interesting research question to examine their influence on the resulting vocabulary.

Apart from tools which are intended to foster the emergence of semantics, the following section highlights ideas to make use of the latter by different kinds of feedback mechanisms.

9.1.2. Feeding back Semantics

Having observed the inverse relation between the strengths and weaknesses of social and semantic annotation (see Section 5.1.3), a further promising application field is use the harvested semantics to address some inherent shortcomings of Social Annotation Systems. Because the latter are often related to homonymy, polysemy and lack of structure, we will present (partially implemented) ideas of semantic search, concept recommendation, vocabulary maintenance and semantic browsing to address these issues.

Semantic Search

Especially the keyword-based retrieval of resources within Social Annotation Systems is negatively affected when different users use varying annotations for the same “concept”. This problem could be tackled effectively by the methods of concept learning presented in Section 7.2: Having identified that, e. g., *oo* and *object-oriented* belong to the same “synset”, using this information e. g., for query expansion would raise the possibility to find relevant content. On the other hand, polysemous keywords lower the usability of the search results, because the returned resources may belong to different (potentially unrelated) meanings. Such cases could be detected using the sense discovery methods mentioned in Section 7.2.3, in the best case completed by a disambiguation step. Finally, in case an overly general or overly specific keyword is used, the number of returned results will be either too high or too low. These problems could be addressed by an automatically learned concept hierarchy (see Section 7.4), which would facilitate the suggestion of broader or narrower search terms, leading to a suitable number of results.

Vocabulary Maintenance

Despite the data model of many Social Annotation Systems is characterized by a “flat” keyword space, some systems allow the user-driven specification of keyword relations. Within Delicious, those are denoted as “bundles”, while BibSonomy facilitates the creation of generic subtag-supertag relationships. A

set of relationships including a particular keyword is denoted a “concept”. Those can be used to structure the keyword and resource collection. Because this process can be expected to be laborious, supporting it by means of a “concept recommender” would be desirable: Similar to tag recommendation engines (as explained in the previous section), the goal would be to come up with concept suggestions, making it easier for users to structure their annotation vocabulary.

Another experience, especially when the number of keywords is growing, is that the vocabulary becomes “noisy”, in the sense that users do not necessarily stick to a consistent keyword usage. As an example, it might happen that synonymous keywords are introduced, because the tagger simply does not remember which variants he used before. While this may even be desired (e. g., by “describers” as explained in Section 3.1.6), the elimination of this noise would surely be useful to a number of users. While the aforementioned suggestion of concepts is one possibility, another one is to provide a facility to “clean” the vocabulary by providing recommendations which keywords should, e. g., be replaced by a canonical variant. Such activities were also subsumed under the term “tag gardening” in the literature (Weller and Peters, 2008). As an example, such a suggestion could be to replace all occurrences of the keywords *ontology* and *ontologies* with the standard singular form *ontology*. Generally speaking, this would tackle the retrieval problems of Social Annotation Systems directly at the annotation level, while the aforementioned methods of semantic search target the retrieval process instead.

Semantic Browsing

Besides their role as a searchable repository, especially the various *browsing* facilities of Social Annotation Systems have been seen as one of their core strengths (Golder and Huberman, 2006). However, typically exploration is possible only along *explicit* links between tags, users and resources. While this already allows for serendipitous discovery of interesting content, the additional possibility of a “semantic” browsing direction would represent an orthogonal and potentially useful aspect. For such purposes, especially the measures of semantic keyword relatedness (Section 7.1) are applicable. More precisely, on each keyword page within BibSonomy, we added an additional facility to browse along “similar tags”, which are computed based on the tag context relatedness (see Figure 9.4). The similarities are recomputed on a daily basis in order to

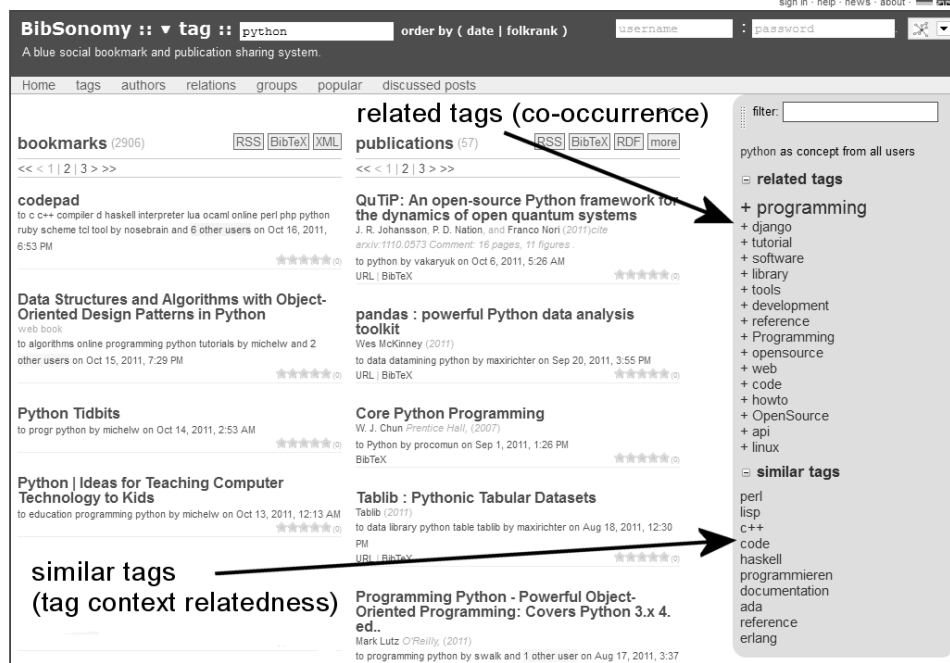


Figure 9.4.: Screenshot of a semantic browsing facility within BibSonomy.

reflect changes in the underlying semantics. As is visible within the example of Figure 9.4, this introduces in fact a new kind of browsing direction: While the most related keywords to *python* are *programming*, *django*, *tutorial* and *software*, its most *similar* tags are *perl*, *lisp*, *c++* and *code*. This allows e. g., to extend the browsing activities also to other kinds of programming languages.

Shifting the focus away from Social Annotation Systems themselves, the following section discusses ideas on how the captured emergent structures can be useful for other kinds of semantic applications.

9.2. Enhancing Semantic Applications

While having a stronger background on the “social” side by running BibSonomy as sharing platform, our ideas to reuse the learned semantic structures for external purposes remain on a more visionary level. However, having observed

the deficiencies of the “ontological” way of capturing knowledge, we expect the strongest benefits in the field of optimizing web search engines, ontology maintenance and rich user profiling – each of which will be briefly explained in the following.

Optimizing Web Search Engines

The idea of optimizing web search engines by exploiting social information is not new; as an example, (Bao et al., 2007) proposed a *SocialPageRank* including information from Delicious. Another example is (Au Yeung et al., 2008), who reported high precision when using semantics from collaborative tagging systems to disambiguate web search queries. These two works exemplify that the meaningfulness of the captured semantic structures within Social Annotation platforms is not limited to the systems themselves, but can also be reused in different contexts. So from a more general point of view, it seems plausible that, e. g., the synonym sets discovered by the concept learning methods from Section 7.2.1 can not only be useful for enhancing search *within* their originating system, but also, e. g., for query expansion in the context of traditional web search engines.

Ontology Maintenance

The knowledge acquisition bottleneck and the high cost of maintaining a consistent and up-to-date ontology effectively hampers their widespread usage. Though the semantic structures learned from Social Annotation data can in their current form surely not be seen as a direct replacement, they can still be very useful to assist knowledge engineers in construction and maintenance tasks. Especially semi-automatic approaches which include collaborative ontology editing tools like *Soboleo* (Braun et al., 2007b) could benefit strongly from, e. g., presenting the creators with a set of concept suggestions derived from Social Annotation data. Once an ontology has been established, it is thinkable to keep it up-to-date (or to foster its “maturing” (Braun et al., 2007a)) by checking on a regular basis, e. g., if there exists a new meaning of a contained term (using the sense discovery methods described in Section 7.2.3). In addition, the synonym resolution methods from Section 7.2.1 could be used to complete the vocabulary by including newly introduced synonymous keywords from Social Annotation Systems.

Rich User Profiling

Providing the user with a *personalized* experience is an important goal in many Web applications. Ranging from shopping recommendations to search results, a concise representation of a user's interest is hereby a great advantage. While semantic techniques are an obvious candidate (e.g., (Middleton et al., 2001)), other studies have explored the value of Social Annotations for web search personalization (Noll and Meinel, 2007). Because within this field, the prompt reaction to changes in user interests is especially important, the adaptivity of usage-driven semantics derived from Social Annotation Systems could be especially useful. One could imagine, e.g., to complete the set of terms which depict a user's interest by learned synonym keywords, or to sharpen his profile by disambiguating the latter.

9.3. Summary

This chapter was intended to convey an impression to the reader how the insights gained from the previous studies can be exploited by concrete applications. These concerned mainly two directions, namely (i) enhancing Social Annotation Systems themselves, and (ii) transferring the learned semantic structures to external applications. For the first class, *integration tools* like a JabRef plugin or a REST-API were described as means to foster the emergence of semantics, similar to *tag recommendation engines* which stimulate the rich annotation. Furthermore, it was explained how inherent weaknesses of Social Annotation Systems could be tackled by applications in the field of semantic search, vocabulary maintenance and semantic browsing. Apart from those methods of "feeding back" the learned semantics into their originating systems, in the following external applications were highlighted which could benefit from the latter. These were mainly found in the fields of optimizing web search engines, creating and maintaining ontologies and supporting personalization by enhanced user profiling techniques.

With these explanations, the chapter on applications is closing. The following chapter summarizes the overall contributions of this dissertation, and discusses which questions could not be answered within its context, before closing with concluding remarks.

Chapter 10.

Conclusions

The core topic of this dissertation was to explore possible synergies between two contrasting paradigms of Knowledge Organization and Engineering on the Web, namely Social Annotations and the Semantic Web. The pursued direction hereby was to explore the possibility to “mediate” between both by means of ontology learning methods which capture emergent semantics. The final chapter is now intended to briefly summarize the results of each major part, and then to clarify the contributions of this dissertation. An outlook to promising further research directions will also be given.

10.1. Summary

The first part of this dissertation was concerned with laying the groundwork by introducing Social Annotations and the Semantic Web as two paradigms of organizing information resources on the World Wide Web. Hereby the individual characteristics were discussed, and especially the respective strengths and weaknesses were highlighted: While Social Annotations represent an immediately useful service, which has demonstrated its potential to engage large user populations into the process of resource annotation, its inherent uncontrolled nature and lack of structure can be problematic. On the other hand, ontologies as a core element of the Semantic Web offer a precise and unambiguous knowledge representation paradigm, but suffer from the knowledge acquisition and annotation bottleneck. This apparently “inverse” relation of strengths and weaknesses was the main motivation behind the subsequently proposed approach to bridge between both worlds by means of ontology learning methods. The main difference between Social Annotation data and “traditional” input of ontology learning algorithms was then explained to consist mainly in (i) the motivation of contributors, (ii) the communication among contributors and (iii) the requirements for contribution. In the following coverage of relevant work

and the state of the art in this field, the existing works were categorized based on a set of comparison dimensions, namely which data source they exploit, what kind of data filtering they pursue, which learning technique is used to address which learning task, and finally which kind of evaluation is performed. An outcome hereby was that a systematic and comparative study of the applicability of ontology learning methods to capture emergent keyword semantics is largely missing. With the goal to address this gap, the precise approach of this dissertation was elaborated.

After introducing which Social Annotation Systems (e. g., BibSonomy and Delicious) were used as objects of investigation, and describing a set of gold-standard ontologies for evaluation purposes, the following core chapter started with a clarification which learning tasks are specifically addressed. For this purpose, an adapted version of the ontology learning layer cake was proposed, consisting mainly of four parts: (i) Measures of semantic relatedness, (ii) concept learning, (iii) measures of semantic generality and (iv) concept hierarchy learning. Then, several methods and approaches for each task were introduced and compared, using mostly a reference-based evaluation paradigm.

For the case of *semantic relatedness*, distributional measures based on different context definitions were compared to co-occurrence and graph-based variants. It turned out that the *tag context relatedness*, which captures the semantics of a keyword based on its co-occurrence distribution with other keywords, is a semantically precise and computationally feasible metric across a variety of datasets. The analysis of alternative aggregation and weighting schemes identified pointwise positive mutual information (PPMI) to have a positive effect. For some datasets, alternative similarity measures like the α -skew divergence showed better results.

These insights were the basis of subsequently introduced methods to discover *concepts*, firstly using synonymy relations among keywords. Hereby hierarchical clustering, together with an appropriate choice of an inconsistency coefficient threshold, seemed to lead to two natural steps of synonym grouping. For the purpose of discovering different senses of a given keyword, different kinds of clustering techniques which partition its context were analyzed. Among them, hierarchical clustering with a distance criterion threshold led to the best results, measured by the ability to reproduce manually defined senses from WordNet and a disambiguated dataset.

As a prerequisite to infer hierarchical structures, the next step was the study of measures of *semantic generality*, which capture the level of “abstractness”

of a given keyword. Hereby we reported mixed results of different measures among datasets, but two consistent observations where (i) that measures based on a co-occurrence network were better suited than those based on similarity networks, and (ii) that computationally less complex measures (like, e. g., the frequency count or degree centrality) already yielded good approximations of semantic generality.

The final step in the layered model of capturing emergent semantics was then to include concept hierarchies from the initially flat keyword space. For this purpose, two hierarchical clustering algorithm variants were compared to generality-based algorithms, which were specifically designed for Social Annotation data. Based on a comparison of the outcomes against several reference taxonomies, the latter showed a better performance. This was also confirmed by a user study, involving human consistency assessments of the learned hierarchical relations. In addition, the consideration of the learned synonyms and polysemous keywords led to a further improvement of the results.

Shifting our attention away from specific methods, the next step was to study which factors lead to the evolution of better (or worse) implicit semantic structures. Here we analyzed a set of keyword properties, finding out that a higher degree of usage as well as a careful normalization with respect to interface artifacts have beneficial effects. From a different perspective, we then analyzed pragmatic aspects, asking the question if the way *how* keywords are used could potentially affect *what* tags mean. Based on a broad distinction of annotators according to their habits into *categorizers* and *describers*, the interesting outcome was that the verbosity of describers led to a faster and clearer emergence of keyword semantics – which even outperformed the semantic precision of the complete data in some cases. This provided evidence for a causal link between tagging pragmatics and semantics, which should be considered when trying to capture emergent semantics. As a last step, because malicious users are attracted by the popularity of Social Annotation Systems, we were interested if their activity could be exploited as well. In general, we observed hereby evidences for both a detrimental (by introducing noisy relations) and beneficial (by providing simply more data) influence. We concluded that a choice respecting the individual characteristics of a specific dataset should be preferred to an exclusion of spammer data as a matter of principle.

Based on the insights gained in the methodological chapters, we then highlighted a set of potential applications. These were broadly found in two areas: First, in the field enhancing Social Annotation Systems themselves, either by

stimulating the emergence of semantics (using, e. g., integration tools like the JabRef Plugin or the REST-API) or feeding back the learned semantics into the originating system (using, e. g., semantic search or browsing facilities). Second, we focused on applications targeted towards improving semantic applications; hereby we mentioned the areas of optimizing web search engines, ontology maintenance and rich user profiling.

In the next section, we will relate these results to the overall contributions of this dissertation.

10.2. Contributions and Outlook

This work has contributed to the advancement of the state of the art in the field of analyzing emergent semantics in various ways. Specifically, we see the contributions hereby mainly in the following fields:

- We presented a *methodology* of semantic grounding, which allows to assess the extent to which semantic structures derived from Social Annotation Systems resemble those which are defined in existing semantic resources like ontologies. This methodology is furthermore suitable to analyze different methods and algorithms to make these structures explicit, especially regarding (i) their ability to capture them in a precise way and (ii) the question which *kind* of relations are preferentially captured by a particular approach. This allows a more principled choice of methods for a given task.
- Based on the aforementioned methodology, we performed extensive *empirical* studies to examine which methods from the field of ontology learning and related areas are applicable to the domain of Social Annotation systems. Those were validated on a variety of datasets which are characterized by different properties, in order to assure that the findings can be generalized. As a result, we identified a set of suitable tools for various purposes related to capturing semantic structures, which can be directly applied in the context of further analyses.
- Finally, we have contributed to a deeper *understanding* of the phenomenon of emergent semantics within Social Annotation data by analyzing several influencing factors. Hereby we provided especially empirical evidence

for a causal link between tagging pragmatics (i. e., different annotation patterns) and the resulting keyword semantics. These insights can on the one hand be used to further fine-tune and optimize the capturing methods, but are on the other hand also a valuable input to operators of Social Annotation Systems which seek to stimulate the emergence of semantic structures within their platforms.

Instead of “concluding” the research direction pursued in this dissertation, our results of course open up several interesting areas for future research. We expect the most promising starting points in the fields described below.

Characterization of Emergent Semantics from Different Sources

While we provided a systematic analysis of a variety of Social Annotation Systems, the phenomenon of emergent semantics is not limited to those, but can be observed in other information systems as well (Aberer et al., 2004). Especially as pervasive technologies and ubiquitous connectivity are entering more and more parts of our professional and personal lives, we think that the resulting growing amount of digital traces of humans interacting with information resources exhibits similar dynamics. Because it would surely be too narrow to treat all such systems the same, the question arises which *kind* of semantics emerges from a particular sort of system. As an example, microblogging platforms like Twitter might be appropriate for mining a more ephemeral kind of knowledge, compared to what we could harvest from, e. g., Delicious. We expect a great benefit from a comparative study to this end (potentially based on the methodology proposed within this dissertation), ultimately envisioning some kind of framework by which emergent semantics from different systems can be characterized.

Further Refinement of Captured Semantics

Although some of our proposed methods reach a remarkable semantic “precision”, there is still room for the overall improvement of the quality of the learned semantics. Apart from further optimizing the methods themselves, an interesting question is if an *iterative* refinement approach is feasible. The iteration steps could hereby take place on various levels: First, an idea would be to, e. g., compute an iteratively refined synsetized folksonomy, i. e., starting with synsets which consist of the original keywords, and then to recompute co-occurrence and the synonymy measures to derive “synsets of synsets” in a second step. The

hope hereby is that the repetition of this process converges to a semantically more precise system, compared to the initial state.

Combination of Evidences of Emergent Semantics

Within the scope of this dissertation, exclusively the network structure of Social Annotation Systems was exploited to derive emergent semantic relations. While this was done in order to ensure the applicability to different kinds of system and resource types, one might expect that the consideration of additional information sources leads to better results. Hereby it would be especially interesting to investigate how *several* information sources can be combined – as an example, in a system which is made up of textual content, and offers the definition of keyword relations, the integration of all these evidences of semantics into a single model might be beneficial.

Tighter Integration of Social Annotation Systems and Semantic Resources

While the approaches presented in this thesis can contribute to “bridge” between Social Annotation Systems and semantic resources, the question how exactly both can be integrated was not elaborated in detail. Interesting issues to this end are, e. g., how to define interfaces by which the user interacts with the learned semantics, or how to ensure quality within the process of ontology emergence. Furthermore it would be interesting to study how the process of “feeding back” the learned semantics influences the further development of a Social Annotation System. While we expect further synergies from a reciprocal improvement of Social Annotations and ontologies, this assumption would have to be validated in further studies.

10.3. Closing Remarks

Social Annotations should not be seen as the “silver bullet” of knowledge organization on the Web, and neither should their emergent semantic structures be understood as an intended ultimate replacement for ontologies. However, their popularity has shown up important requirements of humans interacting with growing amounts of information resources, which were obviously not met by prior approaches. In this light, Social Annotations along with the emergent

10.3. Closing Remarks

semantics should be interpreted as a further mosaic piece, contributing to complete the picture how to design an intelligent Social Semantic Web.

Bibliography

- R. Abbasi and S. Staab. RichVSM: enRiched Vector Space Models for Folksonomies. In *Proceedings of 20th ACM conference on Hypertext and Hypermedia (HT2009)*, 2009.
- K. Aberer, P. Cudré-Mauroux, A. M. Ouksel, T. Catarci, M.-S. Hacid, A. Illarramendi, V. Kashyap, M. Mecella, E. Mena, E. J. Neuhold, O. D. Troyer, T. Risse, M. Scannapieco, F. Saltor, L. D. Santis, S. Spaccapietra, S. Staab, and R. Studer. Emergent semantics principles and issues. In Y.-J. Lee, J. Li, K.-Y. Whang, and D. Lee, editors, *Proceedings of the 9th International Conference on Database Systems for Advanced Applications (DASFAA2004)*, volume 2973 of *Lecture Notes in Computer Science*, pages 25–38. Springer, 2004. ISBN 3-540-21047-4.
- H. S. Al-Khalifa and H. C. Davis. Folksonomies versus automatic keyword extraction: An empirical study. *IADIS International Journal on Computer Science and Information Systems (IJCSIS)*, 1:132–143, Oct. 2006.
- H. S. Al-Khalifa and H. C. Davis. Exploring the value of folksonomies for creating semantic metadata. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 3:13–39, 2007.
- R. B. Allen and Y. Wu. Generality of texts. In E. Lim, S. Foo, C. Khoo, H. Chen, E. Fox, S. Urs, and T. Costantino, editors, *Digital Libraries: People, Knowledge, and Technology*, volume 2555 of *Lecture Notes in Computer Science*, pages 111–116. Springer, Berlin / Heidelberg, 2010. doi: 10.1007/3-540-36227-4_11.
- J. R. Anderson. *Kognitive Psychologie*. Spektrum Akademischer Verlag, 3rd edition, 2001. ISBN 382741024X.
- S. Angeletou. Semantic enrichment of folksonomy tagspaces. In A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, and K. Thirunarayan, editors, *Proceedings of the 7th International Semantic Web*

- Conference (ISWC2008)*, volume 5318 of *Lecture Notes in Computer Science*, pages 889–894. Springer, 2008. ISBN 978-3-540-88563-4.
- S. Angeletou, M. Sabou, and E. Motta. Semantically enriching folksonomies with flor. In *Proceedings of the CISWeb Workshop*, 2008.
- C. M. Au Yeung. *From User Behaviours to Collective Semantics*. PhD thesis, University of Southampton, 2009.
- C. M. Au Yeung, N. Gibbins, and N. Shadbolt. Understanding the semantics of ambiguous tags in folksonomies. In *Proceedings of the International Workshop on Emergent Semantics and Ontology Evolution (ESOE2007)*, 2007.
- C. M. Au Yeung, N. Gibbins, and N. Shadbolt. Web search disambiguation by collaborative tagging. In *Proceedings of the Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR2008)*, pages 48–61, 2008.
- C. M. Au Yeung, N. Gibbins, and N. Shadbolt. Contextualising tags in collaborative tagging systems. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia (HT2009)*, pages 251–260, New York, NY, USA, 2009a. ACM. ISBN 978-1-60558-486-7. doi: 10.1145/1557914.1557958.
- C. M. Au Yeung, M. G. Noll, N. Gibbins, C. Meinel, and N. Shadbolt. On measuring expertise in collaborative tagging systems. In *Proceedings of the 1st Web Science Conference (WebSci2009)*, Mar. 2009b.
- S. Auer and J. Lehmann. What have innsbruck and leipzig in common? extracting semantics from wiki content. In *Proceedings of the 4th European Semantic Web Conference (ESWC2007)*, pages 503–517, 2007.
- K. Bade. *Personalized Hierarchical Structuring*. PhD thesis, Otto-von-Guericke-Universität, Magdeburg, 2009.
- K. Bade and D. Benz. Evaluation strategies for learning algorithms of hierarchical structures. In *Proceedings of the 32nd Annual Conference of the German Classification Society - Advances in Data Analysis, Data Handling and Business Intelligence (GfKI2008)*, Studies in Classification, Data Analysis, and Knowledge Organization, Berlin-Heidelberg, 2008. Springer.

- S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *Proceedings of the 16th International Conference on World Wide Web (WWW2007)*, pages 501–510, New York, NY, USA, 2007. ACM Press. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242640.
- G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Proceedings of the Collaborative Web Tagging Workshop*, Edinburgh, Scotland, May 2006.
- D. Benz. Collaborative ontology learning. Master’s thesis, Albert-Ludwigs-Universität Freiburg, Department of Computer Science, 2007.
- T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.
- C. Biemann. Ontology learning from text: A survey of methods. *LDV Forum*, 20(2):75–93, 2005.
- V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008 (12pp), 2008.
- M. Bork. Webservice API für bibsonomy. Project report, 2006.
- U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.
- J. Brank, M. Grobelnik, and D. Mladenić. A survey of ontology evaluation techniques. In *Proceedings of the Conference on Data Mining and Data Warehouses*, pages 166–169, 2005.
- J. Brank, D. Madenic, and M. Groblenik. Gold standard based ontology evaluation using instance assignment. In *Proceedings of the 4th Workshop on Evaluating Ontologies for the Web (EON2006)*, Edinburgh, Scotland, May 2006.
- S. Braun, A. Schmidt, A. Walter, and V. Zacharias. The ontology maturing approach to collaborative and work-integrated ontology development: Evaluation results and future directions. In *Proceedings of the International Workshop on Emergent Semantics and Ontology Evolution (ESOE2007)*, 2007a.

- S. Braun, A. Schmidt, and V. Zacharias. Soboleo: vom kollaborativen tagging zur leichtgewichtigen ontologie. In T. Gross, editor, *Proceedings of Mensch & Computer - 7. Fachübergreifende Konferenz (M&C2007)*, pages 209–218, München, 2007b. Oldenbourg Verlag. ISBN 978-3-486-58496-7.
- K. Breitman, M. A. Casanova, and W. Truszkowski. *Semantic Web: Concepts, Technologies and Applications (NASA Monographs in Systems and Software Engineering)*. Springer London, 1st edition, 2007. ISBN 184628581X.
- S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, Apr. 1998. doi: 10.1016/S0169-7552(98)00110-X.
- C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proceedings of the 15th International Conference on World Wide Web (WWW2006)*, pages 625–632, New York, NY, USA, 2006. ACM Press.
- A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguists*, 32(1):13–47, 2006.
- J. A. Bullinaria and J. P. Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510 – 526, 2007.
- I. Cantador, M. Szomszor, H. Alani, M. Fernandez, and P. Castells. Enriching ontological user profiles with tagging history for multi-domain recommendations. In *Proceedings of the 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb2008)*, June 2008.
- C. Cattuto. Semiotic dynamics in online social communities. *The European Physical Journal C - Particles and Fields*, 46:33–37, Aug. 2006.
- C. Cattuto, C. Schmitz, A. Baldassarri, V. D. P. Servedio, V. Loreto, A. Hotho, M. Grahl, and G. Stumme. Network properties of folksonomies. *AI Communications*, 20(4):245–262, Dec. 2007. ISSN 0921-7126.
- D. Chandler. *Semiotics: The Basics*. Taylor & Francis, 2nd edition, 2007. ISBN 978-0415363754.

- W. Cheng, M. Rademaker, B. D. Baets, and E. Hüllermeier. Predicting partial orders: Ranking with abstention. In J. L. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, editors, *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD2010)*, volume 6321 of *Lecture Notes in Computer Science*, pages 215–230. Springer, 2010. ISBN 978-3-642-15879-7.
- P. A. Chirita, S. Costache, W. Nejdl, and S. Handschuh. P-tag: large scale automatic generation of personalized annotation tags for the web. In *Proceedings of the 16th International Conference on World Wide Web (WWW2007)*, pages 845–854, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242686.
- R. Cilibrasi and P. M. B. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19:370, 2007.
- P. Cimiano. *Ontology learning and population from text - algorithms, evaluation and applications*. Springer, 2006. ISBN 978-0-387-30632-2.
- P. Cimiano, A. Mädche, S. Staab, and J. Völker. Ontology learning. In S. Staab and R. Studer, editors, *Handbook on Ontologies*, International Handbooks Information System, pages 245–267. Springer Berlin Heidelberg, 2009. ISBN 978-3-540-92673-3.
- J. M. Clark and A. Paivio. Extensions of the paivio, yuille, and madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):371, 2004. ISSN 1554-3528.
- O. Corcho and A. Gomez-Perez. A roadmap to ontology specification languages. In R. Dieng and O. Corby, editors, *Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW2000)*, volume 1937 of *Lecture Notes in Computer Science*, pages 80–96. Springer, 2000. ISBN 3-540-41119-4.
- P. Cudré-Mauroux. Definition of emergent semantics in the encyclopedia of database systems, 2009. URL <http://people.csail.mit.edu/pcm/papers/EmergentSemantics.pdf>.

- J. Cullen and A. Bryman. The knowledge acquisition bottleneck: Time for reassessment? *Expert Systems*, 5(3):216–225, 1988. doi: 10.1111/j.1468-0394.1988.tb00065.x.
- F. de Saussure. *Course in General Linguistics*. Duckworth, London, 1916.
- S. Deerwester. Improving information retrieval with latent semantic indexing. In C. L. Borgman and E. Y. H. Pai, editors, *Proceedings of the 51st ASIS Annual Meeting (ASIS1988)*, volume 25, Atlanta, Georgia, Oct. 1988. American Society for Information Science.
- K. Dellschaft and S. Staab. On how to perform a gold standard based evaluation of ontology learning. In *Proceedings of the 5th International Semantic Web Conference (ISWC2006)*, Athens, GA, USA, Nov. 2006. Springer, LNCS.
- K. Dellschaft and S. Staab. An epistemic dynamic model for tagging systems. In *Proceedings of the 19th ACM conference on Hypertext and hypermedia (HT2008)*, pages 71–80, New York, NY, USA, 2008. ACM. ISBN 978-1-59593-985-2. doi: 10.1145/1379092.1379109.
- I. S. Dhillon, J. Fan, and Y. Guan. Efficient clustering of very large document collections. In R. Grossman, C. Kamath, and R. Naburu, editors, *Data Mining for Scientific and Engineering Applications*. Kluwer Academic Publishers, Heidelberg, 2001.
- B. Dorow and D. Widdows. Discovering corpus-specific word senses. In *Proceedings of the 10th conference on European chapter of the Association for Computational Linguistics*, volume 2 of *EACL '03*, pages 79–82, Morristown, NJ, USA, 2003. Association for Computational Linguistics. ISBN 1-111-56789-0. doi: <http://dx.doi.org/10.3115/1067737.1067753>.
- I. A. Doush and E. Pontelli. Integrating semantic web and folksonomies to improve e-learning accessibility. In K. Miesenberger, J. Klaus, W. Zagler, and A. Karshmer, editors, *Computers Helping People with Special Needs*, volume 6179 of *Lecture Notes in Computer Science*, pages 376–383. Springer Berlin / Heidelberg, 2010.
- T. Eda, M. Yoshikawa, T. Uchiyama, and T. Uchiyama. The effectiveness of latent semantic analysis for building up a bottom-up taxonomy from folksonomy tags. *World Wide Web*, 12(4):421–440, 2009.

- C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998. ISBN 978-0-262-06197-1.
- J. R. Firth. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32, 1957.
- J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.
- B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, Jan. 2007. ISSN 1095-9203. doi: 10.1126/science.1136800.
- B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, 1999.
- A. Garcia-Silva, M. Szomszor, H. Alani, and O. Corcho. Preliminary results in tag disambiguation using dbpedia. In *Proceedings of the 1st International Workshop on Collective Knowledge Capturing and Representation (CKCaR2009)*, Sept. 2009.
- A. Garcia-Silva, O. Corcho, H. Alani, and A. Gomez-Perez. Review of the state of the art: Discovering and associating semantics to tags in folksonomies. *Knowledge Engineering Review*, 26(4), Dec. 2011.
- J. Gemmell, A. Shepitsen, B. Mobasher, and R. Burke. Personalizing navigation in folksonomies using hierarchical tag clustering. In *Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery*, pages 196–205, 2008.
- E. Giannakidou, V. A. Koutsonikola, A. Vakali, and Y. Kompatsiaris. Co-clustering tags and social data sources. In *Proceedings of the 9th International Conference on Web-Age Information Management (WAIM2008)*, pages 317–324. IEEE, 2008. ISBN 978-0-7695-3185-4.
- A. Gómez-Pérez, M. Fernández-López, and O. Corcho. *Ontological Engineering*. Springer, Heidelberg, 2004.
- S. Golder and B. A. Huberman. The structure of collaborative tagging systems. *Journal of Information Sciences*, 32(2):198–208, Apr. 2006.

- A. Gomez-Perez and O. Corcho. Ontology specification languages for the semantic web. *IEEE Intelligent Systems*, 17(1):54–60, 2002. ISSN 1541-1672. doi: 10.1109/5254.988453.
- M. Grahl, A. Hotho, and G. Stumme. Conceptual clustering of social bookmarking sites. In *Proceedings of the 7th International Conference on Knowledge Management (I-KNOW 2007)*, pages 356–364, Graz, Austria, Sept. 2007. Know-Center.
- M. Grineva, M. Grinev, D. Turdakov, and P. Velikhov. Harnessing wikipedia for smart tags clustering. In *Proceedings of the International Workshop on Knowledge Acquisition from the Social Web (KASW2008)*, 2008.
- N. Guarino, editor. *Proceedings of the 1st International Conference on Formal Ontology and Information Systems*, 1998. IOS Press.
- N. Guarino, D. Oberle, and S. Staab. What is an ontology? In S. Staab, D. R. Studer, P. Bernus, J. Błażewicz, G. J. Schmidt, and M. J. Shaw, editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 1–17. Springer Berlin Heidelberg, 2009. ISBN 978-3-540-92673-3.
- H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th International World Wide Web Conference (WWW2007)*, New York, NY, USA, 2007. ACM Press.
- M. Hamasaki, Y. Matsuo, and T. Nisimura. Ontology extraction using social network, 2007.
- T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (i) - a general review. *D-Lib Magazine*, 11(4), Apr. 2005.
- Z. S. Harris. *Mathematical Structures of Language*. Wiley, New York, 1968.
- F. Hayes-Roth, D. A. Waterman, and D. B. Lenat. *Building expert systems*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1983. ISBN 0-201-10686-8.
- M. Heckner, M. Heilemann, and C. Wolff. Personal information management vs. resource sharing: Towards a model of information behaviour in social tagging systems. In *Proceedings of the 3rd International Conference on Weblogs and Social Media (ICWSM2009)*, San Jose, CA, USA, May 2009.

- D. Helic, M. Strohmaier, C. Trattner, M. Muhr, and K. Lerman. Pragmatic evaluation of folksonomies. In *Proceedings of the 20th International World Wide Web Conference (WWW2011)*, 2011.
- J. Hendler. Agents and the semantic web. *IEEE Intelligent Systems*, 16:30–37, Mar. 2001. ISSN 1541-1672. doi: 10.1109/5254.920597.
- U. Hengartner and A. Meier, editors. *Web 3.0 & Semantic Web*. Number 271 in HMD – Praxis der Wirtschaftsinformatik. dpunkt, Heidelberg, 2010. ISBN 978-3-89864-624-6.
- P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Computer Science Department, Stanford University, Apr. 2006.
- P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR2008)*, pages 531–538. ACM, 2008. doi: 10.1145/1390334.1390425.
- P. Heymann, A. Paepcke, and H. Garcia-Molina. Tagging human knowledge. In B. D. Davison, T. Suel, N. Craswell, and B. Liu, editors, *Proceedings of the 3rd International Conference on Web Search and Data Mining (WSDM2010)*, pages 51–60. ACM, 2010. ISBN 978-1-60558-889-6.
- B. Hjørland. What is knowledge organization (KO)? *Knowledge organization*, 35(2-3):86–101, 2008.
- G. M. Hodge. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Commission on Preservation &, 2000. ISBN 1887334769.
- A. Hotho and B. Hoser, editors. *Proceedings of the International Workshop on Bridging the Gap between Semantic Web and Web 2.0*, June 2007.
- A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Emergent semantics in bibsonomy. In C. Hochberger and R. Liskowsky, editors, *Informatik 2006 – Informatik für Menschen. Band 2*, volume P-94 of *Lecture Notes in Informatics*, Bonn, Oct. 2006a. Gesellschaft für Informatik.

- A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In Y. Sure and J. Domingue, editors, *Proceedings of the 3rd European Semantic Web Conference (ESWC2006)*, volume 4011 of *Lecture Notes in Computer Science*, pages 411–426, Heidelberg, June 2006b. Springer.
- A. Hotho, D. Benz, R. Jäschke, and B. Krause, editors. *ECML PKDD Discovery Challenge 2008 (RSDC'08)*, 2008. Workshop at 18th Europ. Conf. on Machine Learning (ECML'08) / 11th Europ. Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'08).
- E. K. Jacob. Classification and categorization: a difference that makes a difference. *Library Trends*, 52(3):515 – 540, 2004.
- A. K. Jain, N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31:264–323, Sept. 1999. ISSN 0360-0300. doi: 10.1145/331499.331504.
- J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference Research on Computational Linguistics (ROCLING)*, Taiwan, 1997.
- R. Jäschke, A. Hotho, C. Schmitz, B. Ganter, and G. Stumme. Discovering shared conceptualizations in folksonomies. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1):38–53, Feb. 2008a. ISSN 1570-8268. doi: 10.1016/j.websem.2007.11.004.
- R. Jäschke, B. Krause, A. Hotho, and G. Stumme. Logsonomy – a search engine folksonomy. In *Proceedings of the 2nd International Conference on Weblogs and Social Media(ICWSM2008)*. AAAI Press, 2008b.
- J. Jung. Matching multilingual tags based on community of lingual practice from multiple folksonomy: A preliminary result. In N. García-Pedrajas, F. Herrera, C. Fyfe, J. Benítez, and M. Ali, editors, *Trends in Applied Intelligent Systems*, volume 6097 of *Lecture Notes in Computer Science*, pages 39–46. Springer, Berlin / Heidelberg, 2010. doi: 10.1007/978-3-642-13025-0_5.
- R. Kammann and L. Streeter. Two meanings of word abstractness. *Journal of Verbal Learning and Verbal Behavior*, 10(3):303 – 306, 1971. ISSN 0022-5371. doi: 10.1016/S0022-5371(71)80058-0.

- L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *Proceedings of the 15th International Conference on Multimedia (MULTIMEDIA2007)*, pages 631–640, New York, NY, USA, 2007. ACM. ISBN 9781595937025. doi: 10.1145/1291233.1291384.
- S. H. Kome. Hierarchical subject relationships in folksonomies. Technical report, School of Information and Library Science, Nov. 2005.
- C. Körner. Understanding the motivation behind tagging. ACM Student Research Competition - Hypertext 2009, July 2009.
- B. Krause, R. Jäschke, A. Hotho, and G. Stumme. Logsonomy - social information retrieval with logdata. In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia (HT2008)*, pages 157–166, New York, NY, USA, 2008a. ACM. ISBN 978-1-59593-985-2. doi: 10.1145/1379092.1379123.
- B. Krause, C. Schmitz, A. Hotho, and G. Stumme. The anti-social tagger - detecting spam in social bookmarking systems. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWEB2008)*, pages 61–68, New York, NY, USA, Apr. 2008b. ACM. ISBN 978-1-60558-159-0. doi: 10.1145/1451983.1451998.
- C. Körner, R. Kern, H.-P. Grahsl, and M. Strohmaier. Of categorizers and describers: an evaluation of quantitative measures for tagging motivation. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HT2010)*, pages 157–166, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0041-4. doi: 10.1145/1810617.1810645.
- H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World wide web (WWW2010)*, pages 591–600, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772751.
- K. Lee, H. Kim, H. Shin, and H.-J. Kim. Tag sense disambiguation for clarifying the vocabulary of social tags. In *Proceedings of the International Conference on Computational Science and Engineering (CSE2009)*, pages 729–734. IEEE Computer Society, 2009.

- L. Lee. Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 25–32, Morristown, NJ, USA, 1999. Association for Computational Linguistics. ISBN 1-55860-609-3. doi: <http://dx.doi.org/10.3115/1034678.1034693>.
- L. Lee. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics*, pages 65–72, 2001.
- S.-S. Lee and H.-S. Yong. Tagplus: A retrieval system using synonym tag in folksonomy. In *Proceedings of the International Conference on Multimedia and Ubiquitous Engineering (MUE2007)*, pages 294–298, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-2777-9. doi: <http://dx.doi.org/10.1109/MUE.2007.201>.
- F. Lehmann and R. Wille. A triadic approach to formal concept analysis. In *Conceptual Structures: Applications, Implementation and Theory*, volume 954 of *Lecture Notes in Computer Science*. Springer, 1995. ISBN 3-540-60161-9.
- J. Lehmann, C. Bizer, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165, 2009. doi: [doi:10.1016/j.websem.2009.07.002](https://doi.org/10.1016/j.websem.2009.07.002).
- J. Leskovec, K. J. Lang, and M. W. Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web (WWW2010)*, 2010.
- M. Levy and M. Sandler. Learning latent semantic models for music from social tags. *Journal of New Music Research*, 37(2):137–150, 2008.
- D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational linguistics*, pages 768–774, Morristown, NJ, USA, 1998. Association for Computational Linguistics. doi: <http://dx.doi.org/10.3115/980691.980696>.
- H. Lin, J. Davis, and Y. Zhou. An integrated approach to extracting ontological structures from folksonomies. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimitano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, editors, *Proceedings of the 8th International Semantic Web Conference*

-
- (ISWC2009), volume 5554 of *Lecture Notes in Computer Science*, pages 654–668. Springer, Berlin / Heidelberg, 2009. doi: 10.1007/978-3-642-02121-3_48.
- X. Lin, J. E. Beaudoin, Y. Bul, and K. Desal. Exploring characteristics of social classification. In *Proceedings 17th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research 17*, 2006.
- M. Lipczak, Y. Hu, Y. Kollet, and E. Milios. Tag sources for recommendation in collaborative tagging systems. In F. Eisterlehner, A. Hotho, and R. Jäschke, editors, *Proceedings of the ECML PKDD Discovery Challenge 2009 (DC09)*, volume 497 of *CEUR-WS.org*, pages 157–172, Sept. 2009.
- B. Lund, T. Hammond, M. Flack, and T. Hannay. Social Bookmarking Tools (II): A Case Study - Connotea. *D-Lib Magazine*, 11(4), Apr. 2005.
- M. Z. Maala, A. Delteil, and A. Azough. A conversion process from flickr tags to rdf descriptions. In D. Flejter and M. Kowalkiewicz, editors, *Proceedings of the BIS Workshop on Social Aspects of the Web*, volume 245, 2007.
- G. Macgregor and E. Mcculloch. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, 55(5):291 – 300, 2006.
- A. Maedche. *Ontology Learning for the Semantic Web*. Kluwer Academic Publishing, Boston, 2002.
- A. Maedche and S. Staab. Measuring similarity between ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web (EKAW2002)*, pages 251–263, London, UK, 2002. Springer-Verlag.
- A. Maedche and S. Staab. Ontology learning for the semantic web. *Intelligent Systems, IEEE*, 16(2):72–79, 2005. ISSN 1541-1672.
- C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, 1999.
- C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

- L. B. Marinho, K. Buza, and L. Schmidt-Thieme. Folksonomy-based collaborative learning. In A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, and K. Thirunarayan, editors, *Proceedings of the 7th International Semantic Web Conference (ISWC2008)*, volume 5318 of *Lecture Notes in Computer Science*, pages 261–276. Springer, 2008. ISBN 978-3-540-88563-4.
- C. Marlow, M. Naaman, D. Boyd, and M. Davis. Position paper, tagging, taxonomy, flickr, article, toread. In *Proceedings of the Collaborative Web Tagging Workshop*, Edinburgh, Scotland, May 2006.
- A. Mathes. Folksonomies - cooperative classification and communication through shared metadata, Dec. 2004. URL <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>.
- D. L. McGuinness. Ontologies come of age. In D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, editors, *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*. MIT Press, 2003.
- M. Meder. Multi-domain klassifikation basierend auf nutzergenerierten metadaten. Master's thesis, Technische Universität Berlin, 2010.
- P. D. Meo, G. Quattrone, and D. Ursino. Exploitation of semantic relationships and hierarchical data structures to support a user in his annotation and browsing activities in folksonomies. *Information Systems*, 34(6):511–535, 2009. ISSN 0306-4379. doi: <http://dx.doi.org/10.1016/j.is.2009.02.004>.
- E. Michlmayr. A case study on emergent semantics in communities. In *Proceedings of the Workshop on Social Network Analysis*, Nov. 2005.
- S. E. Middleton, D. C. De Roure, and N. R. Shadbolt. Capturing knowledge of user preferences: ontologies in recommender systems. In *Proceedings of the International Conference on Knowledge capture (KCAP2001)*, pages 100–107. ACM Press, 2001. ISBN 1-58113-380-4.
- P. Mika. Ontologies are us: A unified model of social networks and semantics. In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *Proceedings of the 4th International Semantic Web Conference (ISWC2005)*, volume 3729 of *Lecture Notes in Computer Science*, pages 522–536. Springer, 2005.
- D. Millen, J. Feinberg, and B. Kerr. Social bookmarking in the enterprise. *Queue*, 3(9):28–35, Nov. 2005.

- G. Mishne. Autotag: a collaborative approach to automated tag assignment for weblog posts. In *Proceedings of the 15th International Conference on World Wide Web (WWW2006)*, pages 953–954, New York, NY, USA, 2006. ACM Press.
- S. Mohammad and G. Hirst. Distributional measures as proxies for semantic distance: A survey. *Computational Linguistics*, 1(1), 2006.
- NCSA. About ncsa mosaic. <http://www.ncsa.illinois.edu/Projects/mosaic.html>, 2011. Retrieved on 2011/08/03.
- M. Noll and C. Meinel. Web search personalization via social bookmarking and tagging. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, G. Schreiber, and P. Cudré-Mauroux, editors, *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC2007)*, volume 4825 of *LNCS*, pages 365–378, Berlin, Heidelberg, Nov. 2007. Springer Verlag.
- B. Omelayenko. Learning of ontologies for the web: the analysis of existent approaches. In *Proceedings of the International Workshop on Web Dynamics*, 2001.
- A. Paivio, J. C. Yuille, and S. A. Madigan. Concreteness, imagery and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76(1, Part 2):1 – 25, 1968. ISSN 0022-1015. doi: 10.1037/h0025327.
- P. Pantel and D. Lin. Discovering word senses from text. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Canada, 2002.
- A. Passant. Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs. In *Proceedings of the 1st International Conference on Weblogs and Social Media (ICWSM2007)*, Boulder, Colorado, Mar. 2007.
- A. Passant and P. Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *Proceedings of Workshop Linked Data on the Web (LDOW2008)*, 2008.
- O. Patashnik. *BibTeXing*, Feb. 1988.

- A. Plangprasopchok and K. Lerman. Constructing folksonomies from user-specified relations on flickr. In J. Quemada, G. León, Y. S. Maarek, and W. Nejdl, editors, *Proceedings of the 18th international conference on World wide web (WWW2009)*, pages 781–790. ACM, 2009. ISBN 978-1-60558-487-4.
- A. Plangprasopchok, K. Lerman, and L. Getoor. A probabilistic approach for learning folksonomies from structured data. In *Proceedings of the 4th ACM Web Search and Data Mining Conference (WSDM2010)*, 2010.
- S. P. Ponzetto and M. Strube. Deriving a large-scale taxonomy from wikipedia. In *Proceedings of the 22nd national conference on Artificial intelligence (AAAI2007)*, pages 1440–1445. AAAI Press, 2007. ISBN 978-1-57735-323-2.
- E. Quintarelli. Folksonomies: power to the people, June 2005.
- E. Quintarelli, L. Rosati, and A. Resmini. Facetag: Integrating bottom-up and top-down classification in a social tagging system. In *Proceedings of the IA Summit*, 2007.
- J. Radelaar, A.-J. Boor, D. Vandic, J.-W. van Dam, F. Hogenboom, and F. Frasincar. Improving the exploration of tag spaces using automated tag clustering. In S. Auer, O. Díaz, and G. Papadopoulos, editors, *Web Engineering*, volume 6757 of *Lecture Notes in Computer Science*, pages 274–288. Springer Berlin / Heidelberg, 2011.
- D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina. Clustering the tagged web. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM2009)*, pages 54–63, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-390-7. doi: 10.1145/1498759.1498809.
- R. Rapp. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of the 19th international conference on Computational linguistics (COLING2002)*, 2002.
- R. Rapp. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the 9th Machine Translation Summit*, pages 315–322, 2003.
- P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th international joint conference on Artificial intelligence (IJCAI1995)*, volume 1, pages 448–453, San Francisco,

- CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-363-8, 978-1-558-60363-9.
- M. Sabou, C. Wroe, C. Goble, and H. Stuckenschmidt. Learning domain ontologies for semantic web service descriptions. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(4):340–365, Dec. 2005.
- G. Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.
- M. Sanderson and W. B. Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR1999)*, pages 206–213, 1999.
- P. Schmitz. Inducing ontology from flickr tags. In *Proceedings of the Workshop on Collaborative Web Tagging*, Edinburgh, Scotland, May 2006.
- M. Shamsfard and A. A. Barforoush. The state of the art in ontology learning: a framework for comparison. *The Knowledge Engineering Review*, 18(04): 293–316, 2003. ISSN 0269-8889.
- K. Shen and L. Wu. Folksonomy as a complex network, Sept. 2005.
- C. Shirky. Ontology is overrated: Categories, links and tags, May 2005. URL http://www.shirky.com/writings/ontology_overrated.html.
- X. Si and M. Sun. Disambiguating tags in blogs. In V. Matousek and P. Mautner, editors, *Text, Speech and Dialogue*, volume 5729 of *Lecture Notes in Computer Science*, pages 139–146. Springer, 2009. ISBN 978-3-642-04207-2.
- S. Siersdorfer and S. Sizov. Social recommender systems for web 2.0 folksonomies. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia (HT2009)*, pages 261–270, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-486-7. doi: 10.1145/1557914.1557959.
- R. Sinha. A cognitive analysis of tagging, 2005. URL http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html.

- L. Specia and E. Motta. Integrating folksonomies with the semantic web. In *Proceedings of the 4th European Semantic Web Conference (ESWC2007)*, volume 4519/2007 of *Lecture Notes in Computer Science*, pages 624–639. Springer Berlin / Heidelberg, 2007.
- M. Strohmaier, D. Helic, D. Benz, C. Körner, and R. Kern. Evaluation of folksonomy induction algorithms. *Transactions on Intelligent Systems and Technology*, 2011. to appear.
- M. Strube and S. P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI2006)*. AAAI Press, 2006.
- R. Studer, R. R. Benjamins, and D. Fensel. Knowledge Engineering: Principles and Methods. *Data Knowledge Engineering*, 25(1-2):161–197, 1998.
- F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web (WWW2007)*, pages 697–706, New York, NY, USA, 2007. ACM Press. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242667.
- M. Tatu and D. I. Moldovan. Inducing ontologies from folksonomies using natural language understanding. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the International Conference on Language Resources and Evaluation (LREC2010)*. European Language Resources Association, 2010. ISBN 2-9517408-6-7.
- L. D. V. Terrientes, A. Moreno, and D. Sánchez. Discovery of relation axioms from the web. In Y. Bi and M.-A. Williams, editors, *Proceedings of the 4th international conference on Knowledge science (KSEM2010)*, volume 6291 of *Lecture Notes in Computer Science*, pages 222–233. Springer, 2010. ISBN 978-3-642-15279-5.
- M. Tesconi, F. Ronzano, A. Marchetti, and S. Minutoli. Semantify del.icio.us: Automatically turn your tags into senses. In *Proceedings of the Workshop Social Data on the Web (SDoW2008)*, 2008.
- P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research (JAIR)*, 37:141–188, 2010.

-
- M. Uschold and M. Grüninger. Ontologies and semantics for seamless connectivity. *SIGMOD Record*, 33(4):58–64, 2004.
- T. Vander Wal. Explaining and showing broad and narrow folksonomies, Feb. 2005. URL <http://vanderwal.net/random/entrysel.php?blog=1635>.
- C. Veres. The language of folksonomies: What tags reveal about user classification. In *Natural Language Processing and Information Systems*, volume 3999/2006 of *Lecture Notes in Computer Science*, pages 58–69, Berlin / Heidelberg, July 2006. Springer.
- J. Voss. Collaborative thesaurus tagging the wikipedia way. Eprint, 2006. URL <http://arxiv.org/abs/cs/0604036>.
- J. Voss. Tagging, folksonomy & co. renaissance of manual indexing? In *Proceedings of the 10th International Symposium for Information Science (ISI2007)*, 2007.
- C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proceedings of the Semantic Search 2010 Workshop (SemSearch2010)*, Apr. 2010.
- S. Wasserman and K. Faust. *Social network analysis: Methods and applications*. Cambridge University Press, 1994.
- J. Weeds, D. Weir, and D. McCarthy. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING2004)*, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics. doi: 10.3115/1220355.1220501.
- K. Weller. *Knowledge Representation in the Social Semantic Web*. K G Saur Verlag, 1st edition, 2010. ISBN 3598251807.
- K. Weller and I. Peters. Seeding, weeding, fertilizing. different tag gardening activities for folksonomy maintenance and enrichment. In S. Auer, S. Schaffert, and T. Pellegrini, editors, *Proceedings of the International Conference on Semantic Systems (I-SEMANTICS2008)*, pages 10–117, 2008.
- R. Wetzker. *Graph-Based Recommendations in Broad Folksonomies*. PhD thesis, Technische Universität Berlin, 2010.

- R. Wetzker, C. Zimmermann, C. Bauckhage, and S. Albayrak. I tag, you tag: translating tags for advanced user models. In B. D. Davison, T. Suel, N. Craswell, and B. Liu, editors, *Proceedings of the third ACM international conference on Web search and data mining (WSDM2010)*, pages 71–80. ACM, 2010. ISBN 978-1-60558-889-6.
- D. Widdows and B. Dorow. A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics (COLING2002)*, 2002.
- H. Wu, M. Zubair, and K. Maly. Harvesting social knowledge from folksonomies. In *Proceedings of the 17th conference on Hypertext and hypermedia (HT2006)*, pages 111–114, New York, NY, USA, 2006a. ACM Press.
- L. Wu, L. Yang, N. Yu, and X. S. Hua. Learning to tag. In *Proceedings of the 18th International Conference on World wide web (WWW2009)*, pages 361–370, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4. doi: 10.1145/1526709.1526758.
- X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *Proceedings of the 15th International Conference on the World Wide Web (WWW2006)*, pages 417–426, New York, NY, USA, 2006b. ACM Press.
- M. Xia, Y. Huang, W. Duan, and A. B. Whinston. Ballot box communication in online communities. *Communications of the ACM*, 52(9):138–142, 2009. ISSN 0001-0782. doi: 10.1145/1562164.1562199.
- L. Zhang, X. Wu, and Y. Yu. Emergent semantics from folksonomies: A quantitative study. In S. Spaccapietra, K. Aberer, and P. Cudré-Mauroux, editors, *Journal on Data Semantics VI*, volume 4090 of *Lecture Notes in Computer Science*, pages 168 – 186. Springer Berlin / Heidelberg, 2006.
- Q. Zhang. Fuzziness - vagueness - generality - ambiguity. *Journal of Pragmatics*, 29(1):13 – 31, 1998. ISSN 0378-2166. doi: DOI:10.1016/S0378-2166(97)00014-3.
- S. Zhong. Efficient online spherical k-means clustering. *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN2005)*, 5: 3180–3185, July 2005. doi: 10.1109/IJCNN.2005.1556436.

- M. Zhou, S. Bao, X. Wu, and Y. Yu. An unsupervised model for exploring hierarchical semantics from social annotations. In *Proceedings of the 7th international Semantic Web Conference (ISWC2008) and 2nd Asian semantic Web Conference (ASWC2008)*, pages 680–693, 2008.