

# The Epistemic Dynamic Model: Developing a Theory of Tagging Systems

Klaas Dellschaft  
klaasd@uni-koblenz.de

Institut für Web Science and Technologies  
Universität Koblenz-Landau

September 2012

Zur Erlangung des akademischen Grades  
Doktor der Naturwissenschaften (Dr. rer. nat.)



# Abstract

Tagging systems are intriguing dynamic systems, in which users collaboratively index resources with the so-called tags. In order to leverage the full potential of tagging systems, it is important to understand the relationship between the micro-level behavior of the individual users and the macro-level properties of the whole tagging system. In this thesis, we present the Epistemic Dynamic Model, which tries to bridge this gap between the micro-level behavior and the macro-level properties by developing a theory of tagging systems. The model is based on the assumption that the combined influence of the shared background knowledge of the users and the imitation of tag recommendations are sufficient for explaining the emergence of the tag frequency distribution and the vocabulary growth in tagging systems. Both macro-level properties of tagging systems are closely related to the emergence of the shared community vocabulary.

With the help of the Epistemic Dynamic Model, we show that the general shape of the tag frequency distribution and of the vocabulary growth have their origin in the shared background knowledge of the users. Tag recommendations can then be used for selectively influencing this general shape. In this thesis, we especially concentrate on studying the influence of recommending a set of popular tags. Recommending popular tags adds a feedback mechanism between the vocabularies of individual users that increases the inter-indexer consistency of the tag assignments. How does this influence the indexing quality in a tagging system? For this purpose, we investigate a methodology for measuring the inter-resource consistency of tag assignments. The inter-resource consistency is an indicator of the indexing quality, which positively correlates with the precision and recall of query results. It measures the degree to which the tag vectors of indexed resources reflect how the users perceive the similarity between resources. We argue with our model, and show it with a user experiment, that recommending popular tags decreases the inter-resource consistency in a tagging system. Furthermore, we show that recommending the user his/her previously used tags helps to increase the inter-resource consistency. Our measure of the inter-resource consistency complements existing measures for the evaluation and comparison of tag recommendation algorithms, moving the focus to evaluating their influence on the indexing quality.



# Zusammenfassung

Tagging-Systeme sind faszinierende dynamische Systeme in denen Benutzer kollaborativ Ressourcen mit sogenannten Tags indexieren. Um das volle Potential von Tagging-Systemen nutzen zu können ist es wichtig zu verstehen, wie sich das Verhalten der einzelnen Benutzer auf die Eigenschaften des Gesamtsystems auswirkt. In der vorliegenden Arbeit wird das Epistemic Dynamic Model präsentiert. Es schlägt eine Brücke zwischen dem Benutzerverhalten und den Systemeigenschaften. Das Modell basiert auf der Annahme, dass der Einfluss des gemeinsamen Hintergrundwissens der Benutzer und der Imitation von Tag-Vorschlägen ausreicht, um die Entstehung der Häufigkeitsverteilungen der Tags und des Wachstums des Vokabulars zu erklären. Diese beiden Eigenschaften eines Tagging-Systems hängen eng mit der Entstehung eines gemeinsamen Vokabulars der Benutzer zusammen.

Mit Hilfe des Epistemic Dynamic Models zeigen wir, dass die generelle Ausprägung der Tag-Häufigkeitsverteilungen und des Wachstums des Vokabulars ihren Ursprung in dem gemeinsamen Hintergrundwissen der Benutzer haben. Tag-Vorschläge können dann dazu genutzt werden, um gezielt diese generelle Ausprägung zu beeinflussen. In der vorliegenden Arbeit untersuchen wir hauptsächlich den Einfluss der von Vorschlägen populärer Tags ausgeht. Populäre Tags sorgen für einen Feedback-Mechanismus zwischen den Vokabularen der einzelnen Benutzer, der die Inter-Indexer Konsistenz der Tag-Zuweisungen erhöht. Wie wird aber dadurch die Indexierungsqualität in Tagging-Systemen beeinflusst? Zur Klärung dieser Frage untersuchen wir eine Methode zur Messung der Inter-Ressourcen Konsistenz der Tag-Zuweisungen. Die Inter-Ressourcen Konsistenz korreliert positiv mit der Indexierungsqualität, und mit der Trefferquote und der Genauigkeit von Suchanfragen an das System. Sie misst inwieweit die Tag-Vektoren die durch Benutzer wahrgenommene Ähnlichkeit der jeweiligen Ressourcen widerspiegeln. Wir legen mit Hilfe unseres Modell dar, und zeigen es auch mit Hilfe eines Benutzerexperiments, dass populäre Tags zu einer verringerten Inter-Ressourcen Konsistenz führen. Des Weiteren zeigen wir, dass die Inter-Ressourcen Konsistenz erhöht wird, wenn dem Benutzer das eigene, bisher genutzte Vokabular vorgeschlagen wird. Unsere Methode zur Messung der Inter-Ressourcen Konsistenz ergänzt bestehende Evaluationsmaße für Tag-Vorschlags-Algorithmen um den Aspekt der Indexierungsqualität.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Research Methodology . . . . .	3
1.2	Why Tagging Models? . . . . .	4
1.3	Structure of the Thesis . . . . .	5
1.4	Contributions and Publications . . . . .	6
<b>2</b>	<b>Foundations and Applications</b>	<b>7</b>
2.1	Folksonomies . . . . .	9
2.1.1	Hypergraph View of Folksonomies . . . . .	10
2.1.2	Stream View of Folksonomies . . . . .	11
2.2	Tag Recommendation . . . . .	13
2.2.1	Graph-based Recommenders . . . . .	13
2.2.2	Content-based Recommenders . . . . .	15
2.2.3	Evaluation of Tag Recommenders . . . . .	15
2.3	Spam Detection . . . . .	16
2.3.1	Spam Classification . . . . .	17
2.3.2	Ranking of Spam Content . . . . .	18
2.4	Retrieval of Resources . . . . .	19
2.4.1	Ranking the Relevance of Resources . . . . .	20
2.4.2	Visualizations of the Tag Space . . . . .	23
<b>3</b>	<b>Macro-Level Properties of Folksonomies</b>	<b>27</b>
3.1	Used Data Sets . . . . .	28
3.2	Tag Frequency Distribution . . . . .	30
3.3	Vocabulary Growth and Size . . . . .	36
3.4	Further Properties . . . . .	39
<b>4</b>	<b>An Epistemic Dynamic Model</b>	<b>41</b>
4.1	Building Blocks of the Model . . . . .	42
4.1.1	Simulating a Stream of Postings . . . . .	43
4.1.2	Simulating Background Knowledge . . . . .	44
4.1.3	Simulating Tag Suggestions . . . . .	50
4.2	Configurations of the Epistemic Model . . . . .	54

4.2.1	The Epistemic Model with Word Frequencies . . . . .	54
4.2.2	The Epistemic Model with Semantic Networks . . . . .	55
4.2.3	The Natural Language Model . . . . .	56
4.3	Related Work . . . . .	57
4.3.1	Influence Factors . . . . .	57
4.3.2	Tagging Models . . . . .	58
4.3.3	Summary . . . . .	65
<b>5</b>	<b>Evaluation of the Epistemic Model</b>	<b>67</b>
5.1	Comparing Simulated and Observed Properties . . . . .	68
5.2	Evaluation Measures . . . . .	70
5.2.1	Comparing Tag Frequency Distributions . . . . .	70
5.2.2	Comparing Vocabulary Size and Growth . . . . .	73
5.3	Evaluation Outline . . . . .	74
5.4	Results . . . . .	76
5.4.1	Fitting of Model Parameters . . . . .	76
5.4.2	The Epistemic Model with Word Frequencies . . . . .	77
5.4.3	The Epistemic Model with Semantic Networks . . . . .	78
5.4.4	The Natural Language Model . . . . .	78
5.4.5	The Semantic Walker Model . . . . .	81
5.4.6	Yule-Simon Model with Memory . . . . .	84
5.5	Discussion . . . . .	85
5.5.1	Reproducing the Tag Frequency Distribution . . . . .	86
5.5.2	Reproducing the Vocabulary Growth . . . . .	89
5.5.3	Influence of Imitating Tag Suggestions . . . . .	94
5.5.4	Influence of the Background Knowledge . . . . .	99
5.6	Conclusions . . . . .	102
<b>6</b>	<b>Recommendations and Indexing Quality</b>	<b>103</b>
6.1	Measures of Indexing Quality . . . . .	104
6.1.1	Inter-Resource Consistency . . . . .	105
6.1.2	Inter-Indexer Consistency . . . . .	110
6.1.3	Further Measures . . . . .	111
6.2	Research Hypotheses . . . . .	112
6.2.1	Increasing the Inter-Indexer Consistency . . . . .	112
6.2.2	Increasing the Inter-Resource Consistency . . . . .	113
6.3	User Experiment . . . . .	115
6.3.1	Phase 1: Tagging of Web Pages . . . . .	115
6.3.2	Phase 2: Grouping of Web Pages . . . . .	119
6.3.3	Recruiting the Participants . . . . .	120
6.4	Results . . . . .	122
6.4.1	Similarity of Topical Clusters . . . . .	122
6.4.2	Measuring the Inter-Resource Consistency . . . . .	125
6.4.3	Measuring the Inter-Indexer Consistency . . . . .	129



6.5	Discussion . . . . .	131
6.5.1	Influence of Comprehension of the Web Pages . . . . .	132
6.5.2	Influence of Learning Effects . . . . .	134
6.6	Conclusions . . . . .	137
<b>7</b>	<b>Conclusions</b>	<b>139</b>
7.1	The Epistemic Dynamic Model . . . . .	139
7.2	Tag Recommendations and Indexing Quality . . . . .	140
7.3	Outlook . . . . .	141
<b>A</b>	<b>Software</b>	<b>143</b>
A.1	Simulation – GUI . . . . .	143
A.2	Simulation – Command Line . . . . .	145
A.3	Generating Plots . . . . .	145
A.4	Applying the Smirnov Test . . . . .	145
A.5	Source Code . . . . .	146
<b>B</b>	<b>Data Sets – Co-occurrence Streams</b>	<b>147</b>
B.1	Delicious . . . . .	148
B.2	Bibsonomy . . . . .	150
B.3	Detailed Plots . . . . .	150
<b>C</b>	<b>Data Sets – User Experiment</b>	<b>157</b>
C.1	Questionnaire . . . . .	157



# List of Figures

2.1	Tagging interface of Delicious. Users can enter free tags in the <i>tags</i> input field. Furthermore, they can reuse tags from the suggestions provided below the input field, i. e. tags from the set of <i>recommended tags</i> , <i>your tags</i> and <i>popular tags</i> (see Section 2.2 for more details). . . . .	8
2.2	Tagging interface of Bibsonomy for bookmarking web pages. Users can enter free tags in the <i>tags</i> input field. Furthermore, they can select tags from the set of recommended tags below the input field. By clicking on the button to the right of the set of recommended tags, users can get a new set of recommended tags for bookmarking the web page. . . . .	9
2.3	Representation of a folksonomy $\mathbb{F}$ as a tripartite, undirected hypergraph (left) and as a tripartite, undirected graph (right). In the folksonomy, <i>user1</i> has tagged <i>res2</i> with the tags <i>tag1</i> and <i>tag2</i> . Furthermore, <i>user2</i> has tagged <i>res1</i> with <i>tag1</i> . . . . .	11
2.4	Example tag cloud from <a href="http://www.bibsonomy.org/">http://www.bibsonomy.org/</a> . The tags are ordered alphabetically. The font size and intensity indicate the frequency with which the respective tags are used in a collection of resources. . . . .	20
2.5	(a) Example of a frequency-ordered tag cloud. (b) Example of a frequency-ordered list. Both examples are taken from <a href="http://www.bibsonomy.org/">http://www.bibsonomy.org/</a> . . . . .	25
3.1	Zipf plots of the occurrence probabilities of tags in dependency on their rank for all streams from Tab. 3.2. Only tag assignments of regular users in the respective streams are taken into account. . . . .	31
3.2	Stabilization of the occurrence probabilities of the 10 most often used tags in the <i>social</i> co-occurrence stream from Tab. 3.2. Only tag assignments of regular users are taken into account. . . . .	32

- 3.3 Zipf plots of the occurrence probabilities of words for text corpora that have the same topical focus as the streams from Tab. 3.2 and Fig. 3.1. The text corpora have been obtained by downloading the documents that are tagged in the corresponding stream. For better comparability, only as many words from the text corpora are taken into account as there are tag assignments in the corresponding stream. . . . . 33
- 3.4 Drop in the relative occurrence probabilities for tags between rank 7 and 10 in resource streams. The relative occurrence probability corresponds to the occurrence probability normalized by the occurrence probability of the most frequent tag. The average relative occurrence probability in the graph has been computed by averaging the relative occurrence probability for 500 randomly selected resource streams from our overall Delicious data set. As a guide for the eye, a line for the best-fitting power-law distribution is also included in the graph. Only resource streams are taken into account to which more than 100 regular users contributed. . . . . 35
- 3.5 Zipf plots of the occurrence probability of tags in dependency on their rank for the *ringtones* and *social* stream pairs from Tab. 3.2 and 3.3. The filtered variant of a stream only contains tag assignments of regular users. The unfiltered variant of a stream contains a mix of tag assignments from regular users and spammers. The detailed plots for the remaining co-occurrence stream pairs from Tab. 3.2 and Tab. 3.3 are available in Appendix B. . . . . 37
- 3.6 Vocabulary growth for all streams from Tab. 3.2 for the first 10,000 tag assignments. Only tag assignments of regular users in the respective streams are taken into account. A high variance in the vocabulary growth rates can be observed. . . 37
- 3.7 Comparison of vocabulary size (left) and tag frequency distribution (right) for the *costs*, the *design* and the *checkbox* co-occurrence streams from Tab. 3.2. The *costs* and the *design* stream have been restricted to their first 4,758 tag assignments so that they contain the same number of tag assignments as the *checkbox* stream. The smaller the vocabulary, the steeper the decline in the occurrence probabilities of tags in the respective stream. . . . . 38

3.8	Vocabulary growth for the <i>ringtones</i> and <i>social</i> stream pairs from Tab. 3.2 and 3.3. The filtered variant of a stream only contains tag assignments of regular users. The unfiltered variant of a stream contains a mix of tag assignments from regular users and spammers. Only the first 20,000 tag assignments of the respective streams are shown. The detailed plots for the remaining co-occurrence stream pairs from Tab. 3.2 and Tab. 3.3 are available in Appendix B. . . . .	39
4.1	Distribution of the posting sizes used for simulating co-occurrence streams. It corresponds to the distribution of posting sizes from regular users in the overall Delicious data set described in Section 3.1. . . . .	44
4.2	Zipf plots of the word frequency distribution empirically measured in the 15 crawled text corpora. The corresponding word frequency distributions will be used in the Epistemic Model for simulating the distributions $p(W r)$ and/or $p(W t)$ . . . .	47
5.1	Methodology for evaluating tagging models that try to explain the emergence of specific properties in tagging systems. It is checked in how far the same properties emerge in the simulated tagging behavior as in the real tagging behavior. .	68
5.2	Zipf plot (left) and empirical distribution function $S(x)$ (right) of the tag frequency distribution of the filtered <i>ringtones</i> and <i>social</i> streams from Tab. 3.2. Zipf plots are the most common representation of the tag frequency distribution in the literature about tagging systems. The empirical distribution function $S(x)$ forms the basis for applying statistical methods during our evaluation. . . . .	71
5.3	Comparing the <i>Epistemic Model with Word Frequencies</i> to other models that include the shared background knowledge as an influence factor. The <i>Epistemic Model with Word Frequencies</i> and the <i>Epistemic Model with Semantic Networks</i> additionally model the influence coming from the collaboration between users. The comparison is based on the $D$ values reported for the respective models in Tab. 5.2, 5.3, 5.4 and 5.6.	86
5.4	Comparing the <i>Epistemic Model with Word Frequencies</i> to models from the literature that only model either the influence of the shared background knowledge (Semantic Walker Model + Watts-Strogatz; Tab. 5.5) or the influence of the collaboration (Yule-Simon Model with Memory; Tab. 5.7). . .	87

- 5.5 Plots of the vocabulary growth in the (a) *ringtones*, (b) *decorative*, (c) *checkbox* and (d) *analysis* co-occurrence streams. Below the plots with the vocabulary growth it is shown at which time of the stream tag assignments of a certain user or resource have been added to the stream. The observable phases of low vocabulary growth correspond to phases during which a single user or resource dominates the tag assignments in the co-occurrence stream. . . . . 91
- 5.6 Influence of the probability of imitation  $I$  on the vocabulary growth in streams simulated with the *Epistemic Model with Word Frequencies*. For each of the streams, 20,000 tag assignments have been simulated. The higher  $I$ , the lower the vocabulary growth. . . . . 96
- 5.7 Plots of how the occurrence probabilities of tags are influenced by the imitation of tag suggestions. In (a), the predictions of the Epistemic Model for different imitation probabilities are shown. This plot is based on the simulation of 20,000 tag assignments. In (b), the average occurrence probabilities of tags are shown which have been observed for the two experimental groups in [13]. In both cases, the imitation of tag suggestions leads to increased occurrence probabilities of the most often used tags. The data shown in (b) is used with kind permission of the authors of [13]. . . . . 97
- 5.8 Simulation of a resource stream in Delicious with the help of the *Epistemic Model with Word Frequencies*. The simulated resource stream contains 5,000 tag assignments and the parameter  $n$  has been adapted to the number of popular tags that are visible in the Delicious user interface (see Fig. 2.1). The relative occurrence probability corresponds to the occurrence probability normalized by the occurrence probability of the most frequent tag. . . . . 98
- 5.9 Relative occurrence probabilities in the resource streams of Bibsonomy. The average relative occurrence probability in the graph has been computed by averaging the relative occurrence probabilities for the 30 resource streams from our Bibsonomy data set (see Tab. 3.1) to which more than 100 users contributed. As a guide for the eye, a line for the best-fitting power-law distribution is also included in the graph. . . . . 99

- 5.10 Influence of the degree  $d$  of the start node on the vocabulary growth in streams simulated with the *Epistemic Model with Semantic Networks*. The vocabulary growth increases with increasing start node degree  $d$ . For each of the streams, 20.000 tag assignments have been simulated. Except the  $d$ -parameter, all other parameters of the model have been kept constant during simulating the different streams. . . . . 100
- 5.11 Influence of the number of meanings of a tag on the vocabulary size in its co-occurrence stream after 1,000 tag assignments. The number of meanings of a tag has been extracted from WordNet. The single vocabulary sizes are shown as a scatter plot in the background. The median vocabulary size and its trend are shown as lines in the foreground. . . . . 101
- 6.1 Example of tag vectors describing three resources. During retrieval, the similarity of the tag vectors influences how close together the corresponding resources get ranked. In order to get high precision and recall, the similarity of the tag vectors needs to correlate with the user perceived similarity of the resources. . . . . 106
- 6.2 Three examples of topical clusters, which illustrate the idea of measuring the Silhouette Coefficient for resource  $r_j$  (white circle). The lengths of the arrows correspond to the distances between the tag vectors of the resources. The average length of the arrows with continuous lines corresponds to  $a_{ij}$ . The average length of the arrows with dotted lines corresponds to  $b_{ij}$ . . . . . 109
- 6.3 Instructions given to the participants of our experiment. . . . 116
- 6.4 The tagging interface used for assigning tags to the 10 web pages. Depending on the experimental condition, a tag cloud with the tag recommendations is displayed below the input field for the tags. Here, the interface for the *Popular Tags* condition is shown. . . . . 117
- 6.5 The user interface for grouping the web pages into topical clusters. Instructions on how to use the interface are given at the top. . . . . 120

6.6	Visualization of the 11 most frequently identified clusters of web pages. Each box in the gray area corresponds to one cluster. Within the box of each cluster it is given by how many participants the cluster has been identified. For example, 28% of all participants put the <i>BBC</i> web page (URL-2) alone into a cluster, leading to cluster <i>cl-1</i> . Another 34% of the participants instead decided to group <i>BBC</i> (URL-2) together with <i>The Onion</i> (URL-1), leading to cluster <i>cl-2</i> . The remaining 38% of the participants have put URL-2 into other, less frequent clusters. Nevertheless, an analysis of the names used for <i>cl-1</i> and <i>cl-2</i> reveals that both clusters are seen as related to the <i>News</i> topic. . . . .	123
6.7	Overview of how the participants of the two language variants of the experiment have clustered the three web pages from Fig. 6.6 that are on the border between two topics. For each of the three web pages the probabilities are given with which they have been put into one of the eleven most popular clusters from Fig. 6.6, or in another of the 140 overall identified clusters. . . . .	133
6.8	Screenshot of <i>The Onion</i> (URL-1) that was shown to the experiment participants. . . . .	134
6.9	Overview of how the participants of the three English experimental conditions have clustered the three web pages from Fig. 6.6 that are on the border between two topics. For each of the three web pages the probabilities are given with which they have been put into one of the eleven most popular clusters from Fig. 6.6, or in another of the 140 overall identified clusters. . . . .	135
6.10	Overview of how the participants of the three German experimental conditions have clustered the three web pages from Fig. 6.6 that are on the border between two topics. For each of the three web pages the probabilities are given with which they have been put into one of the eleven most popular clusters from Fig. 6.6, or in another of the 140 overall identified clusters. . . . .	136
A.1	Screenshot of the simulation GUI. . . . .	144
A.2	Screenshot of the file dialog for extracting the frequency distribution and vocabulary growth from stream files. . . . .	146
B.1	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>ringtones</i> stream pair from Tab. 3.2 and 3.3. . . . .	151



B.2	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>setup</i> stream pair from Tab. 3.2 and 3.3. . . . .	151
B.3	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>boat</i> stream pair from Tab. 3.2 and 3.3. . . . .	151
B.4	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>historical</i> stream pair from Tab. 3.2 and 3.3. . . . .	152
B.5	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>messages</i> stream pair from Tab. 3.2 and 3.3. . . . .	152
B.6	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>decorative</i> stream pair from Tab. 3.2 and 3.3. . . . .	152
B.7	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>costs</i> stream pair from Tab. 3.2 and 3.3. . . . .	153
B.8	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>ff</i> stream pair from Tab. 3.2 and 3.3. . . . .	153
B.9	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>checkbox</i> stream pair from Tab. 3.2 and 3.3. . . . .	153
B.10	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>datawarehouse</i> stream pair from Tab. 3.2 and 3.3. . . . .	154
B.11	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>tools</i> stream pair from Tab. 3.2 and 3.3. . . . .	154
B.12	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>social</i> stream pair from Tab. 3.2 and 3.3. . . . .	154
B.13	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>design</i> stream pair from Tab. 3.2 and 3.3. . . . .	155
B.14	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>analysis</i> stream pair from Tab. 3.2 and 3.3. . . . .	155
B.15	Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the <i>blogs</i> stream pair from Tab. 3.2 and 3.3. . . . .	155

C.1	English question: “Please give your age”. German question: “Bitte geben Sie Ihr Alter an” . . . . .	158
C.2	English question: “Please give your gender”. German question: “Bitte geben Sie Ihr Geschlecht an” . . . . .	158
C.3	English question: “How comprehensible have been the web pages?”. German question: “Wie gut war der Inhalt der Webseiten zu verstehen?” . . . . .	159
C.4	English question: “How often did you already use tagging systems for searching?”. German question: “Wie oft haben Sie schon Tagging-Systeme zum Suchen benutzt?” . . . . .	159
C.5	English question: “How often did you already add content to tagging systems?”. German question: “Wie oft haben Sie schon Inhalte zu Tagging-Systemen hinzugefügt?” . . . . .	160
C.6	English question: “If you already used tagging systems, which did you use?”. German question: “Wenn Sie schon Tagging-Systeme benutzt haben, welche waren das?”. Multiple choices possible. . . . .	160

# List of Tables

3.1	Sizes of the data sets from Delicious and Bibsonomy that are used in this thesis. Both data sets contain the activity of spammers. In the Delicious data set, approximately 1.4% of the users are spammers. In the Bibsonomy data set, approximately 93% of the users are spammers. For more details, see Appendix B. . . . .	28
3.2	Statistics of the filtered co-occurrence streams from Delicious (top) and Bibsonomy (bottom). The filtered streams only contain tag assignments made by regular users. . . . .	29
3.3	Statistics of the unfiltered co-occurrence streams from Delicious (top) and Bibsonomy (bottom). The unfiltered streams contain tag assignments made by regular users as well as tag assignments made by spammers. In the columns $ U $ and $ Y $ , the overall number of users and tag assignments in the respective stream are given. The parentheses contain how many of the overall number of users are spammers and how many of the tag assignments are provided by the spammers. . . . .	30
3.4	Average, standard deviation, minimum and maximum of the $\alpha$ -values if the Zipf plots of the tag frequencies in Fig. 3.1 and of the word frequencies in Fig. 3.3 are approximated with a power-law. . . . .	34
4.1	Comparison of the properties of the Word Association Norms data set from [81] with the properties of networks generated with the Erdős-Rényi Model, the Uncorrelated Configuration Model, the Watts-Strogatz Model and the Growing Network Model. Only the Growing Network Model is able to reproduce all three properties of the Word Association Norms data set.	49
4.2	Parameters of the <i>Epistemic Model with Word Frequencies</i> . .	55
4.3	Parameters of the <i>Epistemic Model with Semantic Networks</i> . .	56
4.4	Parameters of the Natural Language Model. . . . .	57

4.5	Overview of the influence factors implemented in different tagging models described in the literature. Furthermore, it is given against which observable properties the models have been evaluated. . . . .	60
5.1	Evaluation results reported in the literature for the models from Tab. 4.5 for the properties from Chapter 3. The Multinomial Tagging Models are excluded from this overview because they neither target the tag frequency distribution nor the vocabulary growth. For the tag frequency distribution, it has been visually evaluated whether a power-law like distribution with exponential cut-off is reproduced (+) or whether only a plain power-law without the cut-off (o). With regard to reproducing the sublinear vocabulary growth, no information is available from the respective papers about the Organizing Model and the Semantic Imitation Model. . . . .	75
5.2	Evaluation results for the <i>Epistemic Model with Word Frequencies</i> . Highlighted are the $D$ values for the co-occurrence streams from Delicious (top) and Bibsonomy (bottom) where this configuration of the Epistemic Model achieves the lowest $D$ value of all evaluated models. Furthermore, the $p$ -values are highlighted for which it can not be safely rejected that simulated and real tag frequencies are drawn from the same distribution function, i. e. where $p \geq 0.1$ . Finally, all $\Delta_{last}$ and $\Delta_{max}$ values are highlighted where the simulated vocabulary size and growth differs by at most $\pm 10\%$ from the real vocabulary size and growth. . . . .	77
5.3	Evaluation results for the <i>Epistemic Model with Semantic Networks</i> . Highlighted are the $D$ values for the co-occurrence streams from Delicious (top) and Bibsonomy (bottom) where this configuration of the Epistemic Model achieves the lowest $D$ value of all evaluated models. Furthermore, the $p$ -values are highlighted for which it can not be safely rejected that simulated and real tag frequencies are drawn from the same distribution function, i. e. where $p \geq 0.1$ . Finally, all $\Delta_{last}$ and $\Delta_{max}$ values are highlighted where the simulated vocabulary size and growth differs by at most $\pm 10\%$ from the real vocabulary size and growth. . . . .	79

5.4 Evaluation results for the Natural Language Model. For none of the streams from Delicious (top) and Bibsonomy (bottom), the Natural Language Model achieves the lowest  $D$  value, compared to all other models, or stays in the  $\pm 10\%$  band around the real vocabulary growth, i.e. for all streams  $\Delta_{max} > 0.1$ . The significance value  $p$  is 0% for all tested streams, i.e. there are significant differences between the simulated and the real tag frequencies. . . . . 80

5.5 Evaluation results for the Semantic Walker Model in conjunction with the Watts-Strogatz Model for the streams from Delicious (top) and Bibsonomy (bottom). Highlighted is the  $D$  value for the stream where this configuration of the Semantic Walker Model achieves the lowest value of all evaluated models. For all tested streams, the significance value  $p$  is below 10%, i.e. there are significant differences between the simulated and the real tag frequency distributions. Finally, all  $\Delta_{last}$  and  $\Delta_{max}$  values are highlighted where the simulated vocabulary size and growth differs by at most  $\pm 10\%$  from the real vocabulary size and growth. . . . . 82

5.6 Evaluation results for the Semantic Walker Model in conjunction with the Growing Network Model. For none of the streams from Delicious (top) and Bibsonomy (bottom), this configuration achieves the lowest  $D$  value, compared to all other models, or stays in the  $\pm 10\%$  band around the real vocabulary growth, i.e. for all streams  $\Delta_{max} > 0.1$ . The significance value  $p$  is 0% for all tested streams, i.e. there are significant differences between the simulated and the real tag frequencies. . . . . 83

5.7 Evaluation results for the Yule-Simon Model with Memory for the streams from Delicious (top) and Bibsonomy (bottom). Highlighted is the  $D$  value for the stream where the Yule-Simon Model with Memory achieves the lowest value of all evaluated models. The significance value  $p$  is 0% for all tested streams, i.e. there are significant differences between the simulated and the real tag frequencies. The free parameter  $p_{ys}$  has been adapted such that the simulated vocabulary size differs by at most 10% from the real vocabulary size, as it is shown by the  $\Delta_{last}$  values. . . . . 84

5.8 Average, median, minimum and maximum of the  $D$  values reported for the different models in Tab. 5.2–5.7. The *Epistemic Model with Word Frequencies* achieves the best values, closely followed by the *Epistemic Model with Semantic Networks*. . . . . 88

5.9	Average node degree $k$ and clustering coefficient $C$ of the semantic networks that have been generated with the Watts-Strogatz Model and that led to the best fitting vocabulary growth and tag frequency distribution in Tab. 5.5. The values significantly deviate from the average node degree and clustering coefficient empirically observed in the Word Association Norms data set from [81], where $k = 22$ , $C = 0.186$ and $(1 - C) \cdot (k - 1) = 17.91$ (cf. [106]). . . . .	94
6.1	URLs of the 10 web pages used during the experiment. . . . .	117
6.2	Tags used for bootstrapping the <i>Popular Tags</i> condition. . . . .	119
6.3	Sizes of the experimental data sets. Only participants who finished tagging all ten web pages are included. . . . .	121
6.4	Influence of the experimental conditions on the inter-resource consistency and on the inter-indexer consistency. For the results on the left, we have restricted the number of users such that under each of the experimental conditions the same number of users contributed to the tag vectors. For the results on the right, we have restricted the number of tag assignments (TAS) such that the tag vectors for the same resource contain the same number of tag assignments. . . . .	126

# Chapter 1

## Introduction

During the last years, collaborative tagging systems like Flickr, Delicious and Bibsonomy have become more and more popular<sup>1</sup>. Tagging systems allow users to upload their resources, like photos, bookmarks or BibTeX entries, and organize them by assigning keywords to them. In the context of tagging systems, these keywords are called *tags*. Over time, the tag assignments of the different users lead to the emergence of a loose categorization system for resources which is frequently called a *folksonomy* (see [78]).

Folksonomies constitute intriguing dynamic systems constructed by the collaboration and interaction of their users. They offer new possibilities for indexing and searching resources. One key aspect of tagging systems is the uncontrolled nature of the community's vocabulary. This lowers the entry barrier for using tagging systems but also poses challenges during search and navigation of the resources. For example, due to the uncontrolled vocabulary, folksonomies have to cope with problems like ambiguous and/or synonymous tags, which do not arise if centrally controlled vocabularies like thesauri are used for annotating resources [78]. Nevertheless, it has been observed in [38, p. 205] that the aggregated tag assignments of users “give rise to a stable pattern in which the proportions of each tag are nearly fixed”. This is typically taken as an indicator that tagging is successful in collaboratively indexing resources despite of its uncontrolled nature.

Where do these stable patterns come from? How do they emerge from the micro-level behavior of the individual users in absence of a central, coordinating instance? Are we able to influence this process, e. g. in order to increase the indexing quality in tagging systems? In order to leverage the full potential of tagging systems, it is important “to understand the characteristics of user activity in (collaborative) tagging systems” [92, p. 183], e. g. to understand the relationship between the usage of tags on the micro-level of individual users and the emergence of macro-level properties of whole

---

<sup>1</sup>See <http://www.flickr.com/>, <http://www.delicious.com/>, and <http://www.bibsonomy.org/>.

folksonomies. With this thesis, we want to contribute to such a better understanding of the processes that are ongoing in tagging systems. Furthermore, we want to show how a better understanding of the processes can be put into use for selectively influencing and improving them.

In this thesis, we focus on analyzing the behavior and interaction of users in broad folksonomies. The reason for this focus is that in broad folksonomies a resource may be tagged by several users and a tag can be assigned multiple times to the same resource. Examples of tagging systems that produce broad folksonomies are Delicious and Bibsonomy. In contrast, in narrow folksonomies each resource is usually tagged only by a single user and a tag can only be assigned a single time to the same resource (see Section 2.1 for more details). An example of a tagging system that produces a narrow folksonomy is Flickr. All in all, analyzing the behavior and interaction of users in broad folksonomies has two advantages over studying it in narrow folksonomies:

- In broad folksonomies, the behavior of users can directly be compared to each other given that they have at least one resource in common. In that case, it can be compared which tags they have used for describing the same resource. In narrow folksonomies, such a direct comparison is not possible because the system prevents that two users can use the same tag at the same resource. Furthermore, most of the resources in narrow folksonomies like Flickr are only tagged by a single user.
- In broad folksonomies, there is a more direct interaction of users in the context of a resource. For example, in Delicious a user sees amongst others the set of popular tags of the resource he/she is currently tagging. This adds a direct feedback mechanism between the vocabularies of the users who have tagged the same resource, thus leading to a collaborative effort of tagging resources. In contrast, users of narrow folksonomies typically only get in contact with the vocabulary of other users during searching or browsing resources but not during tagging a resource. Thus, the influence of the other users' vocabulary is more indirect in narrow folksonomies.

In this thesis, we concentrate on analyzing properties of broad folksonomies that are related to the emergence of the shared community vocabulary and to the navigability of the resulting folksonomy. Central to the development of the community's vocabulary are the properties of the tag frequency distribution and the size of the used vocabulary (see Chapter 3). For example, the size of the vocabulary influences how many search terms can be used for accessing the resources in a folksonomy. Furthermore, both properties are closely related to the entropy of the used tagging vocabulary for which it has been shown in [21] that the entropy can be used for measuring the navigability of a folksonomy.



## 1.1 Research Methodology

For understanding the connection between the micro-level behavior of individual users in tagging systems and the emergence of certain macro-level properties of the system, we develop a theory about how different factors like the background knowledge of the users and the tag recommendations provided in the user interface of tagging systems interact with each other. We express our theory about the dynamics in tagging systems in form of the Epistemic Dynamic Model (see Chapter 4). Our model can be used for predicting the macro-level properties that we expect to emerge in a folksonomy given that the underlying assumptions of our model hold. It is our objective to identify with our model the influence factors on the users' tagging behavior that are required for explaining the emergence of the properties described in Chapter 3.

During the evaluation of our model in Chapter 5, we use Popper's *critical method* [86, p. 13ff] for comparing our theory about the relevant influence factors to competing theories from the literature, which try to explain the same observations as we do with our Epistemic Dynamic Model (see Section 4.3). Each of these competing theories corresponds to different assumptions about which influence factors are relevant for explaining the emergence of the observed properties. In such a case, Popper's critical method can be used for ruling out some of the competing theories but it can not be used for identifying whether a theory is "true".

According to Popper's *critical method*, theories are evaluated by generating a number of *test statements*. In our case, the test statements correspond to evaluating whether a theory and its corresponding model are able to predict the observable macro-level properties of folksonomies. If several theories pass the test with our test statements then Popper suggests to develop more rigid tests that are able to rule out some of the competing theories.

For example, the majority of the tagging models described in Section 4.3 as well as our own Epistemic Dynamic Model have been designed with the objective to explain the tag frequency distributions in tagging systems. But in the current literature, only quite weak tests are used for determining whether a given tagging model successfully explains the tag frequency distribution. Often, it is only visually compared whether the simulated tag frequency distribution belongs to the family of power-law like, heavy-tailed distributions. In contrast, in Chapter 5 we suggest to use more rigid tests. We use statistical methods for additionally checking whether also the observed exponents of the power-law like distributions can be reproduced. In our evaluation in Chapter 5, the more rigid tests help us to rule out some of the competing tagging models from the literature. Furthermore, we show that one of the competing models can be integrated as a submodel into our own Epistemic Dynamic Model, then leading to equivalent evaluation results.

## 1.2 Why Tagging Models?

In this section, we shortly discuss the potential benefits of having models of the tagging behavior of users. In contrast to most works in computer science, tagging models do not provide an immediate benefit, e.g. by improving the performance of an application. Instead, they are more related to fundamental research, which leads to a better understanding of the complex dynamics in tagging systems by integrating different assumptions about the users' behavior and the influencing factors into a theory of tagging.

Such a better understanding can be used for better exploiting the potential of tagging systems and their paradigm of annotating resources with uncontrolled vocabularies. For example, in Chapter 6 we use the Epistemic Dynamic Model for predicting how a specific kind of tag recommendations, i.e. the suggestion of the popular tags at a resource, influences the indexing quality in a tagging system, and the navigability of the indexed resources. These predictions of the model are confirmed in Chapter 6 by a user experiment. According to Popper's critical method, this user experiment can be seen as a further test of the Epistemic Dynamic Model in how far it allows to make correct predictions about the dynamics in tagging systems. From a more practical point of view, the Epistemic Dynamic Model not only helps us in predicting the outcome of a user experiment but it also provides an explanation for the observations made during the experiment.

Thus, the user experiment in Chapter 6 exemplifies one use case of tagging models according to which the model is used for generating hypotheses about the expected outcome of user experiments prior to designing and conducting them. If the hypotheses are confirmed by the experiment then the model provides an explanation for the observations and the experiment contributes to a more thorough testing of the model. In contrast, if the hypotheses can not be confirmed, this may lead to a modification and/or extension of the model and its corresponding theory of tagging. According to Popper's critical method, the new model, which results from the latest observations, should correct the old model such that it not only explains the latest observations but also all previous observations where the old model was successful [86].

All in all, tagging models can be seen as a way for formulating our assumptions about the dynamics in tagging systems and making them explicit. The tagging model can then be used for showing that the assumptions are plausible by making predictions with the model and evaluating them with the help of available tagging data and/or with user experiments. Upon successful evaluation of a model, it provides an explanation for the observations in the tagging data and/or user experiments. If several observations can be reproduced then the model helps us in connecting these, priorly unrelated observations.

### 1.3 Structure of the Thesis

We start this thesis in Chapter 2 with giving an overview of the foundations and applications of tagging systems. We give a formal definition of folksonomies and define different views on them. In this thesis we take a stream view on a folksonomy, which is especially well suited for studying the dynamics in a folksonomy, i. e. how the folksonomy develops over time. In the stream view, a folksonomy is viewed as a sequence of tag assignments that is ordered by their creation time. Furthermore, we give an overview of the literature about tag recommendations, spam detection and retrieval in folksonomies, which is related to the topic of this thesis.

In Chapter 3, we describe the folksonomy data sets that we use throughout this thesis. Based on the data sets, we then give an overview of the macro-level properties that are generally observable in folksonomies. We especially concentrate on the tag frequency distribution and the vocabulary growth and size in folksonomies. Both properties are related to the emergence of a shared community vocabulary. Furthermore, they influence the navigability of folksonomies, i. e. how easy it is to search and browse the resources in a folksonomy. Finally, we give an overview of further properties discussed in the literature.

In Chapter 4, we define our Epistemic Dynamic Model of tagging systems. It is based on the assumption that at least the shared background knowledge of the users and the imitation of tag recommendations are required for explaining the emergence of the properties from Chapter 3. Each of these two assumed influence factors is modeled by a separate building block in our model. Based on the building blocks, we define different configurations of our Epistemic Dynamic Model that can be used for testing and comparing alternative implementations of the same building block, and for studying the influence of tag recommendations on the emergent properties. Finally, we give an overview of further influence factors and tagging models that are currently discussed in the literature.

In Chapter 5, we evaluate our Epistemic Dynamic Model and compare it to competing tagging models from the literature. For this purpose, we first define evaluation measures that enable us to do a more rigid testing of tagging models than the current methods used in the literature. We use the measures for evaluating in how far really both of our assumed influence factors are required for explaining the emergence of the properties from Chapter 3. Furthermore, we use the measures for comparing two alternative models to each other that simulate the shared background knowledge of users.

In Chapter 6, we analyze the influence of tag recommendations on the indexing quality in tagging systems. The tag assignments in a tagging system have a high indexing quality if they link resources that have aspects in common, thus increasing the recall during retrieval. Additionally, tag

assignments with a high indexing quality should help to discriminate between resources, thus increasing the precision during retrieval. Based on the findings from our evaluation of the Epistemic Dynamic Model in Chapter 5, we predict for two exemplary tag recommenders from Delicious how they influence the indexing quality in tagging systems. We then use a controlled user experiment for evaluating our predictions.

Finally, Chapter 7 summarizes the main contributions of this thesis.

## 1.4 Contributions and Publications

This thesis provides three main contributions to the literature about tagging systems:

First, we propose and describe the Epistemic Dynamic Model of tagging systems, which integrates the influences coming from the shared background knowledge of the users and from the imitation of tag recommendations (see Chapter 4). An initial version of the model has been published in the paper “An Epistemic Dynamic Model for Tagging Systems” on the 19<sup>th</sup> ACM Conference on Hypertext and Hypermedia in 2008 [26]. At the time of its publication, the Epistemic Dynamic Model has been the first model that is able to explain the emergence of the sublinear vocabulary growth in tagging systems. With the paper [26], we have won the *Ted Nelson Newcomer Award*.

Second, we propose and describe a more rigid approach to the evaluation of tagging models than currently applied in the literature (see Chapter 5). A first version of our evaluation approach has been published in the paper “On Differences in the Tagging Behavior of Spammers and Regular Users” on the 2<sup>nd</sup> Web Science Conference in 2010 [27]. The paper contains the comparison of our model to the Yule-Simon Model with Memory. The comparison of our model to the Semantic Walker Model has been published in the paper “Das Epistemic Model – Ein Modell zur Erklärung der Dynamik in Tagging Systemen” on the 2. DGI-Konferenz der Deutschen Gesellschaft für Informationswissenschaften und Informationspraxis in 2012 [25].

Third, we propose and describe a novel approach for evaluating the influence of tag recommenders on the indexing quality in tagging systems (see Chapter 6). We demonstrate the approach in the context of a user experiment for two exemplary tag recommenders. We use the results of the user experiment for evaluating our predictions, which are based on the Epistemic Dynamic Model, about how the two tag recommenders influence the indexing quality in tagging systems. The user experiment and its results have been published in the paper “Measuring the Indexing Quality of Tag Recommenders on the Indexing Quality in Tagging Systems” at the 23<sup>rd</sup> ACM Conference on Hypertext and Social Media in 2012 [28].

## Chapter 2

# Foundations and Applications of Tagging Systems

In tagging systems, users can upload resources and assign arbitrary words to them, the so-called tags. Later, the tags can be used for retrieving and browsing the collection of resources. The collection of all users, resources and tag assignments of a tagging system are called *folksonomy*. Depending on how the users are allowed to assign the tags, folksonomies can be further divided into *broad folksonomies* and *narrow folksonomies*. In broad folksonomies, a tag can be assigned several times to the same resource by different users. In narrow folksonomies, a tag can only be assigned once to a resource. More details about the constituting elements of a folksonomy and the distinction between broad and narrow folksonomies are available in Section 2.1.

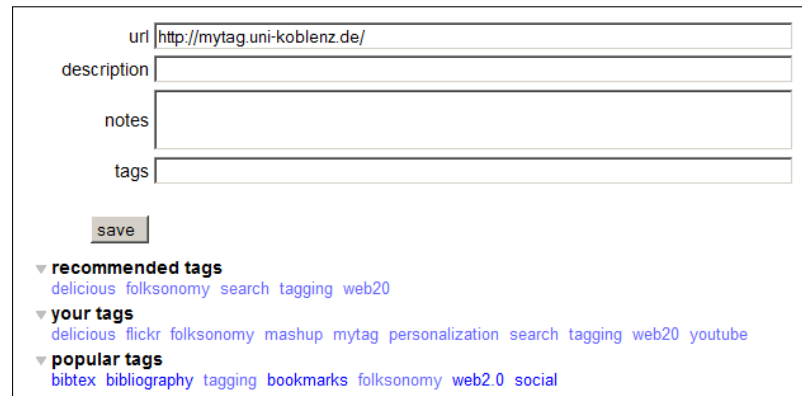
In the following chapters, we focus on modeling and analyzing the behavior and interaction of users in broad folksonomies because they exhibit a higher level of interaction between the users (see Chapter 1). Probably the most prominent tagging system that generates a broad folksonomy is Delicious<sup>1</sup>. In Delicious, users can upload and create bookmarks of arbitrary web pages. In Fig. 2.1, the tagging interface of Delicious is shown as it was in use in January 2008. During tagging, the user sees three sets with tag suggestions, namely the *recommended tags*, the *your tags* and the *popular tags*. By clicking on a suggested tag, the user is able to include it to the set of tags that will be assigned to the bookmark.

Very similar to Delicious is the Bibsonomy<sup>2</sup> system. Bibsonomy is a tagging system in which users can create bookmarks of arbitrary web pages as well as of BibTeX references. Bibsonomy has a smaller user community

---

<sup>1</sup><http://www.delicious.com/>

<sup>2</sup><http://www.bibsonomy.org>



url

description

notes

tags

▼ **recommended tags**  
delicious folksonomy search tagging web20

▼ **your tags**  
delicious flickr folksonomy mashup mytag personalization search tagging web20 youtube

▼ **popular tags**  
bibtex bibliography tagging bookmarks folksonomy web2.0 social

Figure 2.1: Tagging interface of Delicious. Users can enter free tags in the *tags* input field. Furthermore, they can reuse tags from the suggestions provided below the input field, i. e. tags from the set of *recommended tags*, *your tags* and *popular tags* (see Section 2.2 for more details).

than Delicious. Also in Bibsonomy, the user sees a set of recommended tags from which he/she can reuse tags during tagging (see Fig. 2.2).

Delicious and Bibsonomy are just two examples of tagging systems producing broad folksonomies that give tag recommendations. Many other systems also offer such recommendations [56]. It is the objective to “support users in the tagging process and to expose different facets of a resource” [56, p. 506] with the help of tag recommendations. In Section 2.2, we give an overview of the relevant work on tag recommendation algorithms.

One problem of tagging systems is that they are attracting spammers. Spammers are users who use tags in a misleading way for increasing the visibility of some resources or simply for confusing the other users [63]. For example, the owners of Bibsonomy have reported in [65] that 1,411 legitimate users and 18,681 spammers had contributed to the Bibsonomy system until the end of 2007. Thus, in order to retain the usefulness of tagging systems, it became an important research field how to automatically detect and filter spammers in tagging systems. A summary of the related work on spam detection in tagging systems is available in Section 2.3.

Of course, using tagging systems for annotating resources with tags is not an end in itself. An important application of tagging systems is the retrieval of resources, e. g. by searching for resources annotated with specific tags or by browsing the tags and resources contained in a tagging system. In Section 2.4, a summary of the works related to the retrieval of resources is available. We especially concentrate on algorithms for ranking search results and methods for visualizing the tag space of a folksonomy by means of tag clouds.

Figure 2.2: Tagging interface of Bibsonomy for bookmarking web pages. Users can enter free tags in the *tags* input field. Furthermore, they can select tags from the set of recommended tags below the input field. By clicking on the button to the right of the set of recommended tags, users can get a new set of recommended tags for bookmarking the web page.

## 2.1 Folksonomies

The collection of all users, tags, resources and tag assignments in a tagging system is called folksonomy. We formally define a folksonomy as follows (cf. [6, 96]):

**Definition 1** *A folksonomy  $\mathbb{F}$  is a tuple  $\mathbb{F} := (U, T, R, Y, pt)$  where*

- *$U$ ,  $T$ , and  $R$  are finite sets, whose elements are called users, tags and resources, respectively.*
- *$Y$  is a ternary relation between them, i. e.,  $Y \subseteq U \times T \times R$ , called tag assignments (TAS for short).*
- *$pt$  is a function  $pt : Y \rightarrow \mathbb{N}$  that assigns to each tag assignment of  $Y$  a temporal marker  $n \in \mathbb{N}$ . It corresponds to the time at which a user assigned a tag to the resource.*

*The tag assignments can be grouped into several postings. A posting contains all tag assignments made by the same user to the same resource at the same time. The temporal marker of the posting is equal to the temporal marker of each of the contained tag assignments.*

Furthermore, one can distinguish between broad and narrow folksonomies [115]. For broad folksonomies, the Definition 1 from above applies without

any constraints. This means that the same tag can be assigned several times to the same resource by different users. Examples for broad folksonomies are the folksonomies created in Delicious and Bibsonomy (see above). In contrast, for narrow folksonomies, the corresponding tagging system enforces the additional constraint on the set  $Y$  of tag assignments that a tag can only be assigned once to a resource:

**Definition 2** *The folksonomy  $\mathbb{F}$  created by a tagging system is called a narrow folksonomy if the tagging system enforces the following additional constraint on the set  $Y$  of tag assignments:  $\forall (u, t, r) \in Y : \nexists (u', t, r) \in Y : u \neq u'$ .*

An example for narrow folksonomies is the folksonomy created in Flickr<sup>3</sup>, which is a tagging system for sharing photos. Often, narrow folksonomies are used for sharing resources where the initially uploading user is also the owner of the resource, like it is the case for photos (Flickr) or videos (YouTube<sup>4</sup>). In contrast, broad folksonomies are used for sharing resources that are not necessarily owned by one of the users in the tagging system, like it is the case for web pages or bibliographic references.

The above definitions of what to understand under folksonomies only cover the parts that are common to all tagging systems. Depending on the system, additional elements and relations may be available. For example, in Delicious the users can organize a collection of bookmarks in stacks<sup>5</sup>. However, in this thesis, we focus on the core functionalities of tagging systems described in Def. 1 and 2. In the following subsections, we are describing possible representation mechanisms for these core folksonomies, which offer different views on their content.

### 2.1.1 Hypergraph View of Folksonomies

In [50], it has been proposed to use a tripartite, undirected hypergraph view on folksonomies:

**Definition 3** *The hypergraph view for folksonomy  $\mathbb{F}$  is defined as the hypergraph  $G = (V, E)$ , where the set of vertices  $V$  consists of the union of the disjoint sets of users, tags and resources, i. e.  $V = U \cup T \cup R$ . The set of hyperedges  $E$  connects those tags, users and resources that are involved in one of the tag assignments, i. e.  $E = \{\{u, t, r\} | (u, t, r) \in Y\}$ .*

An example for the hypergraph view on a folksonomy is shown in Fig. 2.3. The hypergraph view on a folksonomy can also be transformed to other

---

<sup>3</sup><http://www.flickr.com/>

<sup>4</sup><http://www.youtube.com/>

<sup>5</sup><http://blog.delicious.com/2011/09/a-new-flavor%E2%80%A6still-delicious/>



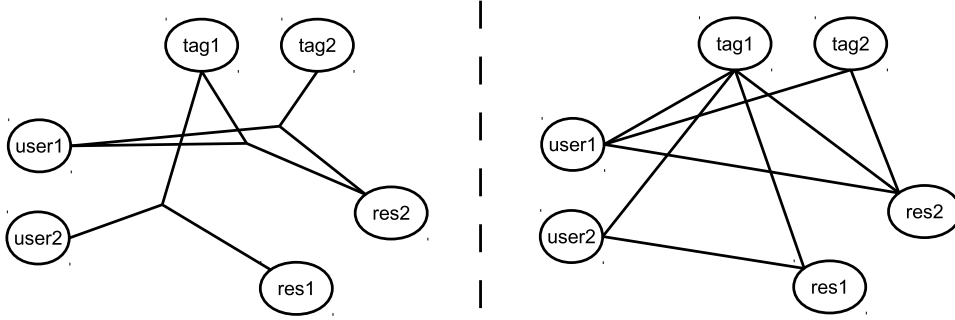


Figure 2.3: Representation of a folksonomy  $\mathbb{F}$  as a tripartite, undirected hypergraph (left) and as a tripartite, undirected graph (right). In the folksonomy, *user1* has tagged *res2* with the tags *tag1* and *tag2*. Furthermore, *user2* has tagged *res1* with *tag1*.

graph views on a folksonomy. For example, in [50] the tripartite, undirected hypergraph is projected to a tripartite, undirected graph by splitting each hyperedge  $\{u, t, r\}$  into three distinct edges  $\{\{u, t\}, \{u, r\}, \{t, r\}\}$  (see Fig. 2.3). This graph view of folksonomies has been used in [50] for defining the FolkRank algorithm, which is summarized in Subsection 2.4.1.

Furthermore, the original hypergraph graph may be projected to bipartite or even unipartite graphs. For example, in [80] it has been proposed to project the tripartite hypergraph into three bipartite graphs with regular edges with edges between either (1) users and tags, (2) users and resources, or (3) tags and resources. An example of how to reduce the tripartite graph to an unipartite graph is given in [95], where the graph of tag co-occurrences is analyzed. In the tag co-occurrence graph described in [95], two tags  $t_1$  and  $t_2$  are connected by an edge if there exist two hyperedges  $\{u_1, t_1, r_1\}$  and  $\{u_2, t_2, r_2\}$  in the original hypergraph for which  $u_1 = u_2$ ,  $r_1 = r_2$  and  $t_1 \neq t_2$ .

In this subsection, we have presented just a few examples for the possible projections of the original hypergraph. For example, until now we have only dealt with unweighted graphs, but during a projection it may happen that two hyperedges are projected to the same edge in the resulting graph. This may be used for defining weighted variants of the projected graphs in which the weight of an edge depends on the number of hyperedges that have been projected to the respective edge (see also [95]).

### 2.1.2 Stream View of Folksonomies

In the previous subsection, we have defined the hypergraph view on a folksonomy. The hypergraph view and projections of it are well suited for representing the relations between users, tags and resources in the folksonomy.

However, the hypergraph view is ignoring the information encoded in the timestamps associated with the different tag assignments, as given by the function  $pt : Y \rightarrow n$ . Thus, if it comes to analyzing the dynamic aspects of a folksonomy, i. e. how it developed over time, one has to use another view. In the following, we define stream views on a folksonomy  $\mathbb{F}$ . A stream view corresponds to a sequence of tag assignments that is ordered by the temporal markers of the tag assignments. We distinguish the *resource stream*, the *co-occurrence stream* and the *user stream*:

**Definition 4** *The resource stream for folksonomy  $\mathbb{F}$  and resource  $r \in R$  is a sequence  $\langle (u_1, t_1, r_1); (u_2, t_2, r_2); \dots; (u_n, t_n, r_n) \rangle$  of all tag assignments for the resource  $r$  such that  $(u_i, t_i, r_i) \in Y$  and  $r_i = r$ . The tag assignments in the resource stream are ordered by their temporal markers such that  $pt((u_i, t_i, r_i)) \leq pt((u_{i+1}, t_{i+1}, r_{i+1}))$ .*

**Definition 5** *The co-occurrence stream for folksonomy  $\mathbb{F}$  and tag  $t \in T$  is a sequence  $\langle (u_1, t_1, r_1); (u_2, t_2, r_2); \dots; (u_n, t_n, r_n) \rangle$  of all tag assignments that co-occur with tag  $t$  in the same posting such that  $(u_i, t_i, r_i) \in Y$  and  $t_i \neq t$  and  $\exists (u', t', r') \in Y : u' = u_i \wedge t' = t \wedge r' = r_i$ . The tag assignments in the co-occurrence stream are ordered by their temporal markers (see Definition 4).*

**Definition 6** *The user stream for folksonomy  $\mathbb{F}$  and user  $u \in U$  is a sequence  $\langle (u_1, t_1, r_1); (u_2, t_2, r_2); \dots; (u_n, t_n, r_n) \rangle$  of all tag assignments of the user  $u$  such that  $(u_i, t_i, r_i) \in Y$  and  $u_i = u$ . The tag assignments in the user stream are ordered by their temporal markers (see Definition 4).*

The analysis of resource streams, co-occurrence streams and user streams gives insights into the dynamics and underlying mechanisms that lead to the development of folksonomies. This helps to better understand the potential and the limits of social tagging systems. For example, the analysis of resource streams gives insights into how users agree on a common description for a certain resource. In this context, it is of especial interest how the size of the vocabulary associated with the resource grows over time. Furthermore, one may analyse the frequency distribution of the tags associated with a resource. From these two observables one may conclude on the degree of consensus between the users how to describe the resource and how the consensus evolves over time.

In co-occurrence streams, one can study the behavior of several users in the context of several resources. The analysis of co-occurrence streams gives insights into how the semantics of a certain tag evolves in a tagging system. Like for the resource streams, the size of the vocabulary co-occurring with a certain tag is of interest as well as the frequency distribution of the co-occurring tags. Both properties influence the navigability of tag clouds, which are often used for browsing the resources of tagging systems (see Subsection 2.4.2).

In user streams, one can study the evolving vocabulary of a single user. The analysis of user streams gives insights into how users organize resources and which interests they have. For example, one may measure the entropy of the user's vocabulary to get insights into how effectively the user is organizing the resources in his/her resource collection. Furthermore, many algorithms for personalized ranking of search results or personalized recommendation of interesting resources make use of the information in the user stream (see Section 2.4).

## 2.2 Tag Recommendation

In Fig. 2.1 and 2.2 it has been shown that tagging systems provide tag recommendations to the users. Such recommendations should support the user in the tagging process by exposing different facets of a resource [56]. Given a user who is about to tag a given resource, e.g. a web page, there are three basic paradigms of recommending tags to this user [70]: One can recommend (1) tags based on the tag assignments of other users (either extracted from the tag assignments associated with the current resource or from all tag assignments), (2) tags based on the previous tag assignments of the current user, and (3) tags based on the content of the current resource, e.g. by extracting keywords from the content or title of a web page.

Tag recommenders that only use a combination of the first two paradigms as a source for their recommendations are subsumed as *graph-based recommenders* (see Subsection 2.2.1) because they only depend on the information that is available in the hypergraph of the folksonomy [31]. Tag recommenders that, amongst others, use the third paradigm as a source for their recommendations are subsumed as *content-based recommenders* [31] (see Subsection 2.2.2). The three recommendation paradigms fulfill different purposes and are used in different ways. In Subsection 2.2.3, we finally summarize how the quality of tag recommendation algorithms is usually evaluated.

### 2.2.1 Graph-based Recommenders

The first paradigm is often used for recommending tags that have been used by other users in the context of the current resource. The most simple example of such a recommender is the *Popular Tags* recommender of Delicious, which is shown in Fig. 2.1. It recommends the seven most popular tags of the current resource. But a recommender that only relies on the set of tags already assigned to the current resource has the disadvantage that it cannot make recommendations for previously unseen resources and that it can not recommend new tags. More sophisticated tag recommenders than the *Popular Tags* recommender from Delicious try to counteract these disadvantages by amending their recommendations based on the first paradigm with tags

based on the other two paradigms. Examples of such more sophisticated recommenders are described in [54, 58, 124].

Another way of using the first paradigm for recommending tags is to not only include tag assignments made in the context of the current resource but also tag assignments made in the context of all other resources. This extended set of tag assignments is used for analyzing the conditional probability  $P(t_2|t_1)$  of observing a tag  $t_2$  together with a tag  $t_1$  in different contexts like a posting, a resource or a user. These conditional probabilities are then used for recommending tags that often co-occur together with tags that are already assigned to the current resource, or that have already been used by the current user.

There exist different techniques for analyzing the co-occurrences of tags like association rules mining [3, 48, 102], collaborative filtering [56, 76, 93], topic models [52, 66, 125] or tensor factorization [89]. From these techniques, the association rule mining can only be used for providing unpersonalized tag recommendations because it does not distinguish between the tag assignments of the current user and those of all other users. In contrast, the other three techniques provide personalized tag recommendations. Thus, they also exploit the second paradigm by taking the previous tag assignments of the current user into account. They are then used for identifying like-minded users and for preferring tags coming from their tagging vocabulary.

Such personalized recommendations, which rely on the second paradigm, have the advantage that they can outperform the best possible non-personalized tag recommender, i. e. a recommender that not only incorporates knowledge about the past behavior of all users but also about their future behavior [89]. Of course, this best possible non-personalized tag recommender can not exist in practice but it only serves as a theoretical upper bound for non-personalized recommenders.

The personalization techniques described above are quite sophisticated because they amalgamate the personalization step, which is based on the second paradigm, with the generation of recommendations based on the first paradigm. This has the disadvantage that these personalized recommenders cannot make recommendations for previously unseen users. This disadvantage does not hold for another kind of personalized recommenders that treat the personalization in an additional step [54, 58, 69, 70]. This has the advantage that the personalization step can be omitted if tag assignments of the current user do not exist in the folksonomy yet. Usually, these recommenders first generate tag recommendations on either the first or the third paradigm, and then filter and rank this initial set of recommendations based on the tagging vocabulary of the current user.

More simple examples of personalized tag recommenders are available in Delicious. Both, the *Your Tags* and the *Recommended Tags* recommender in Fig. 2.1, provide personalized recommendations. The *Your Tags* recommender is only based on the second paradigm because it simply suggests all

previously used tags of the current user. Over time, this set of tags may become very large. In contrast, the *Recommended Tags* recommender suggests the intersection between all previously used tags of the current user and all tags previously used at the current resource. Thus, the recommendations are based on the first and the second paradigm, and they are more focused on the tags that are relevant for the current resource.

### 2.2.2 Content-based Recommenders

Finally, there exist the recommenders that exploit the third paradigm by generating tag recommendations based on the actual content of the resources. This third paradigm has the advantage that it can be used for generating recommendations for previously unseen resources and/or users and that it can even be used for recommending previously unseen tags. One can further distinguish tag recommenders that simply parse the textual content associated with a resource and then rank the extracted words according to some metrics [54, 69, 70, 125], and tag recommenders that analyze the co-occurrence between words in the textual content of a resource and its tags [40, 48, 51, 52, 109, 124]. Although most content-based recommenders assume that the resources have some textual content associated with them, there also exist recommenders for non-textual resources like images [1].

### 2.2.3 Evaluation of Tag Recommenders

After this overview of different tag recommendation algorithms, it is now the question how to measure the quality of the generated tag recommendations and how to compare different algorithms during an evaluation? For this purpose, the methodology proposed by Jäschke et al. [56] has found widespread adoption in the literature about tagging systems. According to this methodology, it should be the objective of tag recommenders to predict as accurate as possible the tags that will finally be assigned by a user. The quality of the set of recommended tags in comparison to the finally assigned tags is then measured in terms of precision, recall and f-measure [114]. With regard to this methodology, one can further distinguish an offline [56] and an online [55] variant of the methodology:

- The **offline variant** takes as input a folksonomy data set, which is then splitted into a training and a test set. It is the objective of tag recommenders to reproduce the tag assignments of the users and resources in the test set based on the information available in the training set. The offline variant of the methodology has the advantage that the quality of the tag recommendations can be automatically computed, and that results of different recommenders can be easily compared given that the same training and test set is used during the evaluation. However, it has also the disadvantage that it does not

take into account that tag recommendations may actually influence and change the tag assignments of a user. For example, the offline evaluation actually penalizes tag recommenders that help the users in providing a more complete or accurate description of resources. Furthermore, it is important that the users in the folksonomy data set have not been influenced by any kind of tag recommenders in order to avoid biasing the evaluation results.

- The **online variant** also takes as input a folksonomy data set but the tag recommendations are presented to actual users who then decide which of the recommendations to pick and which additional tags to assign to a given resource. The online variant has the advantage that it takes into account how the users are influenced by the recommendations. However, it has the disadvantage that it requires a lot more effort in order to evaluate a tag recommender and that the results are harder to compare between different evaluations, e. g. because of differences between the user populations participating in the evaluations.

All in all, the offline variant has the big advantage that it is easy to reproduce, requires few effort and that the results are easy to compare across different evaluations as long as the same data set is used. In contrast, the online variant has the big advantage that it takes into account how the users are influenced by seeing the tag recommendations. Both variants have the disadvantage in common that they can not be used for measuring whether the recommendations help the users in improving the quality of their tag assignments. In case of the offline variant, a tag recommender can not outperform the uninfluenced tag assignments of the users because any deviation from the uninfluenced behavior gets penalized by the evaluation measures. In case of the online variant, it is measured whether users prefer one tag recommender over another during tagging. But this aspect of tag recommendations is distinct from the resulting quality of the tag assignments. In Chapter 6, we are presenting an alternative measure for evaluating tag recommenders that is able to measure in how far the quality of the tag assignments is influenced by a tag recommender.

## 2.3 Spam Detection

Tagging systems are an interesting target for spammers. Spam in tagging systems can be defined as content that legitimate users do not wish to share and content that is tagged in a way to mislead other users [47, 65]. Often, it is the purpose of spam to increase the visibility of specific resources [63]. Tagging systems have to find ways for coping with spam because otherwise the content of legitimate users becomes invisible. For example, in [65] it has been reported that, at the end of 2007, 92% of all registered users in the

Bibsonomy system were spammers. The problem is even bigger than these numbers suggest because often spammers belong to the group of very active users. For example, Wetzker et al. have reported in [121] that 19 of the top 20 most active users in their crawled Delicious data set are spammers. But spam content not only influences the visibility of legitimate resources and users in tagging systems but also important properties of the tagging systems that can be observed at the macro level (see Chapter 3 and [95] for examples). Thus, Wetzker et al. conclude in [24, p. 29–30] that “spam filtering should precede any sophisticated analysis” of tagging systems.

In principle, one can distinguish three different anti-spam techniques [47]: (1) One can manually or automatically classify users as either a spammer or a legitimate user of the system. The spam status of the users is then used for removing their content from the system. (2) One can design the system to reduce the prominence of spam content. For example, this can be achieved by designing spam resistant ranking algorithms. (3) One can try to make contributing spam content more difficult, e. g. by using CAPTCHAs [117] during the creation of a user account. In the following, we concentrate on the former two approaches because only they require algorithms that are specific to tagging systems.

### 2.3.1 Spam Classification

Most of the work on anti-spam techniques in tagging systems is related to the supervised classification of users as either spammers or legitimate users. One of the first supervised spam classifiers has been proposed by Krause et al. in [65]. They found out that especially features related to the co-occurrences of users help to classify users as spammers. In this context, two users are said to co-occur together if they share at least one resource, one tag or one tag-resource pair. Krause et al. have used this co-occurrence information of users together with a set of manually labeled users for training a SVM classifier, which can then be used for automatically labeling the remaining users in the data set.

A further rich source for spam classification is the vocabulary of the users, i. e. how often they have used which tags. For example, the three most successful submissions [20, 37, 59] to the spam detection task of the ECML/PKDD Discovery Challenge 2008 [49] all use this information as their premier feature on which the corresponding classifier gets trained. This success of the tag related features becomes obvious when looking at the statistics of the Bibsonomy data set that has been used during the Discovery Challenge for evaluating the classifiers: In this data set, 92% of all users are spammers and 297,846 of the 359,000 distinct tags are solely used by spammers. Thus, the vast majority of tags directly indicate a spammer.

A further interesting approach for spam classification is described by Neubauer and Obermayer in [82]. The quality of their results is not as good

as those of the previously described supervised classifiers but their approach is nevertheless interesting because it can be used in an unsupervised way, i. e. it does not require a manually labeled training data set. Like Krause et al. in [65], Neubauer and Obermayer also exploit the co-occurrence information of users by partitioning the folksonomy hypergraph into 2-hyperincident connected edge-components. Neubauer and Obermayer have observed that legitimate users are then exclusively contained in the largest connected edge-component. All other edge-components exclusively contain spammers. In [82], this observation is used for creating an unsupervised classifier, but it may also be used as an additional feature in supervised classifiers.

### 2.3.2 Ranking of Spam Content

Until now, we have described methods for automatically classifying users into spammers and legitimate users. In the following, we describe methods that do not assign an explicit spam status to the users but instead try to reduce the prominence of spam content by designing spam resistant ranking algorithms. This is an approach originally used for making ranked results of web search engines more spam resistant (see [24] for an example), which can also be transferred to tagging systems.

One of the first works in this area is the work by Heymann et al. in [47]. They propose a ranking that is based on coincidences between users, i. e. they consider the tag assignments of those users as more reliable who agree more often with other users on the tags for describing a resource. The coincidence based ranking is then compared to a boolean retrieval model and to an occurrence based model (cf. Section 2.4). The quality of the ranked retrieval results is then evaluated with the *SpamFactor* [63]. The SpamFactor of a given ranking increases, the more spam is present in the results and the closer the spam is to the top of the ranking.

The disadvantage of the coincidence based ranking is that it is resistant to uncoordinated spammers but that it can easily be circumvented by collusive attacks by several spammers. To overcome this problem, Wang et al. propose the DSpam tagging system [118]. DSpam also assigns weights to the tag assignments of the different users, but it computes personalized weights based on the social network between users and the similarity of their tag assignments. The evaluation in [118] shows that DSpam is more resistant against spam than the coincidence based ranking of Heymann et al.

Finally, in [83] Noll et al. present the SPEAR algorithm for ranking search results in tagging systems according to the expertise of the users. Such a ranking is more spam resistant because typical spam patterns decrease the expertise of the respective user. For example, users are considered as experts if they often belong to one of the first users to annotate a resource, and if many other users follow their decision to annotate. In contrast, spammers typically belong to one of the last users who annotate



legitimate resources. Furthermore, for resources with spam content, the spammers belong to the first users but they have only very few users who follow them. The evaluation in [83] shows that SPEAR is more successful in demoting spam than comparable algorithms for expertise based ranking like HITS [60].

## 2.4 Retrieval of Resources

In the previous sections, we have summarized works that deal with representing and creating folksonomies. In this section, we summarize works that deal with one of the main applications of tagging systems, namely with the retrieval of resources. In tagging systems, the retrieval of resources is supported by two kinds of interfaces that can either be used for searching resources or for browsing resources. The searching of resources is supported by traditional search interfaces like they are also known from web search, i. e. by an input box that can be used for querying the resource collection with arbitrary search terms. In contrast, the browsing of resources is in most tagging systems supported by visualizations of the tag space [11] like they are provided by tag clouds (see Fig. 2.4). By clicking on one of the tags visualized in the tag cloud, the user implicitly submits a search query with the respective tag as if it would have been entered into the input box of the search interface. Of course, both kinds of interfaces may also be integrated into a single user interface which allows searching and browsing at the same time. Both kinds of interfaces have their merit. It depends on the actual search task of the user which of the two interfaces provides a better support. Typically, two different search tasks are distinguished:

1. On the one hand, there is the simple lookup task during which the user wants to find a specific resource or information [75, 104, 111]. An example for a simple lookup task would be to find articles about the NASA (cf. [104]). For simple lookup tasks, users often prefer traditional search interfaces, which allow to directly enter relevant search terms [104].
2. On the other hand, there is the exploratory search task during which the user has some broader information need that requires multiple searches interwoven with an analysis of the retrieved resources [68, 104, 111]. An example for an exploratory search task would be to identify sport activities that are popular in a certain city like Pittsburgh and then to find resources that provide more details about these activities (cf. [111]). For exploratory search tasks, users often prefer browsing interfaces, which visualize the tag space of the folksonomy [104, 111].

In Subsection 2.4.1, we give an overview of the literature about ranking the search results in tagging systems according to their relevance for the



Figure 2.4: Example tag cloud from <http://www.bibsonomy.org/>. The tags are ordered alphabetically. The font size and intensity indicate the frequency with which the respective tags are used in a collection of resources.

query. Of course, these algorithms can also be used for ranking the results that are shown to the user due to his/her interaction with a tag cloud. Then, in Subsection 2.4.2, we give an overview of the literature about tag space visualizations, more specifically about the visualization of tag clouds.

### 2.4.1 Ranking the Relevance of Resources

With regard to the retrieval of resources, one can distinguish between the *boolean retrieval* and the *ranked retrieval* [74]. In the boolean retrieval, queries can be posed in form of a boolean expression of terms, i. e. terms can be combined with operators like AND, OR, and NOT. For boolean retrieval, resources are typically represented as a set of terms. Thus, it is well suited for the retrieval in narrow folksonomies where a tag can be annotated to a resource at most once (see Section 2.1). In contrast, the ranked retrieval is based on the vector space model in which resources are represented as a vector or a bag of words, i. e. it is well suited for the retrieval in broad folksonomies where a tag can be annotated several times to a resource by different users. In the following, we thus concentrate on the vector space model because in this thesis we are mainly interested in studying broad folksonomies.

In the vector space model, each resource  $r_i$  of a folksonomy is represented by a vector  $v_i$ . In its most simple form, each entry in the vector then corresponds to the frequency with which one of the tags  $t_1, \dots, t_n$  is annotated to  $r_i$ . But the vector space model may also be enriched by taking semantic relations between tags into account [2], or the tag frequencies may

be weighted by measures like TF-IDF (term frequency-inverse document frequency; [74]). But regardless of the actual weight of each tag, the intuition behind the vector space model is that the tag vector captures the relative importance of the different tags for describing a resource.

The vector space model is the fundamental model for several information retrieval tasks like scoring documents on a query, document classification and document clustering [74]. These retrieval tasks have in common that they require to compute the similarity between pairs of resources  $r_i$  and  $r_j$ , or to score the relevance of a resource  $r_i$  for a query  $q$ . A standard way for computing the similarity of resources in the vector space model is the cosine similarity [74]. Given two tag vectors  $v_i$  and  $v_j$ , their similarity is measured as follows:

$$\text{sim}(v_i, v_j) = \cos \Theta = \frac{v_i \cdot v_j}{\|v_i\| \cdot \|v_j\|} \quad (2.1)$$

But the cosine similarity from Equation 2.1 can also be used for scoring the relevance of a resource  $r_i$  for a query  $q$ . For this purpose, the query is also represented as a vector, which can then be compared to the tag vectors of the resources [74], i. e. the relevance of resource  $r_i$  with tag vector  $v_i$  for a query  $q$  corresponds to  $\text{sim}(q, v_i)$ .

In many retrieval systems, the resources do not only have a query specific relevance score, but also a static quality score which is query independent [74]. Probably, the most well-known example of such a static quality score is the PageRank algorithm [14], which is used by Google for ranking web documents based on the link structure of the web. The relevance score of a document for a query then corresponds to a weighted combination of the query dependent score and the static quality score of the document. The weights with which the two scores are combined may be automatically learned [53] in order to achieve an optimal ranking.

Until now, we have mainly concentrated on retrieval and ranking in general. In the following, we concentrate on algorithms that have been specifically proposed for retrieval and ranking in tagging systems, i. e. they make use of the structure of folksonomies. Three research areas are related to the retrieval and ranking in tagging systems: (1) Improving the computation of the query dependent relevance score of resources, (2) improving the computation of the static quality score of resources, and (3) introducing a personalized relevance score of resources into the ranking algorithms.

### Query Dependent Relevance Scores

In folksonomies, one typical problem during computing the query specific relevance score of resources is that an uncontrolled vocabulary is used for annotating the resources. This introduces the typical problem that different users use synonyms for describing the same concept or that tags are pol-

ysemous, i. e. that they have multiple meanings [11]. It is thus a common approach to try to detect the lexical relations between the different tags and then to take these relations into account during the query dependent relevance ranking. For example, the relevance of a resource may be increased if it is annotated with synonyms of one of the search terms.

Examples of approaches for taking the lexical relations between query terms and tags into account are available in [2, 7, 50, 123]. In [2], Abbasi and Staab propose an enriched vector space model in which the original tag vector  $v_i$  of a resource is multiplied with a square matrix  $T_C^S$  that contains similarity values between tags. This square matrix is computed based on the co-occurrence of tags in different contexts  $C$ , e. g. in the context of resources, and the type of similarity  $S$  between tags, e. g. the cosine similarity. The results show that enriching the vector space model is especially useful for queries where only few results are returned.

In [7], Bao et al. show that tagging data from social bookmarking systems like Delicious can be used for optimizing the ranking results in web search. They propose the *Social Similarity Rank* algorithm, which exploits the co-occurrence of tags in the context of the same resource. The inclusion of the Social Similarity Rank as a feature into a baseline web search engine helped to improve the mean average precision of queries.

In [50], Hotho et al. propose the FolkRank algorithm. It uses the same idea as the PageRank algorithm [14], i. e. the relevance of resources for a query is computed based on random walks on the folksonomy hypergraph. In principle, a resource is considered more relevant, the easier it is to reach with random walks on the hypergraph. In FolkRank, it is possible to parametrize the algorithm such that the random walks are more likely starting at specific nodes in the hypergraph. For example, for computing the relevance of resources for a query, the random walks more likely start at the query terms.

In [123], Wu et al. propose to derive the emergent semantics in a folksonomy by representing users, tags and resources in a conceptual space similar to probabilistic topic models [12]. An evaluation with human evaluators has shown that the ranking algorithm retrieves resources that are highly relevant for the search term.

### Static Quality Scores of Resources

With regard to calculating the static quality score of a resource, one important use case is to try to reduce the score of resources that have been uploaded and tagged by spammers. An overview of these ranking algorithms is given in Subsection 2.3.2. Furthermore, one may use ranking algorithms like HITS [60] and PageRank [14], which can be applied on any kind of graph, i. e. also on the folksonomy hypergraph. However, the HITS algorithm has the disadvantage that it is very susceptible to spammers (see [83] and Subsection 2.3) while the PageRank algorithm mainly measures the

degree of the node that represents the resource because the edges in the folksonomy hypergraph are undirected [50]. There also exist adaptations of the PageRank algorithm to folksonomies like FolkRank [50] (see above) and SocialPageRank [7], which both have the idea in common that highly relevant resources are tagged by many users with popular tags. In [7], it has been shown that the inclusion of SocialPageRank as a feature into a baseline web search engine helps to improve the mean average precision of queries.

### Personalized Relevance Scores of Resources

Finally, there exist many algorithms that provide a personalized ranking of resources, i. e. they try to derive which resources a user would likely perceive as interesting. On the one hand, there exist algorithms that personalize the ranking with regard to a query [50, 94, 101, 123]. On the other hand, there exist algorithms which try to recommend the user interesting resources without having a query as a context [15, 41, 61, 85, 112, 120]. Both kinds of algorithms have in common that they analyze which tags a user has previously used and/or which resources they have previously tagged. Then, the algorithms identify like-minded users that have tags and/or resources with the current user in common.

In [92], Santos-Neto et al. predict that like-minded users can be best identified based on the common tags of users because over 90% of the tags are used by more than one user. In contrast, they expect that identifying like-minded users based on common resources is much harder because only 16% of the resources is tagged by more than one user. This prediction of Santos-Neto et al. is confirmed by the evaluation in [120], where it is shown that personalized recommendations based on common tags outperform personalized recommendations based on common resources. The former kind of recommendations only reach the quality of a baseline recommender that simply recommends the most popular resources of the folksonomy. A further source for personalized recommendations are the explicit friendship links between users like they are available in the contact lists of users in a tagging system. But in [94] it is shown that personalized recommendations based on explicit friendship links are outperformed by personalized recommendations based on common tags and/or resources, i. e. users may be friends with each other although they do not have common interests.

#### 2.4.2 Visualizations of the Tag Space

In tagging systems, often a visualization of the tag space is provided to the users. The most prominent example of such a visualization is the tag cloud interface (see Fig. 2.4). According to Rivadeneira et al. [90], tag clouds can be used for (1) *searching* a specific term, e. g. for navigating to the underlying resources, (2) *browsing* the content in a folksonomy, often without a specific

information need, and (3) *impression formation* about the topics covered in a set of resources. Furthermore, Rivadneira et al. mention the task of *recognition*, e. g. for disambiguating the personomies of two users with the same name. But the latter task also requires impression formation, thus we summarize it under *impression formation*.

The support of browsing and impression formation is a unique feature covered by tag clouds. In contrast, when searching for specific resources or information, traditional search interfaces and tag clouds each have their own advantages. For example, traditional search interfaces have the advantage that the users can directly enter relevant search terms without searching for them in a tag cloud. Directly entering search terms is especially useful because tag clouds are often restricted to only visualizing the  $N$  most frequent tags. Thus, users are not able to directly access all resources in a folksonomy from a tag cloud. For example, in [67] Li et al. have shown that in their Delicious data set only approximately 60% of all resources can be directly accessed with the help of the 400 most popular tags. This problem gets even more prevalent, the larger a folksonomy grows [21]. In [67], Li et al. thus propose to introduce hierarchical browsing into tag clouds where the user first browses from the more general tags to the more specific tags. Each browsing step then influences which tags are shown during the next browsing step.

Additionally, the tag cloud also provides benefits during the search task because it can help to mediate between the different vocabularies of the users, i. e. a user may use the tag cloud for locating the terms that have been used by other users for representing a specific concept. Furthermore, in [104] it has been observed that users prefer to click on a tag if it is available in the tag cloud instead of typing it into a search box. Ideally, the traditional search interface is combined with a tag space visualization into a single interface [111].

With regard to the design of a tag cloud, one can distinguish between the design choices for (1) the spatial layout of the tag cloud, i. e. how to arrange the different tags, (2) the visual properties of the single tags, e. g. font size, and (3) the algorithm for selecting the  $N$  tags which are displayed in the tag cloud. The layout of a tag cloud influences for which task a tag cloud is especially well suited. The visual properties of the single tags can be used for influencing which tags get the most attention of the users. Furthermore, the tag selection algorithm can be used for influencing how many resources can be directly accessed from the tag cloud interface.

There exist several approaches for the spatial layout of tag clouds. Probably the most common spatial layout is that of sequentially ordered lines. If each line contains several words, then tag clouds like those in Fig. 2.4 and Fig. 2.5a are generated. If each line only contains a single word, then the tag clouds become vertical lists of words (see Fig. 2.5b). A further visualization might be that of a “bin-packed” cloud where no clear lines are visible

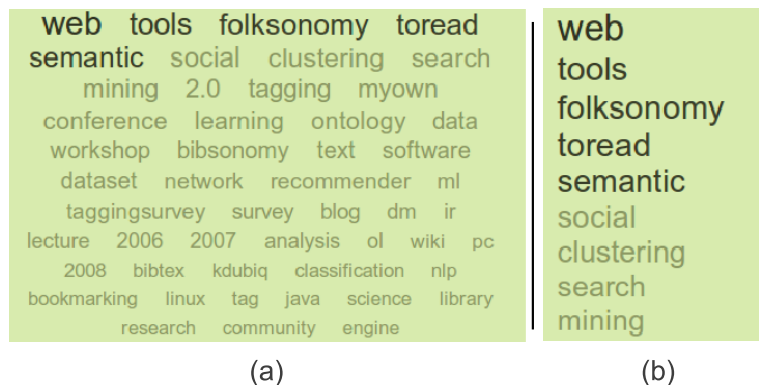


Figure 2.5: (a) Example of a frequency-ordered tag cloud. (b) Example of a frequency-ordered list. Both examples are taken from <http://www.bibsonomy.org/>.

[90], a circular layout [72] or a faceted tag cloud interface [111]. Within the general layout of the tag cloud, a further design decision is the ordering of the tags, like alphabetical ordering (Fig. 2.4), frequency-ordering (Fig. 2.5), or thematic ordering (see [72] for an example).

The choice of the tag cloud visualization depends on the particular task of the user. If the user is looking for a specific term within the cloud, e. g. during a search task, then a sequential layout with alphabetical ordering of the tags is preferred by the users over other layout options. This layout also leads to the fastest selection of the searched term [72]. With regard to looking for tags related to a specific topic, e. g. during a browsing task, there exist contradicting results in the literature. Lohmann et al. report in [72] that a thematic ordering of the tags helped to reduce the time needed for selecting a corresponding term while Schrammel et al. [97] report the fastest selection time for alphabetically ordered tag clouds. It seems that the actual algorithm for creating the thematic ordering influences the time needed by the users [97]. Finally, if the user wants to get an impression of the topics represented in a collection of resources, then a frequency ordered list of tags leads to the best results during user experiments [90].

Not only the spatial layout of tag clouds influences the attention and perception of the users. A further influence comes from the visual properties of the different tags. In [10], Bateman et al. showed that especially the font size, font weight (i. e. bold or normal font) and the intensity have a strong influence on whether and how fast users notice certain tags in a cloud. Furthermore, the position in the tag cloud also has an influence. Tags in the center of the cloud and in the upper left corner are more often noticed and clicked on [10, 97] but this general tendency of the users may be overridden by stronger visual properties [10].

Finally, the algorithm for selecting the  $N$  tags which are displayed may influence the usefulness of a tag cloud for searching, browsing and impression formation. For example, in [43] it has been shown that a selection of tags based on a variation of TF-IDF (term frequency-inverse document frequency; [74]) helps to increase the number of resources which can be directly reached from a tag cloud, if compared to a selection simply based on the frequency of tags. Furthermore, the heterogeneity of the tags in the cloud is increased by a TF-IDF based selection. This increases the navigational efficiency of tag clouds. But the navigational efficiency of tag clouds may also be influenced by which kind of folksonomy is browsed, i. e. a narrow or broad folksonomy. In [46], Helic et al. show that “broad folksonomies create more efficient navigational structures that enable users to find target resources with fewer hops” [46, p. 71].



## Chapter 3

# Macro-Level Properties of Folksonomies

In this chapter, we describe characteristic properties that can be observed at the macro-level of a folksonomy. It is our objective to explain with the Epistemic Dynamic Model how these properties emerge from the micro-level behavior of the individual users (see Chapter 4). We are especially interested in properties that are associated with the emergence of the shared community vocabulary of the users and its effectiveness for navigating the resources indexed in a folksonomy.

With this regard, two central properties are the tag frequency distribution and the growth and size of the used vocabulary. These properties are related as follows to the emergence of a shared vocabulary of the users and the navigability of tagging systems:

1. The **tag frequency distribution** described in Section 3.2 influences how many results are shown for a certain search term. In folksonomies, a power-law like tag frequency distribution can be observed. This means that a few search terms lead to very large result lists while for the majority of search terms only very few results are returned. The exact extent of this effect on the search results is influenced by the exponent of the power-law. The observation of many search terms for which only few results are returned has led to the development of techniques that especially try to improve the browsing and retrieval with these search terms (see Section 2.4).
2. The **size of the vocabulary and its growth** described in Section 3.3 influence how many search terms can be used for accessing the resources in a tagging system. With this regard, a larger vocabulary would be desirable. But on the other hand, the vocabulary growth and size also indicates the level of consensus between the users about which vocabulary to use for annotating resources. The less consensus

Data Set	$ U $	$ T $	$ R $	$ Y $
Delicious	532,938	2,886,015	17,296,850	140,333,714
Bibsonomy	38,920	468,945	1,437,796	16,818,699

Table 3.1: Sizes of the data sets from Delicious and Bibsonomy that are used in this thesis. Both data sets contain the activity of spammers. In the Delicious data set, approximately 1.4% of the users are spammers. In the Bibsonomy data set, approximately 93% of the users are spammers. For more details, see Appendix B.

between the users, the larger and the more diverse the vocabulary. With this regard, a smaller vocabulary would be desirable because it corresponds to a more consistently used vocabulary.

In this thesis, we are focusing on these two properties and how their emergence can be explained with the micro-level behavior of the individual users. They are also the most often discussed properties in the literature about tagging systems (for examples see Tab. 4.5). Nevertheless, also other properties have been discussed in the literature. In Section 3.4, we shortly summarize these further properties.

### 3.1 Used Data Sets

Throughout this thesis, we use two large folksonomy data sets from Delicious and Bibsonomy for studying the properties of folksonomies and for evaluating our Epistemic Dynamic Model. Delicious and Bibsonomy are tagging systems that both create a broad folksonomy. The sizes of the two data sets are summarized in Tab. 3.1. The Delicious data set is based on a crawl of Delicious by the TAGora consortium in November 2006. The Bibsonomy data set is based on a complete dump of the system that has been provided by the owners of Bibsonomy in July 2008. More details about these data sets are available in Appendix B.

As explained in Section 2.1, there exist different views on such folksonomy data sets. In this thesis, we base our analysis on stream views of the folksonomies, more specifically on co-occurrence stream views (see Subsection 2.1.2). We are focusing on stream views because they allow for studying temporal aspects of folksonomies, like the vocabulary growth in Section 3.3. Furthermore, we are focusing on co-occurrence streams because they aggregate the behavior of several users in the context of several resources. This aggregation makes co-occurrence streams less susceptible to the individual behavior of single users or the characteristics of a specific resource, thus

tag	$ U $	$ T $	$ R $	$ Y $
ringtones	3,215	4,458	9,041	74,155
setup	4,176	4,818	5,605	40,689
boat	1,641	5,100	3,907	23,512
historical	1,374	4,664	2,789	16,662
messages	973	3,110	1,326	9,634
decorative	223	1,540	1,057	8,892
costs	709	2,555	1,226	7,359
ff	482	1,544	1,497	5,114
checkbox	869	455	266	4,758
datawarehouse	444	697	814	3,730
tools	183	3,425	4,097	25,437
social	246	3,402	1,765	15,322
design	209	2,797	2,156	14,606
analysis	142	1,855	1,998	12,506
blogs	85	2,250	1,397	8,926

Table 3.2: Statistics of the filtered co-occurrence streams from Delicious (top) and Bibsonomy (bottom). The filtered streams only contain tag assignments made by regular users.

better uncovering the general patterns of user behavior during indexing resources in tagging systems than user or resource stream views.

Throughout this thesis, we will concentrate our analysis on 15 representative tags for which then corresponding co-occurrence streams have been extracted from either the Delicious or Bibsonomy data set. In order to be able to make generalizable observations, we only take tags into account for which at least 1,000 postings of regular users, i. e. non-spammers, are available in the respective data set. For Delicious, we randomly select 10 out of 9,081 possible tags that fulfill this criterion. For Bibsonomy, we randomly select 5 out of 2,064 possible tags.

For each of the 15 tags, we extract a pair of co-occurrence streams from the respective folksonomy data set. The *filtered stream* of such a pair only contains tag assignments that have been made by regular users of the system. In contrast, the corresponding *unfiltered stream* of such a pair contains tag assignments from regular users as well as from spammers.<sup>1</sup> From the *unfiltered stream*, we only use the first  $x$  tag assignments so that it has the same length as its filtered counterpart. The sizes of the extracted filtered

<sup>1</sup>Details about how we identified the spammers in the Delicious and the Bibsonomy data set are available in Appendix B.

tag	$ U $	$ T $	$ R $	$ Y $
ringtones	1,569 (57)	7,953	3,016	74,155 (50,938)
setup	698 (18)	6,498	1,070	40,689 (34,300)
boat	880 (64)	5,404	2,487	23,512 (7,332)
historical	298 (13)	5,430	805	16,662 (13,504)
messages	143 (12)	4,145	178	9,634 (8,669)
decorative	168 (21)	1,462	975	8,892 (697)
costs	84 (3)	3,878	106	7,359 (6,918)
ff	481 (12)	1,578	1,452	5,114 (179)
checkbox	114 (2)	3,596	38	4,758 (4,041)
datawarehouse	426 (2)	723	785	3,730 (147)
tools	283 (169)	3,458	3,652	25,437 (7,708)
social	293 (162)	2,692	1,673	15,322 (6,227)
design	128 (44)	2,108	1,170	14,606 (8,192)
analysis	125 (53)	1,382	1,764	12,506 (2,944)
blogs	146 (96)	1,909	1,002	8,962 (4,158)

Table 3.3: Statistics of the unfiltered co-occurrence streams from Delicious (top) and Bibsonomy (bottom). The unfiltered streams contain tag assignments made by regular users as well as tag assignments made by spammers. In the columns  $|U|$  and  $|Y|$ , the overall number of users and tag assignments in the respective stream are given. The parentheses contain how many of the overall number of users are spammers and how many of the tag assignments are provided by the spammers.

streams are given in Tab. 3.2. The sizes of the unfiltered streams are given in Tab. 3.3 together with information about how many spammers are contained in the respective streams.

We are using the stream pairs for analyzing in how far spammers exhibit a behavior that is different to that of regular users and how this influences the observed properties. Nevertheless, in Chapter 5 during our evaluation of the Epistemic Dynamic Model we are only using the filtered streams because the model is only based on assumptions about the behavior of regular users.

## 3.2 Tag Frequency Distribution

The first characteristic property of co-occurrence streams is the distribution of the tag frequencies. Often, the tag frequency distribution is shown in form of a Zipf plot [126]. In a Zipf plot, the occurrence probability of a tag is shown in dependency on its rank. For example, the most often used tag has rank 1 and the 10th most often used tag has rank 10. One can observe a

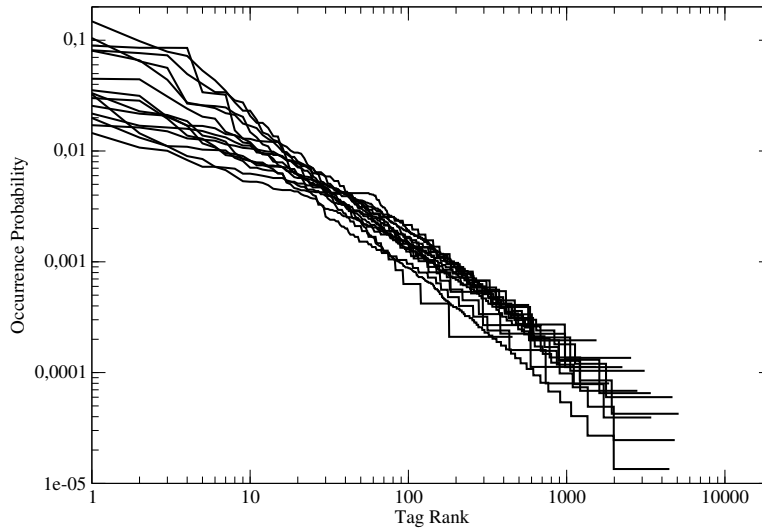


Figure 3.1: Zipf plots of the occurrence probabilities of tags in dependency on their rank for all streams from Tab. 3.2. Only tag assignments of regular users in the respective streams are taken into account.

power-law like decay of the occurrence probabilities for the medium to less frequently used tags in the long tail of the Zipf plot and a flattened slope for the most frequently used tags (cf. [19]).

The characteristic slope of the occurrence probabilities in the Zipf plot, as it has been described by Cattuto et al. in [19], is shown in Fig. 3.1. The plot shows the occurrence probability  $p_n$  of a tag in dependency on its rank  $n$  for all co-occurrence streams from Tab. 3.2. If a logarithmic scaling of the two plot axes is used, the power-law like decay for the medium to less frequently used tags forms a straight line.

It has been observed in [38] that the tag frequencies of the most frequently used tags already stabilize after 100 postings. In Fig. 3.2, an example for this stabilization process is shown for the *social* co-occurrence stream from Bibsonomy (see Tab. 3.2). In [42], the stabilization of the tag frequency distribution has been observed even earlier in a stream, i. e. after 30 postings. Some authors like Golder and Huberman in [38] mainly attribute this stabilization process of the tag frequencies to a consensus between the users about the important tags for indexing a resource. This consensus may form due to seeing the tags of the other users during browsing the system or due to tag recommendations that are based on the tag assignments of the other users (see also Section 2.2). But in [71], Lipczak and Milios found out that this stabilization is misleading because it does not show a consensus between the users but it is caused by the fact that with growing number of tag assignments the influence of the single user on the tag frequencies decreases. Instead of being a sign of collaboration between the users, they

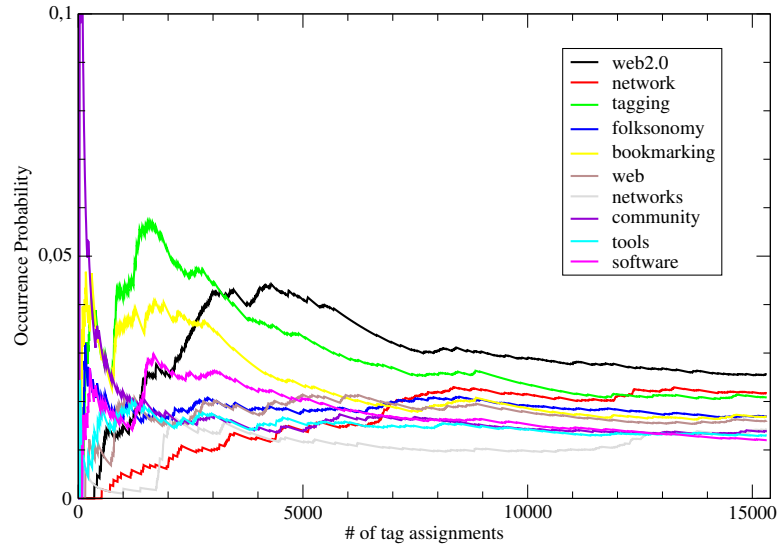


Figure 3.2: Stabilization of the occurrence probabilities of the 10 most often used tags in the *social* co-occurrence stream from Tab. 3.2. Only tag assignments of regular users are taken into account.

found stronger evidence that the stabilization of tag frequencies is rather caused by a shared knowledge between the users.

In text corpora, the distribution of the occurrence probabilities of words is very similar to that of tags in tagging systems. This regularity is commonly known as Zipf's law [126]. It says that the occurrence probabilities of words in natural language texts form a power-law and can be expressed with the following formula:

$$p_n \sim n^{-\alpha}, \text{ with } \alpha > 0 \quad (3.1)$$

Zipf's law states that the slope of the occurrence probabilities will always form a straight line in Zipf plots if both axes are logarithmically scaled. But Zipf's law ignores the flattened slope that can also be observed for the occurrence probabilities of words (see Fig. 3.3). Thus, Benoit Mandelbrot introduced a constant  $m$  into Zipf's law so that it better accounts for the observations made in text corpora [73]:

$$p_n \sim (n + m)^{-\alpha}, \text{ with } \alpha > 0 \quad (3.2)$$

This additional constant  $m$  leads to a flattened slope for the most frequent words. Examples for Zipf plots of the occurrence probabilities of words in text corpora are shown in Fig. 3.3. The used text corpora have the same size and the same topical focus as the streams in Fig. 3.1. The text corpora have been obtained by downloading the resources that are tagged in the

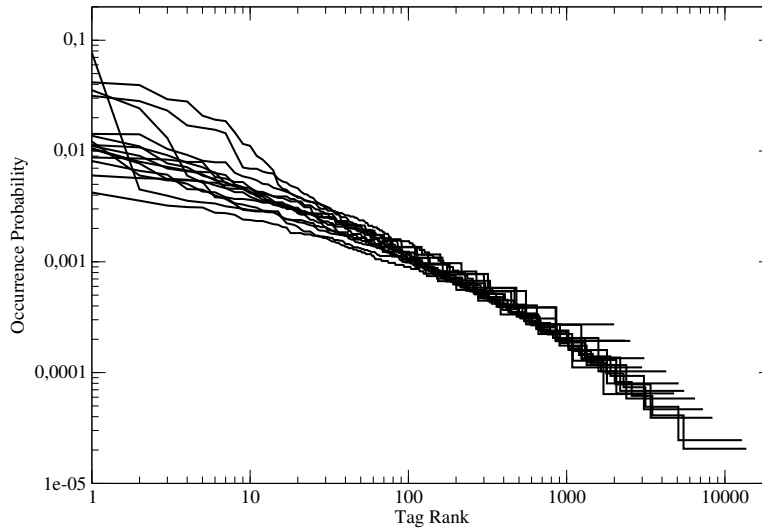


Figure 3.3: Zipf plots of the occurrence probabilities of words for text corpora that have the same topical focus as the streams from Tab. 3.2 and Fig. 3.1. The text corpora have been obtained by downloading the documents that are tagged in the corresponding stream. For better comparability, only as many words from the text corpora are taken into account as there are tag assignments in the corresponding stream.

corresponding stream and taking only as many “word assignments” from the text corpus into account as there are tag assignments in the stream.

Because the Zipf plots of the distributions in Fig. 3.1 and Fig. 3.3 exhibit the same general shape, it is reasonable to assume that they all belong to the same family of heavy-tailed, power-law like distribution functions as they may be approximated with Equation 3.1 or Equation 3.2. But despite of the same general shape it is also important to note that there are significant differences between the plots of tag frequencies in Fig. 3.1 and the word frequencies in Fig. 3.3:

1. There is a much higher probability of observing one of the most frequent tags in a co-occurrence stream than of observing one of the most frequent words in a text corpus. For example, the most frequent tag in the *ringtones* stream is used in 8% of the tag assignments while the most frequent word in the *ringtones* corpus is only used in 0.6% of the “word assignments”.
2. For all analyzed pairs of co-occurrence streams and text corpora, the stream always contains less distinct tags than the corresponding text corpus contains distinct words. The number of the distinct tags or words can be read out from the Zipf plots by looking at tag or word

	Average	Std. Dev.	Minimum	Maximum
Tag Frequencies	0.9266	0.195	0.6572	1.3270
Word Frequencies	0.6398	0.087	0.5514	0.8663

Table 3.4: Average, standard deviation, minimum and maximum of the  $\alpha$ -values if the Zipf plots of the tag frequencies in Fig. 3.1 and of the word frequencies in Fig. 3.3 are approximated with a power-law.

with the maximal rank contained in the plot. For example, the *ring-tones* stream contains 4,458 distinct tags while the *ringtones* corpus contains 12,652 distinct words.

These observed differences point to the fact that if the Zipf plots in Fig. 3.1 and 3.3 are approximated with a power-law (see Equation 3.1) then the best fit is achieved for completely different exponents  $\alpha$ . For example, in case of the tag frequencies in Fig. 3.1 the average of the best-fitting  $\alpha$ -values is 0.9266. In contrast, in case of the word frequencies in Fig. 3.3 the average of the best-fitting  $\alpha$ -values is 0.6398. Furthermore, the distribution of  $\alpha$ -values for word frequencies has a lower standard deviation than for the tag frequencies (see Tab. 3.4).

All in all, it thus seems that the occurrence probabilities of tags and words are influenced by similar mechanisms that cause their common characteristic shape. But there also seems to be an additional influencing factor in case of the tags that causes the overall lower number of tags and the higher probability of observing one of the most frequent tags. Thus, a model for generating the tag frequency distributions in tag streams is likely to extend a model for generating the word frequency distributions in natural language texts. This assumed connection between models for tag frequencies and models for word frequencies is also the underlying assumption of our Epistemic Dynamic Model, which is described in Chapter 4.

Until now, we have considered the tag frequency distributions in co-occurrence streams because they are less susceptible to the individual influences of single users or resources (see Section 3.1). Now, we look at the tag frequency distribution of resource streams. In general, the tag frequency distributions in resource streams also follow the heavy-tailed, power-law like distribution that we observed in co-occurrence streams. But in [42], Halpin et al. noted a significant different behavior for tags between rank 7 and 10 for resource streams in a Delicious data set. Between these ranks, a drop in the occurrence probabilities of the respective tags can be observed (see Fig. 3.4).

In [42, p. 216], Halpin et al. conclude that this drop in the occurrence probabilities is likely “a consistent effect of the way tagging is performed”. Two possible explanations are offered in [42]: (1) It may be related to a



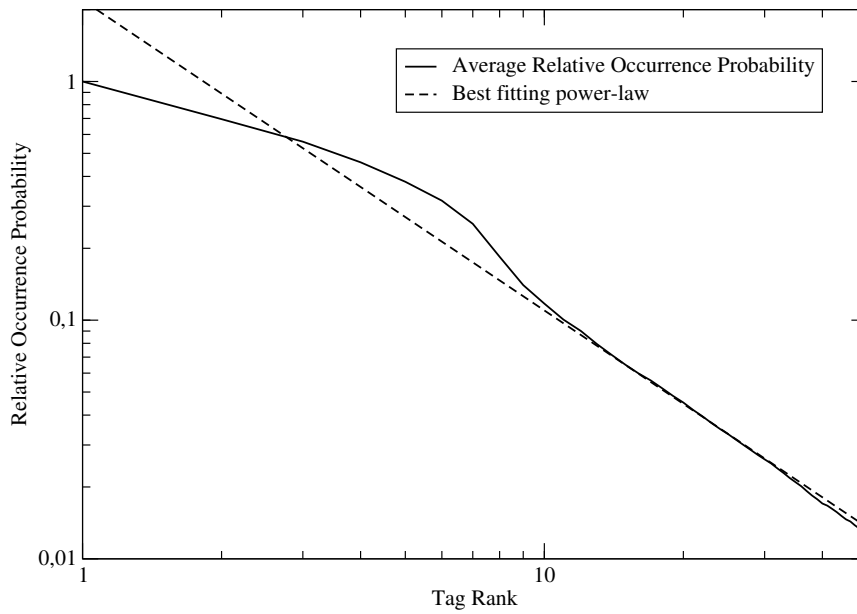


Figure 3.4: Drop in the relative occurrence probabilities for tags between rank 7 and 10 in resource streams. The relative occurrence probability corresponds to the occurrence probability normalized by the occurrence probability of the most frequent tag. The average relative occurrence probability in the graph has been computed by averaging the relative occurrence probability for 500 randomly selected resource streams from our overall Delicious data set. As a guide for the eye, a line for the best-fitting power-law distribution is also included in the graph. Only resource streams are taken into account to which more than 100 regular users contributed.

cognitive effect during tagging, e.g. based on the average number of tags contained in a posting, or (2) it may be an artifact of the user interface specific to Delicious. In Subsection 5.5.3, we show with the help of our Epistemic Dynamic Model from Chapter 4 that this artifact can plausibly be explained as being an artifact of the Delicious user interface.

With regard to how the tag frequency distributions in co-occurrence streams are influenced by the presence of spammers, no single pattern can be identified (see Fig. 3.5). On the one hand, there exist co-occurrence streams like the *ringtones* stream where the presence of spammers in the *unfiltered* version of the stream increases the probability of the infrequent tags. On the other hand, there exist co-occurrence streams like the *social* stream where the presence of spammers rather increases the occurrence probability of the most frequent tags.

All in all, one can conclude from these observations that spammers have a strong influence on the tag frequency distributions in tagging systems. The differences in the tag frequency distributions may be caused by different kinds of spamming strategies like they are described in [63]. Thus, during our evaluation of the Epistemic Dynamic Model in Chapter 5, we have to compare the predictions of our model to the tag frequency distributions in the *filtered* streams because our model only aims at explaining the behavior of regular users.

### 3.3 Vocabulary Growth and Size

The second characteristic property observable in tagging systems is the sub-linear growth of the used vocabulary, i.e. with increasing number of tag assignments in a stream the probability of inventing a new tag decreases. This sublinear growth pattern can be observed in all kind of stream views of a folksonomy as well as in the whole folksonomy [17, 38]. This effect is also well known in linguistics and information retrieval where it is known as Heaps' law [44]. If we transfer Heaps' law to tagging systems, then it states that the number of distinct tags  $|T|$  in a tag stream with  $|Y|$  tag assignments grows according to Equation 3.3.

$$|T| \sim |Y|^\beta \quad (3.3)$$

In Fig. 3.6, the vocabulary growth of all co-occurrence streams from Tab. 3.2 is shown. It can be seen that there is a high variance between the vocabulary growth speeds. For example, after approximately 5,000 tag assignments, the stream with the lowest growth speed contains around 450 distinct tags and the stream with the highest growth speed contains around 1,900 distinct tags. This high variance in the vocabulary growth speeds of single streams can not only be observed for co-occurrence streams but also for resource streams as well as for user streams [17, 38].

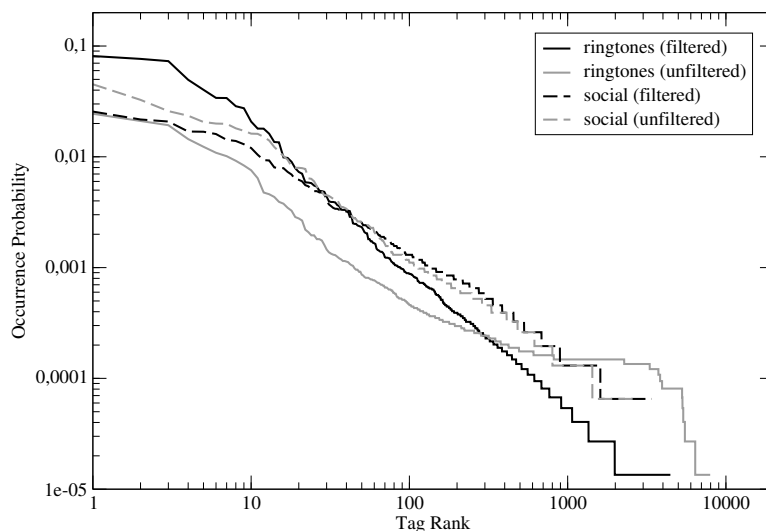


Figure 3.5: Zipf plots of the occurrence probability of tags in dependency on their rank for the *ringtones* and *social* stream pairs from Tab. 3.2 and 3.3. The filtered variant of a stream only contains tag assignments of regular users. The unfiltered variant of a stream contains a mix of tag assignments from regular users and spammers. The detailed plots for the remaining co-occurrence stream pairs from Tab. 3.2 and Tab. 3.3 are available in Appendix B.

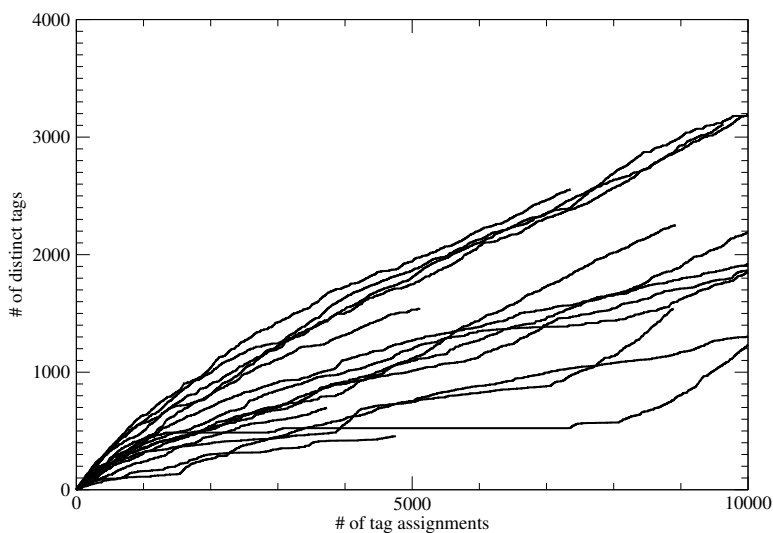


Figure 3.6: Vocabulary growth for all streams from Tab. 3.2 for the first 10,000 tag assignments. Only tag assignments of regular users in the respective streams are taken into account. A high variance in the vocabulary growth rates can be observed.

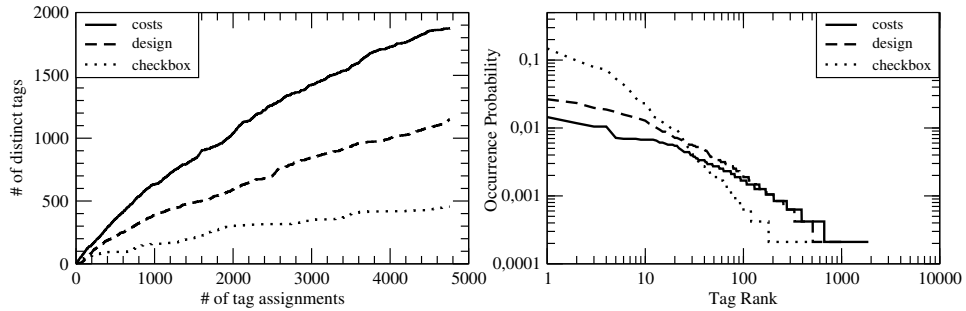


Figure 3.7: Comparison of vocabulary size (left) and tag frequency distribution (right) for the *costs*, the *design* and the *checkbox* co-occurrence streams from Tab. 3.2. The *costs* and the *design* stream have been restricted to their first 4,758 tag assignments so that they contain the same number of tag assignments as the *checkbox* stream. The smaller the vocabulary, the steeper the decline in the occurrence probabilities of tags in the respective stream.

In [5, 113], it has been shown that the vocabulary growth rate according to Heaps' law and the tag frequency distribution according to Zipf's law are correlated with each other. Under the assumption that a tag stream is generated by randomly drawing a number of tags from the tag frequency distribution that adheres to Equation 3.1 or Equation 3.2 with exponent  $\alpha$  then this results in a vocabulary growth according to Heaps' law (see Equation 3.3) with exponent  $\beta = \frac{1}{\alpha}$ . This means that  $\alpha$  and  $\beta$  are negatively correlated with each other.

However, in practice this mathematical dependency does not hold exactly because Zipf's law and Heaps' law can only be used for approximating the real distribution and the real vocabulary growth. Nevertheless, the predicted negative correlation between  $\alpha$  and  $\beta$  can be observed in our tagging data (see Fig. 3.7). Translated to vocabulary size and tag frequency distribution, the negative correlation means that we expect to observe a steeper decline in the occurrence probabilities of tags, i. e. a higher exponent  $\alpha$ , if a smaller vocabulary size is observed, i. e. a lower growth exponent  $\beta$ .

This negative correlation between vocabulary size and tag frequency distribution is also not disturbed by the presence of spammers. In Fig. 3.8, it is shown how the presence of spammers influences the vocabulary size for the *ringtones* and the *social* stream. In case of the *ringtones* stream the presence of spammers leads to a larger vocabulary size. This reflects our observation in Fig. 3.5 where the presence of spammers leads to a less steep decline in the occurrence probabilities of tags. In contrast, in case of the *social* stream, the presence of spammers leads to a slightly smaller vocabulary in Fig. 3.8 and to a slightly steeper decline of the tag's occurrence probabilities in Fig. 3.5.

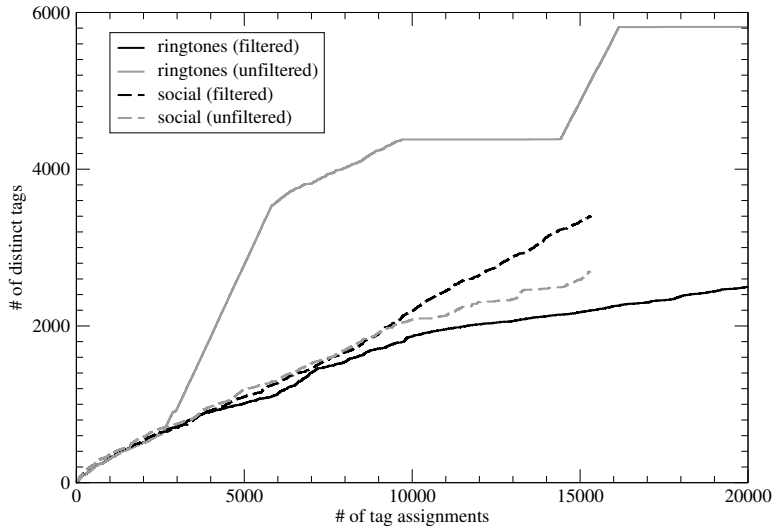


Figure 3.8: Vocabulary growth for the *ringtones* and *social* stream pairs from Tab. 3.2 and 3.3. The filtered variant of a stream only contains tag assignments of regular users. The unfiltered variant of a stream contains a mix of tag assignments from regular users and spammers. Only the first 20,000 tag assignments of the respective streams are shown. The detailed plots for the remaining co-occurrence stream pairs from Tab. 3.2 and Tab. 3.3 are available in Appendix B.

The different effects of spammers on the vocabulary size, and thus also on the tag frequency distribution, may be explained with different spam patterns in Delicious and Bibsonomy. In case of the unfiltered *ringtones* stream from Delicious, the unusual growth pattern between tag assignment 9,700 and 17,300 is caused by two very large postings of a single user, each containing around 4,000 tags. In contrast the unfiltered *social* stream: It is from Bibsonomy where the maximum size of a posting is restricted by the system to 100 tag assignments. Here, spammers can not create such big postings and thus spammers have to use other spam patterns that may even lead to a lower vocabulary growth rate.

### 3.4 Further Properties

In the previous sections, we have discussed two properties that can be observed in tagging systems: The tag frequency distribution, and the vocabulary growth. We discussed these two properties because they are closely related to the emergence of a consensus between the users and to the navigability of the resources in tagging systems. Together with the stabilization of the tag frequencies (see Fig. 3.2), these two properties receive the most

attention in the literature about tagging systems. An overview of where the different properties are discussed in the literature is available in Tab. 4.5 and in Section 4.3.

Nevertheless, in the literature about tagging systems also other properties have been described and discussed. The most influential work with this regard is that of Cattuto et al. in [18] about the Semantic Walker Model (see Subsection 4.3.2 for a summary). They are taking another view on folksonomies, i. e. they are looking at the emergence and the properties of the co-occurrence network of tags. In the co-occurrence network of tags, the nodes represent the different tags in the folksonomy, which are connected by weighted edges. Two tags are connected if there exists at least one posting in which they were used in conjunction. The weight of the edge then corresponds to the number of postings in which the two tags co-occur.

Given such a co-occurrence network of tags, one can then observe and describe the standard properties of networks that are used in the literature about complex networks and the validation of network generating models [9, 84]. For example, in [18], Cattuto et al. study the distributions of the degrees, strengths and weights in the network. Furthermore, they are studying the distribution of the average degree of the nearest neighbors of nodes as well as the average clustering coefficients of nodes in dependency on the nodes' degree.

The view of Cattuto et al. in [18] on folksonomies in form of a co-occurrence network of tags moves the attention towards the internal structure of single postings and how it is influenced by the semantics of tags. In contrast, the stream view that we take in this thesis (cf. Subsection 2.1.2) concentrates more on the dynamics and interactions of users in a folksonomy. But although the two views and the respective observed properties serve different purposes, they are partially correlated with each other. For example, the frequency of a tag influences the maximal weight of each of the edges connected to it in the co-occurrence network. Thus, if we observe in the tag frequency distribution an increased probability of tags that only occur once in the folksonomy then this influences the probability of observing edges whose weight is 1. Similar examples can also be constructed for the other properties of a co-occurrence network and how they are influenced by the tag frequency distribution.

## Chapter 4

# An Epistemic Dynamic Model of Tagging Systems

In this chapter, we describe our Epistemic Dynamic Model of tagging systems. It was first described in [26] and then further refined in [27]. The Epistemic Dynamic Model is based on the assumption that the shared background knowledge and language of the users as well as their exposure to each others tags are relevant for explaining the emergence of the properties that are described in Chapter 3. For example, given the tagging interface of Delicious (see Fig. 2.1), the shared background knowledge and language of users influences which tags are entered into the *tags* input field. Furthermore, the users are exposed to each others tags in form of the *recommended tags* and the *popular tags* suggestions (see also Section 2.2).

Our Epistemic Dynamic Model is designed as a system of building blocks. Each building block corresponds to one of the influence factors that are assumed to be relevant for explaining the emergence of certain properties. In this thesis, we are concentrating on influence factors and building blocks that are required for explaining the properties described in Chapter 3. If further properties should be explained, it might become necessary to extend the Epistemic Dynamic Model with further building blocks, which then cover further influence factors like those described in Subsection 4.3.1.

The rest of this chapter is structured as follows: In Section 4.1, we introduce the building blocks that we assume to be required for explaining the emergence of the properties described in Chapter 3. The two most important building blocks simulate the shared background knowledge of the users (Subsection 4.1.2) and the collaboration between the users due to their exposure to each others tags (Subsection 4.1.3). Then, in Section 4.2 we propose three concrete configurations of the Epistemic Dynamic Model based on these building blocks:

1. The configuration from Subsection 4.2.1 corresponds to the Epistemic Dynamic Model as it has been described in [27]. In this configuration,

the building block for simulating the shared background knowledge of the users is treated as a black box from which we randomly draw words according to word frequency distributions that can be observed in corpora of natural language texts.

2. In the configuration from Subsection 4.2.2, we replace the black box implementation of the shared background knowledge with another implementation that is based on the Semantic Walker Model from Cattuto et al. [18] (see also Subsection 4.3.2). In Chapter 5, we use this configuration for analyzing in how far the Semantic Walker Model provides a plausible explanation of the processes in the black box.
3. In the configuration from Subsection 4.2.3 we deactivate the building block that is used for simulating the collaboration of users due to their exposure to each others tags. In Chapter 5, we use this configuration for analyzing in how far the exposure to each others tags is really required for explaining the properties observed in Chapter 3.

After describing our own Epistemic Dynamic Model, we provide in Section 4.3 an overview of further influence factors and tagging models that are currently discussed in the literature about tagging systems.

## 4.1 Building Blocks of the Epistemic Model

In this section, we present the building blocks that are used in our Epistemic Dynamic Model. Each building block corresponds to one of the influence factors that we assume to be required for explaining how the tag frequency distributions emerge from the micro-level behavior of the individual users. Furthermore, we are interested in explaining the related property of the vocabulary growth. We are especially interested in explaining these two properties because they are closely related to the emergence of the shared community vocabulary and its effectiveness for navigating the resources in a folksonomy (see Chapter 3).

By analyzing the user interface of Delicious in Fig. 2.1, one can identify two important influence factors on a user's tag choice:

- First, a user can add free tags in the *tags* input field. These tags most likely come from a user's background knowledge about the content of the resource. To some extent, a user shares his/her background knowledge, as well as his/her natural language for describing it, with the other users in the tagging system.
- Second, users are exposed to each others tags in form of the tag suggestions provided by the tagging system. In Delicious, the *recommended tags* as well as the *popular tags* are based on the previously used tags of



other users (see Section 2.2). This exposure to each others tags is seen to facilitate a collaboration between users during indexing resources (cf. [38]). Furthermore, users may also be exposed to each others tags in form of tag space visualizations during the retrieval of resources (see Section 2.4).

In the Subsections 4.1.2 and 4.1.3, we present the corresponding building blocks from our Epistemic Dynamic Model that are used for modeling these two influence factors. Prior to that, we describe in Subsection 4.1.1 the building block that is used for simulating the tag assignments as a stream of postings. In Chapter 5, we then show that the three building blocks of our model are sufficient for explaining the emergence of the properties described in Chapter 3.

#### 4.1.1 Simulating a Stream of Postings

In this thesis, we take a stream view on folksonomies (see Subsection 2.1.2). A stream corresponds to a list of postings that is ordered by the creation time of the postings. When simulating such a stream with our Epistemic Dynamic Model, we always start with an empty stream. Then, in each step a new posting is simulated. A posting comprises the information about the set of tags that has been annotated by a specific user to a specific resource. In the following, we describe the building block of our Epistemic Dynamic Model that is used for generating this information about a posting. The actual simulation of how the respective user then selects appropriate tags for annotating a resource is done with the help of the other two building blocks that are described in Subsection 4.1.2 and in Subsection 4.1.3.

As described above, for simulating a posting we first have to predetermine the information about the user, the resource and the size of the posting's tag set. It depends on the kind of stream view, how to predetermine this information: For example, when simulating a resource stream, all postings will be annotated to the same resource. Similarly, when simulating a user stream, all postings will be created by the same user. Only when simulating co-occurrence streams, postings will be created by different users and be annotated to different resources. Depending on the user and resource, a user specific background knowledge or resource specific tag suggestions may be simulated by the other two building blocks of our model.

Nevertheless, in the context of our work, we do not expect that the simulation of user specific background knowledge or resource specific tag suggestions have an important influence on the tag frequency distribution and the vocabulary growth in tagging systems. Thus, in the following all users share the same background knowledge, which can be seen as an aggregation of the (overlapping) background knowledge structures of the users. Furthermore, when simulating the suggestion of popular tags in co-occurrence

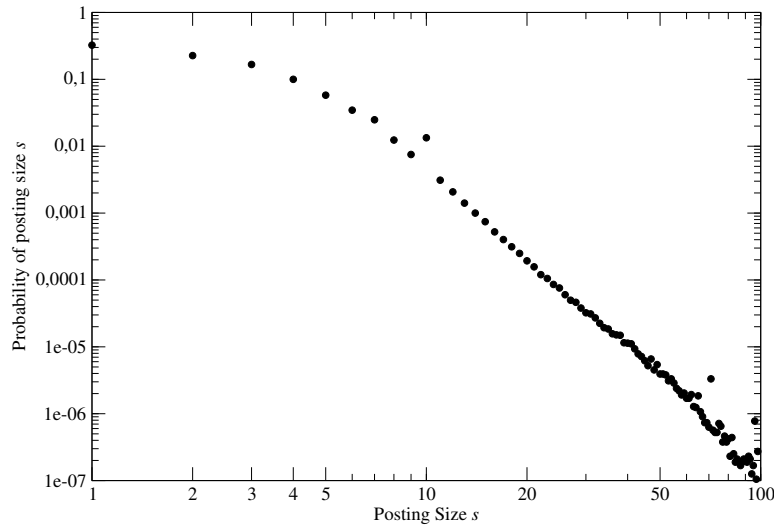


Figure 4.1: Distribution of the posting sizes used for simulating co-occurrence streams. It corresponds to the distribution of posting sizes from regular users in the overall Delicious data set described in Section 3.1.

streams (see Subsection 4.1.3), we also simulate the aggregated popular tags of all resources in the co-occurrence stream. This way, the complexity of the Epistemic Dynamic Model is reduced by a huge amount. Furthermore, more generalizable results are achieved due to abstracting from the concrete influence of a specific user or resource to the aggregated influence of several users or resources.

Besides the information about the user and the resource, a posting also comprises the information about how many tags are assigned by the user to the resource. This information about the number of assigned tags becomes especially important when simulating the background knowledge of users as random walks on semantic networks (see Subsection 4.1.2). In the following, we simulate the size  $s$  of a posting's tag set by randomly drawing it from a probability distribution  $p(s)$ . This parameter of our Epistemic Dynamic Model is a parameter that can be estimated based on our tagging data sets. In [19], it has been discovered that in tagging systems the distribution  $p(s)$  is a heavy-tailed distribution. In Fig. 4.1, the empirical distribution of the posting sizes in our Delicious data set from Section 3.1 is shown. In order to get realistic simulation results, we use this empirically observed distribution of posting sizes during our evaluation in Chapter 5.

#### 4.1.2 Simulating the Shared Background Knowledge

In the following, we describe the building block that simulates with probability  $BK$  how users select appropriate tags from their active vocabulary.

For example, the tags entered into the *tags* input box shown in Fig. 2.1 can be seen to be influenced by the background knowledge of the individual user as well as by the natural language shared with other users. In this thesis, two different implementations of the building block are described and tested: First, we describe why and how word frequency distributions acquired from text corpora can be used for simulating the background knowledge and shared terminology of users. Then we describe how semantic networks can alternatively be used for implementing this building block of the Epistemic Dynamic Model.

### Background Knowledge as Word Frequency Distributions

One way of simulating the selection process of tags from the active vocabulary of users is by using predefined word frequency distributions. For example, when simulating resource streams the word frequency distribution  $p(W|u \cap r)$  may be used. It gives for each word  $w$  from the terminology  $W$  the probability that it will be annotated by user  $u$  to the resource  $r$ . Similarly, when simulating co-occurrence streams the distribution  $p(W|u \cap t)$  may be used. It gives the probability for each word  $w$  that it will be annotated by user  $u$  in the context of the topic  $t$ .

Given these distributions, the tag assignments of a user are simulated by randomly drawing the required number of words from the respective distribution. Obviously, it is not possible to get the active vocabulary and/or frequency distribution for each individual user. Furthermore, the exact distribution may not only be influenced by the current resource  $r$  or topic  $t$  but also by many other factors like the previously visited resources or the purpose for which the user is tagging. Nevertheless, it is reasonable to assume that in any case the occurrence probabilities in the users' active vocabularies adhere to Zipf's law (see [126] and Subsection 3.2). Zipf's law describes an inherent property of human natural language according to which the occurrence probabilities of words in natural language texts form a power-law.

In [36], Gelbukh and Sidorov found out that the concrete power-law exponents of distributions for different topics and authors only have a small standard deviation from an average value. Only the used language (e.g. English) has an influence on the average power-law exponent. Thus, instead of the exact distributions for each user, in the following we approximate them with user independent distributions  $p(W|r)$  and  $p(W|t)$  that give the word probabilities averaged over several users who share the same natural language.

Although it is reasonable to assume that  $p(W|r)$  and  $p(W|t)$  adhere to Zipf's law and that only small deviations between single users can be observed, the question remains about the concrete power-law exponent of the distribution. With this regard it has been shown in [107] that in tagging systems one can distinguish two types of users: (1) The *categorizers* use

tags as a kind of category for organizing resources, and (2) the *describers* describe the content of resources by means of tags. Categorizers choose tags from their active vocabulary so that their tag vectors have a high information value. In contrast, describers choose tag vectors from their active vocabulary so that the tag vectors closely resemble the content of the resource (cf. [107]).

Thus, according to the findings of [107], we expect to observe different power-law exponents for  $p(W|r)$  and  $p(W|t)$ , depending on whether we simulate the tag assignments of a categorizer or describer. Furthermore, it can be concluded that the power-law exponents of  $p(W|r)$  and  $p(W|t)$  for simulating describers are very similar to the power-law exponents observable for the word frequency distributions in continuous texts. Because most users in tagging systems apply a mixture of both tagging styles and because they have a tendency to the descriptive style [107], we assume in the following that  $p(W|r)$  and  $p(W|t)$  can be approximated with word frequency distributions observable in continuous texts. The plausibility of this assumption is evaluated in Chapter 5 by showing that the configuration of the Epistemic Dynamic Model from Subsection 4.2.1, which builds upon this assumption, can be used for reproducing the properties described in Chapter 3.

During the evaluation in Chapter 5, we approximate  $p(W|r)$  and  $p(W|t)$  with empirically measured word frequency distributions from 15 different text corpora consisting of web documents related to the topics of the co-occurrence streams described in Section 3.1. For example, when simulating the co-occurrence stream for a topic  $t$ , we use the word frequency distribution from a text corpus containing all documents that have been annotated in our Delicious data set from Section 3.1 with a tag like *ringtones* that represents the respective topic.

In Fig. 4.2, the frequency-rank plots of the 15 empirically measured word frequency distributions are shown. The plots show that the power-law exponents of the Zipf distributions are very similar across the different topics and the different sizes of the crawled web corpora (i. e. the single plots are rather indistinguishable from each other). This confirms the findings from [36] that the concrete value of the power-law exponent is rather independent of authors and/or topics. Instead, it reflects an inherent property of human natural language. Nevertheless, during evaluating the simulation of co-occurrence streams in Chapter 5, we use the empirically measured word frequency distribution of the corresponding text corpus for simulating  $p(W|t)$ , e. g. the *ringtones* corpus will be used during the simulation of the *ringtones* stream.

All in all, the word frequency distributions from Fig. 4.2 can be seen as a black box that we use for simulating the tag frequency distributions as they would occur in a tagging system if the users are only influenced by their background knowledge. This implementation of the building block for simulating the shared background knowledge of users does not try to explain how the tag frequencies emerge from the structure of natural languages.

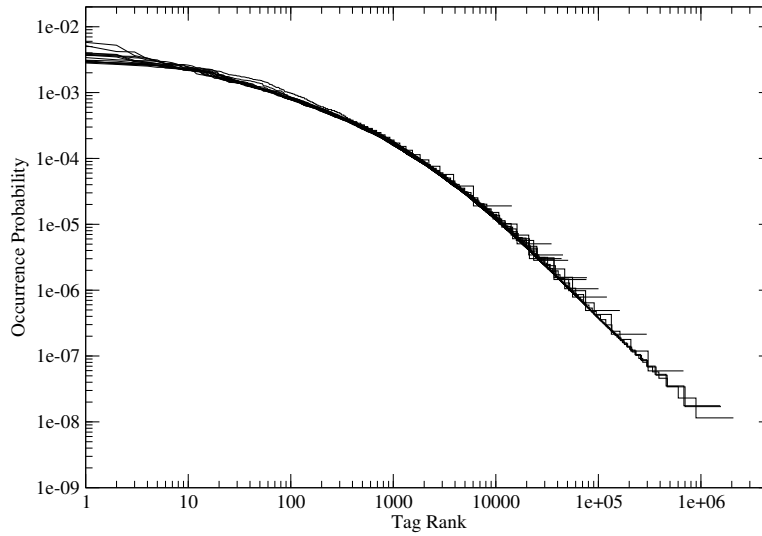


Figure 4.2: Zipf plots of the word frequency distribution empirically measured in the 15 crawled text corpora. The corresponding word frequency distributions will be used in the Epistemic Model for simulating the distributions  $p(W|r)$  and/or  $p(W|t)$ .

### Background Knowledge as Semantic Networks

In the following, we describe an alternative implementation of the building block for simulating the background knowledge and shared terminology of users. Like in the Semantic Walker Model [18], we use an undirected graph  $g$  for modeling a semantic network that represents the background knowledge of an average user. The nodes in the graph represent tags and the vertices represent semantic links between the tags. This approach has the advantage that it provides an explanation how the Zipf distributions  $p(W|r)$  and  $p(W|t)$  may emerge from the memory structure of humans instead of treating them as a black box from which we randomly draw tags.

Given such a graph  $g$ , the simulation of the tag assignments of a user corresponds to a self-avoiding random walk through the graph (see [18]). When simulating an individual co-occurrence stream, all of the random walks for the different simulated users start at the same node in the graph. This node corresponds to the tag that represents the topic for which the co-occurrence stream is constructed. Compared to the original random walk algorithm described in [18], we introduce two modifications:

1. First, we modify how the start node of the random walks is selected at the beginning of the simulation. In [18], the start node is randomly selected among all nodes in the semantic network. In contrast, we introduce a further parameter  $d$ , which allows to restrict this random

selection to nodes in the graph that have the degree  $d$ . The degree  $d$  corresponds to the number of nodes (or topics) to which the start node is directly connected. Thus,  $d$  can be seen as a measure of the semantic breadth of the topic that is simulated. The higher  $d$ , i. e. the more nodes or topics are directly connected to the start node, the more general is the simulated topic.

2. Second, we have to define whether and how the random walks are influenced by imitating one or several of the tag suggestions (see Subsection 4.1.3). We decided that the random walk always resumes at the previously simulated tag of the posting, independent of whether the previous tag has been selected from the tag suggestions or from the semantic network. This simulates how tag suggestions may influence a user's understanding of a resource by triggering word associations. If there does not exist a previous tag in the posting, i. e. if the first tag in a posting is simulated, then the random walk is resumed at the initially selected start node (see above).

The structure of the graph  $g$  is the parameter with the most influence on the properties of the simulated stream. Thus, the structure of the graph used during our simulations should resemble that of real semantic networks. An example for a real semantic network that may be used is the Word Association Norms data set [81], which contains over 5.000 interconnected words. It has been generated by presenting to users a stimulus word and then asking them to write down the first word that came to their mind. In the semantic network, the stimulus word and the responses are connected with each other. The Word Association Norms data set consists of the results of this experiment, repeated for several stimulus words. The resulting graph of this real semantic network has three important properties: (1) It has a small graph diameter, (2) the words form clusters, and (3) a heavy-tailed degree distribution can be observed (see [106]).

Unfortunately, the Word Association Norms data set is too small for simulating co-occurrence streams that may contain more than 5.000 distinct tags. Thus, larger, artificially generated substitutes of this real semantic network have to be used. In [18], the authors show analytically that a small graph diameter is a minimal requirement for the artificial substitutes in order to explain the sublinear growth of the vocabulary. In [18], three different models for producing networks with a small diameter have been tested: (1) The *Erdős-Rényi Model* [32], (2) the *Uncorrelated Configuration Model* [16], and (3) the *Watts-Strogatz Model* [119]. In Tab. 4.1, the properties of the resulting semantic networks are shown.

As can be seen in Tab. 4.1, none of the three graph generation models tested in [18] reproduces all three properties observable in real semantic networks. Thus, they can not be used for realistically simulating the properties

Model Name	Small Diameter	High Clustering	Degree Distrib.
Word Association Norms [81]	✓	✓	heavy-tailed
Erdős-Rényi Model [32]	✓		uniform
Uncorrelated Config. Model [16]	✓		heavy-tailed
Watts-Strogatz Model [119]	✓	✓	normal
Growing Network Model [106]	✓	✓	heavy-tailed

Table 4.1: Comparison of the properties of the Word Association Norms data set from [81] with the properties of networks generated with the Erdős-Rényi Model, the Uncorrelated Configuration Model, the Watts-Strogatz Model and the Growing Network Model. Only the Growing Network Model is able to reproduce all three properties of the Word Association Norms data set.

that emerge in tagging systems due to the structure of the users’ background knowledge. Instead, we propose to use the Growing Network Model described by Steyvers and Tenenbaum in [106]. It generates networks that share all three properties with real semantic networks (see Tab. 4.1). In the following, we assume that semantic networks generated with the Growing Network Model can be used for approximating real semantic networks like the Word Association Norms data set.

The plausibility of this assumption is evaluated in Chapter 5. We show that the configuration of the Epistemic Dynamic Model from Subsection 4.2.2, which builds upon this assumption, reproduces the properties described in Chapter 3. Furthermore, we show that using the Growing Network Model leads to better evaluation results than using one of the other network generation models originally tested in [18]. This is shown by comparing the simulation results achieved with the Growing Network Model to the best performing network model from [18], namely the Watts-Strogatz Model. The Growing Network Model and the Watts-Strogatz Model, which we use during our evaluation in Chapter 5, can be summarized as follows:

- The *Growing Network Model* generates graphs that have a small diameter, a high clustering coefficient, and a heavy-tailed degree distribution [106]. Initially, the graph consists of  $m_{gn}$  fully connected nodes. Then, new nodes are added until the graph contains  $N$  nodes. Each new node  $a_n$  differentiates an already existing node  $a_e$  by connecting to

$m_{gn}$  randomly selected direct neighbors of  $a_e$ . For simulating a graph with properties similar to the Word Association Norms data set, the parameter  $m_{gn}$  has to be fixed to 11 (see [106]). It is the purpose of the Growing Network Model “to capture at an abstract level the relations between the statistics reported ... [for real semantic networks] ... and the dynamics of how semantic structures might grow” [106].

- The *Watts-Strogatz Model* is able to generate graphs of size  $N$  that have the small-world property, i.e. the nodes form clusters and the graph has a small diameter [119]. Initially, all  $N$  nodes in the graph are arranged in a ring in which each node is connected to its  $m_{ws}$  direct neighbors to the left and to its  $m_{ws}$  direct neighbors to the right. Then, each link is rewired with probability  $p_{ws}$  to another random node. The node degree distribution approximates a normal distribution with an average degree of  $2 \cdot m_{ws}$  and a very low variance.

All in all, word frequency distributions and semantic networks provide competing implementations of how to model the tag selection of users according to their background knowledge. But this does not mean that they are contradicting each other. In case of the word frequency distributions, we simply assume that the background knowledge leads to the emergence of a Zipf like tag frequency distribution with an exponent comparable to what can be observed in natural language texts. This treats the background knowledge of the users as a black box from which we randomly draw tags. In contrast, the semantic networks provide a “model in the model” that tries to explain the processes inside of the black box. Furthermore, the Growing Network Model, which is used for generating artificial semantic networks, can be seen to be a “model in the model in the model”. In terms of Popper’s critical method (cf. Section 1.1), if we are able to show that both models of the background knowledge are able to reproduce the properties from Chapter 3 then randomly drawing from  $p(W|r)$  and  $p(W|t)$  can be seen as approximating random walks on semantic networks.

### 4.1.3 Simulating the Tag Suggestions of the User Interface

In this subsection, we describe the building block of our Epistemic Dynamic Model that simulates with probability  $I$  that a user accepts one of the tags suggested by the user interface (see Fig. 2.1). In our model, we are especially interested in analyzing tag suggestions that expose a user to the tags of the other users. This exposure to each others tags introduces a feedback mechanism between the different users in a tagging system. This kind of feedback mechanism is often seen as one of the reasons why tagging works despite of its uncontrolled nature [38].

For example, in case of Delicious (see Fig. 2.1), the users are exposed to each others tags by means of the *recommended tags* and the *popular tags*.



The *recommended tags* consist of the intersection between the current user's tags and all tags already assigned to the current resource. The *popular tags* consist of the 7 most popular tags at the current resource. By clicking on any of the suggested tags the user can easily include it in his/her posting.

If the primary purpose of our Epistemic Dynamic Model would be to simulate as accurately as possible the dynamics in a concrete tagging system like Delicious, then the exact tag recommendation algorithms from the respective system would ideally be simulated. For example, in case of Delicious this would require to simulate the *recommended tags*, the *your tags* and the *popular tags*. But in the following, we are more interested in generalizable results about how users are influenced by being exposed to each others tags. Thus, we are only focusing on modeling the influence coming from the suggestion of the *popular tags* already assigned to the current resource. In case of Delicious, this also models to some extent the influence of the *recommended tags* because they are also based on the tags already assigned to the current resource. According to [116], most existing tagging systems provide tag suggestions that are based on the tags already assigned to the current resource (see also Section 2.2).

All in all, for simulating the influence of the *popular tags* on the user, we introduce two further parameters  $n$  and  $h$  to the model:

- The parameter  $n$  represents the number of popular tags that can be accessed by the user. For example, when simulating tag assignments made in Delicious to a single resource,  $n$  corresponds to the number of popular tags shown by the user interface, i. e.  $n = 7$ .
- The parameter  $h$  may be used for restricting the number of previous tag assignments that are used for determining the  $n$  most popular tags. For example, for  $n = 7$  and  $h = 300$  only the 7 most popular tags during the last 300 tag assignments are suggested to the user. If all previous tag assignments should be taken into account, then  $h$  should be set to the current size of the stream.

### Simulating Popular Tags in Resource Streams

When simulating the tag stream for a single resource, the value of  $n$  is chosen according to what can be observed in the user interface of the tagging system, e. g. in case of Delicious  $n = 7$ . The correct value of  $h$  is unknown but most likely corresponds to the current size of the stream, i. e. the popular tags are computed based on all previous tag assignments. Setting  $h$  to the current size of the stream eliminates this parameter from the Epistemic Model.

### Simulating Popular Tags in Co-occurrence Streams

When simulating the tag assignments of a co-occurrence stream, different values for  $n$  and  $h$  are required. This is because co-occurrence streams aggregate the postings from several resource streams. But as explained in Subsection 4.1.1, in order to reduce the complexity of the Epistemic Model we decided to not simulate each resource and each user separately. Instead, we try to achieve a similar effect by adapting the values of  $n$  and  $h$  as follows:

Co-occurrence streams aggregate the postings from several resources. In principle, for each of these resources the users would see another set of 7 most popular tags. Thus, the users not only have access to the 7 most popular tags in the co-occurrence stream but to more tags. For example, if a co-occurrence stream aggregates 100 resources then users have access to at most 700 most popular tags at a certain point in time. Thus, the number of aggregated resources influences our choice of  $n$ . The more resources are aggregated, the higher  $n$ .

The popular tags of the different aggregated resource streams are not disjunct of each other. Instead, there is a significant overlap between them. This effectively reduces the number of available, distinct popular tags. We assume that for a very specific topic like *datawarehouses*, the overlap between the set of popular tags is higher than for a more general topic like *ringtones*. Thus, the more general the topic of the co-occurrence stream and the more aggregated resources, the higher the value of  $n$ .

The simulation of co-occurrence streams also influences our choice of the parameter  $h$ : In [38] it has been observed that resources typically receive most of the postings within a few days and then the activity at that resource drops off. Thus, a resource typically does not contribute new postings and its set of popular tags during the whole duration of a co-occurrence stream but only in a certain time frame. This kind of “aging” of resources may be simulated with the  $h$  parameter. It corresponds to the number of tag assignments after which a specific resource no longer contributes to the set of popular tags available in a co-occurrence stream.

All in all, in co-occurrence streams, the parameters  $n$  and  $h$  can be seen to also model in an abstract way the influence coming from the number of aggregated resources and from the semantic breadth of its topic. Thus, in contrast to the simulation of resource streams, there is no unique choice of  $n$  and  $h$  for simulating co-occurrence streams. Instead, these parameters have to be fitted to the specific co-occurrence stream in order to best reproduce its concrete properties like the tag frequency distribution.

### Strategies for Choosing from the Suggested Tags

Until now, we have explained how to determine the value of  $n$  and  $h$ , depending on whether a resource stream or a co-occurrence stream is simulated.

In the following, it is now the question how the users choose one of the  $n$  suggested tags. There are several plausible ways how the tag choice process of users can be modeled:

1. Each of the suggested tags is chosen with the same probability. This corresponds to users who randomly choose one of the suggested tags, and users who are not influenced by a tag's semantics. While this may be a plausible strategy for spammers, it is unlikely capturing the behavior of real users.
2. The suggested tags are chosen with a probability that is proportional to the tag's occurrence probability in the previous stream. This corresponds to users who pick up the latest tagging trends, and who are easily influenced by other users. This selection strategy is very similar to the strategy modeled in the Yule-Simon Model with Memory [19] and the model of Golder and Huberman [38].
3. The suggested tags are chosen with a probability that is proportional to the tag's probability in the user's background knowledge. This corresponds to users who choose tags that seem appropriate based on their own background knowledge. Thus, a user unlikely imitates tags that he/she perceives as semantically irrelevant in the current context.

We assume that in the reality a mixture of the second and the third tag selection strategy can be observed, i. e. users are influenced by their own background knowledge and by how often they have seen a certain tag in the context of other resources. Nevertheless, in order to keep the Epistemic Model as simple as possible, we decided to only test the third strategy in the context of our evaluation in Chapter 5. It seems to be the most plausible strategy that users are stronger influenced in their choice by their own background knowledge about the semantics of a tag and only to a lesser extent by the choices of other users.

Nevertheless, we are aware that this may be a simplification of the actual tag choice process of users. For example, the second strategy is required for explaining the observations by Tisselli in [110]. Tisselli analyzed how the tag *thinkflickrthink* emerged as the winner out of several other possible tags for annotating images that protest against censorship in Flickr. Such a winner-takes-all phenomenon can only be explained if the tag choice of a user is influenced by the behavior of the other users (see also [105]). Further evidence for the influence of the second strategy can also be seen in the user study of Held and Cress [45]. They showed that tag suggestions have an influence on the semantic memory structure of users and their information search and learning. But clarifying the question to which extent users apply a mixture of the second and/or third strategy is out of the scope of this thesis and subject to future research.

## 4.2 Configurations of the Epistemic Model

In the previous section, we have described the three building blocks of our Epistemic Dynamic Model that are modeling different influence factors on a user's tagging behavior. These building blocks can be combined in different ways, leading to different configurations of the Epistemic Dynamic Model. In this thesis, we are concentrating on three configurations of the Epistemic Dynamic Model that we summarize in this section:

1. *Epistemic Model with Word Frequencies* (Subsection 4.2.1)
  - Background knowledge is modeled with word frequencies
  - Users get tag suggestions
2. *Epistemic Model with Semantic Networks* (Subsection 4.2.2)
  - Background knowledge is modeled with semantic networks
  - Users get tag suggestions
3. *Natural Language Model* (Subsection 4.2.3)
  - Background knowledge is modeled with word frequencies
  - Users do not get tag suggestions, i. e. the imitation probability  $I$  is set to 0%.

These three configurations serve different purposes during our evaluation in Chapter 5. By comparing the evaluation results of the first and the second configuration, we are able to evaluate in how far random draws from word frequency distributions or random walks on semantic networks are better suitable for modeling the background knowledge of users. Furthermore, by comparing the first two configurations to the third configuration and to the Semantic Walker Model, we get more insights in how the exposure to each others tags in form of tag suggestions influences the macro-level properties from Chapter 3.

### 4.2.1 The Epistemic Model with Word Frequencies

The first configuration of the Epistemic Model uses predefined word frequency distributions for modeling the background knowledge and the shared terminology of the users (see Section 4.1.2). Furthermore, it includes the building block for simulating the influence coming from the imitation of tag suggestions that are shown to the user in the user interface of a tagging system (see Section 4.1.3). Finally, the sizes of the postings, which are added by the users, are modeled by the same distribution as it can also be observed in the Delicious system (see Section 4.1.1).

Parameter	Description
$I$	Probability that a user imitates a previous tag assignment from the stream.
$BK$	Probability that a user selects a word from his/her active vocabulary. Fixed to $BK = 1 - I$ .
$n$	Number of the most popular tags in the previous tag stream that can be imitated. For simulating resource streams, its value corresponds to the number of visible popular tags in the Delicious tagging interface.
$h$	The maximal number of previous tag assignments the system uses to determine rankings for the $n$ popular tags.
$p(W r)$ or $p(W t)$	The word frequency distributions that give for each word $w$ from the terminology $W$ the probability that $w$ will be annotated in the context of resource $r$ or topic $t$ . $p(W r)$ and $p(W t)$ are approximated by the word frequency distributions in Fig. 4.2.
$p(s)$	The probability that a user adds a posting of size $s$ to a stream. $p(s)$ is approximated by the distribution of posting sizes in our Delicious corpus (see Fig. 4.1).

Table 4.2: Parameters of the *Epistemic Model with Word Frequencies*.

A summary of the model parameters is available in Tab. 4.2. Only the parameters related to the imitation of tag suggestions, i. e. the parameters  $I$ ,  $n$  and  $h$ , are free parameters of the *Epistemic Model with Word Frequencies*. All other parameters are fixed to certain, predefined values or distributions. The parameter  $BK$  is fixed to the value  $BK = 1 - I$ . The word frequency distribution  $p(W|r)$  and/or  $p(W|t)$  as well as the posting size distribution  $p(s)$  are fixed to the empirically estimated distributions shown in Fig. 4.2 and Fig. 4.1.

#### 4.2.2 The Epistemic Model with Semantic Networks

The second configuration of the Epistemic Model uses semantic networks for modeling the background knowledge and the shared terminology of the users (see Section 4.1.2). This is the only difference to the previously described *Epistemic Model with Word Frequencies*. Both configurations of the Epistemic Model thus share the parameters  $I$ ,  $BK$ ,  $n$ ,  $h$  and  $p(s)$ . Only the parameters  $g$  and  $d$  for simulating the semantic networks are replacing the previously used word frequency distributions  $p(W|t)$  and  $p(W|r)$ .

Parameter	Description
$I$	Probability that a user imitates a previous tag assignment from the stream.
$BK$	Probability that a user selects a word from his/her active vocabulary. Fixed to $BK = 1 - I$ .
$n$	Number of the most popular tags in the previous tag stream that can be imitated. For simulating resource streams, its value corresponds to the number of visible popular tags in the Del.icio.us tagging interface.
$h$	The maximal number of previous tag assignments the system uses to determine rankings for the $n$ distinct tags.
$g$	Undirected graph representing a semantic network of tags. The semantic network is generated by the Growing Network Model (see Section 4.1.2) with $m_{gn} = 11$ .
$d$	Degree of the node from which all random walks will be initially started.
$p(s)$	The probability that a user adds a posting of size $s$ to a stream. $p(s)$ is approximated by the distribution of posting sizes in our Delicious corpus (see Fig. 4.1).

Table 4.3: Parameters of the *Epistemic Model with Semantic Networks*.

A summary of the model parameters is available in Tab. 4.3. Besides the free parameters  $I$ ,  $n$  and  $h$ , the *Epistemic Model with Semantic Networks* has the additional free parameter  $d$  that corresponds to the degree of the node from which all random walks will be initially started. It can be used for estimating the semantic breadth of the topic  $t$  or the resource  $r$  for which a co-occurrence or resource stream is simulated (see Section 4.1.2). The parameter  $g$  is a fixed parameter because we only use semantic networks generated by the Growing Network Model with  $m_{gn} = 11$  so that  $g$  reproduces all properties of real semantic networks like the Words Association Norms data set (see Section 4.1.2).

### 4.2.3 The Natural Language Model

The third configuration of the Epistemic Model removes the influence coming from imitating tag suggestions. It thus only models the annotation of tags due to the shared background knowledge and the shared terminology of the users (i.e. their natural language). That's why we also call it the *Natural Language Model*. The Natural Language Model corresponds to the *Epistemic Model with Word Frequencies* with  $I = 0.0$ , i.e. the probabil-

Parameter	Description
$p(W r)$ or $p(W t)$	The word frequency distributions that give for each word $w$ from the terminology $W$ the probability that $w$ will be annotated. $p(W r)$ gives the probability for annotating $w$ to a resource $r$ . $p(W t)$ gives the probability for annotating $w$ in the context of the topical area represented by the tag $t$ . $p(W r)$ and $p(W t)$ are approximated with the probability of observing $w$ in a text corpus (see Fig. 4.2).
$p(s)$	The probability that a user adds a posting of size $s$ to a stream. $p(s)$ is approximated by the distribution of posting sizes in our Delicious corpus (see Fig. 4.1).

Table 4.4: Parameters of the Natural Language Model.

ity of imitating tag suggestions is 0%. The Natural Language Model can be used for studying the dynamics in tagging systems as we expect them to be caused by the shared background knowledge and the shared natural language of the users. A summary of the model parameters is available in Tab. 4.4. Both parameters are fixed to empirically observed values. Thus, the Natural Language Model has no free parameters.

### 4.3 Related Work

There are two areas of work related to the Epistemic Dynamic Model. First, there is the work related to which influence factors are relevant for explaining the dynamics in tagging systems. Besides the shared background knowledge and the exposure to each others tags, there are two further influence factors discussed in the literature, namely the influence coming from the personal organization objective of the users and the content of the resources (see Subsection 4.3.1). Second, there exist quite a number of tagging models in the literature that also have the objective to explain the emergence of macro-level properties of the tagging system with the influence coming from the micro-level behavior of users. The tagging models from the literature are summarized in Subsection 4.3.2.

#### 4.3.1 Influence Factors

Tagging is a complex process during which a user's tagging decision is influenced by several factors. In Section 4.1, we have described the building blocks of our Epistemic Dynamic Model that each correspond to an assumed influence factor, namely the shared background knowledge of the users and the exposure to each others tags, which facilitates the collaboration be-

tween users. Besides these influence factors, two further influence factors are discussed in the literature about tagging systems [71], (1) the personal organization objective of the users, and (2) the content of a resource.

The **personal organization objective** of a user is an important motivation of users for contributing to a tagging system. It corresponds to the motivation to use tagging for organizing a personal collection of resources for later retrieval [77]. In [62], Körner et al. show that this motivation may result in two distinctive tagging styles according to which the users either categorize or describe a resource. Often, users apply a mixture of both styles. Most users have a tendency to the descriptive style. Categorizers can be distinguished from describers by looking at the vocabulary size and the tag frequency distribution in the personal vocabulary of a user, i. e. by looking at the *user stream view* (see Definition 6). Categorizers have a smaller vocabulary than describers and the tags have more equally distributed frequencies [62]. In the Epistemic Dynamic Model, we model users that have a tendency to the descriptive style (cf. Subsection 4.1.2).

The **content of a resource** may either have an indirect or a direct influence on the tag assignments of a user. The indirect influence of the content is via the background knowledge of a user. The indirect influence affects which parts of our background knowledge get activated. In Subsection 4.1.2, the indirect influence is covered by the building block responsible for simulating the background knowledge of the users “about the content of the resource”. The indirect influence can occur for any kind of resource, i. e. for textual resources as well as for non-textual resources like images. Related to this indirect influence is the assumed, non-observable variable of the semantic breadth of a resource (cf. Subsection 4.1.2 and [19, 34]).

In contrast, the direct influence can only occur if the resource has textual content associated with it, like a title or description. In [71], Lipczak and Milios found out that textual content influences the probabilities with which users decide on which synonym out of a list of possible synonyms to use for describing a certain topical aspect. The most simple example for the direct influence of textual content is that on the probability of selecting either the singular or plural form of a word. In [71], it has been shown that the plural form of a word in the title of a resource makes it more likely that a user also chooses the plural form as a tag and vice versa. According to Lipczak and Milios, 15-26% of the tag assignments also occur in the title of a resource.

### 4.3.2 Tagging Models

Our Epistemic Dynamic Model, initially described in [26] in 2008, has not been the first nor the last tagging model described in the literature. Several authors have developed models that can be used for simulating the tagging process, and that try to increase our understanding of it. In the following, we describe the most important tagging models currently discussed in the



literature. Tab. 4.5 summarizes which influence factors they model, and against which observable properties they have been evaluated. The table also includes our own Epistemic Dynamic Model.

### **Polya Urn and/or Simon Model**

One of the earliest works in the field of tagging models is the work of Golder and Huberman in [38]. They assume that there are two basic factors that influence users during assigning tags to a resource: On the one hand, they assume that a user selects a tag based on his/her background knowledge about the content of the resource. On the other hand, the user is exposed to previous tag assignments of other users that he/she may simply imitate in order to reduce the effort required for assigning own tags. Nevertheless, the model of Golder and Huberman only includes the imitation of previous tag assignments. They have not modeled the influence coming from the shared background knowledge of the users.

The model of Golder and Huberman corresponds to the stochastic *Polya Urn Model* originally described in [30]. The model is best explained by the metaphor of an urn containing balls with different colors. In each step of the simulation, a ball is selected from the urn and then it is put back together with a second ball of the same color. Transferred to the simulation of tag streams, the different colors of the balls correspond to the distinct tags in a stream. In [30], it has been shown analytically that the fraction of balls of a given color stabilizes over time. The model may thus be suitable for explaining the emergence of stable tag frequencies (see Section 3.2).

One major drawback of the Polya Urn Model is that it assumes a fixed vocabulary size because no balls with previously unknown colors are added during the experiment. This restriction of the original Polya Urn Model is removed in the *Simon Model* [103], which assumes that with a low probability of  $p$  balls with new colors are added to the urn. Thus, the Simon Model leads to a linear growth of the vocabulary size. Furthermore, it has been shown in [103] that for the Simon Model the frequency distribution converges to a plain power-law. Thus, it is not able to explain the cut-off for the most frequent tags that can be observed in the Zipf plot of the tag frequency distribution (see Section 3.2).

All in all, Golder and Huberman conclude from their findings that the collaboration of users in form of imitating one another's tags is an important driving factor behind the stability observed in tagging systems. Nevertheless, they also state that shared knowledge may also contribute to the stability because the stability also persists for less common tags that are not suggested to the users in the user interface of Delicious.

Model Name	Influence Factors				Modeled Properties			
	Collab.	Shared Knowl.	Personal Organiz.	Resource Content	Tag Freq. Distrib.	Freq. Stabil.	Vocab. Growth	Others
Polya Urn Model [30, 38]	✓				✓	✓		
Simon Model [103]	✓				✓	✓		
Yule-Simon Model w. Memory [19]	✓				✓			
Halpin et al. Model [42]	✓		✓		✓	✓		
Organizing Model [88]		✓	✓		✓			✓
Semantic Walker Model [18]		✓					✓	✓
Semantic Imitation Model [34]	✓	✓	✓	✓	✓	✓		
Multinomial Tagging Model [98, 99]	✓							✓
Epistemic Dynamic Model	✓	✓			✓		✓	

Table 4.5: Overview of the influence factors implemented in different tagging models described in the literature. Furthermore, it is given against which observable properties the models have been evaluated.

### Yule-Simon Model with Memory

In [19], Cattuto et al. propose the Yule-Simon Model with Memory, which is a variation of the Simon Model. Like the Simon Model, it assumes that with a low probability of  $p_{ys}$  a new tag is invented by the users. Thus, also the Yule-Simon Model with Memory leads to a linear growth of the vocabulary size. But when imitating a previously used tag, the Yule-Simon Model with Memory takes the order of the tags in the stream into account. Instead of imitating all previous tag assignments with the same probability it introduces a kind of long-term memory. The long-term memory provides a fat-tailed access to the previous tag assignments, i.e. the probability of selecting a tag assignment located  $x$  steps into the past is given by a function  $Q_t(x)$  that returns a power-law distribution of the selection probabilities:

$$Q_t(x) = \frac{a(t)}{x + \tau} \quad (4.1)$$

In this formula,  $t$  is the number of tag assignments already simulated and  $a(t)$  is a normalization factor so that  $\sum_{x=1}^t Q_t(x) = 1$ . The parameter  $\tau$  is the characteristic time scale over which recently added tag assignments have comparable selection probabilities. Cattuto et al. suggest in [19] that the parameter  $\tau$  can be interpreted to model the semantic breadth of the topic currently simulated with the Yule-Simon Model with Memory. The more general a simulated topic, the higher  $\tau$  should be.

All in all, the Yule-Simon Model with Memory is the first model that is able to reproduce the characteristic slope of the Zipf plot of the tag frequency distributions with its exponential cut-off for the most frequent tags (see Section 3.2). Thus, Cattuto et al. conclude from their findings that the cut-off in the observed tag distributions should be attributed to the long-term memory of the users.

### Halpin et al. Model

In [42], Halpin et al. propose a model that offers an information theoretic perspective on the tag selection process of a user. In this model, users not only collaborate by imitating one another's previous tag assignments but they also select tags based on their expected information theoretic value, i.e. whether the tags help to quickly find the tagged resource. The information theoretic value is used for modeling the personal organization objective of users.

In the Halpin et al. Model, the imitation of the previous tag assignments of other users is modeled as a Simon Model (see above). The Simon Model is then amended with a model for selecting tags based on their information theoretic value. The information theoretic value of a tag is 1 if it can be used for selecting an cognitively appropriate number of resources, such as the number of resources that fit on the screen. In practice, the number may

vary between the users depending on which number of resources is perceived as being appropriate. In terms of the tagging styles described in [62], this modeled behavior corresponds to a categorizing user (cf. Section 4.3.1).

All in all, Halpin et al. show in their evaluation that the model leads to a plain power-law distribution, i. e. it does not reproduce the exponential cut-off for the most frequent tags in the Zipf plot of the tag frequency distribution. Furthermore, it only leads to a linear growth of the vocabulary size. Thus, with regard to explaining these two observable properties, the Halpin et al. Model has no advantages over the plain Simon Model. Thus, the selection of tags based on their information theoretic value only seems to have a minor influence on these properties.

### Organizing Model

In [88], Rader and Wash propose the *Organizing Model*, which integrates the influence coming from a user's personal organization objective, and the influence coming from the shared background knowledge of the users. The personal organization objective is modeled by a preference of the users to reuse tags from their own vocabulary. In the Organizing Model, users have a 50% chance that they reuse tags that they previously applied to another web page. If a user does not reuse previous tags then the tags are chosen from the user's background knowledge.

The shared background knowledge is simulated by randomly drawing a tag from a power-law distribution. This approach is very similar to what is described in Subsection 4.1.2. But instead of drawing from an empirically observed distribution, like we do in our Epistemic Dynamic Model, Rader and Wash draw from an artificially generated power-law distribution. In [88], it is not explained which exponent has been used for generating this artificial power-law distribution.

In [88], Rader and Wash have evaluated their model for two different observable properties, namely the tag frequency distribution and the inter-user agreement. The inter-user agreement measures how often different users chose the same tag for describing a resource. The inter-user agreement is influenced by the tag frequency distribution. For example, given a power-law distribution, the inter-user agreement depends on the exponent of the distribution.

Rader and Wash show in their evaluation that the Organizing Model is able to reproduce a power-law like tag frequency distribution as well as the level of inter-user agreement, which can be observed in tagging systems. Overall, Rader and Wash conclude from these results that it is plausible to assume that idiosyncratic processes, as they are driven by a personal organization objective, influence the tag choice strategy of users. Based on the evaluation results, the shared background knowledge of users seems to explain that the tag frequency distribution belongs to the family of power-

law distributions. The influence from the personal organization objective of the users then leads to the concrete power-law exponent that is typically observed in tagging systems.

### Semantic Walker Model

In [18], Cattuto et al. propose the *Semantic Walker Model*. It only models the influence coming from the shared background knowledge of the users. The general idea of the Semantic Walker Model is to model the tagging process as an exploration of a semantic network that represents the shared background knowledge of the users. In Subsection 4.1.2, we have used the Semantic Walker Model for providing an alternative implementation of the building block in our Epistemic Dynamic Model that is used for simulating the background knowledge of the users. More details about the Semantic Walker Model are available in Subsection 4.1.2 on page 47.

During their evaluation of the Semantic Walker Model, Cattuto et al. consider the sublinearity of the vocabulary growth (see Section 3.3) and several properties that are suitable for characterizing the co-occurrence network generated during the tagging process (see Section 3.4). The Semantic Walker Model is the second model, together with our Epistemic Dynamic Model, which is able to explain the sublinear vocabulary growth. Cattuto et al. also analyze how the semantic network influences the simulated properties. They conclude that the nodes in the semantic network need to have a finite average degree and that the semantic network needs to have a small diameter in order to reproduce the observed properties, including the sublinearity of the vocabulary growth. These results suggest that most of the properties emerging in tagging systems are mainly caused by the structure of the users' background knowledge.

### Semantic Imitation Model

In [34], Fu et al. propose the *Semantic Imitation Model*. The Semantic Imitation Model models the tag assignment process by means of a probabilistic topic model [12]. According to the model, users try to infer the topics of a resource by means of the words contained in the resource's content and by means of the tags already assigned by other users. The user then assigns those tags to the resource that will help him/her in the future to reconstruct the previously inferred topics. The Semantic Imitation Model thus models the influence of the shared background knowledge as well as of the collaboration and the personal organization objective of the users.

In the Semantic Imitation Model, the background knowledge of users corresponds to two probability distributions  $p(c|w)$  and  $p(w|c)$ .  $p(c|w)$  is used for inferring a resource's topics  $c$  from the words  $w$  in the content of the resource and from the tags already assigned by other users.  $p(w|c)$  is

then used for determining the words  $w$  that should be assigned as tags to the resource. The collaboration is modeled by applying  $p(c|w)$  not only on the content of the resource but also on the already assigned tags. The personal organization objective is modeled by selecting tags according to  $p(w|c)$  that maximize the probability that the user can later correctly reconstruct the previously inferred topics of the resource.

The Semantic Imitation Model is the only model that explicitly models the content of a resource as a distribution of words. But also the Semantic Imitation Model only covers the indirect influence of the content (see Subsection 4.3.1) because its influence is mediated through the background knowledge of the users. Nevertheless, it is the first model that uses an observable variable, i. e. the word distribution of the resource's content, for modeling this influence. All other models only use non-observable variables corresponding to the semantic breadth of the content, e. g.  $\tau$  in case of the Yule-Simon Model with Memory, or  $d$  in case of our extended version of the Semantic Walker Model (see Subsection 4.1.2 on page 47).

All in all, the evaluation in [34] shows that the Semantic Imitation Model is able to reproduce the power-law like tag frequency distribution with exponential cut-off for the most frequent tags. Furthermore, it also predicts a stabilization of the tag frequencies that is similar to what can be observed in tagging systems. Later, in [35], Fu and Dong have shown for a slightly modified version of the Semantic Imitation Model that it can predict the vocabulary size of users in an artificial tagging experiment. The evaluation results suggest that the internal knowledge structures  $p(c|w)$  and  $p(w|c)$  are the parameters with the most influence on the tagging process.

### **Fuzzy Trace Multinomial Model**

In [99], Seitlinger and Ley propose a Fuzzy Trace Multinomial Model that hypothesizes on the cognitive backgrounds of how users are influenced in their tagging decision by being exposed to the tags of other users. In general, the Fuzzy Trace Multinomial Model distinguishes between the explicit and implicit processing and/or imitation of tags that the user has seen prior to his/her tagging decision: If a user explicitly processes a tag then it increases the probability that the user reminds the exact verbal form of the tag during his/her tagging decision. In contrast, if a user implicitly processes a tag then this leads to an activation of the tag itself as well as of semantically related tags in the background knowledge of the user.

The influence of the explicit and implicit processing of tags is comparable to the direct and indirect influence of the resources' content discussed in Subsection 4.3.1. Seitlinger and Ley show in [99] that it is possible to distinguish between the explicit and implicit processing with the help of a user experiment. The data from the experiment can be used for estimating the probabilities with which users either explicitly or implicitly process the

tags. Seitlinger and Ley conclude that the explicit processes have a stronger influence on a user's tagging decision than the implicit processes.

### Models Similar to the Epistemic Model

Shortly after we have published our initial idea of the Epistemic Dynamic Model in June 2008 in [26], Rader and Wash have presented their *Imitation-Popular Model* in [88] together with their Organizing Model (see above). The Imitation-Popular Model is very similar to the configuration of the Epistemic Dynamic Model from Subsection 4.2.1 that uses empirically observed word frequency distributions for simulating the shared background knowledge of users. The most important difference to our configuration of the Epistemic Dynamic Model from Subsection 4.2.1 is how the shared background knowledge is simulated. We use empirically observed word frequency distributions. In contrast, Rader and Wash use artificially generated power-law distributions for which they do not specify the used exponent. From their description it does not become clear whether this exponent is a free parameter of the model that can be adapted to the simulated co-occurrence streams or whether it is fixed to some specific value.

Furthermore, Vojnovic et al. have presented their *User's Choice Model* in [116]. Again, it is very similar to the configuration of our Epistemic Dynamic Model from Subsection 4.2.1. Its main difference to our model is with regard to which frequency distribution is used for simulating the shared background knowledge of the users. In [116], the frequency distribution is not restricted to a power-law. Instead, any kind of distribution can be used. But the work in [116] also has another objective than our work: We use the model for reproducing and explaining observable properties. In contrast, in [116] Vojnovic et al. analyze the mathematical properties of the User's Choice Model, e.g. whether one can reconstruct the "true rankings" of tags when users are influenced by the suggestion of popular tags. Vojnovic et al. define the true rankings to be the rankings of the tags as they would emerge from the background knowledge of the users. This analysis becomes important in the context of Chapter 6 where we evaluate how the suggestion of popular tags influences the indexing quality in tagging systems.

#### 4.3.3 Summary

In this section, we have presented an overview of further influence factors on a user's tagging decision that are discussed in the literature. Furthermore, we have presented an overview of the most important tagging models from the literature. For the tagging models, we have described their general idea as well as (1) which influence factors on the users' tagging behavior they model, and (2) against which properties of observed tagging behavior they have been evaluated. A summary of this information is available in Tab. 4.5.

Historically, the research on tagging models started with very simple models, like the Polya Urn Model suggested by Golder and Huberman in [38], which only focus on modeling the collaboration between users. The intuition behind these models is that the collaboration is able to explain the emergence of a consensus between users about how to describe resources. Consequently, the models have been used for analyzing how the collaboration affects properties like the tag frequency distribution or its stabilization over time.

But the results of more recent models suggest that the background knowledge of the users is the factor that most influences the tagging behavior of the users and the principal nature of the observed properties. For example, only models that include the background knowledge, i. e. the Semantic Walker Model as well as our Epistemic Dynamic Model, are successful in reproducing the sublinear vocabulary growth in tagging systems. Furthermore, it seems that the exponential cut-off for the most frequent tags is also caused by the structure of the users' background knowledge (cf. the Semantic Walker Model, the Semantic Imitation Model and our Epistemic Dynamic Model). The collaboration between the users only seems to be able to change details of the properties, e. g. the exponent of the power-law like tag frequency distributions, but not their principal nature.

An alternative explanation for the exponential cut-off of the tag frequency distribution is provided by the Yule-Simon Model with Memory. It attributes this effect to the long-term memory of the users. But during our evaluation in Chapter 5, we show that the Yule-Simon Model with Memory can not as good reproduce the observed tag frequency distribution as our own Epistemic Dynamic Model. In this light, the explanation provided by our Epistemic Dynamic Model seems more plausible than that provided by the Yule-Simon Model with Memory.

With regard to which influence factors are required for explaining the observable properties of tagging systems it seems reasonable that all four influence factors currently discussed in the literature are required for covering the complete dynamics in tagging systems. But it depends on the observable properties and also on the specific view on tagging systems, which of the influence factors are more important. The evaluation results of the current tagging models suggest that especially the shared background knowledge, which also includes the indirect influence of the resources' content, and the collaboration between users are important for explaining the emergence of the tag frequency distribution and the vocabulary growth in co-occurrence and resource streams. The influence of the personal organization objective of the users may become more important when looking at the properties in user streams.



## Chapter 5

# Evaluation of the Epistemic Model

In the previous chapter, we have presented our Epistemic Dynamic Model of tagging systems. It is based on the assumption that users are influenced in their tagging decision by (1) their shared background knowledge and language, and (2) by being exposed to the tags of other users. The model corresponds to a theory about how these influence factors interact with each other and how their interaction leads to the emergence of the macro-level properties of tagging systems described in Chapter 3.

In Section 4.3, we have presented competing models that have been designed with the same objective as the Epistemic Dynamic Model, i. e. to explain how the interaction of influence factors leads to the emergence of specific macro-level properties. Given such a list of competing models, how can we evaluate whether it is rational to prefer one model over another model? In such a case, Popper [86] suggests to develop test statements whose truth or falsity can be checked for the different models (see also Section 1.1). We shall then prefer those of the competing models “whose falsity has not been established” [86, p. 8] by the test statements.

In case of tagging models, natural test statements are whether the models are able to reproduce properties that can be observed in tagging systems. Such test statements are natural because they are connected to the initial objective of using the models to explain the emergence of specific properties. In Fig. 5.1, the general idea of this evaluation methodology is shown. It can be used for comparing models with each other that try to explain the same properties. Thus, our Epistemic Dynamic Model is in principle comparable to models that also target the tag frequency distribution or the sublinear vocabulary growth.

The remainder of this chapter is structured as follows: In Section 5.1, we start with an overview of different methods for evaluating the truth or falsity of our test statements. Then, in Section 5.2, we explain the different

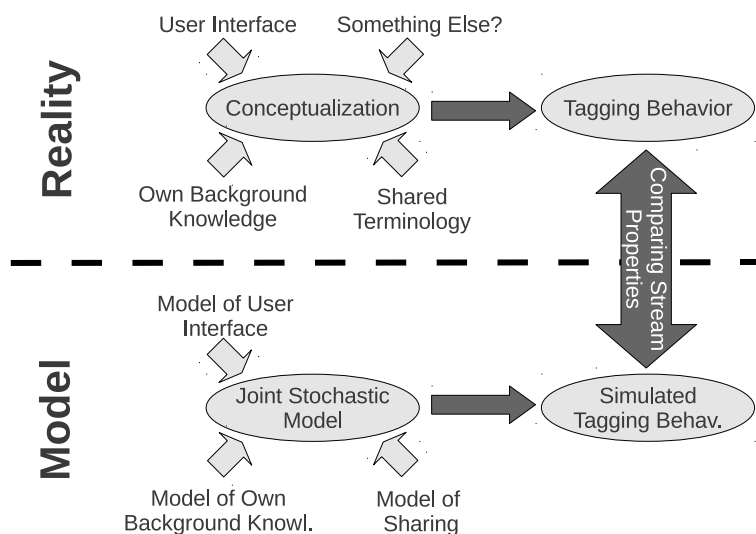


Figure 5.1: Methodology for evaluating tagging models that try to explain the emergence of specific properties in tagging systems. It is checked in how far the same properties emerge in the simulated tagging behavior as in the real tagging behavior.

measures that we use during our evaluation. In Section 5.3, we present the outline of our evaluation. The outline gives an overview of the compared models and the objectives of the comparison. In Section 5.4, we present the evaluation results of the different configurations of our Epistemic Dynamic Model, of the Yule-Simon Model with Memory and the Semantic Walker Model. Finally, in Section 5.5 we discuss our evaluation results.

## 5.1 Comparing Simulated and Observed Properties

In this thesis, we evaluate tagging models by testing whether they are able to reproduce properties that can be observed in tagging systems (cf. Fig. 5.1). This kind of test is also commonly used in the literature about tagging models. The intuition of this test is, the more accurate a model can reproduce the observed properties, the better is the model capturing the processes that lead to the emergence of the observed property.

Although the evaluations of the models described in Section 4.3 are all based on this method, they still differ in the properties against which the models are evaluated. The choice of the property depends on the original objective of the respective model. For example, our Epistemic Dynamic Model and most of the other tagging models from Tab. 4.5 have been designed for explaining the emergence of the tag frequency distribution, and how it is

influenced by different factors. Two tagging models are only comparable if they target the same property.

Furthermore, the evaluations differ in how the simulated and the observed properties are compared to each other. Three different methods are used in the literature about tagging models:

1. **Analytical Evaluation:** This method is used for showing mathematically that the simulated tagging behavior has certain properties. For example, for the Polya Urn and the Simon Model it has been shown in [30, 103] that the models are able to explain power-law tag frequency distributions. But the problem of the analytical evaluation is that it can only be used for showing properties of the simulated tagging behavior. For comparing the simulated and the observed tagging behavior, one has still to use one of the other two evaluation methods described in the following.
2. **Visual Comparison:** For the visual comparison, the selected property of the simulated and the real tagging behavior is visually plotted. For example, the simulated and the observed tag frequency distribution may be plotted together in a Zipf plot (see Section 3.2). The closer together the plots are visually, the better is the model in reproducing and explaining the emergence of the respective property. The problem with this technique is that it only provides a rather subjective and very coarse measure of the goodness-of-fit between the simulated and the observed property.
3. **Goodness-of-Fit Tests:** Goodness-of-fit tests consist of two consecutive steps. In a first step, an objective measure of distance between the simulated and the observed property is defined and applied on the data. The outcome of this first step already provides a relative measure of the goodness-of-fit to an observed property, i. e. it may be used for objectively deciding which of two competing tagging models provides a better fit. Then, in a second step, this distance may be used for computing an absolute measure of the goodness-of-fit, i. e. it can objectively be decided whether the simulated and the observed property are statistically indistinguishable or not. While the first step can be applied on any property, e. g. on the vocabulary growth, the second step can only be applied on properties that can be represented as probability distributions, e. g. on the tag frequency distribution.

In this thesis, we use a goodness-of-fit test for evaluating the different configurations of our Epistemic Dynamic Model and for comparing them to models described in the literature. The reason for this decision is that only goodness-of-fit tests can be used for objectively comparing two competing models. In contrast, the visual comparison of simulated and real

properties can only be used for getting a first intuitive impression of the ability of a model to reproduce the respective property. Nevertheless, most of the tagging models reported in the literature have been evaluated with the visual comparison method, sometimes in combination with an analytical evaluation. Only Rader and Wash have used goodness-of-fit tests in their evaluation [88].

## 5.2 Evaluation Measures

Most of the models described in the literature, including our own Epistemic Dynamic Model, have been designed with the objective to explain the emergence of the tag frequency distribution in tagging systems (see Tab. 4.5). Thus, we primarily use the tag frequency distribution for comparing our Epistemic Dynamic Model to the other models. Only for the Semantic Walker Model we also include the sublinear vocabulary growth into our evaluation and comparison to the Epistemic Dynamic Model. Accordingly, we distinguish between measures of the goodness-of-fit of the simulated tag frequency distributions (Subsection 5.2.1), and measures of the difference between simulated and observed vocabulary size and growth (Subsection 5.2.2).

### 5.2.1 Comparing Tag Frequency Distributions

When we speak of tag frequency distributions, we imply a certain stochastic view on the process of assigning tags to resources. In this stochastic view, tag assignments are modeled as a stochastic experiment that consists of randomly drawing for each of the distinct tags its usage frequency in the stream. The frequencies of the tags then sum up to the length of the stream. Modeling the tag frequencies in a stream as a stochastic experiment is a prerequisite for applying statistical methods like goodness-of-fit tests.

During our stochastic experiment, the frequencies of the tags can be described in form of their probability function  $f(x)$ , which gives the probability of drawing a tag that has the frequency  $x$ . Because tag frequencies are measured on an ordinal scale, i. e. we can arrange them from the smallest to the largest value, we can furthermore define the distribution function  $F(x) = \sum_{t \geq x} f(t)$ , which describes the accumulated probability of observing a tag that has a frequency of  $x$  or higher.

Of course, if we use tagging models for simulating tag streams, we are not really drawing the tag frequencies from an explicitly defined distribution function  $F(x)$ . The distribution function is only implicitly defined by the tagging model and the chosen parameter values. Instead, we empirically observe tag frequencies in the simulated tag stream. The same holds for the real tag streams, where we also only empirically observe tag frequencies. These empirically observed tag frequencies can be used for constructing the empirical distribution function  $S(x)$ . For a tag stream with  $n$  distinct

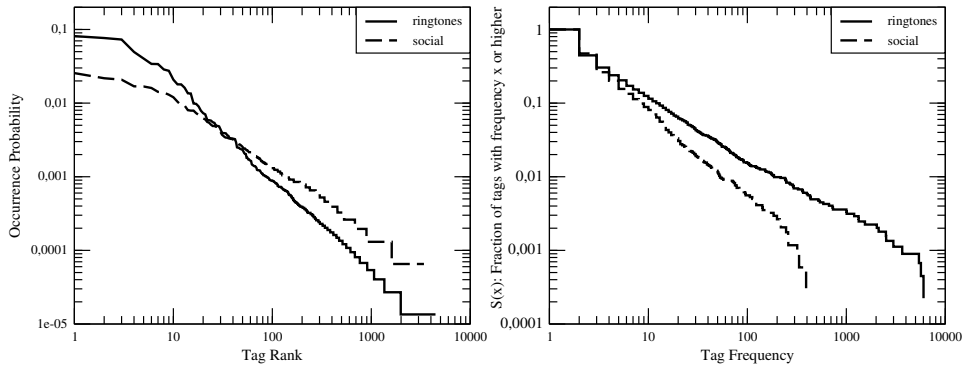


Figure 5.2: Zipf plot (left) and empirical distribution function  $S(x)$  (right) of the tag frequency distribution of the filtered *ringtones* and *social* streams from Tab. 3.2. Zipf plots are the most common representation of the tag frequency distribution in the literature about tagging systems. The empirical distribution function  $S(x)$  forms the basis for applying statistical methods during our evaluation.

tags and the tag frequencies  $x_i \in \mathbb{N}, i = 1, \dots, n$ , the empirical distribution function  $S(x)$  is a step function defined by Equation 5.1.

$$S(x) = \frac{\text{number of } x_i \geq x}{n} \quad (5.1)$$

The empirical distribution function gives the probability of observing a tag that occurs at least  $x$  times. In Fig. 5.2, this view of the tag frequency distribution in form of its empirical distribution function is compared to the previously used representation in form of a Zipf plot (cf. Fig. 3.5). Like for the Zipf plots, we also use a logarithmic scaling of the axes for the plots of the empirical distribution function so that a power-law distribution results in a straight line.

Given this view on the tag frequency distributions in form of their empirical distribution functions, we can now use statistical goodness-of-fit tests for comparing a simulated to a real tag frequency distribution. In the following, we use the Smirnov test [23, p. 456ff] for this purpose. The Smirnov test is a nonparametric test for comparing in how far two empirical distribution functions  $S_1(x)$  and  $S_2(x)$  can possibly be drawn from the same unknown distribution function  $F(x)$ . In our case,  $S_1(x)$  and  $S_2(x)$  are based on the tag frequencies observed in a simulated and a real tag stream. The Smirnov test has to be preferred over other tests if non-normal distributions are tested, like the heavy-tailed tag frequency distribution, and if at least an ordinal measurement scale is used. Examples of the usage of the Smirnov test are available in [22, 88]. In [22], it is used for evaluating whether observed distributions follow a power-law. In [88], it is used for evaluating the Organizing Model (see Subsection 4.3).

Given two empirical distribution functions  $S_1(x)$  and  $S_2(x)$ , the Smirnov test first measures the maximum distance  $D$  between the two functions:

$$D = \max_{-\infty < x < \infty} |S_1(x) - S_2(x)| \quad (5.2)$$

This distance  $D$  can be used in two ways: First, we can compare the distances achieved by different tagging models with each other. One model provides a better explanation for the emergence of the tag frequencies than another model if it achieves lower distances between the simulated and the real tag frequency distributions. Second, we can also compute the level of significance  $p$  of a concrete  $D$  value. A low  $p$ -value provides evidence of a discrepancy between the two compared distributions that goes beyond what would be expected if  $S_1(x)$  and  $S_2(x)$  are two independent empirical observations of the same, unknown distribution function  $F(x)$ . In contrast, a high  $p$ -value means that the compared distributions are consistent with the hypothesis that they are empirical observations of the same, unknown distribution function. Given the distance  $D$  between two mutually independent empirical distribution functions  $S_1(x)$  and  $S_2(x)$ , the level of significance  $p$  can be computed as follows [87, p. 624]:

$$p = Q_S([\sqrt{N_e} + 0.12 + 0.11/\sqrt{N_e}]D) \quad (5.3)$$

$Q_S$  is a monotonic function with the limiting values  $Q_S(0) = 1$  and  $Q_S(\infty) = 0$ .  $N_e$  is a normalization factor that accounts for the size of the data sets. If  $S_1(x)$  contains  $n$  distinct tags and  $S_2(x)$  contains  $m$  distinct tags, then  $Q_S(\lambda)$  and  $N_e$  are defined as follows:

$$Q_S(\lambda) = 2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2\lambda^2} \quad (5.4)$$

$$N_e = \frac{n \cdot m}{n + m} \quad (5.5)$$

Given this  $p$ -value, one would typically reject the hypothesis that  $S_1(x)$  and  $S_2(x)$  are empirical observations of the same  $F(x)$  if  $p < 0.1$ . For  $p \geq 0.1$  the hypothesis would be accepted (cf. [79]). If  $S_1(x)$  and  $S_2(x)$  are not mutually independent, e. g. if the simulation parameters have been fitted according to the real distribution, then Equation 5.3 will only give an upper bound for the actual  $p$ -value (cf. [39]). In that case, one can use the  $p$ -value from Equation 5.3 only for rejecting the hypothesis that they are drawn from the same  $F(x)$ , i. e. if  $p < 0.1$ . In order to accept the hypothesis, we would have to numerically compute the exact  $p$ -value with a Monte Carlo like approach [22].

### 5.2.2 Comparing Vocabulary Size and Growth

In the introduction to this section, we have said that we primarily use the tag frequency distribution for comparing our Epistemic Dynamic Model to the other models. But in [5, 113] it has been shown that there exists a relationship between the tag frequency distribution and the vocabulary size and growth (see Section 3.3 and Fig. 3.7). Thus, depending on the tag frequency distribution  $F(x)$ , we expect to observe a certain vocabulary size  $|T|$  in a tag stream with  $|Y|$  tag assignments. For example, in [5, 113] it has been shown that  $|T| \sim |Y|^{1/\alpha}$  if  $F(x) \sim x^{-\alpha}$ .

How  $|T|$  and  $|Y|$  are exactly correlated with each other depends on the actual distribution function  $F(x)$ . In our case,  $F(x)$  is unknown. Nevertheless, due to this correlation, we expect to observe comparable vocabulary sizes in two streams that contain the same number of tag assignments, given that the tag frequencies are empirical observations of the same  $F(x)$ . The larger the difference in the final vocabulary size of two streams, the more unlikely they are empirical observations of the same  $F(x)$ .

In this thesis, we use two different measures that give an impression of how well the different models explain the emergence of a certain vocabulary size and growth. They are both based on Equation 5.6 for calculating the difference  $\Delta_i$  between the real and the simulated vocabulary size at the time of tag assignment  $i$ :

$$\Delta_i = \frac{|T_i^s| - |T_i^o|}{|T_{last}^o|} \quad (5.6)$$

In Equation 5.6,  $|T_i^s|$  is the vocabulary size after  $i$  tag assignments in the simulated stream.  $|T_i^o|$  is the vocabulary size that can be observed after  $i$  tag assignments in the real stream.  $|T_{last}^o|$  is the final vocabulary size observed at the end of the real stream. During our evaluation, we report the following two concrete measures:

- First, we report  $\Delta_{last}$  for all evaluated models. It gives the difference between the simulated and the real vocabulary size at the end of the stream.
- Second, we report  $\Delta_{max}$  for all evaluated models that are able to explain the emergence of a sublinear vocabulary growth.  $\Delta_{max}$  corresponds to the  $\Delta_i$  value where  $|\Delta_i|$  reaches its maximum.

The sign of  $\Delta_{last}$  and  $\Delta_{max}$  indicates whether the simulated vocabulary size is above (+) or below (−) the real vocabulary size at that point in time.  $\Delta_{last}$  is reported for all evaluated models, irrespective of whether they simulate a sublinear vocabulary growth or whether they approximate it with a linear vocabulary growth. Even if a linear vocabulary growth is simulated, it is reasonable that the simulation should approximate the

vocabulary size observed in real streams. In contrast,  $\Delta_{max}$  is only reported for models that simulate a sublinear vocabulary growth because only there we can reasonably expect that the vocabulary size is not only approximated at the end of the stream but over the whole stream.

In case of the tag frequency distribution (see Subsection 5.2.1), we apply statistical goodness-of-fit tests. Such goodness-of-fit tests are not only able to report the raw difference, i. e.  $D$  in case of the Smirnov test, but also the level of significance  $p$ . For  $\Delta_{last}$  and  $\Delta_{max}$  we are not able to compute such a level of significance because neither the final vocabulary size nor the vocabulary growth can be viewed as a stochastic experiment, which is a prerequisite for computing the level of significance.

### 5.3 Evaluation Outline

In Section 5.4, we evaluate our Epistemic Dynamic Model and compare it to models described in the literature (see Tab. 4.5). The evaluation can be divided into three steps, which fulfill different objectives:

In a first step, we evaluate in how far we can explain the emergence of the tag frequency distribution and of the sublinear vocabulary growth (see Chapter 3) with the influence coming from the shared background knowledge of the users and the imitation of tag recommendations. This is done by evaluating the *Epistemic Model with Word Frequencies* from Subsection 4.2.1. The evaluation of the *Epistemic Model with Word Frequencies* is used as a benchmark against which we compare the evaluation results of the other models that are evaluated in the subsequent steps.

In a second step, we evaluate in how far the black box implementation of the background knowledge as random draws from word frequencies is equivalent to the implementation of the background knowledge that is based on random walks in semantic networks. For this purpose, we evaluate whether the *Epistemic Model with Semantic Networks* (see Subsection 4.2.2) achieves evaluation results comparable to those of the *Epistemic Model with Word Frequencies*.

In a third step, we evaluate in how far the *Epistemic Model with Word Frequencies* and the *Epistemic Model with Semantic Networks* are minimal models, i. e. we evaluate whether really both of their influence factors are required for explaining the emergence of the tag frequency distribution and the sublinear vocabulary growth. For this purpose we compare the results from the first two steps to evaluation results achieved for the *Natural Language Model* (see Subsection 4.2.3), the *Semantic Walker Model* [18] and the *Yule-Simon Model with Memory* [19]. The Natural Language Model and the Semantic Walker Model are representative for models that only include the influence from the shared background knowledge of the users. The Yule-Simon Model with Memory is representative for models that only include



Model Name	Tag Frequencies	Vocabulary Growth
Polya Urn Model [30, 38]	◦	fixed size
Simon Model [103]	◦	linear
Yule-Simon Model with Memory [19]	+	linear
Halpin et al. Model [42]	◦	linear
Organizing Model [88]	+	N/A
Semantic Walker Model [18]	+	sublinear
Semantic Imitation Model [34]	+	N/A

Table 5.1: Evaluation results reported in the literature for the models from Tab. 4.5 for the properties from Chapter 3. The Multinomial Tagging Models are excluded from this overview because they neither target the tag frequency distribution nor the vocabulary growth. For the tag frequency distribution, it has been visually evaluated whether a power-law like distribution with exponential cut-off is reproduced (+) or whether only a plain power-law without the cut-off (◦). With regard to reproducing the sublinear vocabulary growth, no information is available from the respective papers about the Organizing Model and the Semantic Imitation Model.

the influence of imitating tags of other users. We do not explicitly compare our Epistemic Model to the Polya Urn Model [30, 38] and the Simon Model [103], which are also only modeling the influence of imitating tags, because already based on the available literature it can be seen that they are outperformed by the Yule-Simon Model with Memory with regard to explaining the tag frequency distribution (see Tab. 5.1).

In Tab. 5.1, three further models are listed that have not yet been considered in our three step evaluation plan, namely the Halpin et al. Model [42], the Organizing Model [88] and the Semantic Imitation Model [34]. All three models include, amongst others, the personal organization objective of the users (see Subsection 4.3 and Tab. 4.5) but only the Organizing Model and the Semantic Imitation Model are able to explain power-law like tag frequency distributions with exponential cut-off. Comparing our Epistemic Dynamic Model to these two models would be a reasonable next step for showing whether the personal organization objective of users is able to explain additional observations whose emergence can not be explained with our Epistemic Dynamic Model. But this step is out of the scope of this thesis and is subject to future research.

## 5.4 Results

In the following, we report the evaluation results for our Epistemic Model and for models from the literature. The structure of this section follows the evaluation outline described in Section 5.3. For each model, we evaluate its ability to reproduce the tag frequencies and the vocabulary growth of the co-occurrence streams that we extracted from Delicious and Bibsonomy (see Tab. 3.2 in Section 3.1).

### 5.4.1 Fitting of Model Parameters

Most of the evaluated models have free parameters whose values can not be determined a-priori. Instead, they have to be fitted to the observations in the respective co-occurrence stream. Thus, during the evaluation we systematically test the parameter space created by a model's free parameters in order to find the best fitting parameter combination. For each parameter combination, several repeated simulations are done. This is necessary because due to the stochastic nature of the models there might be larger fluctuations between the single simulations with the same parameter combination.

From the measurements for the repeated simulations with the same parameter combination we report in the following the median value. The number of simulation runs per parameter combination is dynamically adapted to the variance between the single results. The simulations and measurements for a parameter combination are stopped when the upper and lower bound of the median's 95% confidence interval differs by less than 2%. In the following, we do not report the upper and lower bound of the confidence interval but only the single, empirically computed median value. The confidence interval is computed with the method described in [23, p. 143f].

The more free parameters a model has, the larger is the search space for the best fitting parameter combination. The search space can be narrowed down by using the relationship between the vocabulary size and the tag frequency distribution (see Subsection 5.2.2). This relationship implies that we can only reproduce the tag frequency distribution if we also reproduce the vocabulary size and/or growth. Accordingly, we first identify the region in the search space where the simulated and real vocabulary size and/or growth are comparable to each other. In this region, we then search for the parameter combination with the best fitting tag frequency distribution. During our evaluation, we use a 10% difference between simulated and real vocabulary size and/or growth as a threshold for identifying the relevant region. For models that reproduce a sublinear vocabulary growth, we check this threshold for the whole stream by using  $|\Delta_{max}| \leq 0.1$ . For models that reproduce a linear vocabulary growth, we check only the final vocabulary size with  $|\Delta_{last}| \leq 0.1$ . If no such region can be identified, then only parameter combinations are tested where  $|\Delta_{max}|$  or  $|\Delta_{last}|$  is minimized.

stream	Tag Frequencies		Vocab. Growth		Free Parameters		
	$D$	$p$	$\Delta_{max}$	$\Delta_{last}$	$I$	$n$	$h$
ringtones	<b>0.0497</b>	0.0	0.48	0.46	0.97	1,500	11,000
setup	<b>0.0288</b>	0.03	<b>-0.09</b>	<b>0.08</b>	0.905	1,000	5,000
boat	<b>0.0181</b>	<b>0.37</b>	<b>-0.10</b>	<b>0.02</b>	0.685	1,000	3,000
historical	<b>0.0123</b>	<b>0.86</b>	<b>0.10</b>	<b>0.09</b>	0.505	2,000	5,000
messages	<b>0.0152</b>	<b>0.85</b>	<b>0.07</b>	<b>0.04</b>	0.51	500	1,000
decorative	<b>0.0546</b>	0.03	-0.20	-0.20	0.96	200	1,000
costs	<b>0.0118</b>	<b>0.99</b>	<b>0.09</b>	<b>0.08</b>	0.475	2,000	5,000
ff	<b>0.0244</b>	<b>0.72</b>	<b>0.10</b>	<b>0.10</b>	0.635	1,000	3,000
checkbox	<b>0.0492</b>	<b>0.51</b>	0.64	0.62	0.97	150	1,000
datawarehouse	<b>0.0687</b>	0.06	<b>-0.10</b>	<b>0.02</b>	0.9	200	1,000
tools	<b>0.0631</b>	0.0	<b>0.08</b>	<b>-0.06</b>	0.94	1,000	5,000
social	0.0568	0.0	<b>0.09</b>	<b>-0.02</b>	0.765	2,000	7,000
design	<b>0.0394</b>	0.03	<b>0.10</b>	<b>0.00</b>	0.84	750	3,000
analysis	0.0805	0.0	0.34	<b>-0.07</b>	0.97	200	1,000
blogs	0.0374	0.09	<b>0.10</b>	<b>-0.08</b>	0.775	750	3,000

Table 5.2: Evaluation results for the *Epistemic Model with Word Frequencies*. Highlighted are the  $D$  values for the co-occurrence streams from Delicious (top) and Bibsonomy (bottom) where this configuration of the Epistemic Model achieves the lowest  $D$  value of all evaluated models. Furthermore, the  $p$ -values are highlighted for which it can not be safely rejected that simulated and real tag frequencies are drawn from the same distribution function, i. e. where  $p \geq 0.1$ . Finally, all  $\Delta_{last}$  and  $\Delta_{max}$  values are highlighted where the simulated vocabulary size and growth differs by at most  $\pm 10\%$  from the real vocabulary size and growth.

#### 5.4.2 The Epistemic Model with Word Frequencies

Tab. 5.2 contains the evaluation results for *Epistemic Model with Word Frequencies* (see Subsection 4.2.1). This configuration of the Epistemic Dynamic Model has three free parameters  $I$ ,  $n$  and  $h$ , whose values can not be determined a-priori but which have to be fitted to the observations in the respective co-occurrence stream. Especially the two parameters  $n$  and  $h$  can not be determined a-priori because they also model in an abstract way the influence coming from the number of resources aggregated in the co-occurrence stream and from the semantic breadth of the topic (see Subsection 4.1.3). Additionally, this configuration has the empirically estimated parameters  $p(s)$  and  $p(W|t)$  for which the distributions from Fig. 4.1 and Fig. 4.2 are used (see also Tab. 4.2).

For finding the best fitting parameter combination, we have systematically tested different parameter combinations in the three-dimensional parameter space spanned by the three free parameters of the model. For the  $I$

parameter we have tested values from  $I = 0$  to  $I = 0.97$  with a step width of 0.005. For the  $h$  parameter, we have tested values from  $h = 1000$  to  $h = 31000$  with a step width of 1,000. For the  $n$  parameter, we have tested values from  $n = 100$  to  $n = 5000$  with a variable step width: From 100 to 500, we have used a step width of 50. From 500 to 1,000 we have used a step width of 250. From 1,000 onward we have used a step width of 1,000. The concrete step widths have been selected during a pretest so that the measured evaluation results only slightly change between two neighboring parameter combinations. The search for the best fitting parameter combination has been restricted to the region where  $\Delta_{max} \leq 0.1$  or, if no such region exists, where  $|\Delta_{max}|$  reaches its minimum.

### 5.4.3 The Epistemic Model with Semantic Networks

Tab. 5.3 contains the evaluation results of our *Epistemic Model with Semantic Networks* (see Subsection 4.2.2). Compared to the *Epistemic Model with Word Frequencies*, this configuration additionally has the free parameter  $d$  (see Tab. 4.3). For the semantic network  $g$ , we have used the Growing Network Model with  $m_{gnm} = 11$  for simulating a network with 50,000 nodes that has properties comparable to the Word Association Norms data set (see Subsection 4.1.2).

Like for the evaluation in Subsection 5.4.2, we have systematically tested the four-dimensional parameter space spanned by the free parameters of the model. For the  $I$ ,  $n$  and  $h$  parameters we have used the same parameter range and step width as in Subsection 5.4.2. For the  $d$  parameter, we have tested values from  $d = 11$  to  $d = 101$  with a step width of 10. The search for the best fitting parameter combination has been restricted to the region where  $\Delta_{max} \leq 0.1$  or, if no such region exists, where  $|\Delta_{max}|$  reaches its minimum.

### 5.4.4 The Natural Language Model

Tab. 5.4 contains the evaluation results for the configuration of our Epistemic Dynamic Model that corresponds to the Natural Language Model (see Subsection 4.2.3). In the Natural Language Model, the influence coming from the tag suggestions is deactivated. The Natural Language Model has no free parameters, which can be tuned according to the observed tag frequency distributions. The only empirically estimated parameters of the Natural Language Model are the parameters  $p(s)$  and  $p(W|t)$  (see Tab. 4.4) for which the distributions from Fig. 4.1 and Fig. 4.2 have been used. Thus, no parameter fitting is necessary for the Natural Language Model.

stream	Tag Frequencies		Vocab. Growth		Free Parameters			
	$D$	$p$	$\Delta_{max}$	$\Delta_{last}$	$I$	$n$	$h$	$d$
ringtones	0.1595	0.0	0.63	0.63	0.805	100	3,000	41
setup	0.0506	0.0	<b>-0.07</b>	<b>0.03</b>	0.85	150	1,000	31
boat	<b>0.0163</b>	<b>0.50</b>	0.11	<b>0.07</b>	0.435	200	1,000	71
historical	<b>0.0115</b>	<b>0.92</b>	<b>-0.03</b>	-0.02	0.42	750	3,000	101
messages	0.0225	<b>0.42</b>	<b>0.07</b>	<b>0.04</b>	0.35	1,000	5,000	11
decorative	0.1440	0.0	0.17	-0.11	0.82	100	5,000	21
costs	0.0165	<b>0.87</b>	<b>-0.05</b>	<b>0.05</b>	0.35	1,000	3,000	91
ff	<b>0.0202</b>	<b>0.90</b>	<b>-0.05</b>	<b>0.04</b>	0.46	300	1,000	51
checkbox	0.1324	0.0	0.82	0.8	0.825	100	1,000	11
datawareh.	0.1552	0.0	-0.11	<b>-0.04</b>	0.815	100	1,000	41
tools	0.2096	0.0	0.12	<b>-0.03</b>	0.79	100	1,000	11
social	<b>0.0424</b>	0.01	<b>0.10</b>	<b>-0.04</b>	0.65	300	1,000	71
design	<b>0.0339</b>	<b>0.10</b>	<b>0.10</b>	<b>0.01</b>	0.76	1,000	7,000	11
analysis	0.1248	0.0	0.44	0.12	0.825	100	1,000	11
blogs	<b>0.0229</b>	<b>0.61</b>	0.12	<b>-0.03</b>	0.59	750	3,000	31

Table 5.3: Evaluation results for the *Epistemic Model with Semantic Networks*. Highlighted are the  $D$  values for the co-occurrence streams from Delicious (top) and Bibsonomy (bottom) where this configuration of the Epistemic Model achieves the lowest  $D$  value of all evaluated models. Furthermore, the  $p$ -values are highlighted for which it can not be safely rejected that simulated and real tag frequencies are drawn from the same distribution function, i. e. where  $p \geq 0.1$ . Finally, all  $\Delta_{last}$  and  $\Delta_{max}$  values are highlighted where the simulated vocabulary size and growth differs by at most  $\pm 10\%$  from the real vocabulary size and growth.

stream	Tag Frequencies		Vocab. Growth	
	$D$	$p$	$\Delta_{max}$	$\Delta_{last}$
ringtones	0.0827	0.0	4.03	4.03
setup	0.0815	0.0	2.14	2.14
boat	0.0606	0.0	1.04	1.04
historical	0.0770	0.0	0.71	0.71
messages	0.0531	0.0	0.70	0.70
decorative	0.0968	0.0	2.23	2.23
costs	0.1198	0.0	0.71	0.71
ff	0.1648	0.0	1.24	1.23
checkbox	0.1334	0.0	5.24	5.24
datawarehouse	0.2000	0.0	2.59	2.59
tools	0.2083	0.0	2.34	2.34
social	0.1979	0.0	1.35	1.35
design	0.1644	0.0	1.78	1.78
analysis	0.1327	0.0	2.74	2.74
blogs	0.1949	0.0	1.40	1.40

Table 5.4: Evaluation results for the Natural Language Model. For none of the streams from Delicious (top) and Bibsonomy (bottom), the Natural Language Model achieves the lowest  $D$  value, compared to all other models, or stays in the  $\pm 10\%$  band around the real vocabulary growth, i. e. for all streams  $\Delta_{max} > 0.1$ . The significance value  $p$  is 0% for all tested streams, i. e. there are significant differences between the simulated and the real tag frequencies.

### 5.4.5 The Semantic Walker Model

The Semantic Walker Model, originally described in [18], has the empirically estimated parameter  $p(s)$ , which corresponds to the distribution of posting sizes in Fig. 4.1. Furthermore, it has the free parameter  $d$ , which corresponds to the degree of the start node of the random walks in the semantic network, and the free parameter  $g$ , which corresponds to the semantic network. For the Semantic Walker Model, two different configurations are tested, which use different models for simulating the semantic network.

#### Using the Watts-Strogatz Model

In the first configuration, we use the Watts-Strogatz Model [119] for simulating semantic networks with 500,000 nodes. The Watts-Strogatz Model has the free parameters  $m_{ws}$  and  $p_{ws}$  (see Subsection 4.1.2), which influence the regularity of the semantic network and the average node degree (see [119]). In [18], the Semantic Walker Model in conjunction with the Watts-Strogatz Model has led to the best evaluation results. In Tab. 5.5, our evaluation results for this configuration of the Semantic Walker Model are shown.

For finding the best fitting parameter combination, we have systematically tested the three-dimensional parameter space spanned by the three free parameters of the model. For the  $m_{ws}$  parameter, we have tested values between 4 and 8 with a step width of 1. For the  $p_{ws}$  parameter, we have tested values between 0.01 and 0.2 with a step width of 0.005. In this value range, the Watts-Strogatz Model generates small-world networks [119] with a small diameter and high clustering coefficient. For the  $d$  parameter, we have tested values around the average node degree in the simulated semantic network, i. e. we have tested values between  $2 \cdot m_{ws} - 2$  and  $2 \cdot m_{ws} + 2$  with a step width of 1. The search for the best fitting parameter combination has been restricted to the region where  $\Delta_{max} \leq 0.1$  or, if no such region exists, where  $|\Delta_{max}|$  reaches its minimum.

#### Using the Growing Network Model

In the second configuration, we use the Growing Network Model [106] for simulating semantic networks with 100,000 nodes. Because we want to use the Growing Network Model for simulating semantic networks that have the same properties as empirically observable semantic networks, we fix the parameter  $m_{gnm}$  of the Growing Network Model to 11 (see [106] and Subsection 4.1.2). The Semantic Walker Model in conjunction with the Growing Network Model has only the free parameter  $d$ . For finding the best fitting parameter value, we have systematically tested different values over the range of the node degrees in the semantic network, i. e. from  $d = 11$  to  $d = 101$  with a step width of 5. In Tab. 5.6, our evaluation results for this configuration of the Semantic Walker Model are shown.

stream	Tag Frequencies		Vocab. Growth		Free Parameters		
	$D$	$p$	$\Delta_{max}$	$\Delta_{last}$	$d$	$p_{ws}$	$m_{ws}$
ringtones	0.1050	0.0	-0.21	-0.14	18	0.02	7
setup	0.1326	0.0	<b>-0.10</b>	<b>-0.04</b>	20	0.04	8
boat	0.1598	0.0	0.11	<b>0.07</b>	19	0.085	7
historical	0.1756	0.0	<b>-0.04</b>	<b>-0.04</b>	16	0.11	6
messages	0.1382	0.0	<b>-0.09</b>	<b>-0.09</b>	12	0.14	4
decorative	0.0649	0.01	-0.25	-0.25	14	0.035	6
costs	0.1956	0.0	-0.11	<b>-0.03</b>	10	0.185	4
ff	0.1762	0.0	<b>0.08</b>	<b>0.06</b>	14	0.105	5
checkbox	0.0798	0.09	0.20	0.17	14	0.02	7
datawarehouse	0.1062	0.0	0.13	<b>0.10</b>	15	0.035	6
tools	0.2122	0.0	-0.17	-0.17	15	0.045	6
social	0.2425	0.0	0.12	<b>0.03</b>	20	0.065	8
design	0.1808	0.0	0.14	<b>0.02</b>	15	0.065	6
analysis	<b>0.0560</b>	0.04	-0.45	-0.45	18	0.015	8
blogs	0.2100	0.0	-0.20	-0.19	18	0.055	8

Table 5.5: Evaluation results for the Semantic Walker Model in conjunction with the Watts-Strogatz Model for the streams from Delicious (top) and Bibsonomy (bottom). Highlighted is the  $D$  value for the stream where this configuration of the Semantic Walker Model achieves the lowest value of all evaluated models. For all tested streams, the significance value  $p$  is below 10%, i.e. there are significant differences between the simulated and the real tag frequency distributions. Finally, all  $\Delta_{last}$  and  $\Delta_{max}$  values are highlighted where the simulated vocabulary size and growth differs by at most  $\pm 10\%$  from the real vocabulary size and growth.



stream	Tag Frequencies		Vocab. Growth		Free Parameters	
	$D$	$p$	$\Delta_{max}$	$\Delta_{last}$	$d$	$m_{gnm}$
ringtones	0.1564	0.0	3.19	3.19	26	11
setup	0.1702	0.0	1.85	1.85	11	11
boat	0.1608	0.0	0.84	0.84	51	11
historical	0.1779	0.0	0.42	0.41	21	11
messages	0.1573	0.0	0.37	0.37	11	11
decorative	0.2019	0.0	1.87	1.79	31	11
costs	0.2085	0.0	0.40	0.40	16	11
ff	0.2153	0.0	0.74	0.74	16	11
checkbox	0.2250	0.0	4.79	4.79	51	11
datawarehouse	0.2641	0.0	1.74	1.74	16	11
tools	0.2832	0.0	1.64	1.64	11	11
social	0.2759	0.0	1.06	1.06	11	11
design	0.2377	0.0	1.17	1.16	16	11
analysis	0.2067	0.0	2.04	2.04	16	11
blogs	0.2541	0.0	0.90	0.90	21	11

Table 5.6: Evaluation results for the Semantic Walker Model in conjunction with the Growing Network Model. For none of the streams from Delicious (top) and Bibsonomy (bottom), this configuration achieves the lowest  $D$  value, compared to all other models, or stays in the  $\pm 10\%$  band around the real vocabulary growth, i. e. for all streams  $\Delta_{max} > 0.1$ . The significance value  $p$  is 0% for all tested streams, i. e. there are significant differences between the simulated and the real tag frequencies.

stream	Tag Frequencies		Vocab. Size $\Delta_{last}$	Free Parameters		
	$D$	$p$		$p_{ys}$	$\tau$	$n_0$
ringtones	0.1487	0.0	<b>0.10</b>	0.066	400	100
setup	0.1471	0.0	<b>0.09</b>	0.129	500	100
boat	0.1280	0.0	<b>0.10</b>	0.239	500	100
historical	0.0926	0.0	<b>0.10</b>	0.308	500	100
messages	0.1096	0.0	<b>0.09</b>	0.352	480	100
decorative	0.1556	0.0	<b>-0.08</b>	0.159	500	100
costs	0.0509	0.0	<b>0.10</b>	0.382	500	100
ff	0.0705	0.0	<b>0.10</b>	0.332	500	100
checkbox	0.1753	0.0	<b>0.08</b>	0.103	500	100
datawarehouse	0.0899	0.0	<b>0.06</b>	0.198	500	100
tools	<b>0.0620</b>	0.0	<b>0.08</b>	0.145	460	100
social	0.0655	0.0	<b>0.09</b>	0.242	500	100
design	0.0835	0.0	<b>0.06</b>	0.203	500	100
analysis	0.1559	0.0	<b>-0.08</b>	0.136	500	100
blogs	0.0565	0.0	<b>0.04</b>	0.262	500	100

Table 5.7: Evaluation results for the Yule-Simon Model with Memory for the streams from Delicious (top) and Bibsonomy (bottom). Highlighted is the  $D$  value for the stream where the Yule-Simon Model with Memory achieves the lowest value of all evaluated models. The significance value  $p$  is 0% for all tested streams, i. e. there are significant differences between the simulated and the real tag frequencies. The free parameter  $p_{ys}$  has been adapted such that the simulated vocabulary size differs by at most 10% from the real vocabulary size, as it is shown by the  $\Delta_{last}$  values.

#### 5.4.6 Yule-Simon Model with Memory

Tab. 5.7 contains the evaluation results for the Yule-Simon Model with Memory [19]. As described in Subsection 4.3, the Yule-Simon Model with Memory has the two free parameters  $p_{ys}$  and  $\tau$ . Furthermore, it has the free parameter  $n_0$ , which is used for initializing the simulated co-occurrence stream with  $n_0$  tag assignments from the corresponding real co-occurrence stream. Because a pretest has shown that  $n_0$  has no influence on the vocabulary size or the tag frequency distribution in the simulated co-occurrence streams, we have fixed this parameter to  $n_0 = 100$ .

The Yule-Simon Model with Memory, as it is described in [19], can originally only simulate postings with size  $s = 1$ . In order to make the results of the Yule-Simon Model with Memory better comparable to the results achieved for the other models, in this thesis we instead use the empirically observed distribution of posting sizes from Fig. 4.1. If the simulation would lead to a duplicate tag within a posting, the duplicate tag is removed and the corresponding simulation step is repeated.

For finding the best fitting parameter combination, we have systematically tested different parameter combinations in the two-dimensional parameter space spanned by the two free parameters  $p_{ys}$  and  $\tau$  of the model. The free parameter  $n_0$  has been fixed to  $n_0 = 100$  (see above). First, we have identified the region in the parameter space where  $|\Delta_{last}| \leq 0.1$ . Because the Yule-Simon Model with Memory only reproduces a linear vocabulary growth, we have only checked the  $\pm 10\%$  threshold for  $\Delta_{last}$ , i. e. for the final vocabulary size, but not for  $\Delta_{max}$ . Then, within this region we have searched for the minimal distance  $D$  between simulated and real tag frequencies.

Given the size  $|T|$  of the vocabulary in a real co-occurrence stream and the number of tag assignments  $|Y|$  in it, we have constrained the parameter range of the  $p_{ys}$  parameter by Equation 5.7. This constraint on  $p_{ys}$  enforces that  $|\Delta_{last}| \leq 0.1$ . The step width for  $p_{ys}$  has been adapted such that 20 different values have been tested within this range of  $p_{ys}$ .

$$0.9 \cdot \frac{|T|}{|Y|} \leq p_{ys} \leq 1.1 \cdot \frac{|T|}{|Y|} \quad (5.7)$$

For the parameter  $\tau$  we have tested values between 20 and 500 with a step width of 20. The parameter range has been chosen according to the numbers in [19] where values between  $\tau = 40$  and  $\tau = 120$  are reported. We have increased this range in order to be sure to not miss the best fitting parameters.

## 5.5 Discussion

In the following, we discuss and compare the evaluation results from Section 5.4 for the different models. In Subsection 5.5.1 and 5.5.2, we discuss the ability of the different models to reproduce the tag frequency distribution and the vocabulary growth in co-occurrence streams. Then, in Subsection 5.5.3, we discuss for our Epistemic Dynamic Model which influence it predicts for the imitation of tag suggestions on the tag frequencies and the vocabulary growth. Together, these three parts of the discussion helps us in answering whether our Epistemic Dynamic Model can explain the tag frequencies and the vocabulary growth in tagging systems, and whether really background knowledge and imitation are required for explaining the emergence of these properties (see the first and third step in our evaluation outline in Section 5.3). Finally, in Subsection 5.5.4, we discuss in how far the predictions of our *Epistemic Model with Word Frequencies* differ from the predictions of our *Epistemic Model with Semantic Networks*. This part of the discussion is related to the second step in our evaluation outline in Section 5.3.

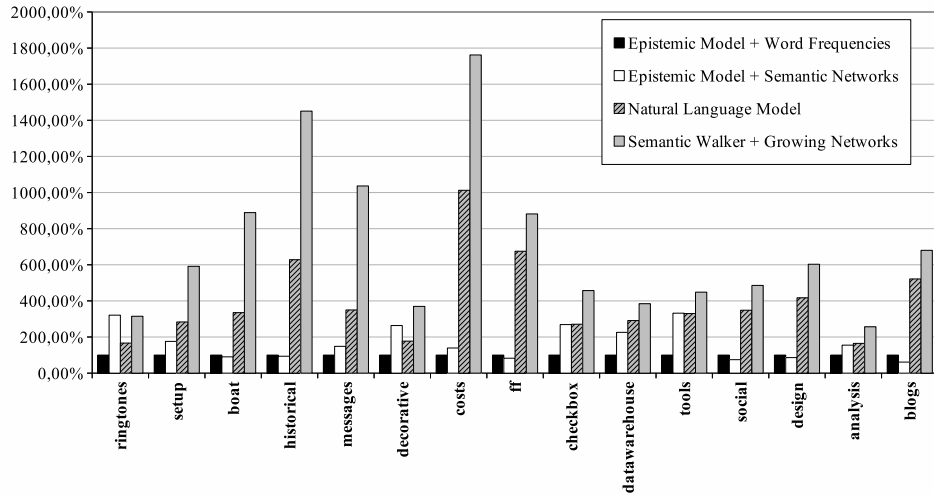


Figure 5.3: Comparing the *Epistemic Model with Word Frequencies* to other models that include the shared background knowledge as an influence factor. The *Epistemic Model with Word Frequencies* and the *Epistemic Model with Semantic Networks* additionally model the influence coming from the collaboration between users. The comparison is based on the  $D$  values reported for the respective models in Tab. 5.2, 5.3, 5.4 and 5.6.

### 5.5.1 Reproducing the Tag Frequency Distribution

In Fig. 5.3 and 5.4 the evaluation results from Tab. 5.2–5.7 are summarized. The figures compare for the different models how good they are in reproducing the tag frequency distributions in co-occurrence streams. For this comparison, we use the results of the *Epistemic Model with Word Frequencies* as a baseline, i. e. we use the  $D$  values reported in Tab. 5.2 as a baseline by defining them as 100%. Thus, only if one of the other models achieves a value  $< 100\%$  in Fig. 5.3 or Fig. 5.4 then it is better in reproducing the tag frequency distribution of the respective co-occurrence stream than our Epistemic Dynamic Model. In Fig. 5.3, all differences between the  $D$  values achieved by the models are significant except the difference between the *Epistemic Model with Word Frequencies* and the *Epistemic Model with Semantic Networks* in case of the *boat*, *historical*, *ff* and *design* stream. In Fig. 5.4, all differences are significant except the difference between the *Epistemic Model with Word Frequencies* and the Yule-Simon Model with Memory in case of the *tools* stream. The difference between two  $D$  values is considered significant if their 95% confidence intervals do not overlap. The confidence interval is computed with the method described in [23, p. 143f].

From the compared models, the *Epistemic Model with Word Frequencies* is the model that best explains the emergence of the tag frequency distributions observed in the co-occurrence streams. For 12 out of the 15 tested

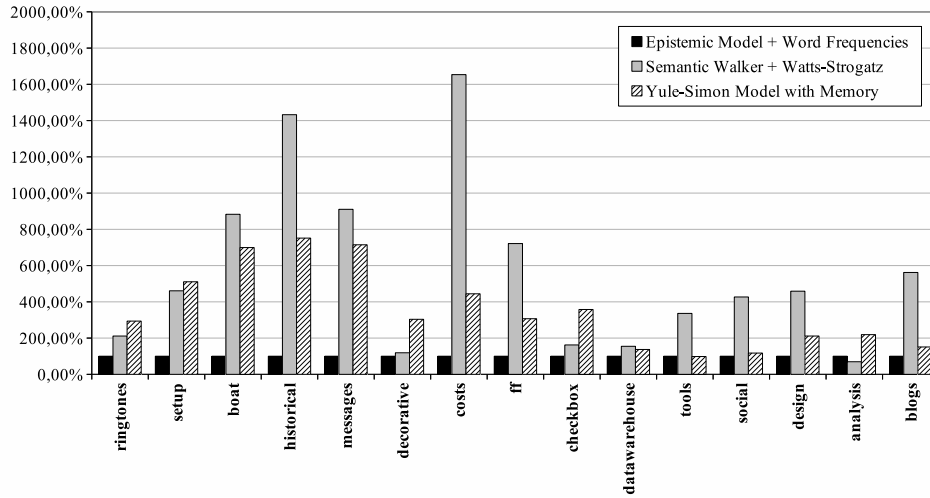


Figure 5.4: Comparing the *Epistemic Model with Word Frequencies* to models from the literature that only model either the influence of the shared background knowledge (Semantic Walker Model + Watts-Strogatz; Tab. 5.5) or the influence of the collaboration (Yule-Simon Model with Memory; Tab. 5.7).

co-occurrence streams, it is among the models that achieve the lowest  $D$  value.<sup>1</sup> The second best model is the *Epistemic Dynamic Model with Semantic Networks*. For 8 of the tested co-occurrence streams, it is among the models that achieve the lowest  $D$  value. In contrast, the Semantic Walker Model in conjunction with the Watts-Strogatz Model and the Yule-Simon Model with Memory are only for 1 co-occurrence stream among the models with the lowest  $D$  value. All other models are never among the models with the lowest  $D$  value. Based on the median  $D$  value over all tested co-occurrence stream (see Tab. 5.8), one gets the following ranked order of the different models with regard to their ability to reproduce the tag frequency distributions in co-occurrence streams: (1) *Epistemic Model with Word Frequencies*, (2) *Epistemic Model with Semantic Networks*, (3) Yule-Simon Model with Memory, (4) Natural Language Model, (5) Semantic Walker Model in conjunction with the Watts-Strogatz Model, (6) Semantic Walker Model in conjunction with the Growing Networks Model.

After giving this general overview on the evaluation results of the different models, we now discuss how the evaluation results are related to our initial outline of the evaluation (see Section 5.3). According to this outline, our evaluation should fulfill three objectives:

<sup>1</sup>In some cases, two models both achieve the lowest  $D$  value because the difference between the two  $D$  values is not significant. This is the case for the *boat*, *historical*, *ff*, *design* and *tools* stream (see above).

name	Average	Median	Min	Max
Epistemic Model + Word Frequencies	0.0407	0.0394	0.0118	0.0805
Epistemic Model + Semantic Networks	0.0775	0.0424	0.0115	0.2096
Natural Language Model	0.1312	0.1327	0.0531	0.2083
Semantic Walker + Growing Networks	0.2130	0.2085	0.1564	0.2832
Semantic Walker + Watts Strogatz	0.1490	0.1598	0.0560	0.2425
Yule-Simon Model with Memory	0.1086	0.0943	0.0526	0.1763

Table 5.8: Average, median, minimum and maximum of the  $D$  values reported for the different models in Tab. 5.2–5.7. The *Epistemic Model with Word Frequencies* achieves the best values, closely followed by the *Epistemic Model with Semantic Networks*.

The first objective has been to evaluate in how far the *Epistemic Model with Word Frequencies* is able to explain the emergence of the properties from Chapter 3. According to our summary of the evaluation results (see above), we can already say that it provides the best explanation compared to all other tested models. But the Smirnov test from Subsection 5.2.1 not only allows to have a relative comparison to other models. The level of significance  $p$  of the evaluation results can be used for assessing whether the simulated tag frequency distributions can be considered statistically indistinguishable from the observed tag frequency distributions. With this regard, the  $p$ -values achieved by the *Epistemic Model with Word Frequencies* (see Tab. 5.2) show that it is still possible to improve it. In case of 9 tested co-occurrence streams we can directly reject the hypothesis that simulated and observed tag frequencies are indistinguishable. In case of 6 streams, only a more precise measurement of the  $p$ -value would be able to finally decide whether they are indistinguishable.

The second objective has been to evaluate in how far random walks on semantic networks, as they have been originally proposed by the Semantic Walker Model, are suitable for replacing the black box implementation of the shared background knowledge that uses random drawings from word frequency distributions. For this purpose, we have to compare the evaluation results of the *Epistemic Model with Word Frequencies* to the *Epistemic*

*Model with Semantic Networks.* The two models only differ in how the background knowledge of the users is simulated. As can be seen in Fig. 5.3, using random walks on semantic networks for simulating the background knowledge leads to minor improvements of the  $D$  value for the *social* and the *blogs* stream. In case of 4 streams, the  $D$  values of the two models do not differ significantly. In case of 8 streams, the implementation with semantic networks leads to a worse  $D$  values. Nevertheless, the *Epistemic Model with Semantic Networks* is the second best model of all tested models and there is only a minor difference in the median of the  $D$  values if compared to the *Epistemic Model with Word Frequencies* (see Tab. 5.8). All in all, both configurations of the Epistemic Model can be considered to reproduce almost equally well the tag frequency distributions in co-occurrence streams.

The third objective has been to evaluate in how far the background knowledge of users as well as the collaboration between users have to be part of a tagging model in order to explain the tag frequencies. With this regard, our evaluation shows that neither models solely based on the influence of the background knowledge of users, i. e. the Natural Language Model and the two configurations of the Semantic Walker Model, nor models solely based on the influence of the collaboration between users, i. e. the Yule-Simon Model with Memory, are sufficient for explaining the emergence of the tag frequency distributions in co-occurrence streams. The integration of both factors into a unified model like the *Epistemic Model with Word Frequencies* leads to significantly better results. For example, by deactivating the influence of the collaboration between users in the *Epistemic Model with Word Frequencies*, thus leading to the Natural Language Model, the average  $D$  value gets three times higher (see Tab. 5.8). The same increase in the  $D$  value can also be observed when deactivating the influence of the collaboration in the *Epistemic Model with Semantic Networks*, thus leading to the Semantic Walker Model that uses the Growing Network Model. Also the best models from the literature that only include one of the two influence factors, i. e. the Semantic Walker Model in conjunction with the Watts-Strogatz Model and the Yule-Simon Model with Memory, can not explain the emergence of the tag frequencies nearly as good as the two configurations of our Epistemic Model that integrate both factors.

### 5.5.2 Reproducing the Vocabulary Growth

In the previous subsection, we have shown that the *Epistemic Model with Word Frequencies* and the *Epistemic Model with Semantic Networks* are by far the best performing models with regard to reproducing the tag frequencies in co-occurrence streams. In this subsection, we discuss in how far the different tested models are able to additionally reproduce the correct vocabulary growth. Reproducing both properties at the same time is an important requirement for a valid model of co-occurrence streams. In the

following, we do not take the Yule-Simon Model with Memory into account because it only simulates a linear vocabulary growth.

With regard to the sublinear vocabulary growth, by far the best performing model is the *Epistemic Model with Word Frequencies*. In Tab. 5.2, we can see that it can be parametrized such that the simulated vocabulary growth stays in the  $\pm 10\%$  band around the real vocabulary growth for 11 out of 15 streams, i. e.  $|\Delta_{max}| \leq 0.1$ . The second best model is the *Epistemic Model with Semantic Networks* (7 out of 15 streams; Tab. 5.3), closely followed by the Semantic Walker Model in conjunction with the Watts-Strogatz Model (4 out of 15 streams; Tab. 5.5). By far the worst performing models are the Natural Language Model (Tab. 5.4) and the Semantic Walker Model in conjunction with the Growing Network Model (Tab. 5.6). They exceed the observed growth in all cases. For example, in case of the Natural Language Model the median deviation from the observed vocabulary growth is +178% and in case of the Semantic Walker Model in conjunction with the Growing Network Model it is +117%. And even in the best case, the two models exceed the observed vocabulary growth by +70% and +37% respectively. In the following, we discuss two observations with regard to these evaluation results in more detail:

First, by comparing the  $\Delta_{max}$  values of the different models from Tab. 5.2–5.6 it is noticeable that none of the tested models can be parametrized such that is able to reproduce the vocabulary growth of the *ringtones*, *decorative*, *checkbox* and *analysis* streams. These are the only streams for which also the *Epistemic Model with Word Frequencies* fails to reproduce the vocabulary growth. We show that the failure to reproduce the vocabulary growth for these streams is caused by the assumption common to all models that the average number of contributing users and resources does not change over the time of the stream. This assumption is violated in case of these four streams.

Second, by comparing the  $\Delta_{max}$  values of the two configurations of the Semantic Walker Model in Tab. 5.5 and 5.6 it is noticeable that they differ in which vocabulary growth speeds they can explain. We show that these differences in the predicted growth speeds are caused by the topology of the semantic networks as they are generated by the Watts-Strogatz Model and the Growing Network Model. We discuss implications of these findings for modeling the background knowledge of users with the help of semantic networks.

### **Influence of the Average Number of Resources and Users**

As mentioned before, none of the tested models is able to reproduce the vocabulary growth of the *ringtones*, *decorative*, *checkbox* and *analysis* streams. In Fig. 5.5, the detailed plots of the vocabulary growth for these streams are shown. One can observe alternating phases of high and low vocabulary



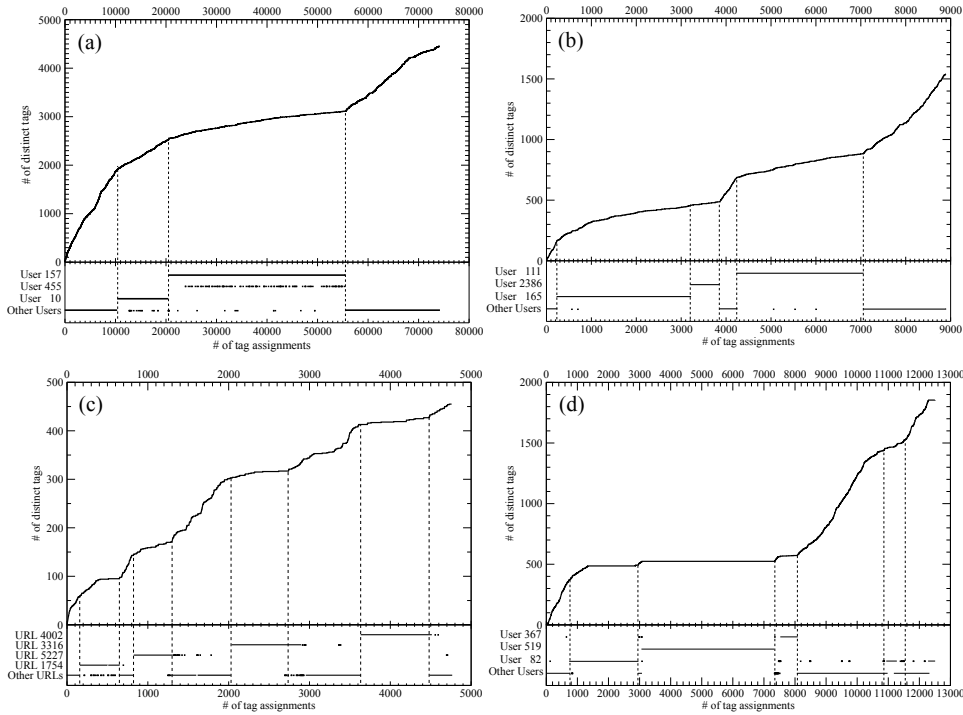


Figure 5.5: Plots of the vocabulary growth in the (a) *ringtones*, (b) *decorative*, (c) *checkbox* and (d) *analysis* co-occurrence streams. Below the plots with the vocabulary growth it is shown at which time of the stream tag assignments of a certain user or resource have been added to the stream. The observable phases of low vocabulary growth correspond to phases during which a single user or resource dominates the tag assignments in the co-occurrence stream.

growth speeds. But all of our tested models can only explain the emergence of a sublinear growth where the growth speed is constantly decreasing. None of the models is able to explain such irregular patterns of alternating high and low vocabulary growth speeds.

A closer inspection of the four streams with the irregular vocabulary growth patterns reveals that the phases with the lower vocabulary growth speeds correspond to long phases where either only single users or single resources are contributing to the postings of the co-occurrence stream. For example, in case of the *ringtones* stream, between tag assignment 10,000 and tag assignment 55,000 only 3 users are contributing new tag assignments to the stream. All in all, the 3 users are contributing 44,811 or 60.4% of the tag assignments to the *ringtones* stream. Also for the *decorative* and the *analysis* stream, three users can be identified who are the only users that contribute to the co-occurrence stream during such a phase of low vocabulary growth speed. In contrast, in case of the *checkbox* stream, the

phases of low vocabulary growth speeds are not caused by single users but by four resources that dominate the tag assignments during these phases.

In case of the *ringtones* stream and the *decorative* stream, the closer inspection reveals that the three dominating users are spammers that haven't been detected by our heuristics described in Appendix B in Section B.1. They neither used many words from the blacklist of spam tags nor they have used extremely large postings. But in case of the *analysis* stream, the dominating users are regular users who did a mass import of bookmarks from which many have been annotated with the tag *analysis*. In case of the *checkbox* stream, the dominating resources are regular resources that have been tagged by several regular users within a short period of time.

Thus, in all four cases, the sudden changes in the vocabulary growth speed corresponds to phases where the respective co-occurrence stream either degenerates to a stream of a single user or a single resource. For this observation it seems to be irrelevant whether the tag assignments are provided by regular users, as in the case of the *analysis* and the *checkbox* stream, or by spammers, as in the case of the *ringtones* and the *decorative* stream. With this regard, it would be subject to future research whether the usage patterns of spammers may cause such a degeneration more often than the usage patterns of regular users. If this is the case, one may use it as a feature for detecting spammers in tagging systems. But this is out of the scope of this thesis.

### **Influence of the Topology of Semantic Networks**

By comparing the  $\Delta_{max}$  values in Tab. 5.5 and 5.6 it can be seen that the Semantic Walker Model predicts completely different vocabulary growth speeds, depending on the used semantic network. If the Watts-Strogatz Model is used for generating the semantic network then the predicted vocabulary growth is close to what can be observed in real co-occurrence streams. In contrast, if the Growing Network Model is used for generating the semantic network then the predicted vocabulary growth is in average more than twice as high as the observable vocabulary growth. With this regard, the Semantic Walker Model in conjunction with the Growing Network Model is very similar to the Natural Language Model (see Tab. 5.4).

How can these differences in the predicted vocabulary growth speeds be explained? According to the supporting information of [18], the speed of the predicted vocabulary growth depends on two factors: (1) On the distribution of posting lengths  $p(s)$ , and (2) on the number of nodes reachable by random walks with a certain length. The more nodes can be reached by a random walk of length  $l$ , the higher is the predicted vocabulary growth. Because the distribution of posting lengths  $p(s)$  has been fixed in all our experiments to the distribution shown in Fig. 4.1, the differences between the simulations with the Growing Network Model and the Watts-Strogatz Model can only

be explained with different numbers of reachable nodes for random walks of length  $l$ . In [18], it has not been further analyzed how the topology of a network influences this number of reachable nodes. But in the following we argue that the average node degree  $k$  and the average clustering coefficient  $C$  can be used for estimating this influence:

The number of reachable nodes for a random walk of length  $l$  corresponds to the sum of the nodes contained in the rings  $N_i$  around the start node with  $i \leq l$  [18]. A node is contained in ring  $N_i$  if its shortest path to the start node has length  $i$  [8]. For example, the ring  $N_0$  contains the start node itself and the ring  $N_1$  contains all nodes that are directly connected with the start node, i. e.  $N_0 = 1$  and  $N_1 = d$ . Then, given an average node degree  $k$ , the ring  $N_{l+1}$  may contain in average up to  $N_l \cdot (k - 1)$  nodes. Thus, the higher the average node degree  $k$ , the more nodes are reachable for a random walk of length  $l + 1$ .

But not all connections from ring  $N_l$  are also to nodes in ring  $N_{l+1}$ . Instead, there also exist connections to other nodes in ring  $N_{l-1}$  and  $N_l$ . The number of connections to the previous ring or within the same ring is measured by the average clustering coefficient  $C$ . Altogether, the number of nodes in ring  $N_{l+1}$  can be approximated with  $N_l \cdot (1 - C) \cdot (k - 1)$ . Thus, the higher the value of  $(1 - C) \cdot (k - 1)$ , the higher the increase in the number of reachable nodes and the higher the vocabulary growth.

Because of this dependency between  $(1 - C) \cdot (k - 1)$  and the vocabulary growth, it is important to align this value with empirical evidence for reproducing a vocabulary growth as it might also be caused by empirically observed semantic networks. For example, in case of the Words Association Norms data set we have  $C = 0.186$  and  $k = 22$  (see [106]) and thus  $(1 - C) \cdot (k - 1) = 17.91$ . In case of our experiments with the Growing Network Model reported in Tab. 5.6, all networks have an average clustering coefficient of  $C = 0.164$  and an average node degree of  $k = 22$  and thus  $(1 - C) \cdot (k - 1) = 17.56$ . In conclusion, the networks simulated with the Growing Network Model lead to a realistic vocabulary growth given that random walks on semantic networks are appropriate for modeling how the users' tagging behavior is influenced by the users' background knowledge.

In contrast, the clustering coefficient  $C$ , the average node degree  $k$  and the  $(1 - C) \cdot (k - 1)$  values of the networks simulated with the Watts-Strogatz Model significantly deviate from those of the Words Association Norms data set (see Tab. 5.9). During our experiments, semantic networks with  $(1 - C) \cdot (k - 1)$  values between 4.27 and 6.75 have led to the best-fitting vocabulary growth, i. e. values that are approximately one-third of the value observed in the Word Association Norms data set. Thus, the semantic networks simulated with the Watts-Strogatz Model lead to a much lower vocabulary growth than expected based on empirically observable semantic networks.

These findings suggest that it is important to align the topology of artificially generated semantic networks with the topology of real semantic

stream	Parameters		Network Properties		$(1 - C) \cdot (k - 1)$
	$p_{ws}$	$m_{ws}$	$k$	$C$	
ringtones	0.02	7	14	0.64	4.68
setup	0.04	8	16	0.60	6.00
boat	0.085	7	14	0.51	6.37
historical	0.11	6	12	0.46	5.94
messages	0.14	4	8	0.39	4.27
decorative	0.035	6	12	0.60	4.40
costs	0.185	4	8	0.34	4.62
ff	0.105	5	10	0.46	4.86
checkbox	0.02	7	14	0.64	4.68
datawarehouse	0.035	6	12	0.60	4.40
tools	0.045	6	12	0.58	4.62
social	0.065	8	16	0.55	6.75
design	0.065	6	12	0.54	5.06
analysis	0.015	8	16	0.66	5.10
blogs	0.055	8	16	0.57	6.45

Table 5.9: Average node degree  $k$  and clustering coefficient  $C$  of the semantic networks that have been generated with the Watts-Strogatz Model and that led to the best fitting vocabulary growth and tag frequency distribution in Tab. 5.5. The values significantly deviate from the average node degree and clustering coefficient empirically observed in the Word Association Norms data set from [81], where  $k = 22$ ,  $C = 0.186$  and  $(1 - C) \cdot (k - 1) = 17.91$  (cf. [106]).

networks like the Words Association Norms data set. Otherwise, the Semantic Walker Model can be tuned to predict an arbitrary vocabulary growth. Thus, only the results of the Semantic Walker Model in conjunction with the Growing Network Model (see Tab. 5.6) allow to conclude on the vocabulary growth as it originates from the shared background knowledge of users. In this case, the Semantic Walker Model in conjunction with the Growing Network Model predicts a vocabulary growth that is similar to what is predicted by the Natural Language Model, which is also aligned with empirical evidence. Thus, if we align any of the two available models of the users' shared background knowledge with empirical evidence, both models agree in the prediction that one would expect a much higher vocabulary growth in co-occurrence streams if the users are only influenced by their shared background knowledge.

### 5.5.3 Influence of Imitating Tag Suggestions

In this section, we predict with the help of our Epistemic Dynamic Model how the imitation of tag suggestions effects the properties observable in

tagging systems. We derive hypotheses how tag suggestions influence the vocabulary growth and the tag frequency distribution. We are evaluating these hypotheses by looking into our tagging data sets whether we can observe the predicted effects, and by looking into results of independent user experiments that are reported in the literature.

### Vocabulary Growth

First, we are predicting how the imitation of tag suggestions influences the speed of the vocabulary growth in tagging systems. With this regard, we have shown in Subsection 5.5.2 that models predict a higher vocabulary growth if they are solely based on the users' shared background knowledge and not also on the imitation of tag suggestions. Thus, we expect that the imitation of tag suggestions reduces the vocabulary size in co-occurrence streams:

**Hypothesis 1** *The imitation of tag suggestions leads to a reduced vocabulary size compared to systems without tag suggestions. The higher the imitation probability, the smaller the vocabulary and the lower the vocabulary growth.*

In Fig. 5.6, the correlation between vocabulary growth and imitation probability  $I$  is shown, as we would expect it based on the *Epistemic Model with Word Frequencies*. Thus, the model not only helps us in determining the general influence, as it is expressed in Hypothesis 1, but it also helps us in quantifying the expected effect on the vocabulary growth and size given that we know the approximate probability with which users imitate tag suggestions. For example, after 5,000 tag assignments the *Epistemic Model with Word Frequencies* predicts a vocabulary size of approximately 3,300 tags if the users do not imitate tag suggestions ( $I = 0.0$ ). In contrast, a vocabulary size of approximately 1,700 tags is predicted if 60% of the tag assignments are imitations of tag suggestions ( $I = 0.6$ ). Thus, the Epistemic Dynamic Model predicts a reduction of the vocabulary size by 48% in case of  $I = 0.6$ .

Such a quantified prediction of the Epistemic Dynamic Model can be evaluated with the help of a user experiment like it is reported by Kowatsch and Maass in [64]. In this experiment, the users were split into two groups. Both groups were asked to assign any number of tags to different resources but only the second user group also got a set of popular tags suggested. Thus, the users in the first group were only influenced by their own background knowledge, and the users in the second group were additionally influenced by tag suggestions.

In [64], the average imitation probability of the second user group was around 60%, i. e.  $I = 0.6$ . Given this average imitation probability, the

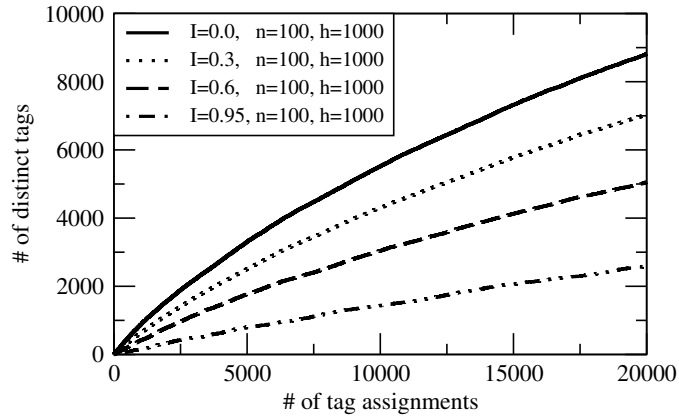


Figure 5.6: Influence of the probability of imitation  $I$  on the vocabulary growth in streams simulated with the *Epistemic Model with Word Frequencies*. For each of the streams, 20.000 tag assignments have been simulated. The higher  $I$ , the lower the vocabulary growth.

tag suggestions have led to a reduction of the vocabulary size by 30–50%. Thus, the user experiment not only confirms the general expectation that imitation reduces the vocabulary size (see Hypothesis 1) but also the more specific expectation that a reduction of the vocabulary size by 48% can be expected if  $I = 0.6$ , as it has been predicted by the Epistemic Dynamic Model in Fig. 5.6. This experiment shows that the Epistemic Dynamic Model is well capturing the influence of tag suggestions on the vocabulary growth.

### Tag Frequency Distribution

The imitation of tag suggestions not only influences the vocabulary growth but also the tag frequency distribution. In the following, we distinguish between the influence on the tag frequency distribution in co-occurrence streams and in resource streams. In Fig. 5.7a, the expected correlation between the tag frequency distribution and the imitation probability is shown for co-occurrence streams. In co-occurrence streams, the parameters  $n$  and  $h$  are abstract parameters that have no concrete meaning in the user interface. Instead, they capture amongst others the influence of the number of resources in a co-occurrence stream (see Subsection 4.1.3). In general, based on the Epistemic Dynamic Model we expect the following influence of tag suggestions on the tag frequency distribution:

**Hypothesis 2** *The imitation of tag suggestions leads to an increased probability of the most frequent tags in a stream and a decreased probability for the infrequent tags.*

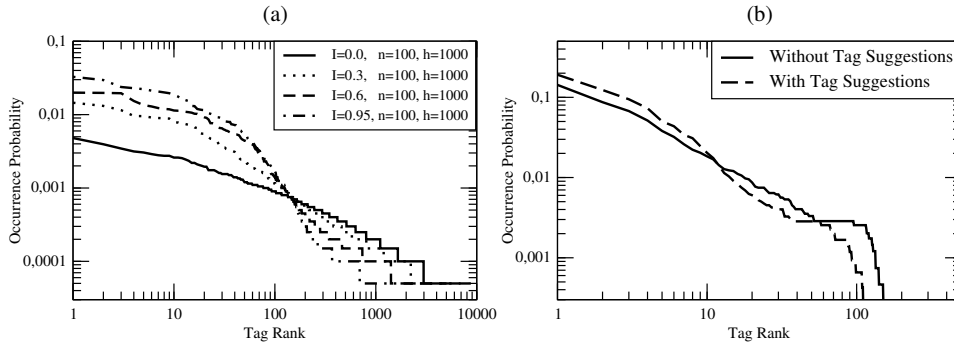


Figure 5.7: Plots of how the occurrence probabilities of tags are influenced by the imitation of tag suggestions. In (a), the predictions of the Epistemic Model for different imitation probabilities are shown. This plot is based on the simulation of 20,000 tag assignments. In (b), the average occurrence probabilities of tags are shown which have been observed for the two experimental groups in [13]. In both cases, the imitation of tag suggestions leads to increased occurrence probabilities of the most often used tags. The data shown in (b) is used with kind permission of the authors of [13].

The user experiment by Bollen and Halpin [13] can be used for evaluating Hypothesis 2. In this experiment, two user groups were asked to assign tags to 11 web pages. The web pages were randomly selected from web pages that are annotated with the tag *lifestyle* in Delicious. The first group of users got no tag suggestions, and the second group of users got 7 tag suggestions for each of the web pages. Overall, the first group made 3,556 tag assignments and the second group 3,694. The resulting tag frequencies are shown in Fig. 5.7b. Overall, the experiment confirms Hypothesis 2 of how tag suggestions influence the tag frequencies.

In contrast, if we use the Epistemic Dynamic Model for predicting the influence of tag suggestions in resource streams, we have to adapt the parameter  $n$  to the number of tags actually suggested in the user interface. Furthermore, we eliminate the parameter  $h$  by setting it such that all previous tag assignments are taken into account for computing the set of suggested tags (see Subsection 4.1.3). If we now parametrize the Epistemic Dynamic Model for simulating the user interface of Delicious by setting  $n = 7$ , we get the tag frequency distribution shown in Fig. 5.8. A sudden drop in the relative occurrence probabilities for tags around rank 7 is visible.

This sudden drop in the relative occurrence probabilities is very similar to what has been discovered by Halpin et al. in [42] for resource streams in Delicious (see Section 3.2 and Fig. 3.4). In [42], it has been speculated that this sudden drop (1) may be related to cognitive effects during tagging, or (2) it may be an artifact of the Delicious user interface. The simulation with the Epistemic Dynamic Model shows that the artifact can plausibly be

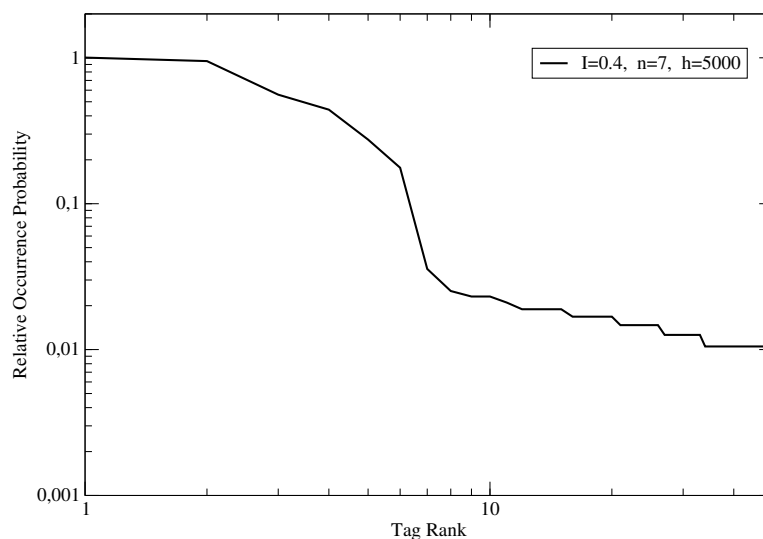


Figure 5.8: Simulation of a resource stream in Delicious with the help of the *Epistemic Model with Word Frequencies*. The simulated resource stream contains 5,000 tag assignments and the parameter  $n$  has been adapted to the number of popular tags that are visible in the Delicious user interface (see Fig. 2.1). The relative occurrence probability corresponds to the occurrence probability normalized by the occurrence probability of the most frequent tag.

explained as being an artifact of the Delicious user interface. Thus, we favor the following hypothesis:

**Hypothesis 3** *The sudden drop in the occurrence probabilities for tags between rank 7 and 10, as it has been observed by Halpin et al. [42] in Delicious' resource streams, is an artifact of the number of popular tags shown in the user interface of Delicious.*

As a consequence, we expect to not observe this artifact if a tagging system does not recommend a set of  $n$  popular tags. An example for a tagging system that does not recommend a set of  $n$  popular tags is the Bibsonomy system (see Fig. 2.2). Instead of a set of popular tags, it presents a set of recommended tags that can be generated by various recommendation algorithms (see [55] and the *Online Tag Recommendation Task* in [31]). Fig. 5.9 shows the average relative occurrence probability that can be observed in Bibsonomy's resource streams. As expected, no sudden drop in the relative occurrence probabilities for tags between rank 7 and 10 is visible. This confirms Hypothesis 3.



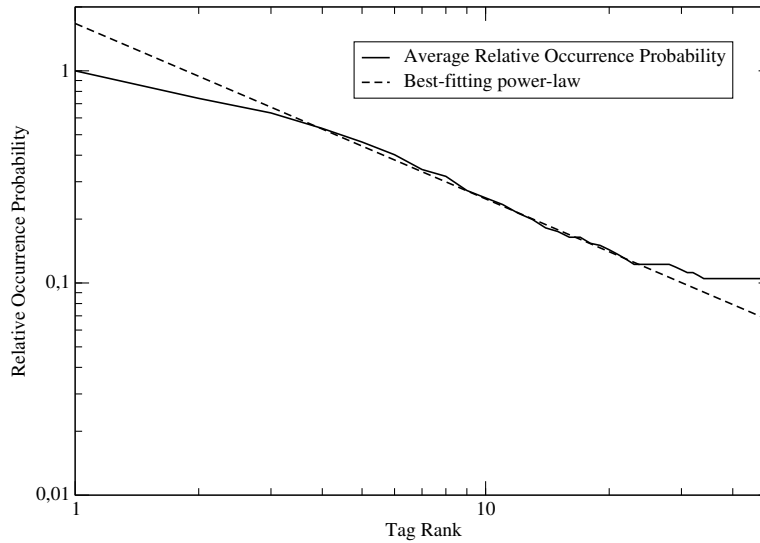


Figure 5.9: Relative occurrence probabilities in the resource streams of Bibsonomy. The average relative occurrence probability in the graph has been computed by averaging the relative occurrence probabilities for the 30 resource streams from our Bibsonomy data set (see Tab. 3.1) to which more than 100 users contributed. As a guide for the eye, a line for the best-fitting power-law distribution is also included in the graph.

#### 5.5.4 Influence of the Background Knowledge

In the following, we discuss in how far the background knowledge of the users about the topic of a stream influences the speed of the vocabulary growth in co-occurrence streams. Depending on whether we use the *Epistemic Model with Semantic Networks* or the *Epistemic Model with Word Frequencies*, we have different expectations with regard to this influence. In case of the *Epistemic Model with Semantic Networks*, we expect that for broader topics a higher vocabulary growth can be observed than for narrower topics. In case of the *Epistemic Model with Word Frequencies*, we do not expect such influence (see below). The reason for this difference between the models is the additional parameter  $d$ , which is introduced in the *Epistemic Model with Semantic Networks*. Thus, showing the expected influence of the broadness of a topic on the vocabulary growth would support the introduction of the additional parameter  $d$  into the Epistemic Model.

In case of the *Epistemic Model with Semantic Networks*, the influence of the users' background knowledge is simulated by random walks on an artificially generated semantic network. Depending on the topic of a stream, the random walks start at different nodes in the network. In Subsection 4.1.2, we have proposed to introduce the degree  $d$  of the start node as a further parameter into our model. We have interpreted  $d$  as a measure of the se-

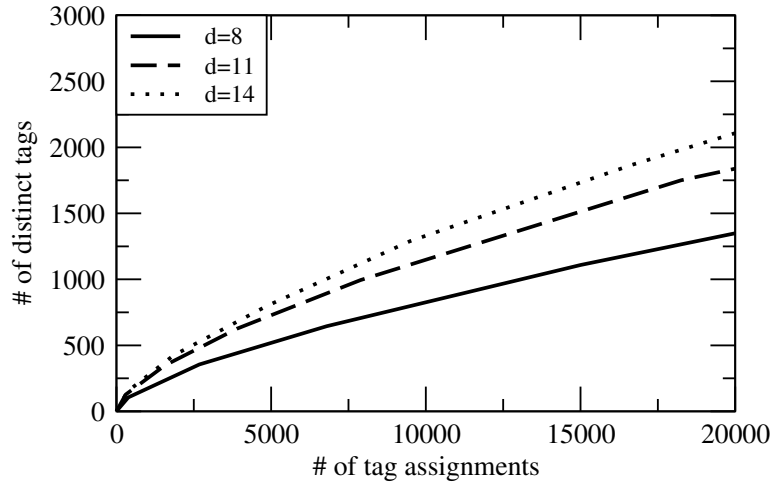


Figure 5.10: Influence of the degree  $d$  of the start node on the vocabulary growth in streams simulated with the *Epistemic Model with Semantic Networks*. The vocabulary growth increases with increasing start node degree  $d$ . For each of the streams, 20.000 tag assignments have been simulated. Except the  $d$ -parameter, all other parameters of the model have been kept constant during simulating the different streams.

semantic breadth of the topic that is described by the start node. The higher  $d$ , the broader is the topic of the simulated co-occurrence stream.

In Fig. 5.10, it is shown how the parameter  $d$  influences the speed of the simulated vocabulary growth. Thus, according to the *Epistemic Model with Semantic Networks* we expect to observe a higher vocabulary growth speed for co-occurrence streams that cover broad topics. The explanation for this influence of  $d$  is that with increasing value of  $d$  a higher number of nodes in the network can already be reached with very short random walks thus leading to a faster vocabulary growth (cf. Subsection 5.5.2).

This assumed influence of a stream's broadness on the speed of the vocabulary growth can be confirmed with the help of our Delicious data set from Tab. 3.1. For this purpose, we have extracted from it all co-occurrence streams for tags that have a direct correspondence in WordNet. WordNet<sup>2</sup> is a lexical database of English nouns, verbs, adjectives and adverbs. These words are grouped into sets of cognitive synonyms, the so-called synsets, each expressing a distinct concept. If a word belongs to several synsets, then it has multiple meanings. In the following, we take this number of meanings of a word as an indicator of the number of topics covered in the corresponding tag's co-occurrence stream. The more meanings a word or tag has, the broader the topical area covered in the co-occurrence stream.

<sup>2</sup><http://wordnet.princeton.edu/>

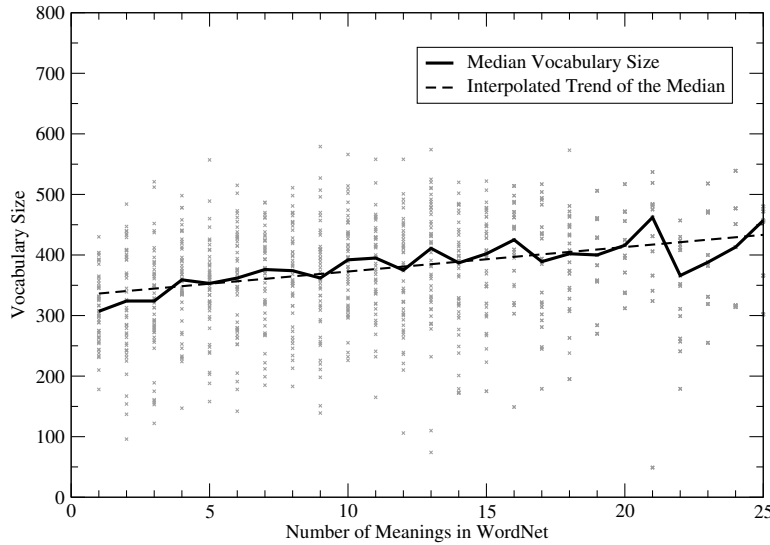


Figure 5.11: Influence of the number of meanings of a tag on the vocabulary size in its co-occurrence stream after 1,000 tag assignments. The number of meanings of a tag has been extracted from WordNet. The single vocabulary sizes are shown as a scatter plot in the background. The median vocabulary size and its trend are shown as lines in the foreground.

In Fig. 5.11, the vocabulary sizes of the extracted co-occurrence streams after 1,000 tag assignments is shown in dependency on the number of meanings of the corresponding tag in WordNet. It can be seen that the median vocabulary size of the co-occurrence streams increases with the number of meanings of a tag. This confirms our expectation based on simulations with the *Epistemic Model with Semantic Networks* that the broader a topic, the higher the vocabulary growth in its co-occurrence stream. This influence coming from the broadness of a co-occurrence stream’s topic may also explain to some extent the high variance in the vocabulary growth speeds that can be observed in Fig. 3.6.

In how far can the observation from Fig. 5.11 also be explained by the *Epistemic Model with Word Frequencies*? In this case, the influence of the users’ background knowledge is simulated by randomly drawing from the word frequency distributions  $p(W|t)$  shown in Fig. 4.2. In [5] it has been shown that the exponent of the power-law distribution, which can be used for approximating the distributions in Fig. 4.2, influences the speed of the simulated vocabulary growth (see also Section 3.3). Thus, the word frequency distributions  $p(W|t)$  may indeed cause different speeds of vocabulary growth given that their exponent changes for different topics. But as can be seen in Fig. 4.2, there only exists a small variance in the exponents of the different  $p(W|t)$  distributions. Thus, this effect can only explain to a minor extent the variance in the vocabulary growth speeds from Fig. 3.6.

## 5.6 Conclusions

In this chapter, we have evaluated the ability of our Epistemic Dynamic Model to explain the emergence of the tag frequency distribution and the vocabulary growth in co-occurrence streams. Our evaluation has shown that the combined influence of shared background knowledge and imitation, as modeled by our Epistemic Dynamic Model, leads to improved evaluation results with regard to these two properties, if compared to models that only rely on one of these two influence factors.

According to the Epistemic Dynamic Model, the general shape of the tag frequency distribution as well as the sublinearity of the vocabulary growth emerge from the shared background knowledge and terminology of the users. This can be shown by deactivating the influence of the imitation in the Epistemic Dynamic Model by setting  $I = 0.0$ , thus leading to the Natural Language Model and/or the Semantic Walker Model. Both models are still able to reproduce the general shape of the tag frequency distribution and a sublinear vocabulary growth. The imitation of tag suggestions leads to (1) a reduced vocabulary growth speed, to (2) an increased probability of the most frequent tags, and to (3) a decreased probability of the infrequent tags (see Subsection 5.5.3).

With regard to modeling the influence of the shared background knowledge, we have tested two alternative implementations. On the one hand, we have tested a black box implementation that is based on randomly drawing from empirically observed word frequency distributions. On the other hand, we have tested an implementation that models the background knowledge as random walks on semantic networks. If we use semantic networks that are aligned with empirical evidence, i. e. if we use the Growing Network Model for generating them, then both implementations of the background knowledge are very similar with regard to the simulated tagging behavior. For example, both implementations agree in their prediction that a much higher vocabulary growth would be expected if the tag assignments of the users are only influenced by the shared background knowledge (see Subsection 5.5.2).

All in all, the implementation that is based on the empirically observed word frequencies can be considered to provide a simple black box model of the users' background knowledge that approximates the more sophisticated implementation that is based on semantic networks. The black box implementation has the advantage that it is slightly better in reproducing the exact shape of the tag frequency distribution and of the vocabulary growth (see Subsection 5.5.1 and 5.5.2). In contrast, the implementation based on semantic networks has the advantage that it is better suitable for explaining the high variance in the vocabulary growth speeds as it might be caused by the differences in the broadness of a stream's topic (see Subsection 5.5.4).

## Chapter 6

# Tag Recommendations and Indexing Quality

In this chapter, we analyze how tag recommendations influence the indexing quality in tagging systems. The indexing quality is related to the retrieval of resources, i. e. to searching and browsing resources (see Section 2.4). The question is: In how far are users better able to find and discover resources in tagging systems in which the tagging decision of users is influenced by tag recommendations?

In the previous chapters, we have analyzed how tag recommendations that are based on popular tags of a resource change the tag frequency distribution and the vocabulary size. Both properties are related to the emergence of the shared community vocabulary, and indirectly also to the indexing quality. For example, the size of the vocabulary associated with a resource influences how many query terms can be used for accessing it, and how easy it can be reached when browsing the tagging system. But although a large vocabulary may be an indicator for a high indexing quality, additional requirements have to be met for a high indexing quality, like that only relevant tags are annotated to a resource. Without this requirement, in the extreme case, spammers who maximize the number of tags at a resource would be perceived to produce tag assignments with a high indexing quality.

This is where the tag frequency distribution comes into play, which can be used for measuring the inter-indexer consistency of the tag assignments. The inter-indexer consistency corresponds to the degree to which the users have agreed on a common vocabulary for describing the single resources. In the literature [33, 38, 64, 77, 100], it is assumed that increasing the inter-indexer consistency is important for dealing with the uncontrolled nature of the vocabulary in tagging systems, and thus for getting better indexed resources. But even the inter-indexer consistency is not directly correlated with the indexing quality. The additional requirement has to be met that the tag assignments are not only consistent for the single resources but also

across several resources. In the following, we call the consistency across several resources the inter-resource consistency (see [122] and Subsection 6.1.1).

In [127], it has been pointed out that the inter-indexer consistency is positively correlated with the inter-resource consistency if the indexing terms are “selected individually and independently by each of the indexers”. But what happens if the users are influenced in their tagging decision by tag recommenders? In this chapter, we show that one can then not automatically assume their positive correlation. Thus, one has to always use measures of the inter-resource consistency instead of measures of the inter-indexer consistency for analyzing how tag recommenders influence the indexing quality in tagging systems.

This chapter is structured as follows: In Section 6.1, we present two concrete measures of the inter-resource and inter-indexer consistency in tagging systems. Furthermore, we discuss additional aspects of good tag recommendations that might be measured during an evaluation. Then, in Section 6.2, we derive for two of the tag recommenders from Delicious, how they influence the inter-resource and inter-indexer consistency in a tagging system. According to our hypotheses we expect that the inter-indexer and inter-resource consistency are not positively correlated with each other for these two exemplary tag recommenders. In Section 6.3, we describe the user experiment that we use for evaluating our hypotheses. The results of the experiment are presented and discussed in Section 6.4 and 6.5.

## 6.1 Measures of Indexing Quality

One typical question with regard to tag recommendation algorithms is how to measure the quality of the generated tag recommendations and how to compare different algorithms to each other (see Subsection 2.2.3). Two complementary dimensions of tag recommendations can be identified: (1) On the one hand, one may see tag recommendations as a tool for improving the quality of the tag assignments, thus leading to better indexed resources. (2) On the other hand, one may see tag recommendations as a tool for reducing the effort that is required by a user for indexing a resource. Depending on the dimension, one has to choose other evaluation measures. Ideally, tag recommendation algorithms should generate tag recommendations that score high in both dimensions.

In the following, we concentrate on measures for evaluating the influence of tag recommenders on the quality of the tag assignments. In Subsection 6.1.1, we introduce a measure and methodology that can be used for measuring the inter-resource consistency of annotations in information systems that use the vector-space model during retrieval, like it is the case for tagging systems that produce broad folksonomies. Then, in Subsection 6.1.2, we discuss existing measures of the inter-indexer consistency in

broad folksonomies and select one of the existing measures for our evaluation. Finally, in Subsection 6.1.3 we discuss for the evaluation measures from Subsection 2.2.3 how they are connected to measuring the influence of tag recommenders on the indexing quality.

### 6.1.1 Inter-Resource Consistency

In general, the inter-resource consistency measures in how far indexers are successful in linking similar resources by indexing their common aspects with common terms. Assuming the match between indexing terms and query term, it has been argued in [127, 122] that an improved inter-resource consistency leads to an improved precision and recall for the results of a query. For example, only if similar resources are linked to each other by indexing them with terms that express their common aspects, one gets a high recall when searching or browsing a tagging system with the respective terms. Furthermore, for achieving a high precision, the indexing terms have to be discriminative, i.e. they have to link similar resources but not also dissimilar resources. It has been argued in [122] that measuring the inter-resource consistency leads to a critique not of the total retrieval system, which includes the ranking algorithms et cetera, but of one of its modifiable components, the indexing.

All in all, measuring the inter-resource consistency of tag assignments corresponds to comparing the tag vector based similarity of the resources to an independent indicator of resource similarity. The inter-resource consistency in a system correlates with the indexing quality if the independent indicator of resource similarity reflects how potential users of the system perceive the similarity of the resources. In case of tagging systems, the independent indicator of resource similarity should thus reflect how the indexers perceive the similarity of resources because the indexers are at the same time users of the tagging system and vice versa. This distinguishes tagging systems from other information systems like library catalogs where indexers and users of the system are mostly disjoint from each other.

In Fig. 6.1, an example is shown that illustrates this comparison of the tag vector based similarity of resources to the user perceived similarity of resources: In Fig. 6.1 three resources are shown together with their tag vectors. Each of the resources has been tagged by 10 indexers. According to its tag vector, resource  $r_2$  is implicitly linked to the other two resources. 8 of the 10 indexers assigned the tag *news* to  $r_2$ , thus creating an implicit link between  $r_2$  and  $r_1$ . Furthermore, 6 of the 10 indexers implicitly linked  $r_2$  to  $r_3$  via the term *humor*. The inter-resource consistency measures in how far the strength of the implicitly created links correlates with the strength of the links that can be acquired by explicitly asking the indexers. For example, based on the cosine similarities  $sim(v_1, v_2)$  and  $sim(v_2, v_3)$  (see Equation 2.1 in Subsection 2.4.1), we expect that resource  $r_2$  is perceived

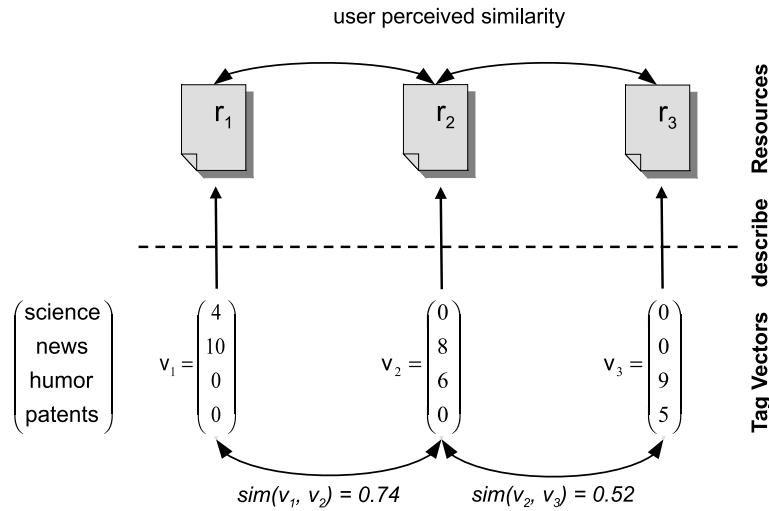


Figure 6.1: Example of tag vectors describing three resources. During retrieval, the similarity of the tag vectors influences how close together the corresponding resources get ranked. In order to get high precision and recall, the similarity of the tag vectors needs to correlate with the user perceived similarity of the resources.

by a majority of the indexers as more similar to  $r_1$  than to  $r_3$ . Otherwise, their tag assignments are inconsistent with how they perceive the similarity of the resources. In consequence, when querying with the terms that cause such an inconsistency, the indexers likely get results that are inconsistent with their expectations, thus decreasing precision and/or recall of the query results.

According to this general idea, measuring the inter-resource consistency requires two-steps: In a first step, we have to acquire the resource similarities as they are explicitly perceived by the indexers of the resources. In a second step, we can then compare in how far the implicit resource similarities, as they are given by the cosine similarities of the tag vectors, are consistent with the explicit resource similarities from the first step.

### Acquiring Explicit Resource Similarities

For measuring the inter-resource consistency we need pairwise similarities between the resources that are based on explicitly asking the indexers for their perception of the resources' similarities. But collecting the pairwise similarities from each indexer is not feasible already for very small collections of resources because the number of pairwise similarities increases quadratically with the number of resources. For example, already for 10 resources we would have to collect 45 pairwise similarities from each indexer. Thus,



a more efficient method is required that can be used for concluding on the explicit, pairwise similarities of resources:

In the following, we propose to ask each indexer to group the resources according to their similarity. As a result, we get from each indexer  $u_i$  a set of topical clusters  $C_i$ . By putting two resources into the same cluster  $c \in C_i$ , the indexer  $u_i$  indicates that he/she perceives the two resources as similar to each other with regard to at least one aspect. Furthermore, we add the constraint that each resource can only be contained in exactly one cluster. Thus, if an indexer sees a resource related to more than one topical cluster, he/she has to decide in which of the potential clusters to put the resource. Due to this constraint, the amount of data to be acquired from each indexer increases only linearly with the number of resources.

By acquiring from each indexer  $u_i \in U$  such a set of topical clusters, we are then able to reconstruct the relative strength of the pairwise similarities between the resources. The more indexers have put two resources together in one topical cluster, the higher the perceived similarity between the two resources. The tag assignments of the indexers are consistent with this independent indicator of resource similarities if the probability of two resources to be in the same cluster is correlated with the probability of the two resources having a tag in common. For example, the tag assignments in Fig. 6.1 are consistent with the user perceived similarity of the resources if more users cluster  $r_2$  together with  $r_1$  than with  $r_3$ .

A very similar approach for acquiring the explicit resource similarities has been used by White and Griffith in [122] for measuring the inter-resource consistency of annotations that have been created by professional indexers in online bibliographic data bases. White and Griffith compare the annotations of the professional indexers to topical clusters that are based on co-citations of the annotated bibliographic references. By citing two bibliographic references together in a paper, the author of the paper indicates that the two references are subject-related to each other and should thus also have annotations in common that identify this common subject. It makes sense to use the co-citations as an independent indicator of resource similarities because the authors of papers are at the same time potential users of online bibliographic data bases. Thus, the annotations should be consistent with their judgment about the relatedness of two bibliographic references.

### Comparing Explicit and Implicit Resource Similarities

Given a set of resources  $R = \{r_1, \dots, r_m\}$  and a set of topical clusters  $C_i$  from indexer  $u_i$ , the idea of inter-resource consistency is as follows: (1) If two resources are contained in the same topical cluster  $c \in C_i$  then this should be reflected by a high cosine similarity of their tag vectors. (2) If two resources are contained in different topical clusters  $c$  and  $c'$  then this should be reflected by a low cosine similarity of their tag vectors.

Technically, this idea of inter-resource consistency may be implemented by computing the ratio between the similarities in the same cluster and the similarities across the clusters. In the following, we propose to use the Silhouette Coefficient for computing this ratio, and thus for measuring the inter-resource consistency. The Silhouette Coefficients has been first introduced in [91] for evaluating in how far a clustering algorithm identifies clusters that are consistent with the distances between the vectors that describe resources. We use it the other way round for evaluating in how far the distances between the tag vectors are consistent with the clusters identified by an indexer.

The Silhouette Coefficient is individually computed for each of the resources contained in the set of topical clusters  $C_i = \{c_1, \dots, c_k\}$  identified by indexer  $u_i$ . Instead of similarities between tag vectors, the Silhouette Coefficient uses distances between the tag vectors of the resources in  $C_i$ . In the following, we use the angle  $\Theta = \arccos(\text{sim}(v_1, v_2))$  (cf. Equation 2.1) for measuring the distance between two tag vectors.  $\Theta$  is the complement to the cosine similarity on which most ranked retrieval tasks are based (see Subsection 2.4.1). Given a resource  $r_j$  and a corresponding set of topical clusters  $C_i$ , its Silhouette Coefficient  $s_{ij}$  is computed as follows:

First, the average distance  $a_{ij}$  of  $r_j$  to all other resources in its cluster  $c \in C_i$  is computed. Second, from all clusters in  $C_i$  that do not contain  $r_j$ , we identify the cluster  $c' \in C_i$  whose resources have in average the lowest distance to  $r_j$ . We call this minimal average distance  $b_{ij}$ . The distance of two resources corresponds to the angle  $\Theta$  between the tag vectors describing the two resources. Finally,  $a_{ij}$  and  $b_{ij}$  are set into relation to each other as follows:

$$s_{ij} = \frac{b_{ij} - a_{ij}}{\max(a_{ij}, b_{ij})} \quad (6.1)$$

The Silhouette Coefficient  $s_{ij}$  ranges between  $-1$  and  $+1$ .  $s_{ij}$  takes a positive value if resource  $r_j$  is closer to the resources in the same cluster  $c$  than to resources in the closest other cluster  $c'$ . It reaches its maximal value if the average distance  $a_{ij}$  to the resources in  $c$  is 0. If  $s_{ij} > 0$ , we say that the tag vector of resource  $r_j$  is consistent with regard to the topical clusters in  $C_i$  and the tag vectors of the other resources (see Fig. 6.2). In contrast,  $s_{ij}$  takes a negative value if  $r_j$  is farther away from the resources in  $c$  than from the resources in  $c'$ . It reaches its minimal value if the average distance  $b_{ij}$  to the resources in cluster  $c'$  is 0. If  $s_{ij} < 0$ , we say that the tag vector of resource  $r_j$  is inconsistent with regard to the topical clusters in  $C_i$  and the tag vectors of the other resources. Furthermore,  $s_{ij}$  is undefined if resource  $r_j$  is in a cluster of size 1 in  $C_i$ . The reason is that in such a case the average distance  $a_{ij}$  to all other resources in  $r_j$ 's cluster is not well defined (cf. [91]).

In order to get a global measure of the inter-resource consistency, we propose to average the Silhouette Coefficient over all resources in  $R =$

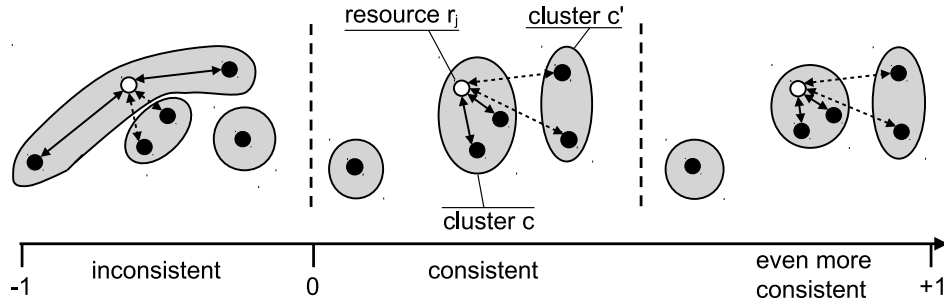


Figure 6.2: Three examples of topical clusters, which illustrate the idea of measuring the Silhouette Coefficient for resource  $r_j$  (white circle). The lengths of the arrows correspond to the distances between the tag vectors of the resources. The average length of the arrows with continuous lines corresponds to  $a_{ij}$ . The average length of the arrows with dotted lines corresponds to  $b_{ij}$ .

$\{r_1, \dots, r_m\}$  and over all sets of topical clusters of the indexers in  $U = \{u_1, \dots, u_n\}$ . The average Silhouette Coefficient  $E(s_{ij})$  (see Equation 6.2) measures the inter-resource consistency of the tag vectors  $V = \{v_1, \dots, v_m\}$  of the resources in  $R$  with regard to how the indexers in  $U$  perceive in average the similarity of the resources.

$$E(s_{ij}) = \frac{1}{n \cdot m} \cdot \sum_{i=1}^n \sum_{j=1}^m s_{ij} \quad (6.2)$$

The higher the  $E(s_{ij})$ -value, the higher the inter-resource consistency of the tag vectors. The  $E(s_{ij})$ -values for two sets of tag vectors  $V_1$  and  $V_2$  can be compared given that they describe the same set of resources  $R$  and given that they are compared to the same sets of topical clusters provided by the indexers in  $U$ . Only if these preconditions are fulfilled, we can be sure that a difference in the two  $E(s_{ij})$ -values indicates a difference in the inter-resource consistency for  $V_1$  and/or  $V_2$ . Otherwise, two  $E(s_{ij})$ -values are incomparable.

In Equation 6.2 it is assumed that none of the  $s_{ij}$ -values is undefined, i. e. that none of the resources is in a cluster of size 1 (see above). If there exist undefined  $s_{ij}$ -values then they are excluded from the calculation of the average Silhouette Coefficient  $E(s_{ij})$ . Nevertheless, the tag vector of resource  $r_j$ , for which  $s_{ij}$  is undefined, still takes part in the computation of the  $b_{ij}$ -values of the remaining resources clustered in  $C_i$ .

### 6.1.2 Inter-Indexer Consistency

In general, the inter-indexer consistency measures in how far the indexers have agreed on a common vocabulary for describing the relevant aspects of a resource. It is assumed that the inter-indexer consistency of a tag vector is an indicator of how good the tag vector describes the resource. In Fig. 6.1, the inter-indexer consistency is related to the connection between a tag vector and its corresponding resource. It is not checking the relations between several tag vectors and/or resources. Nevertheless, according to [127], one can assume a positive correlation between the inter-indexer consistency and the inter-resource consistency of the tag vectors if the tags are “selected individually and independently by each of the indexers”.

In the literature about tagging systems, it is a common assumption that a high inter-indexer consistency indicates a high indexing quality, also if the indexers are influenced by tag recommendations. Accordingly, measures of inter-indexer consistency are used by many authors for concluding on the indexing quality and how it is influenced by certain kinds of tag recommendations [33, 38, 64, 77, 100]. Two different measures are available in the literature:

The first measure analyzes how the vocabulary size in a tagging system is influenced by giving tag recommendations [33, 57, 64, 77]. Given the same number of indexers, a smaller vocabulary size is taken as an indicator of a less idiosyncratic vocabulary of each indexer, and thus of a higher inter-indexer consistency. But the vocabulary size is not a very robust measure of inter-indexer consistency because it is not only influenced by the overlap of the users’ individual vocabularies, which indicates the inter-indexer consistency, but also by the average size of the users’ individual vocabularies. Furthermore, a large vocabulary may also have a positive influence on the indexing quality in a tagging system because more search terms can be used for accessing the resources during retrieval.

The second measure analyzes how the tag reuse rate is influenced by giving tag recommendations. In [100], the tag reuse rate is defined as “the average number of users who apply a tag”. If an indexer reuses the tag of another indexer then this is seen as an indicator that the tag is perceived as relevant for describing the resource. In the following, we use the tag reuse rate for measuring the inter-indexer consistency of the tag vectors. We define the tag reuse rate as follows:

Given a set of resources  $R = \{r_1, \dots, r_m\}$ , we first measure the tag reuse rate  $tr_j$  for each of the resources  $r_j \in R$  individually. The tag reuse rate  $tr_j$  corresponds to the number of tag assignments aggregated in  $r_j$ ’s tag vector divided by the number of distinct tags in the tag vector (see Equation 6.3). Given the tag reuse rates for the individual resources in  $R$ , the global tag reuse rate over all resources is measured by the average  $E(tr_j)$  over the individual reuse rates (see Equation 6.4).

$$tr_j = \frac{|\{(u, t, r_j) \in Y\}|}{|\{t \in T | \exists u \in U : (u, t, r_j) \in Y\}|} \quad (6.3)$$

$$E(tr_j) = \frac{1}{m} \cdot \sum_{j=1}^m tr_j \quad (6.4)$$

The higher the  $E(tr_j)$ -value, the higher the inter-indexer consistency of the tag vectors that describe the resources in  $R$ . But the tag reuse rate is also influenced by the number of tag assignments for which it is measured: By dividing the number of the tag assignments through the vocabulary size, the tag reuse rate is basically measuring the speed of the vocabulary growth. Because of the sublinearity of the vocabulary growth, the ratio between the two values changes over time, even if the inter-indexer consistency remains unchanged. Thus, one can only conclude from the tag reuse rate on a changed inter-indexer consistency if one compares it between two tag vectors  $v_1$  and  $v_2$  describing the same resource  $r_j$  and aggregating the same number of tag assignments. Nevertheless, with increasing number of tag assignments, the tag reuse rate becomes more and more robust with regard to smaller differences in the number of tag assignments aggregated in two tag vectors.

### 6.1.3 Further Measures

In the literature about tag recommenders, additional measures and methodologies are proposed for evaluating and comparing recommendation algorithms to each other. In Subsection 2.2.3, an overview of the two most widespread evaluation methodologies are given. These two methodologies have in common that the quality of tag recommenders is measured by the precision and recall of the set of recommended tags. Precision and recall are used for comparing the given tag recommendations against a gold standard. It depends on the chosen gold standard how to interpret the measurements:

In case of the offline evaluation methodology (see Subsection 2.2.3 and [56]), the uninfluenced tag assignments of the individual users are used as the gold standard. According to this methodology, the perfect tag recommender knows exactly which tags the user would assign to a resource without seeing the recommendations. By seeing the recommendations, the user no longer has to type tags but can simply accept the recommendations. In consequence, tag recommendations are primarily seen as a tool for reducing the effort that is required by a user for indexing a resource. A further consequence is that the methodology judges any deviation from the uninfluenced behavior of users as negative, i. e. it neglects that a tag recommender might be able to improve the quality of the tag assignments. This point of view is also taken in [108, 116] where it is studied in how far users deviate from

their true preferences, i. e. from their uninfluenced behavior, when seeing tag recommendations.

In case of the online evaluation methodology (see Subsection 2.2.3 and [55]), the gold standard consists of the influenced tag assignments of the individual users after seeing the respective tag recommendations. Thus, it measures in the context of a user experiment how often users really accept one of the recommendations. In [55], it is even proposed to measure it live in the Bibsonomy system. According to the online evaluation methodology, the perfect tag recommender knows exactly which tags the user assigns to a resource after seeing the recommendations. Thus, also the online methodology sees tag recommendations primarily as a tool for reducing the indexing effort. But in contrast to the offline methodology, it is able to cope with a changed behavior of the users due to seeing the recommendations. Nevertheless, it is not able to distinguish a positive change in the behavior, e. g. due to better indexed resources, from a negative change in the behavior.

## 6.2 Research Hypotheses

In this section, we analyze for two of the tag recommenders from Delicious (see Fig. 2.1 and Section 2.2) how we expect that they influence the inter-resource and inter-indexer consistency in a tagging system. For the two tag recommenders we derive the hypotheses that in their case the inter-resource consistency and the inter-indexer consistency are not positively correlated. If we are able to show with the help of our user experiment in Section 6.3 that our hypotheses hold then only the inter-resource consistency and not also the inter-indexer consistency can be used for evaluating the influence of tag recommenders on the indexing quality.

### 6.2.1 Increasing the Inter-Indexer Consistency

One important way for increasing the inter-indexer consistency in a tagging system is to add a feedback mechanism that exposes the users to each others tags. An example of a tag recommender that adds this feedback mechanism is the *Popular Tags* recommender of Delicious (see Fig. 2.1). It is an example of a tag recommender that is based on the first paradigm of recommending tags, i. e. it recommends tags based on the tag assignments of other users (see Section 2.2). The *Popular Tags* recommender suggests the seven most popular tags of a resource.

In several studies [33, 38, 57, 64, 77, 100], it has been shown that the recommendation of popular tags leads to a slower vocabulary growth in a tagging system. With the help of our Epistemic Model, we are able to reproduce these results (see Subsection 5.5.3). The decrease in the vocabulary growth speed depends on the probability with which the users select one of the recommended popular tags (see Fig. 5.6).

In this thesis, we use the tag reuse rate  $tr_j$  (see Equation 6.3) for measuring the influence of a tag recommender on the inter-indexer consistency. It corresponds to the ratio between the number of tag assignments and the number of distinct tags. Thus, if the suggestion of popular tags decreases the vocabulary growth speed in a tagging system, i. e. less distinct tags occur in the same number of tag assignments, we expect that the inter-indexer consistency increases:

**Hypothesis 4** *Suggesting the users a list with the most popular tags at a resource increases the inter-indexer consistency in a tagging system.*

How does the *Popular Tags* recommender influence the inter-resource consistency and the indexing quality in a tagging system? According to [127], a positive correlation between inter-indexer consistency and inter-resource consistency can be expected if the indexing terms are “selected individually and independently by each of the indexers”. But what happens if users are influenced by popular tags, and thus if their tag assignments are no longer completely independent of each other? Again, we can use tagging models for predicting the influence of popular tags, but this time on the inter-resource consistency:

In [116], it has been argued with the *User’s Choice Model*, which is very similar to our *Epistemic Model with Word Frequencies* (see Subsection 4.3.2 on page 65), that a recommender based on popular tags may distort the true tagging preferences of a user. Thus, the user applies different tags than without seeing the recommendations. But distorting the true tagging preferences is not necessarily a negative thing: Increasing as well as decreasing the inter-resource consistency of the tag assignments of users requires distorting the true tagging preferences. However, in [38] it has been argued that in case of the *Popular Tags* recommender, the tag frequencies converge to a random limit. In consequence, the tag frequencies no longer only express the importance of an aspect for describing the resource but they are also influenced by a random process. In the following, we expect that the influence of this random process decreases the inter-resource consistency of the tag assignments:

**Hypothesis 5** *Suggesting the users a list with the most popular tags at a resource decreases the inter-resource consistency in a tagging system.*

### 6.2.2 Increasing the Inter-Resource Consistency

One important way for increasing the inter-resource consistency in a tagging system is to remember a user which aspects he/she has identified before for other resources and which tag has been used for describing the corresponding aspect. This increases the probability that (1) a user does not forget to describe aspects that span several resources in his/her collection, and

that (2) the same tag instead of a synonym or slightly different spelling is used for describing the corresponding aspect. From the three recommenders available in Delicious, the *User Tags* recommender offers this kind of support to the user (see the *Your Tags* suggestions in Fig. 2.1). The *User Tags* recommender is an example of a tag recommender that is based on the second paradigm of recommending tags, i. e. it recommends tags based on the previous tag assignments of the current user (see Section 2.2). The *User Tags* recommender simply recommends all previously used tags of the user. We expect that by helping the individual user in establishing a consistent tagging vocabulary and in consistently applying it in his/her resource collection, we are able to not only increase the inter-resource consistency of the tag assignments of the single user but also the inter-resource consistency in the whole tagging system:

**Hypothesis 6** *Suggesting the user his/her own previously used tags increases the inter-resource consistency and indexing quality in a tagging system.*

How does the *User Tags* recommender influence the inter-indexer consistency of the tag assignments in a tagging system? For this purpose, we discuss how the Epistemic Model has to be adapted in order to simulate the influence of the *User Tags* recommender. When tagging the first resource in a user's collection, the user does not get any recommendations because no previous tag assignments are available. Thus, when describing the aspects of the first resource, the user picks tags according to the probabilities in his/her background knowledge. In consequence, the tag assignments at the first resource reflect the level of inter-indexer consistency that is naturally emerging from the background knowledge of the users. Subsequently, when tagging the second resource, the user can either choose tags from his/her background knowledge or from the tags previously used at the first resource. Both sources for tag assignments reflect the naturally emerging level of inter-indexer consistency. Thus, independent of whether a user picks tags from the background knowledge or from the recommendations of the *User Tags* recommender, the resulting tag assignments for the second resource can also be assumed to reflect the level of inter-indexer consistency that is naturally emerging from the background knowledge of the users. This argument can be repeated for all subsequent resources in the collection of a user, thus leading to the following hypothesis:

**Hypothesis 7** *Suggesting the user his/her own previously used tags does not lead to a significant increase or decrease of the inter-indexer consistency in a tagging system.*



## 6.3 User Experiment

In this section, we describe the web-based user experiment that we use for evaluating our hypotheses from Section 6.2. The experiment is divided into two phases:

During the first phase (see Subsection 6.3.1), we distinguish three experimental conditions. Under all three conditions, the participants are asked to assign tags to a set of ten web pages. But depending on the experimental condition, the participants get different kinds of tag recommendations or no recommendations at all. Under the first experimental condition, the participants do not get any tag recommendations. The tag assignments of these participants are used for determining the level of inter-resource consistency and inter-indexer consistency that naturally emerge in tagging systems due to the shared background knowledge of the participants. Under the second experimental condition, tag recommendations of the *Popular Tags* recommender are shown to the participants. By comparing the results of the first and second experimental condition, we are able to evaluate Hypothesis 4 and 5. Under the third experimental condition, tag recommendations of the *User Tags* recommender are shown to the participants. By comparing the results of the first and the third experimental condition, we are able to evaluate Hypothesis 6 and 7.

During the second phase (see Subsection 6.3.2), the participants are asked to group the previously seen web pages into topical clusters. The clusters indicate which of the web pages are perceived by the participants to be about similar topics, and should thus also have similar tag vectors. The groupings are used as the independent indicator of the resources' similarity to each other that is required for computing the inter-resource consistency of the tag assignments from the first phase of the experiment (see Subsection 6.1.1).

Finally, in Subsection 6.3.3, we describe our strategy for recruiting participants and the sizes of the data set that has been collected during the experiment.

### 6.3.1 Phase 1: Tagging of Web Pages

During the first phase of the experiment, the participants should assign tags to ten web pages that are shown sequentially in a random order. Fig. 6.3 shows the instructions given to the participants prior to starting the first phase. The number of shown web pages is restricted to 10 in order to restrict the effort required for the experiment to approximately 15 minutes. This time restriction is important for avoiding high drop-out rates of the voluntary participants (see Subsection 6.3.3), and for retaining the participants' level of motivation over the duration of the experiment.

**Background of the Experiment**

This experiment is part of my PhD thesis in which I'm studying tagging systems ([What are tagging systems?](#)). The experiment helps to better understand how keywords are used for organizing collections of web pages. **Effort: ~15 minutes.**

**Running the Experiment**

- 10 web pages will be shown to you, one after another.
- Assign any number of keywords to each web page.
- Keywords are like categories and/or they describe the content of a page. **Example:** *You may use the keyword "work" for grouping web pages relevant for your work.*
- The keywords are primarily for yourself, to find your way in your own collection of web pages.

Figure 6.3: Instructions given to the participants of our experiment.

For the experiment, we use the same set of web pages as in an experiment reported by Bollen and Halpin in [13]. For their experiment, Bollen and Halpin randomly selected 11 web pages that are tagged with the tag *lifestyle* on Delicious. With this selection strategy, Bollen and Halpin wanted to ensure that the web pages appeal to the general public, and that no specialized background is required by the participants for understanding the pages. This helps to avoid an influence of the participants' familiarity with a specialist subject matter on the experimental results [13]. Nevertheless, from Bollen and Halpin's set of 11 web pages, we remove one web page because a pretest has shown that participants have problems in understanding the topic of the web page based on a screenshot of it. The URLs of the remaining 10 web pages that are used in our experiment are given in Tab. 6.1.

In Fig. 6.4, the tagging interface is shown that is used during the first phase of our experiment for tagging the web pages from Tab. 6.1. The web pages are shown to the participants in sequential random order. At the top of the tagging interface, the participants see a progress bar that gives an impression how many pages they still have to tag. It is our objective to reduce the drop-out rate of participants by showing them their progress in the experiment. Below the progress bar, the input field for the tags is available. The input field is scrollable such that no restriction on the number of tags is imposed. When tagging the first web page, the participants see a pop-up notice for the input field, which instructs them to separate multiple tags by comma, and to leave the input field blank if the shown web page is incomprehensible to them.

Depending on the experimental condition, the input field for the tags is followed by a tag cloud, which gives tag recommendations to the participant.

ID	URL
1	http://www.theonion.com/
2	http://news.bbc.co.uk/2/hi/uk_news/6057734.stm
3	http://uk.moo.com/
4	http://www.tvtrip.com/
5	http://www.panoramas.dk/
6	http://www.sleeptracker.com/
7	http://blisstree.com/feel/what-happens-to-your-body-if-you-drink-a-coke-right-now/
8	http://www.patentlysilly.com/
9	http://www.whfoods.com/
10	http://www.webmd.com/balance/features/your-guide-to-never-feeling-tired-again/

Table 6.1: URLs of the 10 web pages used during the experiment.

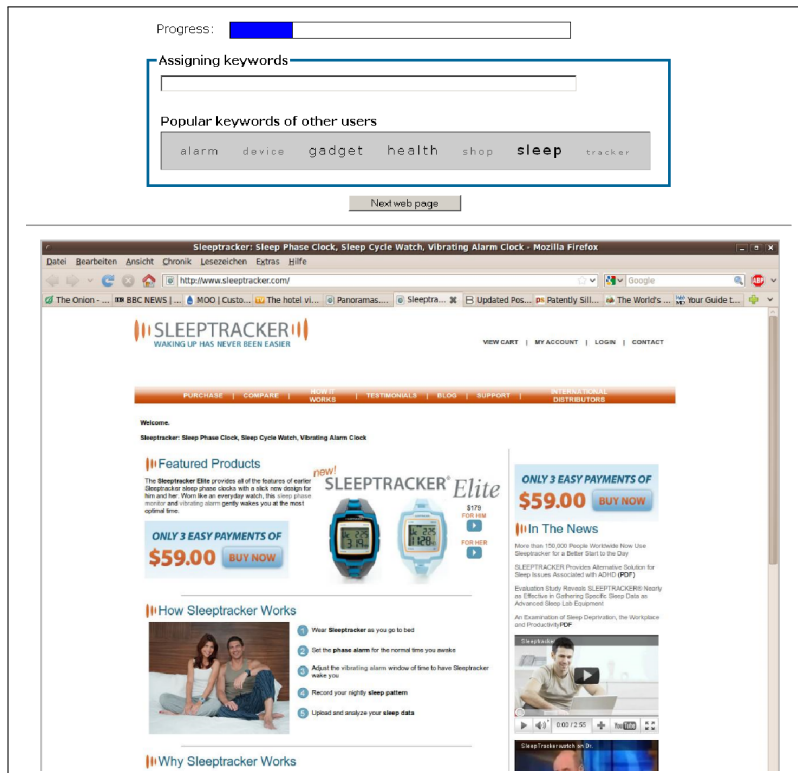


Figure 6.4: The tagging interface used for assigning tags to the 10 web pages. Depending on the experimental condition, a tag cloud with the tag recommendations is displayed below the input field for the tags. Here, the interface for the *Popular Tags* condition is shown.

By clicking on one of the recommended tags, the participant adds it to the input field from where the tag can also be removed again. In our experiment, we distinguish the following three conditions:

- **No Suggestions:** Under this condition, the participants do not get any tag recommendations while tagging the ten web pages. This participant group is the control group to which we compare the results of the other two experimental conditions.
- **Popular Tags:** Under this condition, the seven most popular tags for the current web page are shown to the participants. The participants are informed that the recommended tags correspond to “Popular keywords of other users” (see Fig. 6.4). The most popular tags are based on the tag assignments of the previous participants of the same experimental condition for the same web page. Prior to the experiment, each of the web pages has been initialized with the tags of a random user from Delicious for the same web page (see Tab. 6.2). The initialization with a random posting from Delicious is necessary in order to introduce a comparable level of randomness to the tag assignment process as in a real system like Delicious. In a real system, the web pages would also be first tagged by different users. Without this initialization with a random posting, the first tag assignments would all come from the first participant of the *Popular Tags* condition.
- **User Tags:** Under this condition, a participant sees all tags that he/she previously used in the experiment. For the first web page, no tag recommendations are shown to the participant. The participant is informed that the recommended tags correspond to “Your previously used keywords”.

Together with the tagging interface, the experiment participants see a screenshot of the respective web page. Showing a screenshot instead of the live version of the web page has two reasons: The most important reason is to ensure that all participants get the same stimulus, i. e. that they see exactly the same version of the web page. This is especially important for web pages that are portals with fast changing content like *The Onion* (URL-1), *Panoramas.dk* (URL-5) or *Patently Silly* (URL-8). The second reason is to avoid that the participants get distracted from their tagging task by starting to browse the linked web pages. Because the title and the URL of the web page may provide important information that influence the tag assignments of the participants [71], the screenshot shows the complete browser window in which the web page is opened. The used screenshots of all web pages are available in the materials accompanying this thesis (see Appendix C). Furthermore, the user experiment is still accessible under <http://userpages.uni-koblenz.de/~klaasd/experiment/>.

ID	URL
1	theonion, news, america
2	bbc, news, evolution, human
3	moo, business cards, post cards, printing
4	tvtrip, travel, hotels, reviews
5	panorama, background image
6	sleep, alarm, shop
7	health, coke, diet
8	funny, patents
9	health, food
10	sleep, health, guide

Table 6.2: Tags used for bootstrapping the *Popular Tags* condition.

### 6.3.2 Phase 2: Grouping of Web Pages

During the second phase of the experiment, the participants should group the previously tagged web pages into topical clusters. The topical clusters are required for computing the inter-resource consistency of the tag assignments (see Subsection 6.1.1). As can be seen in Fig. 6.5, the instructions for the second phase are given to the participants together with the user interface for grouping the web pages. Thus, the participants are not aware of the second phase until it is started (see the initial instructions in Fig. 6.3). This avoids that a participant may guess the purpose of the experiment, thus leading to a possibly adapted tagging behavior during the first phase of the experiment.

When entering the second phase, the participant sees in the left column of the user interface the screenshots of the previously tagged web pages. In order to recall the details of a web page, the participant may enlarge a screenshot by clicking on it. On the right side, the participant may create an arbitrary number of clusters. Initially, no clusters are shown on the right side of the user interface in order to avoid biasing the participant to a certain number of clusters. Furthermore, when creating a cluster, the participant is asked to provide a name for it. The name of the cluster fulfills two purposes: First, the name is important for the participant in order to be able to keep track of his/her clusters and their content. Second, during our evaluation of the experiment in Section 6.4 and 6.5 we can use the name for analyzing the intention of the participants, i. e. with regard to which topic they see a connection between the web pages in a cluster.

The grouping phase can not be finished until all web pages are assigned to a cluster. A web page is assigned to a cluster by dragging its screenshot from the left column to the area of the cluster on the right. Afterwards,



Figure 6.5: The user interface for grouping the web pages into topical clusters. Instructions on how to use the interface are given at the top.

the screenshot of the web page is removed from the left column, i. e. a web page can only be assigned to one cluster. Thus, if a participant thinks that a web page is possibly related to different clusters, then the participant is forced to decide to which of the clusters the web page is more related (see Subsection 6.1.1 for the rationale of this restriction).

### 6.3.3 Recruiting the Participants

We used several channels for recruiting participants for the experiment: (1) We approached colleagues and friends. (2) We promoted the experiment during the poster session of the Web Science Conference 2011<sup>1</sup>. (3) We published the call for participation on Twitter.<sup>2</sup> (4) We published the call on several public mailing lists that address the information retrieval or the information science community.<sup>3</sup> (5) We distributed the call in an internal news group of the University of Koblenz. The participation in the experiment was completely voluntary, no incentives were given to finish the experiment.

All in all, 877 users accessed the web page of the experiment. 639 of these users started the first phase of the experiment and 582 users also

<sup>1</sup><http://www.websci11.org>

<sup>2</sup><https://twitter.com/ststaab/status/83170417975635968>

<sup>3</sup>Example: <http://mail.asis.org/pipermail/asis-l/2011-July/005953.html>

<b>German</b>	U	T	Y	Y / U
No Suggestions	74	706	2,134	28.84
User Tags	79	466	1,507	19.08
Popular Tags	78	531	2,228	28.56
<b>English</b>	U	T	Y	Y / U
No Suggestions	115	973	3,150	27.39
User Tags	118	819	2,919	24.74
Popular Tags	118	550	3,003	25.45

Table 6.3: Sizes of the experimental data sets. Only participants who finished tagging all ten web pages are included.

finished the first phase. 530 of the users who finished the first phase also clustered the web pages during the second phase of the experiment. In Section 6.4, we only use the tag assignments and clusters of the 582 users who finished at least the first phase. According to a questionnaire at the end of the experiment, approximately 53% of the participants use tagging systems for searching regularly or sometimes. The rest tried it either once or not all. Furthermore, 45% of the participants upload content to tagging systems regularly or sometimes. More detailed results of the questionnaire are available in Appendix C in Section C.1.

Due to our recruiting strategy, we expected to observe a homogeneous subgroup of native German speakers. Thus, we decided to not only offer an English variant of our experiment but also a German variant. In both variants, the same English web pages are shown but in the German variant we instruct the participants to preferably use German keywords. Thus, German participants are able to use their larger and more accurate active German vocabulary during tagging. Each participant decided on his/her own whether to participate in the German or English variant. All in all, 231 participants finished the experiment in the German variant and 351 participants in the English variant (see Tab. 6.3).

After choosing between the German or the English variant of the experiment, each participant has been randomly assigned to one of the three conditions described in Subsection 6.3.1. The experimental condition with the most participants was excluded from the random assignment, if it already contained at least 5 participants more than the condition with the fewest participants. This ensured a balanced distribution of participants over the experimental conditions. The participants were not aware that different experimental conditions exist and that they have to create topical clusters at the end of the experiment. They were only told that the experiment analyses how keywords are used for organizing collections of web pages (see Fig. 6.3).

## 6.4 Results

In this section, we are presenting the results of our user experiment. The results help us in evaluating our hypotheses from Section 6.2 about how the two tested recommenders influence the inter-resource and the inter-indexer consistency in a tagging system. In a first step, we evaluate in Subsection 6.4.1 in how far the participants of the different experimental conditions have in average the same perception of the resources' similarity, like it is expressed by their topical clusters. Only if the participants under two experimental conditions perceive the similarity of the resources in the same way, we can compare the level of inter-resource consistency between the sets of tag vectors created under the two experimental conditions. The comparison of the inter-resource consistency of the tag vectors created under the different experimental conditions, and thus the evaluation of Hypothesis 5 and 6, is then available in Subsection 6.4.2. Finally, in Subsection 6.4.3 we compare the inter-indexer consistency of the tag vectors created under the different experimental conditions. This comparison is used for evaluating Hypothesis 4 and 7.

### 6.4.1 Similarity of Topical Clusters

In Subsection 6.1.1, we have described how to use the average Silhouette Coefficient  $E(s_{ij})$  for measuring the inter-resource consistency of the tag assignments. But before we can apply this method on our data, we have to verify that the participants of the different experimental conditions have in average identified the same topical clusters during the second phase of the experiment (see Subsection 6.3.2). Otherwise, the differences in the  $E(s_{ij})$ -values may not only be caused by the influence of the respective experimental condition but also by differences in the topical clusters.

During the second phase of the experiment, we received feedback from 530 of our participants. A participant was only able to finish the second phase if every web page was assigned to one cluster. The participants were allowed to provide a name for each cluster in order to make it easier for them to keep track of their clusters. On average, each participant separated the 10 web pages into 4.76 clusters, i. e. 2,521 clusters have been created. Together, the participants identified 140 distinct clusters. Two topical clusters are considered as equal if they contain the same web pages.

In Fig. 6.6, the probabilities of the eleven most frequently identified clusters are shown. The probabilities are based on data from all 530 participants who completed the second phase of the experiment. Altogether, the eleven clusters from Fig. 6.6 represent 70.25% of all identified topical clusters. According to the names of the clusters, the 10 web pages are roughly related to 6 different topics. URL-1, URL-5 and URL-6 are each on the border between two topics. For example, the web page *The Onion* (URL-1) pub-



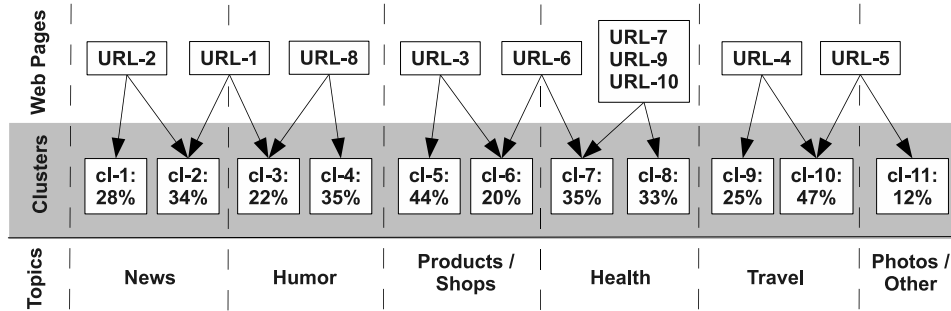


Figure 6.6: Visualization of the 11 most frequently identified clusters of web pages. Each box in the gray area corresponds to one cluster. Within the box of each cluster it is given by how many participants the cluster has been identified. For example, 28% of all participants put the *BBC* web page (URL-2) alone into a cluster, leading to cluster *cl-1*. Another 34% of the participants instead decided to group *BBC* (URL-2) together with *The Onion* (URL-1), leading to cluster *cl-2*. The remaining 38% of the participants have put URL-2 into other, less frequent clusters. Nevertheless, an analysis of the names used for *cl-1* and *cl-2* reveals that both clusters are seen as related to the *News* topic.

lishes satirical news articles. 34% of the participants think that it is more related to the *News*-topic and thus they group it with an article from the *BBC* web page (URL-2), leading to cluster *cl-2*. In contrast, 22% of the participants emphasize more the *Humor*-topic and thus group it with *Patently Silly* (URL-8), which lists funny and strange patents, leading to cluster *cl-3*.

In the following, we evaluate in how far significant differences can be observed between the cluster probabilities if the probabilities are only based on the topical clusters of participants from a single experimental condition. This evaluation helps us to answer in how far the participants of the different experimental conditions have in average the same perception of the resources' similarity. During our evaluation, we use the  $\chi^2$ -Test [23, p. 199ff]. Only if the  $\chi^2$ -Test tells us that in two experimental conditions the clusters have been identified with the same probabilities, we can compare the  $E(s_{ij})$ -values between the two experimental conditions (cf. Subsection 6.1.1).

### The $\chi^2$ -Test

The  $\chi^2$ -Test [23, p. 199ff] is a nonparametric test that can be applied on nominal data, i. e. on data where each observation can be categorized into exactly one of several categories. For the  $\chi^2$ -Test, all observations are arranged in a contingency table with  $r$  rows and  $c$  columns. Each row corresponds to one random sample of observations, and each column corresponds to one of the categories. In our case, a row in the contingency table contains how

often participants from one of the experimental conditions identified one of the 140 distinct clusters. Given a contingency table with  $r$  rows and  $c$  columns, we denote the number of observations in row  $i$  and column  $j$  with  $O_{ij}$ . Furthermore, we define  $n_i = \sum_{j=1}^c O_{ij}$  and  $c_j = \sum_{i=1}^r O_{ij}$ . Given these definitions, the test statistic  $T$  of the  $\chi^2$ -Test is computed as follows:

$$T = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \text{where} \quad E_{ij} = \frac{n_i \cdot c_j}{\sum_{i=1}^r n_i} \quad (6.5)$$

The lower  $T$ , the more similar to each other are the probabilities with which participants identified the same cluster. The level of significance  $p$  of a concrete  $T$  value corresponds to the probability of the  $\chi^2$  distribution with  $(r - 1) \cdot (c - 1)$  degrees of freedom to exceed the observed value of  $T$  [23, p. 201]. For  $p \geq 0.1$  we accept the hypothesis that there exist no significant differences between the probabilities of the different clusters (cf. [79]).

In [23, p. 201f], it has been pointed out that the validity of the  $\chi^2$ -Test may be endangered if not most of the  $E_{ij}$ -values in Equation 6.5 are greater than 1.0. In our case, we observe  $E_{ij}$ -values less than 1.0 for columns that only contain 1 observation. In [23], it is suggested to merge the observations of two or more columns in order to eliminate such small  $E_{ij}$ -values. In our case, we merge the observations of all columns that only contain 1 observation into a single column. The single column then contains how often a participant in one of the compared experimental conditions identifies a cluster that none of the other participants has identified.

### Comparing the Experimental Conditions

After defining the  $\chi^2$ -Test, we can now apply it for comparing the cluster probabilities across the different experimental conditions. In a first test, we compare the cluster probabilities from the English variant of the experiment with those from the German variant of the experiment. The test reveals that the clusterings differ significantly ( $T = 161.69$ ,  $n_1 = 1519$ ,  $n_2 = 1002$ ,  $p < 0.01$ ). Thus, we cannot compare the  $E(s_{ij})$ -values across the two language variants of the experiment. But for evaluating our hypotheses from Section 6.2, it is more important whether we can compare the  $E(s_{ij})$ -values from the *No Suggestions* condition to the  $E(s_{ij})$ -values of the other two experimental conditions within the same language variant:

**No Suggestions vs. Popular Tags** Only for the German variant of the experiment the clusterings from the *No Suggestions* condition and from the *Popular Tags* condition can be considered as equal. For the English variant, the clusterings from the two conditions differ significantly. Possible explanations for the significant differences are discussed in Section 6.5.

German:  $T = 39.25$ ,  $n_1 = 339$ ,  $n_2 = 323$ ,  $p = 0.75$ ;  
 English:  $T = 63.04$ ,  $n_1 = 489$ ,  $n_2 = 515$ ,  $p = 0.06$

**No Suggestions vs. User Tags** For the English variant of the experiment as well as for the German variant the clusterings from the *No Suggestions* condition and from the *User Tags* condition can be considered as equal. For the German variant, the differences between the clusterings are smaller than for the English variant.

German:  $T = 35.03$ ,  $n_1 = 339$ ,  $n_2 = 340$ ,  $p = 0.86$ ;  
 English:  $T = 51.99$ ,  $n_1 = 489$ ,  $n_2 = 515$ ,  $p = 0.36$

All in all, the results in this subsection show that there are only minor differences in the cluster probabilities between the three German experimental conditions. Thus, we can compare the  $E(s_{ij})$ -values between all three German experimental conditions. In contrast, in the English experiment variant we can only compare the *No Suggestions* and the *User Tags* condition. The English *No Suggestions* and the English *Popular Tags* condition cannot be compared because of the differences in the identified topical clusters.

#### 6.4.2 Measuring the Inter-Resource Consistency

In this subsection, we evaluate Hypothesis 5 and 6, which are related to the influence of tag suggestions on the inter-resource consistency. In the following, we use the average Silhouette Coefficient  $E(s_{ij}^x)$  from Subsection 6.1.1 for measuring the inter-resource consistency for a tagging system  $X$ . We compare the  $E(s_{ij}^n)$ -value of the *No Suggestions* condition to the  $E(s_{ij}^p)$ -value of the *Popular Tags* condition. Furthermore, we compare  $E(s_{ij}^n)$  to the  $E(s_{ij}^u)$ -value of the *User Tags* condition. Based on Hypothesis 5 and 6, we expect the following relations between the  $E(s_{ij}^x)$ -values:

$$E(s_{ij}^p) < E(s_{ij}^n) \quad \text{and} \quad E(s_{ij}^n) < E(s_{ij}^u) \quad (6.6)$$

For computing the  $E(s_{ij}^x)$ -values, we have to compare the tag vectors from the different experimental conditions to a set of topical clusters. For the German experiment variant, we compare the tag vectors to the union of all topical clusters that have been given by participants of the German variant. This way, we ensure that differences between the  $E(s_{ij}^x)$ -values are only caused by differences in the tag vectors and not also by slight differences in the cluster probabilities between the experimental conditions. Creating the union of all topical clusters from the German experiment variant is valid because we have shown in Subsection 6.4.1 that the slight differences in the probabilities are not significant. In contrast, for the English experiment variant, we only compare the tag vectors from the *No Suggestions* and the *User Tags* condition to the union of the topical clusters from these two conditions. The clusters and tag vectors from the English *Popular Tags*

<b>German</b>	Controlled for Users		Controlled for TAS	
	$E(s_{ij})$	$E(tr_{ij})$	$E(s_{ij})$	$E(tr_{ij})$
No Suggestions	0.1847	2.44	0.1670	2.07
User Tags	0.2367	2.39	0.2470	2.34
Popular Tags	0.1474	3.60	0.1478	3.23
<b>English</b>				
No Suggestions	0.1713	2.76	0.1682	2.57
User Tags	0.1915	2.68	0.1946	2.71
Popular Tags	N/A	4.67	N/A	4.44

Table 6.4: Influence of the experimental conditions on the inter-resource consistency and on the inter-indexer consistency. For the results on the left, we have restricted the number of users such that under each of the experimental conditions the same number of users contributed to the tag vectors. For the results on the right, we have restricted the number of tag assignments (TAS) such that the tag vectors for the same resource contain the same number of tag assignments.

condition have to be excluded from the evaluation because of the significant differences in the cluster probabilities as it is shown in Subsection 6.4.1.

A summary of the experimental results is shown in Tab. 6.4. For the results, we have restricted the number of participants such that under each of the experimental conditions the same number of participants contributed to the tag vectors. For the German variant, we have restricted it to the first 74 participants of each of the experimental conditions. For the English variant, we have restricted it to the first 115 participants. Thus, we control that different numbers of participants do not cause the differences in the results. Additionally, we control that different numbers of tag assignments do not cause the differences in the results. For this purpose, we have restricted the number of tag assignments such that under each of the experimental conditions the tag vectors for the same resource contain the same number of tag assignments.

### Significance of Results

For checking whether the results reported in Tab. 6.4 are significant, we apply a two-tailed Mann-Whitney Test [23, p. 272ff]. The Mann-Whitney Test is a nonparametric test that can be applied on ordinal data, i.e. on data where the observed values can be arranged from smallest to largest. The two-tailed version of the test compares for two random samples  $X$  and  $Y$  of observed values whether they tend to the the same average value, i.e.  $E(X) = E(Y)$ , or whether they tend to different average values, i.e.  $E(X) \neq E(Y)$ .

Given a random sample  $X$  of size  $n$  and a random sample  $Y$  of size  $m$ , the Mann-Whitney Test first assigns the ranks 1 to  $n+m$  to the observed values from smallest to largest. If several sample values are equal to each other, then the average of the ranks is assigned that would have been assigned to them had there been no ties (cf. [23, p. 272]). The test statistic  $T$  of the Mann-Whitney Test then corresponds to the sum of the ranks assigned to the values in  $X$ . If the observed values from  $X$  and  $Y$  tend to the same average value, then  $T$  is close to  $\frac{n \cdot (n+m+1)}{2}$ . Otherwise,  $T$  deviates from this expected value. If there are many ties, the value of  $T$  has to be corrected to the value  $T_1$  in order to account for the ties. Details on how to calculate  $T_1$  are available in [23, p. 272ff].

In case of the  $E(s_{ij})$ -values in Tab. 6.4, we have to calculate the corrected test statistic  $T_1$  because there are many tied values in the data. If the observed values from  $X$  and  $Y$  tend to the same average value, then the expected value of  $T_1$  is 0. Otherwise,  $T_1$  deviates from this expected value. If  $T_1$  is positive then this means that the values from  $X$  have in average a higher rank than the values in  $Y$ . In consequence, the larger  $T_1$ , the more likely  $E(X) > E(Y)$ . If  $T_1$  is negative then this means that the values from  $X$  have in average a lower rank than the values in  $Y$ . In consequence, the lower  $T_1$ , the more likely  $E(X) < E(Y)$ .

Based on the  $T$  or the  $T_1$  value, we can calculate the level of significance  $p$ . In the two-tailed version of the test,  $p$  can be used for accepting or rejecting the hypothesis  $E(X) = E(Y)$ . If  $p \geq 0.1$ , then we can accept the hypothesis that there are no significant differences between  $E(X)$  and  $E(Y)$ , i. e. we can accept that  $E(X) = E(Y)$ . Otherwise, if  $p < 0.1$ , we have to accept the alternative hypothesis that there are significant differences, i. e. we have to accept that  $E(X) \neq E(Y)$ .

### Effect Size

In addition to the two-tailed Mann-Whitney Test, we also apply the Hodges-Lehmann Estimator of Shift [23, p. 281f] on our results. The Hodges-Lehmann Estimator of Shift can be used for measuring the effect size, i. e. it can be used for determining the 95% confidence interval for the difference  $E(X) - E(Y)$  between the average values for two random samples  $X$  and  $Y$  of observed values. The Hodges-Lehmann Estimator of Shift can be applied on ordinal data.

Given a random sample  $X$  of size  $n$  and a random sample  $Y$  of size  $m$ , we first have to compute the pairwise differences of all possible pairs of a value from  $X$  and a value from  $Y$ . As a result, we get  $n \cdot m$  differences, which we arrange from smallest to largest. From these differences, the  $k^{\text{th}}$  smallest difference and the  $k^{\text{th}}$  largest difference correspond to the lower and upper limit of the confidence interval for the difference  $E(X) - E(Y)$  between the average values of  $X$  and  $Y$ . The value of  $k$  depends on the

values of  $n$ ,  $m$  and on which confidence interval we want to calculate. More details about how to calculate  $k$  are available in [23, p. 281f]. In our case, we calculate a 95% confidence interval, i. e. with a probability of 95%, the true difference of  $E(X) - E(Y)$  is between the calculated lower and upper limit of the confidence interval.

### No Suggestions vs. Popular Tags

In the following, we use the Mann-Whitney Test and the Hodges-Lehmann Estimator of Shift for evaluating the effect of suggesting popular tags on the inter-resource consistency as it is measured by the average Silhouette Coefficient. For this purpose, we compare the  $E(s_{ij}^n)$ -value of the *No Suggestions* condition to the  $E(s_{ij}^p)$ -value of the *Popular Tags* condition as they are shown in Tab. 6.4. According to Hypothesis 5 we expect  $E(s_{ij}^p) < E(s_{ij}^n)$ . Because of the differences in the perception of the resources' similarity for the English variant of the experiment (see Subsection 6.4.1) we can test Hypothesis 5 for the German experiment variant only.

For the German experiment variant, we can confirm that  $E(s_{ij}^p) < E(s_{ij}^n)$ . The two-tailed Mann-Whitney Test shows that the difference between the corresponding  $E(s_{ij}^x)$ -values in Tab. 6.4 are significant. If we control for the influence of the number of participants, we get  $T_1 = 7.42$ ,  $n = m = 1798$ , and  $p < 0.01$ . According to the Hodges-Lehmann Estimator of Shift, suggesting popular tags decreases the average Silhouette Coefficient by 0.0472 with a 95% confidence interval of [0.0337, 0.0582]. If we control for the influence of the number of tag assignments, we get  $T_1 = 4.84$ ,  $n = m = 1798$  and  $p < 0.01$ . In this case, suggesting popular tags decreases the average Silhouette Coefficient by 0.0262 with a 95% confidence interval of [0.0164, 0.0360].

*Thus, our experimental results show that recommending the seven most popular tags of a resource has a significant influence on the inter-resource consistency and the indexing quality in tagging systems. The results support Hypothesis 5 that recommending the popular tags decreases the inter-resource consistency in tagging systems.*

### No Suggestions vs. User Tags

Now, we test the effect of suggesting the user his/her own previously used tags on the inter-resource consistency as it is measured by the average Silhouette Coefficient. We compare the  $E(s_{ij}^n)$ -value for the *No Suggestions* condition to the  $E(s_{ij}^u)$ -value for the *User Tags* condition as they are shown in Tab. 6.4. According to Hypothesis 6 we expect that  $E(s_{ij}^n) < E(s_{ij}^u)$ . For both language variants of the experiment, we can confirm that this relation holds.

If we control for the influence of the number of participants, the two-tailed Mann-Whitney Test shows that the differences between the corresponding  $E(s_{ij}^x)$ -values in Tab. 6.4 are significant (German:  $T_1 = -8.11, n = m = 1796, p < 0.01$ ; English:  $T_1 = -3.0563, n = m = 1721, p < 0.01$ ). For the German variant, suggesting the user his/her own previously used tags increases the average Silhouette Coefficient by 0.0775 with a 95% confidence interval of [0.0584, 0.0955]. For the English experiment variant, the average Silhouette Coefficient increases by 0.0306 with a 95% confidence interval of [0.0106, 0.0434].

If we control for the influence of the number of tag assignments, the differences in Tab. 6.4 are also significant (German:  $T_1 = -9.53, n = m = 1798, p < 0.01$ ; English:  $T_1 = -4.24, n = m = 1721, p < 0.01$ ). For the German variant, suggesting the user his/her own previously used tags increases the average Silhouette Coefficient by 0.1259 with a 95% confidence interval of [0.1051, 0.1437]. For the English experiment variant, the average Silhouette Coefficient increases by 0.0401 with a 95% confidence interval of [0.0218, 0.0583].

*Thus, our experimental results show that suggesting the user his/her own previously used tags has a significant influence on the indexing quality. The results support Hypothesis 6 that recommending the user's tags increases the inter-resource consistency in tagging systems.*

### 6.4.3 Measuring the Inter-Indexer Consistency

In this subsection, we evaluate Hypothesis 4 and 7, which are related to the influence of tag suggestions on the inter-indexer consistency. In the following, we use the average tag reuse rate  $E(tr_{ij}^x)$  from Subsection 6.1.2 for measuring the inter-indexer consistency for a tagging system  $X$ . We compare the  $E(tr_{ij}^n)$ -value of the *No Suggestions* condition to the  $E(tr_{ij}^p)$ -value of the *Popular Tags* condition. Furthermore, we compare  $E(tr_{ij}^n)$  to the  $E(tr_{ij}^u)$ -value of the *User Tags* condition. Based on Hypothesis 4 and 7, we expect the following relations between the  $E(tr_{ij}^x)$ -values:

$$E(tr_{ij}^u) = E(tr_{ij}^n) \quad \text{and} \quad E(tr_{ij}^n) < E(tr_{ij}^p) \quad (6.7)$$

A summary of the results is available in Tab. 6.4. Like in Subsection 6.4.2, we control for the influence of different number of participants and for the influence of different number of tag assignments (TAS). For controlling the significance of the results and for determining the effect size, we use the two-tailed Mann-Whitney Test and the Hodges-Lehmann Estimator of Shift described in Subsection 6.4.3.

### No Suggestions vs. Popular Tags

In the following, we evaluate the effect of suggesting popular tags on the inter-indexer consistency. For this purpose, we compare the  $E(tr_{ij}^n)$ -value of the *No Suggestions* condition to the  $E(tr_{ij}^p)$ -value of the *Popular Tags* condition as they are shown in Tab. 6.4. According to Hypothesis 4 we expect that  $E(tr_{ij}^n) < E(tr_{ij}^p)$ . For both language variants of the experiment, we can confirm that this relation holds.

If we control for the influence of the number of participants, the two-tailed Mann-Whitney Test shows that the differences between the corresponding  $E(tr_{ij}^x)$ -values are significant (German:  $T = 58, n = m = 10, p < 0.01$ ; English:  $T = 55, n = m = 10, p < 0.01$ ). For the German variant, suggesting popular tags increases the average tag reuse rate by 1.2274 with a confidence interval of  $[0.6912, 1.6111]$ . For the English variant, the average tag reuse rate increases by 1.7955 with a 95% confidence interval of  $[1.2760, 2.4027]$ .

If we control for the influence of the number of tag assignments, the differences in Tab. 6.4 are also significant (German:  $T = 57, n = m = 10, p < 0.01$ ; English:  $T = 55, n = m = 10, p < 0.01$ ). For the German variant, suggesting popular tags increases the average tag reuse rate by 1.2384 with a confidence interval of  $[0.7922, 1.5386]$ . For the English variant, the average tag reuse rate increases by 1.7038 with a 95% confidence interval of  $[1.2727, 2.1866]$ .

*Thus, our results show that recommending the seven most popular tags of a resource has a significant influence on the inter-indexer consistency. The results support Hypothesis 4 that recommending the popular tags increases the inter-indexer consistency in tagging systems.*

### No Suggestions vs. User Tags

Now, we test the effect of suggesting the user his/her own previously used tags on the inter-indexer consistency. We compare the  $E(tr_{ij}^n)$ -value for the *No Suggestions* condition to the  $E(tr_{ij}^u)$ -value for the *User Tags* condition as they are shown in Tab. 6.4. According to Hypothesis 7 we expect that  $E(tr_{ij}^n) = E(tr_{ij}^u)$ . For both language variants of the experiment, we can confirm that this relation holds.

If we control for the influence of the number of participants, the two-tailed Mann-Whitney Test shows that the differences between the corresponding  $E(tr_{ij}^x)$ -values in Tab. 6.4 are not significant (German:  $T = 104, n = m = 10, p = 0.97$ ; English:  $T = 113, n = m = 10, p = 0.57$ ). Accordingly, the 95% confidence interval for  $E(tr_{ij}^u) - E(tr_{ij}^n)$  is  $[-0.392, 0.3437]$  for the German variant and  $[-0.392, 0.2804]$  for the English variant.

If we control for the influence of the number of tag assignments, the differences in Tab. 6.4 are also not significant (German:  $T = 85.5, n = m =$



10,  $p = 0.15$ ; English:  $T = 93, n = m = 10, p = 0.41$ ). Accordingly, the 95% confidence interval for  $E(tr_{ij}^u) - E(tr_{ij}^n)$  is  $[-0.0964, 0.6021]$  for the German variant and  $[-0.2531, 0.4738]$  for the English variant.

*Thus, our results show that recommending the user his/her own previously used tags has no significant influence on the indexing quality. The results support Hypothesis 7 that recommending the user's tags does not lead to a significant increase or decrease of the inter-indexer consistency.*

## 6.5 Discussion

In Section 6.4, we have presented the results how the *User Tags* recommender and the *Popular Tags* recommender influence the inter-resource consistency and the inter-indexer consistency in tagging systems. The results are in agreement with our hypotheses from Section 6.2 according to which we have expected that the inter-resource consistency and inter-indexer consistency are not necessarily positively correlated with each other if tag recommendations are given to users. Thus, we have been able to show that one can not use the less complex measures of inter-indexer consistency for concluding on the influence of tag recommendations on the indexing quality in tagging systems. Instead one has to use measures of the inter-resource consistency, which check an additional requirement for a high indexing quality, namely that the users not only agree on how to annotate single resources but also on how to annotate a complete set of resources.

Our methodology for measuring the inter-resource consistency, which is described in Subsection 6.1.1, is able to evaluate in how far the tag assignments of the users are consistent with how the users perceive the similarity of resources. It can be used for comparing the indexing quality of the tag assignments in different tagging systems, given that the users in both tagging systems perceive the similarity of the resources in average in the same way. If this precondition of similar perception is not fulfilled, then it can not be said to which extent the differences in the tag assignments are reflecting different levels of inter-resource consistency or simply a different perception of the similarity of the resources.

Due to this restriction of our methodology, we have only been able to evaluate Hypothesis 5 for the German experiment variant. For the English variant, the levels of inter-resource consistency are incomparable because of differences in the perceived similarity of resources between the English *No Suggestions* and the English *Popular Tags* condition. In the following, we discuss possible explanations for the differences, and in how far they may be caused by the influence of the tag recommendations.

We start in Subsection 6.5.1 with discussing possible explanations for the differences between the English and the German experiment variant. We show that these differences are caused by the fact that the participants of the

English and the German experiment variant are random samples from two different populations, which differ in their level of comprehension of the web pages. Then, in Subsection 6.5.2, we show that the differences between the English *No Suggestions* condition and the English *Popular Tags* condition can not be explained in the same way. Instead, the results of our experiment point into the direction that seeing the popular tags at a resource may lead to learning effects that change how the users perceive the similarity between resources.

### 6.5.1 Influence of Comprehension of the Web Pages

In Subsection 6.4.1, we have shown that there exist significant differences between the participants of the German and English experiment variant how they perceive the similarity between the resources. In the following, we take a closer look at these differences and discuss in how far they can be caused by different levels of comprehension of the shown web pages. This level of comprehension especially plays an important role when judging for one of the web pages, namely *The Onion* (URL-1), whether it is more related to the topic *News* or to the topic *Humor*. *The Onion* is one of the three web pages that are seen by the participants on the border between two topical clusters (see Fig. 6.6).

In Fig. 6.7, we give an overview of how the participants of the two language variants have clustered the three web pages that are seen on the border between two topical clusters. It can be seen that the two groups of participants mainly differ in their decision of how to cluster *The Onion* (URL-1), which publishes satirical news articles. In the German variant, the vast majority of participants perceive the web page *The Onion* as related to the *News* topic, which is represented by cluster *cl-2* from Fig. 6.6. In contrast, the participants of the English variant see it more related to the *Humor* topic, which is represented by cluster *cl-3*. For the web pages *Sleep-tracker* (URL-6) and *Panoramas.dk* (URL-5), no such significant differences between the two language variants can be observed.

As can be seen in Fig. 3.6, at the first glance *The Onion* (URL-1) gives the impression of a regular news web page. Even the headline of the main article about the Yellowstone National Park may be perceived as a regular news article if not carefully reading and fully understanding it. In the following, we show that the participants' decision of how to cluster *The Onion* correlates with their level of comprehension of the web pages. The lower the level of comprehension, the higher the probability that a participant clusters *The Onion* according to its *News* aspect into cluster *cl-2*. In contrast, the higher the level of comprehension, the higher the probability to cluster it according to its *Humor* aspect into cluster *cl-3*.

An indicator of the level of comprehension is the self-assessment of the participants in the questionnaire from the end of the experiment. The de-

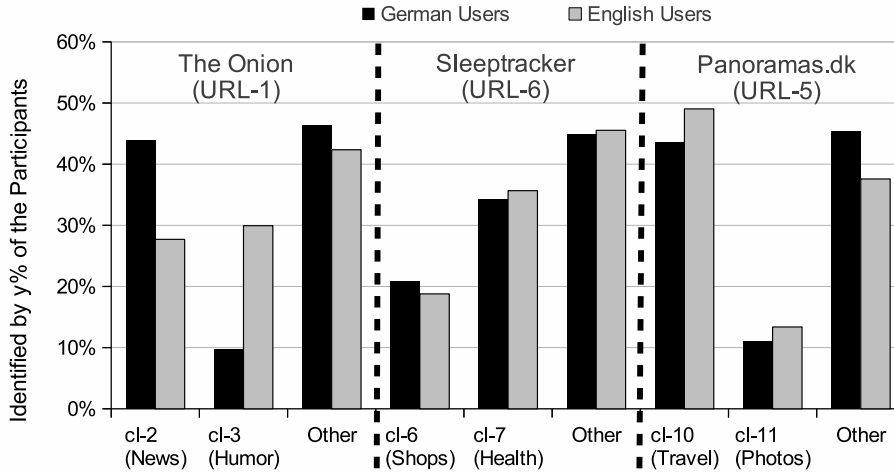


Figure 6.7: Overview of how the participants of the two language variants of the experiment have clustered the three web pages from Fig. 6.6 that are on the border between two topics. For each of the three web pages the probabilities are given with which they have been put into one of the eleven most popular clusters from Fig. 6.6, or in another of the 140 overall identified clusters.

tailed questions and results of the questionnaire are available in Appendix C in Section C.1. In the questionnaire, the participants were amongst others asked to rate the comprehensiveness of the web pages on a scale from 1 (poor) to 6 (good). In this self-assessment, the participants who clustered *The Onion* according to its *News* aspect have rated the comprehensiveness of the web pages in average with 5.23. In contrast, the participants who clustered *The Onion* according to its *Humor* aspect have rated the comprehensiveness of the web pages in average with 4.80. A two-tailed Mann-Whitney Test (see Subsection 6.1.1) shows that the difference between the two average ratings is significant ( $T_1 = -2.31, n = 182, m = 115, p = 0.02$ ). This correlation between the self-assessed level of comprehensiveness and the clustering decision can also be confirmed by only looking at the average ratings of either of the two experiment variants (German: 5.00 vs. 4.66; English: 5.29 vs. 4.95).

These findings suggest that a larger number of participants had problems in comprehending the content of *The Onion*. A lower level of comprehension increases the probability that a participant emphasizes the *News* aspect of *The Onion* by clustering it with a *BBC* web page into cluster *cl-2*. It is plausible that this level of comprehension is also correlated with the English language skills of a user. This additional correlation would also explain why in the German variant more participants had problems than in the English



Figure 6.8: Screenshot of *The Onion* (URL-1) that was shown to the experiment participants.

variant: In the German variant, presumably only non-native English speakers participated. In the English variant, due to our recruiting strategy (see Subsection 6.3.3) it can be assumed that the participants are a mixture of (1) non-native English speakers who likely have language skills comparable to the participants in the German variant, and (2) the native English speakers.

## 6.5.2 Influence of Learning Effects

In Subsection 6.4.1 it has been shown that also significant differences exist between the English *No Suggestions* condition and the English *Popular Tags* condition. In Fig. 6.9 it can be seen that these differences are also mainly caused by the decision of the participants of how to cluster *The Onion* (URL-1). In the following, we discuss in how far these differences can be caused by learning effects of the participants due to being exposed to the tags of other users, as it is also discussed in [35].

In Subsection 6.5.1, we have shown that the self-assessed level of comprehension correlates with the decision of the participants of how to cluster *The Onion*. Thus, before we can discuss whether the observations in Fig. 6.9 can be explained with learning effects, we first have to be sure that they

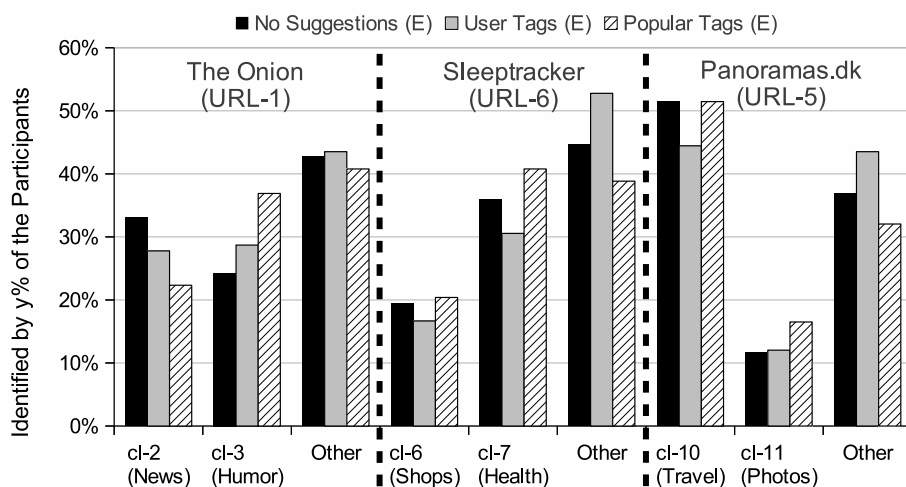


Figure 6.9: Overview of how the participants of the three English experimental conditions have clustered the three web pages from Fig. 6.6 that are on the border between two topics. For each of the three web pages the probabilities are given with which they have been put into one of the eleven most popular clusters from Fig. 6.6, or in another of the 140 overall identified clusters.

may not again be caused by the level of comprehension of the participants. For this purpose, we compare the self-assessed levels of comprehension between the English *No Suggestions* and the English *Popular Tags* condition. In case of the English *No Suggestions* condition the self-assessed level of comprehension is in average 5.09, and in case of the English *Popular Tags* condition it is 4.95. According to a two-tailed Mann-Whitney Test (see Subsection 6.1.1), the difference between the two conditions is just barely significant ( $T_1 = 1.71, n = 101, m = 102, p = 0.09$ ) but pointing to a slightly higher level of comprehension under the *No Suggestions* condition. Due to these results, we can exclude that the differences in the topical clusters between the two experimental conditions are caused by the level of comprehension because this would require a higher level of comprehension under the *Popular Tags* condition, and not the other way round.

Instead, it seems more plausible to explain the differences in Fig. 6.9 for *The Onion* with the influence of the tag recommendations under the *Popular Tags* condition. Indeed, under the English *Popular Tags* condition, for 107 of the 118 participants the list of recommended tags contained the tag “satire”. Additionally, the tag “fun” was contained 104 times in the list, and “humor” 89 times. Thus, for 89 participants almost half of the recommendations were pointing to the humorous aspects of *The Onion*. We

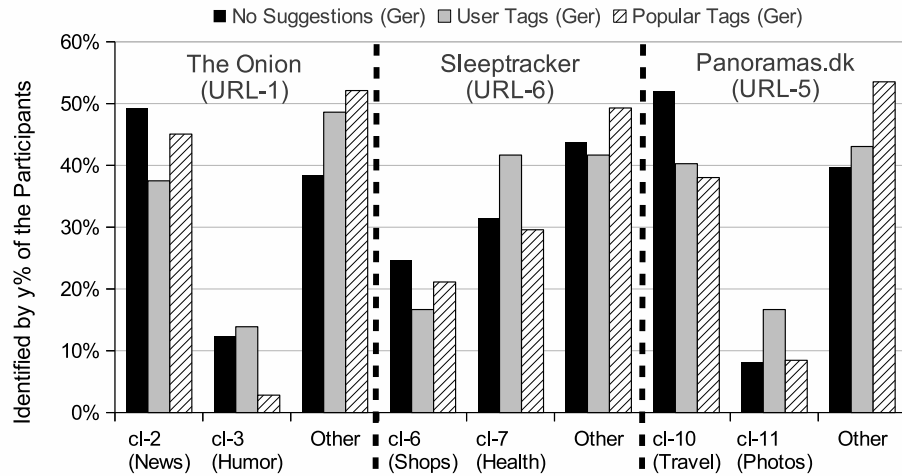


Figure 6.10: Overview of how the participants of the three German experimental conditions have clustered the three web pages from Fig. 6.6 that are on the border between two topics. For each of the three web pages the probabilities are given with which they have been put into one of the eleven most popular clusters from Fig. 6.6, or in another of the 140 overall identified clusters.

assume that seeing these tags helped to increase the likelihood of recognizing this aspect and finally clustering it with *Patently Silly* (URL-8) into cluster *cl-3*.

Why can a similar effect not be observed for the German *Popular Tags* condition? Instead, in Fig. 6.10 we can even see a decreased probability of clustering *The Onion* according to its humorous aspects under the German *Popular Tags* condition. It seems that in the German experiment variant not enough participants recognized the humorous aspects of *The Onion* in order to push corresponding tags into the list of popular tags. Indeed, under the German *Popular Tags* the list of popular tags contains only for 1 participant a tag related to the humorous aspects, namely the tag “lustig” (=funny). Consequently, no increased probability of clustering *The Onion* with *Patently Silly* into cluster *cl-3* can be observed. Quite contrary, the dominance of news related tags in the list of popular tags for the German *Popular Tags* condition even decreases the probability of cluster *cl-3* from 13% for the other two German experimental conditions to 3%. But this decrease of *cl-3*’s probability has no significant influence on the overall distribution of topical clusters (see Subsection 6.4.1).

All in all, it thus seems that suggesting the popular tags has the potential to not only influence the tag vectors but also the perception of the users. But our experiment also suggests that certain preconditions have to be fulfilled

for this learning effect to occur because in our experiment it can only be observed for a single web page. It would be subject to further research to identify these preconditions. Knowing them is required for discussing in how far the effect occurs regularly in tagging systems, or in how far *The Onion* is a rather isolated case, but this discussion is out of the scope of this thesis. Nevertheless, our methodology of measuring the inter-resource consistency has proved to be useful for spotting and isolating the learning effects, which affect the perception of the users. It helps to distinguish the learning effects from changes that only affect the tag assignments of the users.

## 6.6 Conclusions

In this chapter, we have discussed how to measure the influence of tag recommenders on the indexing quality of tagging systems. We have proposed to use the inter-resource consistency as the main target parameter to be optimized by tag recommenders because it influences the precision and recall of queries in a tagging system [127]. In contrast, improving the inter-indexer consistency should only be a secondary target of tag recommenders. We have applied our methodology for measuring the inter-resource and inter-indexer consistency for two exemplary baseline recommenders: (1) The *Popular Tags* recommender, which recommends the seven most popular tags of a resource, and (2) the *User Tags* recommender, which recommends a user his/her previously used tags.

During our user experiment with 582 participants, we have contrasted our measure of the inter-resource consistency with a measure of the inter-indexer consistency. In the literature about tagging systems, the inter-indexer consistency is often used as a measure of indexing quality. But we have shown that the inter-indexer consistency is not positively correlated with the inter-resource consistency and the indexing quality if users are influenced by tag recommendations. In case of the *Popular Tags* recommender, the recommendations have increased the inter-indexer consistency and decreased the inter-resource consistency in our experiment. In case of the *User Tags* recommender, the recommendations didn't have an influence on the inter-indexer consistency while they have increased the inter-resource consistency.

From these results of the user experiment one can conclude that the tag vectors of related resources get more dissimilar to each other if the *Popular Tags* recommender is used in a tagging system like Delicious. In contrast, the tag vectors of related resources get more similar to each other if the *User Tags* recommender is used. Thus, the *User Tags* recommender not only helps a user to better organize his/her own collection of resources but it also helps to improve the global indexing quality in a tagging system, as it is measured by the inter-resource consistency.





# Chapter 7

## Conclusions

Two central questions have guided the research reported in this thesis. The first question has been how the micro-level behavior of the individual users leads to the emergence of certain properties on the macro-level of a tagging system? With this regard, we have formulated our assumptions about the dynamic processes in tagging systems in form of the Epistemic Dynamic Model (Chapter 4). Our evaluation (Chapter 5) has shown that our model can explain the emergence of the tag frequency distribution and the sublinear vocabulary growth in tagging systems. The second question has been whether it is possible to control and selectively influence the dynamic processes in tagging systems in order to achieve a certain desired behavior? With this regard, we have used the findings from the Epistemic Dynamic Model for predicting for two tag recommenders how they influence the indexing quality in tagging systems (Chapter 6). Our predictions have been confirmed by a user experiment (also Chapter 6). In the following, we list the most important findings of this thesis in more detail.

### 7.1 The Epistemic Dynamic Model

Our Epistemic Dynamic Model is based on the assumption that the combined influence of the shared background knowledge of users and the imitation of tag recommendations are sufficient for explaining the emergence of the tag frequency distributions and the sublinear vocabulary growth in tagging systems. In this thesis, we have concentrated on these two properties of tagging systems because they are closely related to the emergence of the shared community vocabulary in a tagging system (see Chapter 3).

We have used the Epistemic Dynamic Model for evaluating our assumption about the relevance of the two influence factors and for studying how their interaction leads to the emergence of the observed macro-level properties of tagging systems. In Chapter 5, we have shown in our evaluation that the general shape of the tag frequency distribution and of the vocabulary

growth likely have their origin in the shared background knowledge of the users. In our evaluation, we have used two alternative implementations for modeling the shared background knowledge of users (see Subsection 4.1.2): One implementation based on word frequency distributions in corpora of natural language texts, and one implementation based on semantic networks. Both implementations agree in their predictions, given that they are aligned with empirical evidence about the background knowledge of users. Thus, the two implementations can be seen to be equivalent to each other.

The imitation of tag recommendations alters the general shape of the two properties in certain directions. For example, in case of recommending a set of popular tags, the imitation leads (1) to a reduced vocabulary growth speed, (2) to an increased probability of the most frequent tags, and (3) to a decreased probability of the infrequent tags. The Epistemic Dynamic Model can not only be used for predicting the direction of how the general shape is altered but also for quantifying the size of the effect, like we have done it for the vocabulary growth in Subsection 5.5.3.

## 7.2 Tag Recommendations and Indexing Quality

All in all, the findings from the evaluation of the Epistemic Dynamic Model suggest that the dynamic processes in tagging systems are primarily driven by the shared background knowledge of the users. The recommendation of tags can then be used for influencing these processes into a specific direction. In Chapter 6, we have predicted with the help of the Epistemic Dynamic Model for two exemplary tag recommenders in which direction they alter the indexing quality in tagging systems. We have expected that the recommendation of a set of popular tags decreases the indexing quality in tagging systems, and that the recommendation of a user's previously used tags increases the indexing quality. We have been able to confirm these predictions with the help of a user experiment.

Our findings with regard to the influence of recommending a set of popular tags contradict a commonly found assumption in the literature about tagging systems that such recommendations are beneficial for the quality of the tag assignments [33, 38, 64, 77, 100]. The reason for this common assumption is that several authors have shown that recommending popular tags adds a feedback mechanism between the different users. This feedback mechanism then increases the inter-indexer consistency of the tag assignments. But we have shown in Chapter 6 that one can not automatically conclude from an increased inter-indexer consistency on an increased indexing quality if the users are influenced by tag recommendations.

Instead, we have proposed to use a more direct measure of the indexing quality, namely the inter-resource consistency of the tag assignments. The inter-resource consistency measures in how far the users are successful in

linking resources that they perceive as similar to each other by indexing their common aspects with common terms. Given a match between indexing terms and query terms, which can be assumed in tagging systems [108], the inter-resource consistency influences the precision and recall of queries [127]. For our experiments in Chapter 6, we have defined a measure of the inter-resource consistency in information systems that apply the vector space model during retrieval, like it is the case for tagging systems that produce broad folksonomies. Our measure of the inter-resource consistency complements existing measures for the evaluation and comparison of tag recommendation algorithms. It moves the focus from evaluating whether tag recommendations reduce the tagging effort of users to evaluating whether tag recommendations increase the quality of the tag assignments.

### 7.3 Outlook

There are two interesting directions of future research that arise from this thesis: First, it would be interesting to further study under which conditions one can observe the learning effects that we discovered in our user experiment (see Subsection 6.5.2). Knowing the conditions would help us in even better understanding and modeling the dynamic processes in tagging systems, and in possibly uncovering further ways of how to selectively influence the behavior of the users. Second, it would be interesting to test in how far the Epistemic Dynamic Model can be used for improving spam detection algorithms in tagging systems. With the Epistemic Dynamic Model, we have a model of regular users. By spotting users who deviate from this modeled behavior, we might be able to identify a considerable amount of the spammers in a tagging system.



# Appendix A

## Software

This thesis is accompanied by a DVD that contains the software that is required for reproducing the results reported in Chapter 4 and 5. The software is available in the jar-file `dissertation-dellschaft.jar` in the main directory of the accompanying DVD. The software can be started using Java SE 7. Java can be downloaded at <http://java.oracle.com/>. In the following, a short description is given of the different Java applications that are contained in the jar-file.

### A.1 Simulation – GUI

There exists a Java application with a graphical user interface for simulating the different tagging models described in this thesis. A screenshot of the user interface is shown in Fig. A.1. The Java application can be started from the command line with the following call:

```
java -Xmx512m -jar dissertation-dellschaft.jar
```

The application can be used for simulating the different configurations of the Epistemic Model described in Chapter 4. The Natural Language Model from Subsection 4.2.3 is simulated by using the *Epistemic Model with Word Frequencies* and setting the imitation probability to  $I = 0.0$ . Furthermore, the Yule-Simon Model with Memory and the Semantic Walker Model can be simulated (see Subsection 4.3.2). The Semantic Walker Model can be used in conjunction with the Watts-Strogatz Model [119], the Erdős-Rényi Model [32], the Uncorrelated Scale-Free Network Model [29] and the Growing Network Model [106]. For reproducing the results reported in this thesis, the following steps are required:

1. Selection of the model and the stream for which to reproduce the results (*Stream to be simulated*).

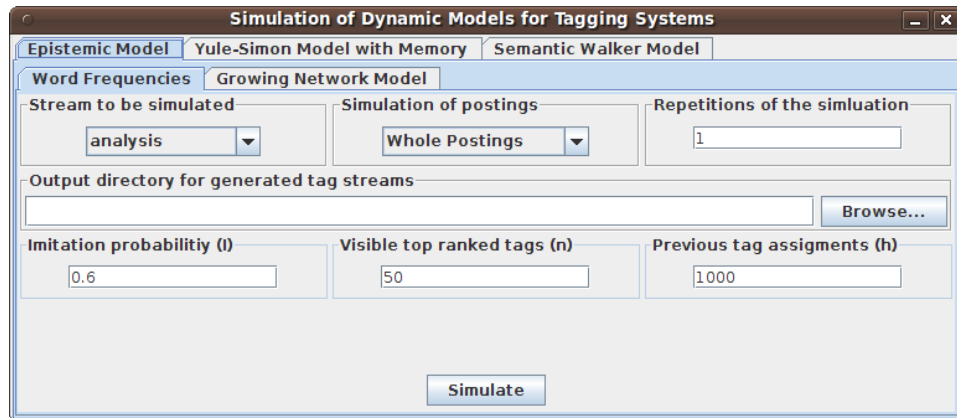


Figure A.1: Screenshot of the simulation GUI.

2. Selection of the simulation of whole postings (*Simulation of postings*).
3. Entering how often the simulations should be repeated (*Repetitions of the simulation*).
4. Setting the parameters of the model to be simulated.
5. Choosing an output directory where the simulated streams should be saved. More details about the file format used for saving the simulated streams is available in the file `README.txt` in the main directory of the accompanying DVD. Depending on the used simulation model, the conventions for the automatically generated file names are:
  - `epistemicWF_IValue_hValue_nValue_runNr.stream`  
(Epistemic Model + Word Frequencies)
  - `epistemicGNM_IValue_hValue_nValue_nsValue_dValue_mValue_runNr.stream` (Epistemic Model + Growing Network Model)
  - `ysm_pValue_tauValue_n0Value_runNr.stream`  
(Yule-Simon Model with Memory)
  - `semWalkWS_nValue_dValue_mValue_pValue_runNr.stream`  
(Semantic Walker Model + Watts-Strogatz Model)
  - `semWalkGNM_nValue_dValue_mValue_runNr.stream`  
(Semantic Walker Model + Growing Network Model)
  - `semWalkER_nValue_dValue_pValue_runNr.stream`  
(Semantic Walker Model + Erdős-Rényi Model)
  - `semWalkUCM_nValue_dValue_γValue_runNr.stream`  
(Semantic Walker Model + Uncorrelated Scale-Free Networks)

## A.2 Simulation – Command Line

The simulations can also be started on the command line. This option gives more flexibility with regard to the simulation, e.g. other distributions of posting sizes may be used than in the GUI, and streams with arbitrary length may be generated. The command line can be started with the following call:

```
java -Xmx512m -cp dissertation-dellschaft.jar
    de.unikold.isweb.TagStreamSimulator
```

More details about the available options are given on the command line if no further parameter is given to the above call.

## A.3 Generating Plots

The two applications from above for doing the actual simulations only save the raw simulated stream. If one wants to extract the tag frequency distribution or the vocabulary growth for one or more of the streams, one has to use another application. It is started with the following call:

```
java -Xmx512m -cp dissertation-dellschaft.jar
    de.unikold.isweb.TagStreamSaver
```

It opens a file chooser where one or more streams can be selected (see Fig. A.2). The extracted plots of the tag frequency distribution as well as the vocabulary growth will be saved in the same directory as the file that contains the stream. The saved files will be named *\*.edf.freq*, *\*.fr.freq* and *\*.growth*. More details about the file format used for saving the files is available in the file `README.txt` in the main directory of the accompanying DVD.

## A.4 Applying the Smirnov Test

For comparing the tag frequency distributions of two co-occurrence streams with the help of the Smirnov Test (see Subsection 5.2.1), the following application can be used:

```
java -Xmx512m -cp dissertation-dellschaft.jar
    de.unikold.isweb.StreamComparison
```

If the application is called without any further parameters, a file chooser dialog appears where one can first select the file which contains the original stream, i.e. a *\*.stream*-file. Details about the used file format are available in the file `README.txt` in the main directory of the accompanying DVD.

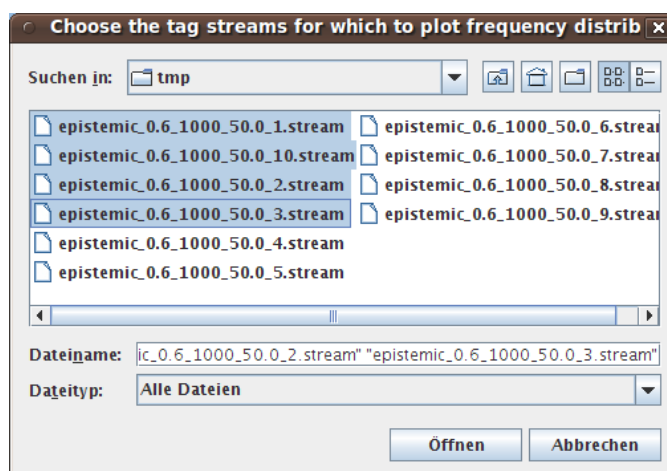


Figure A.2: Screenshot of the file dialog for extracting the frequency distribution and vocabulary growth from stream files.

Then, a second file chooser appears where one or more simulated streams can be selected, which should be compared with the original stream. There is also the option to use the application as a pure command line tool. In that case, the first parameter passed to the application is interpreted as the file name of the original stream. All further parameters are then interpreted as the file names of the simulated streams.

After selecting the *\*.stream*-files that should be compared, the application compares the stream selected with the first file chooser to each of the streams selected with the second file chooser. The result of the comparison is printed to the standard output of the command line. For each of the compared files, the maximum distance  $D$  (see Equation 5.2) between the original and the simulated tag frequency distributions is printed. Additionally, the level of significance  $p$  (see Equation 5.3) is printed.

## A.5 Source Code

The source code of all classes needed for the applications from above is contained in the jar file itself. It can be extracted with any zip utility and is contained in the directory *de/\**. For compiling the source code, the libraries *jopt-simple-2.3.2.jar*<sup>1</sup>, *commons-math-1.2.jar*<sup>2</sup> and *peersim-1.0.5.jar*<sup>3</sup> are required.

<sup>1</sup><http://jopt-simple.sourceforge.net/>

<sup>2</sup><http://commons.apache.org/math/>

<sup>3</sup><http://peersim.sourceforge.net/>



## Appendix B

# Data Sets – Co-occurrence Streams

The accompanying DVD contains the 15 co-occurrence streams and the word occurrence probabilities from the 15 crawled web corpora that are required for reproducing our evaluation reported in Chapter 5. The data is organized in the following directories on the accompanying DVD:

- **streams/\*** This directory contains all 15 filtered and unfiltered co-occurrence streams, which have been extracted from Delicious and Bibsonomy. A more detailed description of the co-occurrence streams is available in Section 3.1.
- **webcorpora/\*** This directory contains the files with the occurrence probabilities in the 15 web corpora, which have been crawled for simulating the background knowledge in the Epistemic Model (see Fig. 4.2).
- **streams/average.posting** This file contains the probabilities of observing postings with a certain size in the overall Delicious data set (see Fig. 4.1). This file is used for simulating the posting distributions in the different models described in Chapter 4.

A description of the used file formats is available in the file `README.txt` in the main directory of the accompanying DVD. In the following, we give a short description of how we acquired the Delicious data set (see Section B.1) and the Bibsonomy data set (see Section B.2) from which we extracted the co-occurrence streams. Furthermore, detailed plots of the vocabulary growth and the tag frequency distributions in the co-occurrence streams are available in Section B.3.

## B.1 Delicious

Our Delicious data set has been crawled by the TAGora consortium in 2006 from November 10 till 24. The data set is available from the homepage of the TAGora project.<sup>1</sup> The crawler was designed in such a way that the data set contains the complete history of the crawled users. However, it is not guaranteed that the history of all users in Delicious has been acquired. The crawling strategy was as follows [4]:

Prior to the actual crawling of the users' history, a central coordinating server monitored the 'recent posts page' of Delicious over a longer period. This monitoring activity resulted in a constantly updated list of user names. Then, in November 2006 small chunks of this list were distributed over several PCs for downloading the complete history of each user on the list from his/her user page in Delicious. If the respective user had tagged more than 5.000 web pages, then also the follow up pages have been crawled.

Due to the used crawling strategy, the Delicious data set is likely incomplete. For example, it doesn't contain data from users who were inactive during the monitoring period of the 'recent posts page'. But even active users may be missed if their postings already disappeared from the recent post lists prior to downloading the next snapshot of it. Nevertheless, for all users that are contained in our Delicious data set, it is guaranteed that the data set contains their complete tagging history as it was publicly available in November 2006.

The crawled data set also contains activity of spammers. The number of spammers can be estimated on a random sample from the overall data set. For this purpose, a single human evaluator manually classified a random sample of 500 users from the Delicious data set. A user has been classified as a spammer if the main purpose of his/her tag assignments seems to be the promotion of a single domain or of a collection of domains (cf. [121]). In most cases, the URLs of the corresponding domain(s) suggest a commercial background (e.g. <http://www.newyorkrealestate.realestateacme.com/>). In most cases, also a very high number of tag assignments per resource has been observed (cf. [121]).

From the random sample of 500 users, only 7 have been identified to be spammers. Thus, based on the random sample we estimate that 1.4% or 7,461 of the users are spammers. The 95% confidence interval for the number of spammers in the Delicious data set is between 1,972 and 12,949 users. This low number of spammers suggests that already a spam filtering has been applied by Delicious. The remaining spammers either passed this spam filter or the user was discovered by the Delicious crawler before being classified as spammer by the Delicious spam filter. Thus, also high-profile spammers may still be contained in the data set.

---

<sup>1</sup><http://www.tagora-project.eu/data/#delicious>

For removing the remaining spammers from the data set, it is thus not our objective to do a full fledged spam filtering of the Delicious data set, e. g. with techniques described in Section 2.3. Instead, the main objective of our spam filtering is to remove the remaining high-profile spammers, which lead to a serious disturbance of the macro-level properties of tagging systems. As shown in Section 3.2 and 3.3, especially the very large postings of high-profile spammers, not seldom containing up to 5,700 tag assignments, have a serious influence on the vocabulary growth and the tag frequency distribution of co-occurrence streams.

Based on the experience from manually classifying the random sample of 500 users, and based on the findings in [121], we applied the following three heuristics for spotting the remaining high-profile spammers in the Delicious data set:

1. We labeled all users as spammers who have (1) at least two postings with more than 20 tags, and (2) at least 1% of their tag assignments use tags from a blacklist of 12,327 spam tags. The blacklist contains tags like *sex*, *porn*, *girls* plus tags that contain these tags as substrings, e. g. *pornstar*. This heuristic marked 2,488 users as spammers. Based on a random sample from the marked users, we estimate that this heuristic achieves a precision of approximately 85%.
2. Regular users seldom assign more than 100 tags in a single posting. For example, the Bibsonomy system restricts the maximal size of a posting to 100 tag assignments. Thus, we marked all users as spammers who have at least one posting with more than 100 tag assignments. This heuristic marked 458 users as spammers and achieves a precision of approximately 85%.
3. Often, spammers create several user accounts for posting the same resource several times in order to increase its popularity [121]. Nevertheless, the tagged resources are still unpopular in the overall system. We thus marked all users as spammers who have more than 20 resources in common with another user and each of the resources has been tagged by at most 5 users. This heuristic marked 736 users as spammers and achieves a precision of approximately 80%.

Altogether, the three heuristics marked 3,340 users, i. e. some users have been spotted by more than one heuristic. Based on a random sample of 80 users from the set of 3,340 marked users, we estimate that between 2,048 and 2,711 of the marked users are really spammers. Thus, the precision of the combined heuristics is between 61% and 81%. By applying the heuristics we have been able to remove a considerable amount of the high-profile spammers in the data set while not removing too many regular users.

## B.2 Bibsonomy

The Bibsonomy data set used in this thesis is a dump of the complete history of the Bibsonomy system until June 30, 2008. The data set is publicly available. It has been provided by the owners of the Bibsonomy system for the ECML PKDD Discovery challenge.<sup>2</sup> Unlike for the Delicious data set, it is guaranteed that the data set contains *all* publicly available data from the Bibsonomy system.

Also the Bibsonomy data set contains a considerable amount of spamming activity. In the data set, for all users the information is available whether the user has been manually classified by the administrators of the Bibsonomy system as a spammer or a regular user. During the ECML PKDD Discovery Challenge, this information has been used for training and testing spam detection algorithms. According to the manual classification of the Bibsonomy administrators, 92% or 36,282 of the 38,920 users are spammers.

Originally, the Bibsonomy data set not only contains bookmarks for web pages but also bookmarks for BibTeX references. But throughout this thesis, we only use the tagging data from the bookmarks for web pages. This way, we ensure that the results achieved for Bibsonomy are better comparable to our results achieved for Delicious.

In the Bibsonomy data set, all spammers have been manually identified by the administrators of the system (see Section B.2). Thus, for the Bibsonomy data set no further spam filtering has to be applied. According to the numbers of the Bibsonomy administrators, 92% of the users in Bibsonomy are spammers and 94% of the tag assignments have been created by them.

## B.3 Detailed Plots

In the following, the detailed plots of the vocabulary growth and the tag frequency distributions are available for the 10 co-occurrence streams from Delicious and the 5 co-occurrence streams from Bibsonomy that have been used in this thesis (see Tab. 3.2 and 3.3).

---

<sup>2</sup><http://www.kde.cs.uni-kassel.de/ws/rsdc08/>

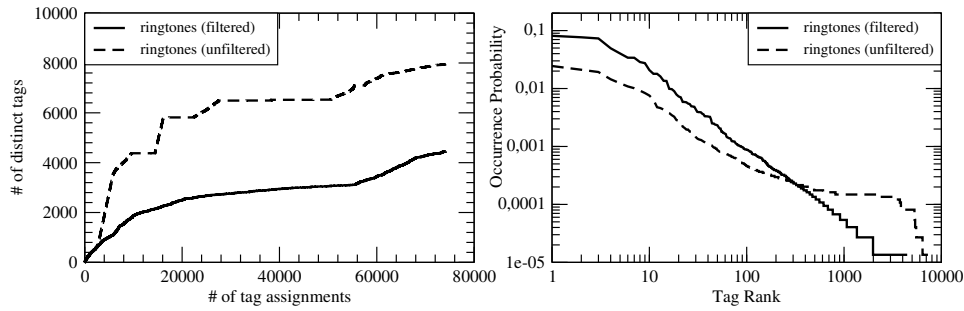


Figure B.1: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *ringtones* stream pair from Tab. 3.2 and 3.3.

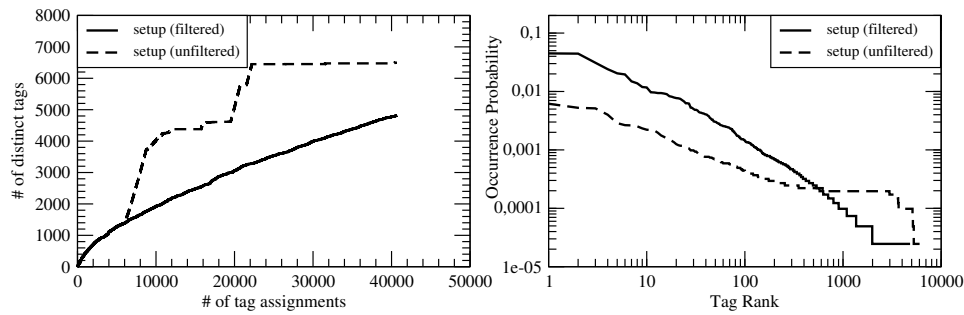


Figure B.2: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *setup* stream pair from Tab. 3.2 and 3.3.

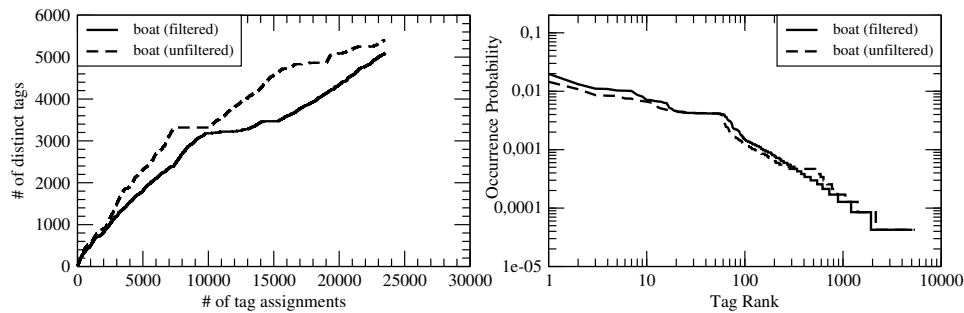


Figure B.3: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *boat* stream pair from Tab. 3.2 and 3.3.

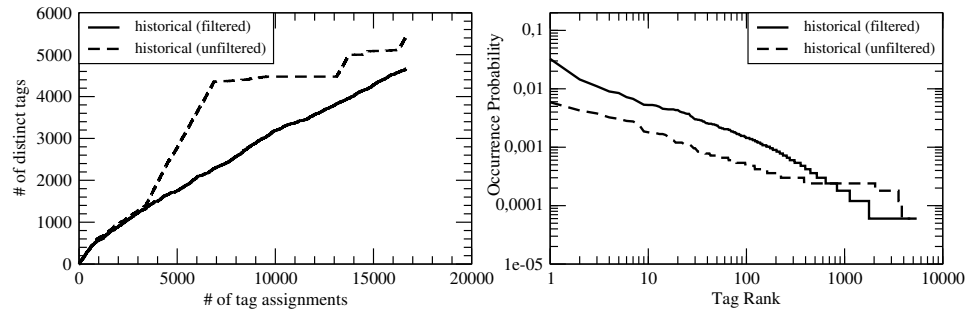


Figure B.4: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *historical* stream pair from Tab. 3.2 and 3.3.

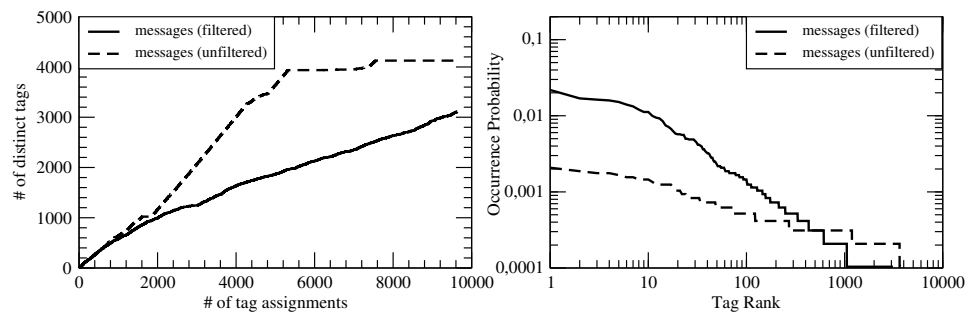


Figure B.5: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *messages* stream pair from Tab. 3.2 and 3.3.

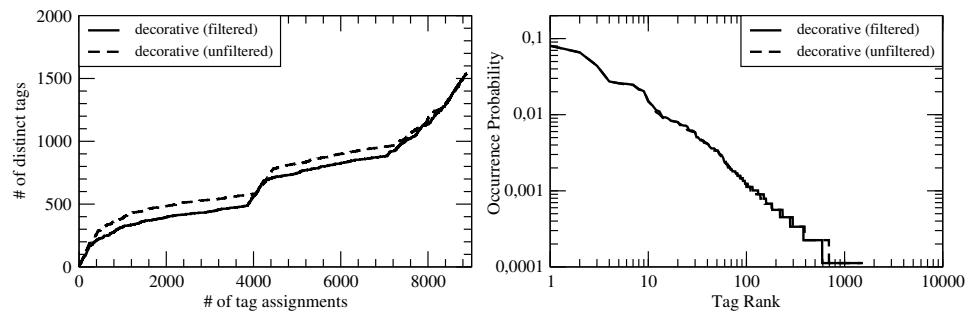


Figure B.6: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *decorative* stream pair from Tab. 3.2 and 3.3.

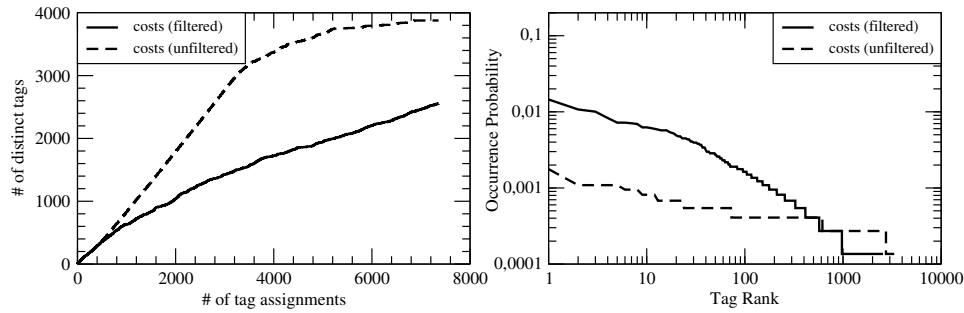


Figure B.7: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *costs* stream pair from Tab. 3.2 and 3.3.

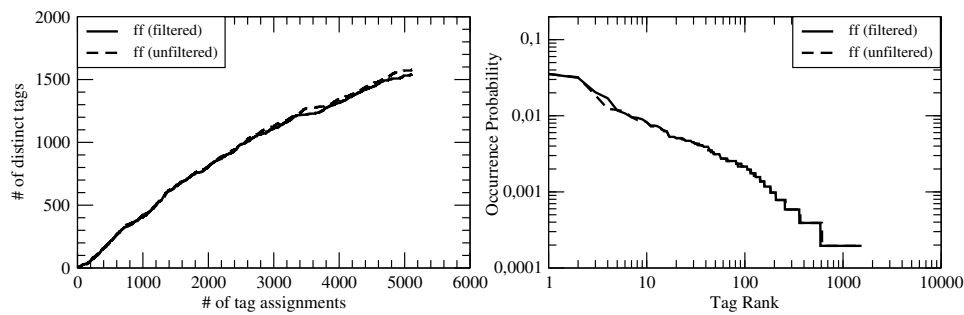


Figure B.8: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *ff* stream pair from Tab. 3.2 and 3.3.

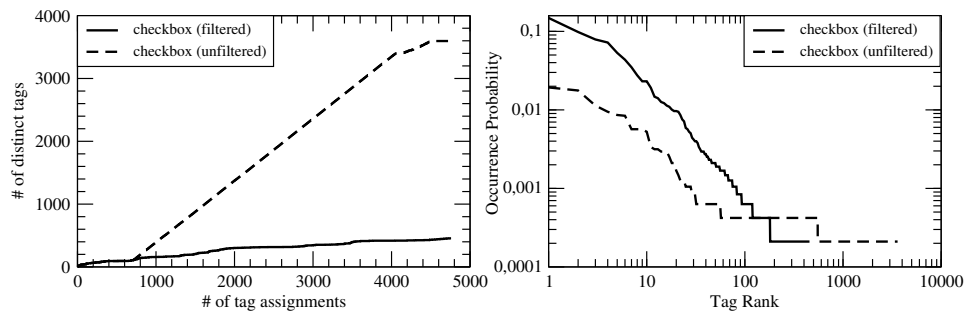


Figure B.9: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *checkbox* stream pair from Tab. 3.2 and 3.3.

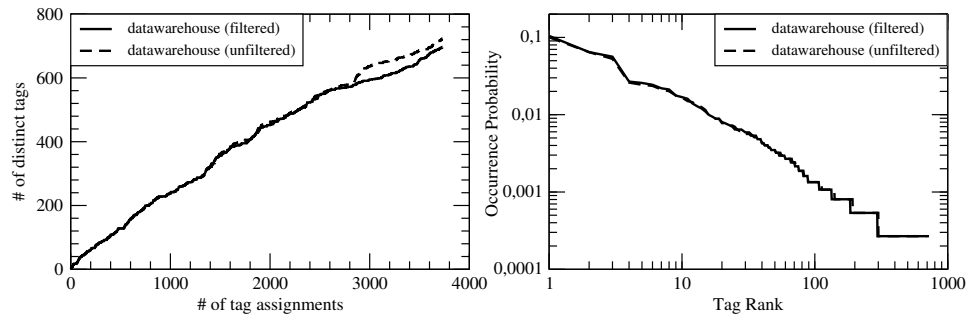


Figure B.10: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *datawarehouse* stream pair from Tab. 3.2 and 3.3.

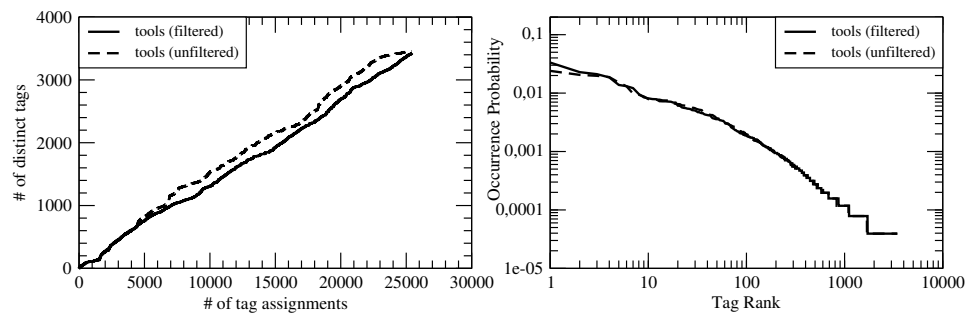


Figure B.11: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *tools* stream pair from Tab. 3.2 and 3.3.

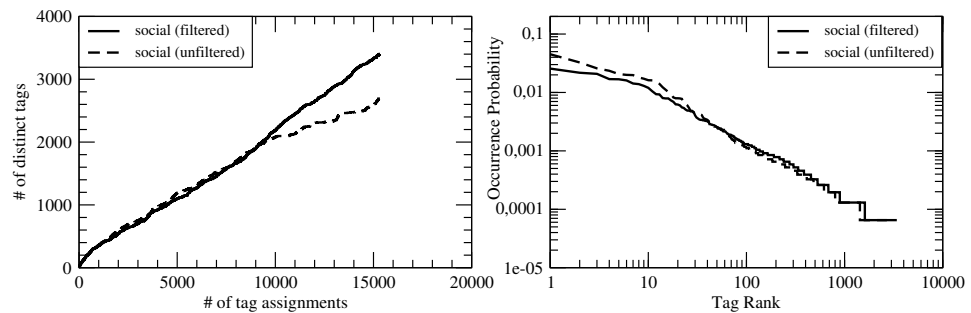


Figure B.12: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *social* stream pair from Tab. 3.2 and 3.3.



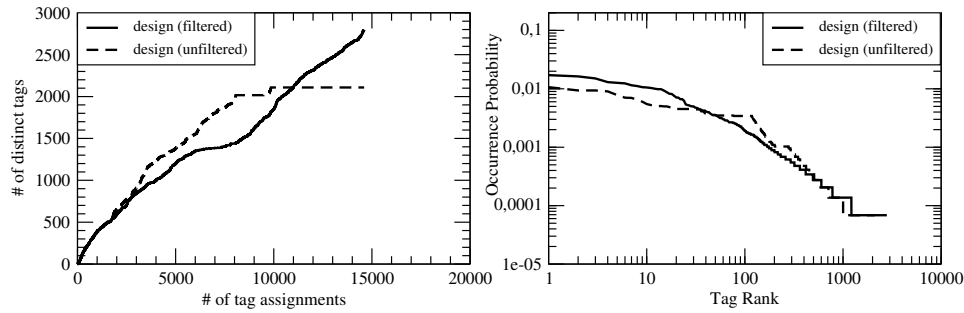


Figure B.13: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *design* stream pair from Tab. 3.2 and 3.3.

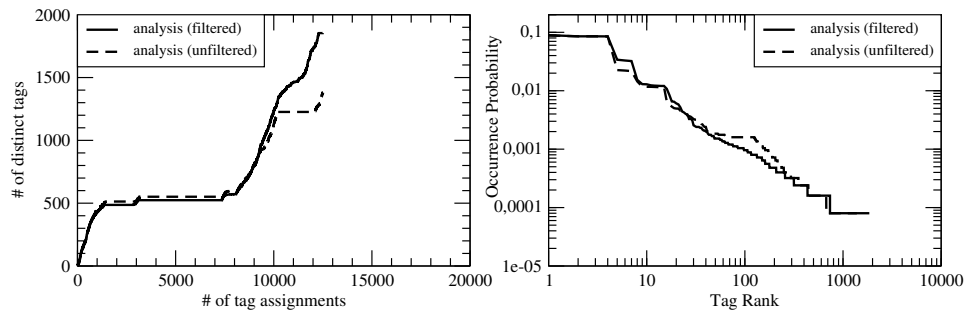


Figure B.14: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *analysis* stream pair from Tab. 3.2 and 3.3.

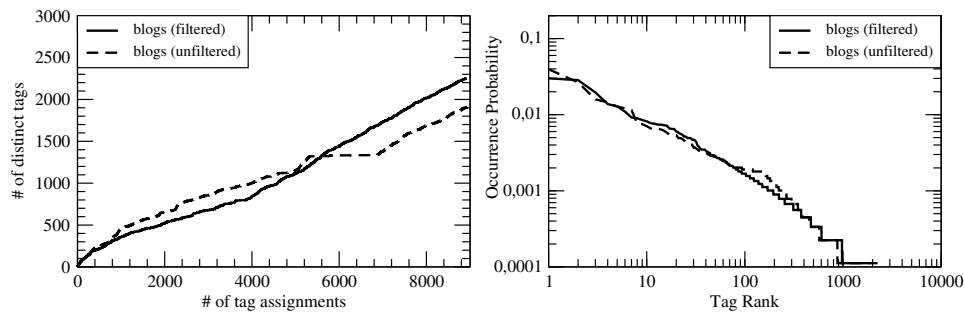


Figure B.15: Plot of the vocabulary growth (left) and Zipf plot of the occurrence probabilities of tags (right) for the *blogs* stream pair from Tab. 3.2 and 3.3.



## Appendix C

# Data Sets – User Experiment

The accompanying DVD contains the complete data set that has been collected during the user experiment described in Chapter 6. Furthermore, the DVD also contains the screenshots of the web pages from Tab. 6.1, which have been shown to the participants of the experiment. The data and the screenshots are organized in the following directories on the DVD:

- `userexperiment/data/*` This directory contains dumps of the tables in the database that has been used for collecting the data of the user experiment. Each table of the data base is dumped into a separate file. See the `README.txt` for a documentation of the files and the format of the files.
- `userexperiment/data/README.txt` A documentation of the different files in the data directory. It especially contains information about the mapping between the three phases of the user experiment (see Section 6.3) and the different database tables.
- `userexperiment/data/schema.sql` This file contains the SQL schema of the tables in the database. It especially contains information about the order of the columns in the table and which information is saved in the columns.
- `userexperiment/screenshots/*` This directory contains the screenshots of the web pages from Tab. 6.1. The file name of each screenshot starts with the ID of the web page in Tab. 6.1 followed by the domain name of the web page.

### C.1 Questionnaire

In the following, the participants' answers to the questions in the questionnaire are documented. The raw data for each of the participants is available on the accompanying DVD (see above).

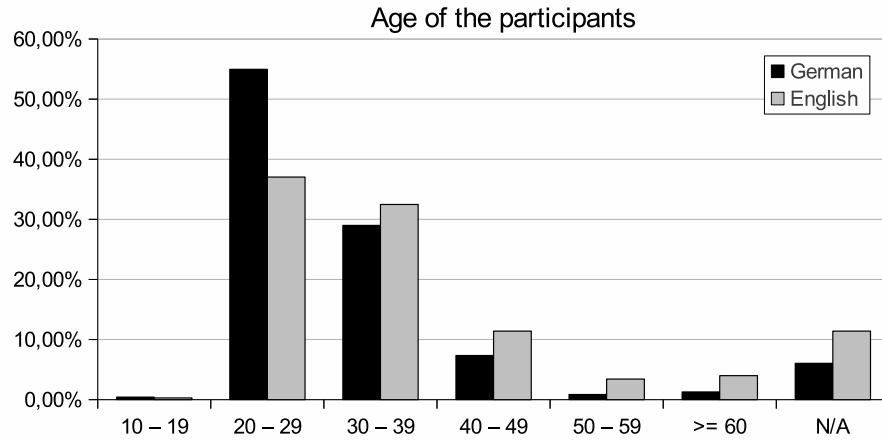


Figure C.1: English question: “Please give your age”. German question: “Bitte geben Sie Ihr Alter an”

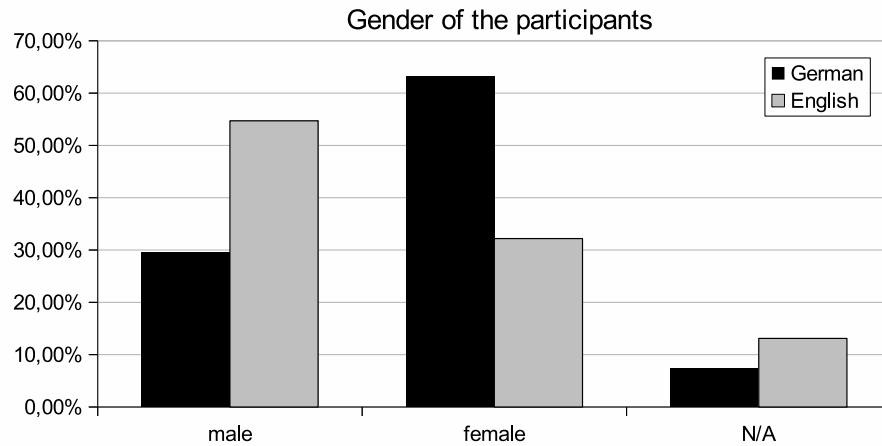


Figure C.2: English question: “Please give your gender”. German question: “Bitte geben Sie Ihr Geschlecht an”

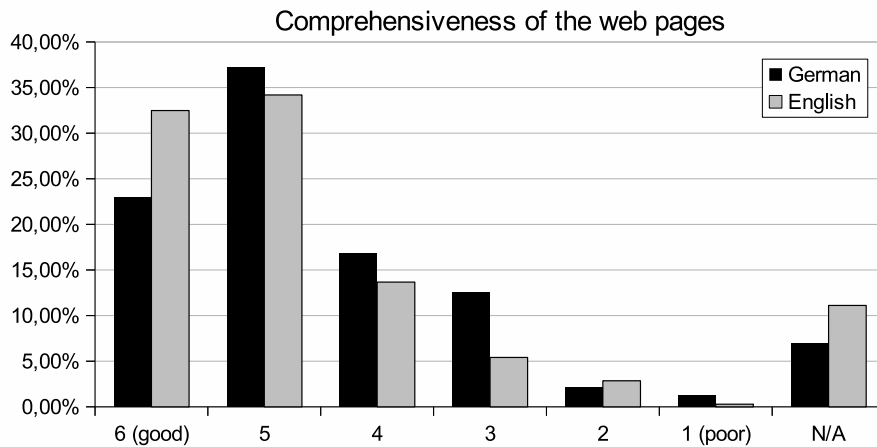


Figure C.3: English question: “How comprehensible have been the web pages?”. German question: “Wie gut war der Inhalt der Webseiten zu verstehen?”

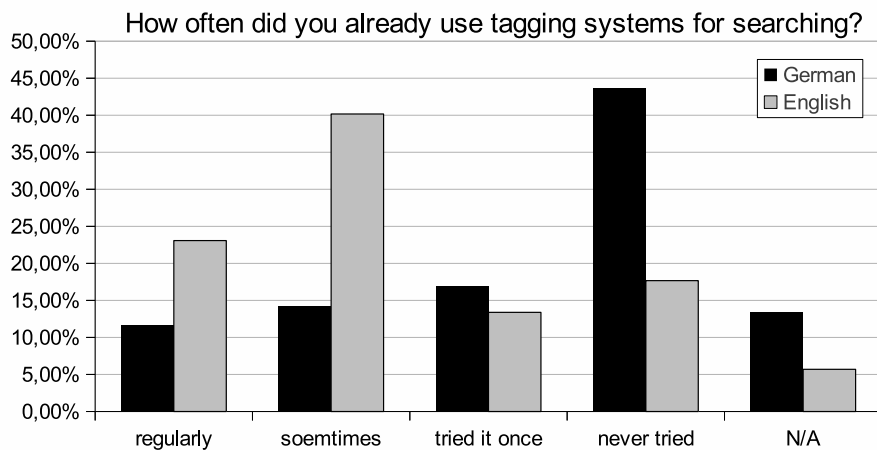


Figure C.4: English question: “How often did you already use tagging systems for searching?”. German question: “Wie oft haben Sie schon Tagging-Systeme zum Suchen benutzt?”

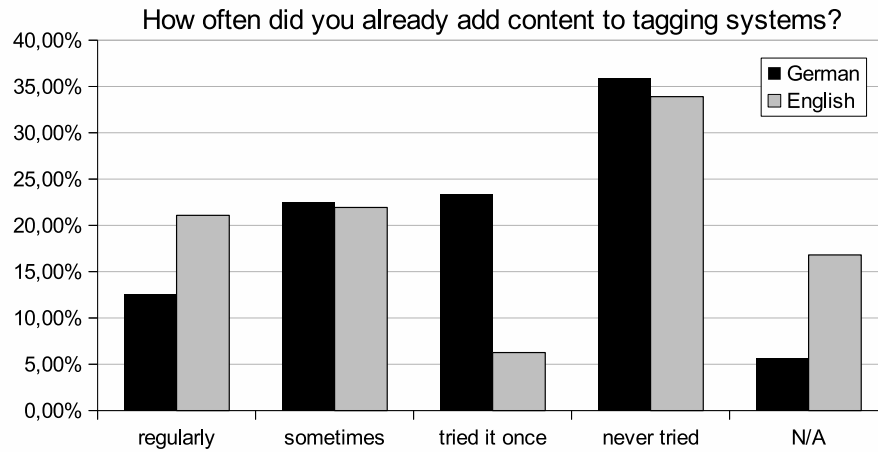


Figure C.5: English question: “How often did you already add content to tagging systems?”. German question: “Wie oft haben Sie schon Inhalte zu Tagging-Systemen hinzugefügt?”

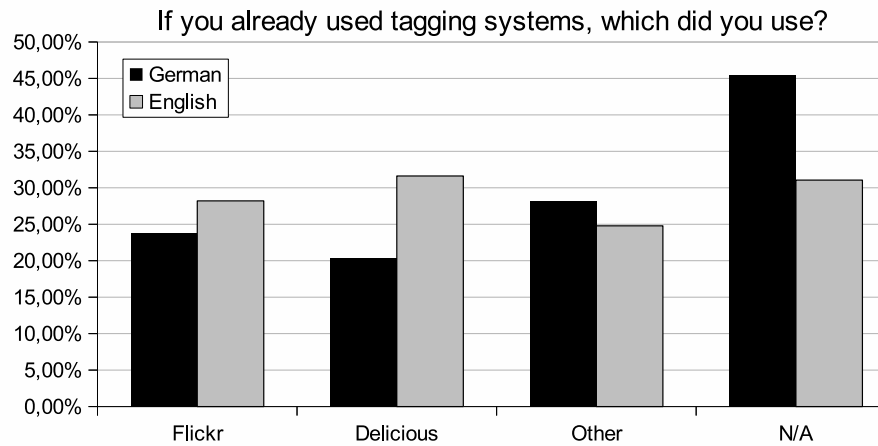


Figure C.6: English question: “If you already used tagging systems, which did you use?”. German question: “Wenn Sie schon Tagging-Systeme benutzt haben, welche waren das?”. Multiple choices possible.

# Bibliography

- [1] R. Abbasi, M. Grzegorzec, and S. Staab. Large Scale Tag Recommendation Using Different Image Representations. In T.-S. Chua, Y. Kompatsiaris, B. Mérialdo, W. Haas, G. Thallinger, and W. Bailer, editors, *Semantic Multimedia*, volume 5887 of *Lecture Notes in Computer Science*, pages 65–76. Springer Berlin / Heidelberg, 2009.
- [2] R. Abbasi and S. Staab. RichVSM: enRiched vector space models for folksonomies. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, HT '09, pages 219–228, New York, NY, USA, 2009. ACM.
- [3] R. Agrawal, T. Imieliński, and A. Swami. Mining Association Rules Between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93, pages 207–216, New York, NY, USA, 1993. ACM.
- [4] H. Alani, C. Cattuto, K. Dellschaft, O. Görlitz, M. Grahl, P. Hanappe, V. Loreto, and G. Stumme. Data Delivery from Selected Folksonomy Sites. Deliverable D1.1, TAGora Project, 2007. <http://www.tagora-project.eu/>.
- [5] R. Baeza-Yates and G. Navarro. Block Addressing Indices for Approximate Text Retrieval. In *CIKM '97: Proceedings of the 6th International Conference on Information and Knowledge Management*, pages 1–8, New York, NY, USA, 1997. ACM.
- [6] A. Baldassarri, C. Cattuto, K. Dellschaft, V. Loreto, V. Servedio, and G. Stumme. Theoretical Tools for Modeling and Analyzing Collaborative Social Tagging Systems – A Stream View. Deliverable D4.1, TAGora Project, 2007. <http://www.tagora-project.eu>.
- [7] 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*, WWW '07, pages 501–510, New York, NY, USA, 2007. ACM.

- [8] A. Baronchelli and V. Loreto. Ring structures and mean first passage time in networks. *Phys. Rev. E*, 73(2):026103, Feb 2006.
- [9] A. Barrat, M. Barthlemy, and A. Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.
- [10] S. Bateman, C. Gutwin, and M. Nacenti. Seeing Things in the Clouds: The Effect of Visual Features on Tag Cloud Selections. In *Proceedings of the 19th ACM Conference on Hypertext and Hypermedia*, HT '08, pages 193–202, New York, NY, USA, 2008. ACM.
- [11] 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 at WWW2006*, 2006.
- [12] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [13] D. Bollen and H. Halpin. An Experimental Analysis of Suggestions in Collaborative Tagging. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, pages 108–115, 2009.
- [14] 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, 1998.
- [15] I. Cantador, A. Bellogín, and D. Vallet. Content-based Recommendation in Social Tagging Systems. In *Proceedings of the 4th ACM Conference on Recommender Systems*, RecSys '10, pages 237–240, New York, NY, USA, 2010. ACM.
- [16] M. Catanzaro, M. Boguñá, and R. Pastor-Satorras. Generation of Uncorrelated Random Scale-free Networks. *Phys. Rev. E*, 71(2):027103, Feb 2005.
- [17] C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto. Vocabulary Growth in Collaborative Tagging Systems. Arxiv e-print, 2007. <http://arxiv.org/abs/0704.3316>.
- [18] C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, and V. Loreto. Collective Dynamics of Social Annotation. *Proceedings of the National Academy of Sciences*, 106(26):10511–10515, 2009.
- [19] C. Cattuto, V. Loreto, and L. Pietronero. Semiotic Dynamics and Collaborative Tagging. *Proceedings of the National Academy of Sciences*, 104(5):1461, 2007.



- [20] J.-F. Chevalier and P. Gramme. RANK for Spam Detection ECML-Discovery Challenge. In *Proceedings of ECML PKDD Discovery Challenge*, 2008.
- [21] E. H. Chi and T. Mytkowicz. Understanding the Efficiency of Social Tagging Systems Using Information Theory. In *Proceedings of the 19th ACM Conference on Hypertext and Hypermedia*, HT '08, pages 81–88, New York, NY, USA, 2008. ACM.
- [22] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law Distributions in Empirical Data. Arxiv e-print, 2007. <http://arxiv.org/abs/0706.1062v1>.
- [23] W. J. Conover. *Practical Nonparametric Statistics*. John Wiley, 3rd edition, 1999.
- [24] A. L. da Costa Carvalho, P. A. Chirita, E. S. de Moura, P. Calado, and W. Nejdl. Site Level Noise Removal for Search Engines. In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, pages 73–82, New York, NY, USA, 2006. ACM.
- [25] K. Dellschaft. Das Epistemic Model - Ein Modell zur Erklärung der Dynamik in Tagging-Systemen. In *Proceedings der 2. DGI-Konferenz der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis*, pages 263–278, 2012.
- [26] K. Dellschaft and S. Staab. An Epistemic Dynamic Model for Tagging Systems. In *Proceedings of the 19th ACM Conference on Hypertext and Hypermedia*, HT '08, pages 71–80, New York, NY, USA, 2008. ACM.
- [27] K. Dellschaft and S. Staab. On Differences in the Tagging Behavior of Spammers and Regular Users. In *Proceedings of the Web Science Conference 2010*, 2010.
- [28] K. Dellschaft and S. Staab. Measuring the Influence of Tag Recommenders on the Indexing Quality in Tagging Systems. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, HT '12, pages 73–82, New York, NY, USA, 2012. ACM.
- [29] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Size-dependent Degree Distribution of a Scale-free Growing Network. *Phys. Rev. E*, 63(6):062101, May 2001.
- [30] F. Eggenberger and G. Polya. Über die Statistik verketteter Vorgänge. *Zeitschrift für Angewandte Mathematik und Mechanik*, 1:279–289, 1923.

- [31] F. Eisterlehner, A. Hotho, and R. Jäschke. ECML/PKDD Discovery Challenge 2009. Workshop, September 2009. <http://www.kde.cs.uni-kassel.de/ws/dc09/>.
- [32] P. Erdős and A. Rényi. On Random Graphs I. *Publicationes Mathematicae*, 6:290–297, 1959.
- [33] F. Floeck, J. Putzke, S. Steinfeld, and K. Fisch. Imitation and Quality of Tags in Social Bookmarking Systems – Collective Intelligence Leading to Folksonomies. In T. Bastiaens, U. Baumöl, and B. Krämer, editors, *On Collective Intelligence*, pages 75–91. Springer Berlin / Heidelberg, 2010.
- [34] W. Fu, T. Kannampallil, and R. Kang. A Semantic Imitation Model of Social Tag Choices. In *International Conference on Computational Science and Engineering*, pages 66–73. IEEE, 2009.
- [35] W.-T. Fu and W. Dong. Collaborative Indexing and Knowledge Exploration: A Social Learning Model. *IEEE Intelligent Systems*, 27(1):39–46, 2012.
- [36] A. Gelbukh and G. Sidorov. Zipf and Heaps Laws’ Coefficients Depend on Language. In *Proceeding of the Conference on Intelligent Text Processing and Computational Linguistics*, pages 332–335, 2001.
- [37] A. Gkanogiannis and T. Kalamboukis. A Novel Supervised Learning Algorithm and Its Use for Spam Detection in Social Bookmarking Systems. In *Proceedings of the ECML PKDD Discovery Challenge*, 2008.
- [38] S. Golder and B. Huberman. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32(2):198–208, 2006. <http://www.hpl.hp.com/research/idl/papers/tags>.
- [39] M. Goldstein, S. Morris, and G. Yen. Problems with Fitting to the Power-law Distribution. *The European Physical Journal B-Condensed Matter*, 41(2):255–258, 2004.
- [40] Z. Guan, J. Bu, Q. Mei, C. Chen, and C. Wang. Personalized Tag Recommendation Using Graph-based Ranking on Multi-type Interrelated Objects. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’09*, pages 540–547, New York, NY, USA, 2009. ACM.
- [41] Z. Guan, C. Wang, J. Bu, C. Chen, K. Yang, D. Cai, and X. He. Document Recommendation in Social Tagging Services. In *Proceedings of the 19th International Conference on World Wide Web, WWW ’10*, pages 391–400, New York, NY, USA, 2010. ACM.

- [42] H. Halpin, V. Robu, and H. Shepherd. The Complex Dynamics of Collaborative Tagging. In *Proceedings of the 16th International Conference on World Wide Web, WWW '07*, pages 211–220, New York, NY, USA, 2007. ACM.
- [43] Y. Hassan-Montero and V. Herrero-Solana. Improving Tag-clouds as Visual Information Retrieval Interfaces. In *International Conference on Multidisciplinary Information Sciences and Technologies*, pages 25–28, 2006.
- [44] H. S. Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, 1978.
- [45] C. Held and U. Cress. Learning by Foraging: The Impact of Social Tags on Knowledge Acquisition. In *Proceeding of the 4<sup>th</sup> European Conference on Technology Enhanced Learning*, 2009.
- [46] D. Helic, C. Körner, M. Granitzer, M. Strohmaier, and C. Trattner. Navigational Efficiency of Broad vs. Narrow Folksonomies. In *Proceedings of the 23rd ACM conference on Hypertext and Social Media, HT '12*, pages 63–72, New York, NY, USA, 2012. ACM.
- [47] P. Heymann, G. Koutrika, and H. Garcia-Molina. Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges. *Internet Computing, IEEE*, 11(6):36–45, 2007.
- [48] 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, SIGIR '08*, pages 531–538, New York, NY, USA, 2008. ACM.
- [49] A. Hotho, D. Benz, R. Jäschke, and B. Krause. ECML/PKDD Discovery Challenge 2008. Workshop, September 2008. <http://www.kde.cs.uni-kassel.de/ws/rsdc08/>.
- [50] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information Retrieval in Folksonomies: Search and Ranking. In Y. Sure and J. Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *Lecture Notes in Computer Science*, pages 411–426. Springer Berlin/Heidelberg, 2006.
- [51] J. Illig, A. Hotho, R. Jäschke, and G. Stumme. A Comparison of Content-Based Tag Recommendations in Folksonomy Systems. In K. Wolff, D. Palchunov, N. Zagoruiko, and U. Andelfinger, editors, *Knowledge Processing and Data Analysis*, volume 6581 of *Lecture Notes in Computer Science*, pages 136–149. Springer Berlin / Heidelberg, 2011.

- [52] Y. Jin, R. Li, Y. Cai, Q. Li, A. Daud, and Y. Li. Semantic Grounding of Hybridization for Tag Recommendation. In L. Chen, C. Tang, J. Yang, and Y. Gao, editors, *Web-Age Information Management*, volume 6184 of *Lecture Notes in Computer Science*, pages 139–150. Springer Berlin / Heidelberg, 2010.
- [53] T. Joachims. Optimizing Search Engines Using Clickthrough Data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA, 2002. ACM.
- [54] S. Ju and K.-B. Hwang. A Weighting Scheme for Tag Recommendation in Social Bookmarking Systems. In *Proceedings of the ECML/PKDD Discovery Challenge*, 2009.
- [55] R. Jäschke, F. Eisterlehner, A. Hotho, and G. Stumme. Testing and Evaluating Tag Recommenders in a Live System. In *Proceedings of the 3rd ACM Conference on Recommender Systems*, RecSys '09, pages 369–372, New York, NY, USA, 2009. ACM.
- [56] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag Recommendations in Folksonomies. In *Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 506–514, 2007.
- [57] T. G. Kannampallil and W.-T. Fu. Trail Patterns in Social Tagging Systems: Role of Tags as Digital Pheromones. In *Proceedings of the International Conference of Human-Computer Interaction*, 2009.
- [58] I. Katakis, G. Tsoumakos, and I. Vlahavas. Multilabel Text Classification for Automated Tag Suggestion. In *Proceedings of the ECML/PKDD Discovery Challenge*, 2008.
- [59] C. Kim and K.-B. Hwang. Naive Bayes Classifier Learning with Feature Selection for Spam Detection in Social Bookmarking. In *Proceedings of ECML PKDD Discovery Challenge*, 2008.
- [60] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.
- [61] I. Konstas, V. Stathopoulos, and J. M. Jose. On Social Networks and Collaborative Recommendation. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 195–202, New York, NY, USA, 2009. ACM.

- [62] 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*, HT '10, pages 157–166, New York, NY, USA, 2010. ACM.
- [63] G. Koutrika, F. A. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina. Combating Spam in Tagging Systems: An Evaluation. *ACM Transactions on the Web*, 2(4):1–34, 2008.
- [64] T. Kowatsch and W. Maass. The Impact of Pre-Defined Terms on the Vocabulary of Collaborative Indexing Systems. In *Proceedings of the 16th European Conference on Information Systems (ECIS)*, 2008.
- [65] B. Krause, 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*, 2008.
- [66] R. Krestel, P. Fankhauser, and W. Nejdl. Latent Dirichlet Allocation for Tag Recommendation. In *Proceedings of the 3rd ACM Conference on Recommender Systems*, RecSys '09, pages 61–68, New York, NY, USA, 2009. ACM.
- [67] R. Li, S. Bao, Y. Yu, B. Fei, and Z. Su. Towards Effective Browsing of Large Scale Social Annotations. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 943–952, New York, NY, USA, 2007. ACM.
- [68] Y. Lin, J.-W. Ahn, P. Brusilovsky, D. He, and W. Real. Imagesieve: Exploratory Search of Museum Archives with Named Entity-based Faceted Browsing. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–10, 2010.
- [69] M. Lipczak. Tag Recommendations for Folksonomies Oriented towards Individual Users. In *Proceedings of the ECML/PKDD Discovery Challenge*, 2008.
- [70] M. Lipczak, Y. Hu, Y. Kollet, and E. Milios. Tag Sources for Recommendation in Collaborative Tagging Systems. In *Proceedings of the ECML PKDD Discovery Challenge*, 2009.
- [71] M. Lipczak and E. Milios. The Impact of Resource Title on Tags in Collaborative Tagging Systems. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, HT '10, pages 179–188, New York, NY, USA, 2010. ACM.

- [72] S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration. In T. Gross, J. Gulliksen, P. Kotzé, L. Oestreicher, P. Palanque, R. Prates, and M. Winckler, editors, *Human-Computer Interaction – INTERACT 2009*, volume 5726 of *Lecture Notes in Computer Science*, pages 392–404. Springer Berlin / Heidelberg, 2009.
- [73] B. Mandelbrot. An Informational Theory of the Statistical Structure of Language. In W. Jackson, editor, *Communication Theory – Papers read at the Symposium on Applications of Communication Theory*, pages 486–502. Butterworths, 1953.
- [74] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [75] G. Marchionini. Exploratory Search: From Finding to Understanding. *Communications of the ACM*, 49(4):41–46, Apr. 2006.
- [76] L. Marinho, C. Preisach, and L. Schmidt-Thieme. Relational Classification for Personalized Tag Recommendation. In *Proceedings of the ECML/PKDD Discovery Challenge*, 2009.
- [77] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the 17th Conference on Hypertext and Hypermedia*, HT '06, pages 31–40, New York, NY, USA, 2006. ACM.
- [78] A. Mathes. Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Website, December 2004. <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>.
- [79] D. G. Mayo and D. R. Cox. Frequentist Statistics as a Theory of Inductive Inference. *Lecture Notes-Monograph Series*, 49:77–97, 2006.
- [80] P. Mika. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In Y. Gil, E. Motta, V. Benjamins, and M. Musen, editors, *The Semantic Web – ISWC 2005*, volume 3729 of *Lecture Notes in Computer Science*, pages 522–536. Springer Berlin / Heidelberg, 2005.
- [81] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber. The University of South Florida Word Association, Rhyme, and Word Fragment Norms. Web Resource, 1998. <http://w3.usf.edu/FreeAssociation/>.
- [82] N. Neubauer and K. Obermayer. Hyperincident Connected Components of Tagging Networks. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, HT '09, pages 229–238, New York, NY, USA, 2009. ACM.

- [83] M. G. Noll, C.-m. Au Yeung, N. Gibbins, C. Meinel, and N. Shadbolt. Telling Experts from Spammers: Expertise Ranking in Folksonomies. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 612–619, New York, NY, USA, 2009. ACM.
- [84] R. Pastor-Satorras and A. Vespignani. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, 2004.
- [85] J. Peng, D. D. Zeng, and Z. Huang. Latent Subject-centered Modeling of Collaborative Tagging: An Application in Social Search. *ACM Trans. Manage. Inf. Syst.*, 2(3):15:1–15:23, Oct. 2008.
- [86] K. Popper. Conjectural Knowledge: My Solution of the Problem of Induction. In *Objective Knowledge – An Evolutionary Approach*, pages 1–31. The Clarendon Press, 1972.
- [87] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*, chapter Statistical Description of Data, pages 609–655. Cambridge University Press, 2nd edition, 1992.
- [88] E. Rader and R. Wash. Influences on Tag Choices in Delicio.us. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, pages 239–248, New York, NY, USA, 2008. ACM.
- [89] S. Rendle, L. Balby Marinho, A. Nanopoulos, and L. Schmidt-Thieme. Learning Optimal Ranking with Tensor Factorization for Tag Recommendation. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 727–736, New York, NY, USA, 2009. ACM.
- [90] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen. Getting our Head in the Clouds: Toward Evaluation Studies of Tag-clouds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 995–998, New York, NY, USA, 2007. ACM.
- [91] P. J. Rousseeuw. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20(0):53 – 65, 1987.
- [92] E. Santos-Neto, D. Condon, N. Andrade, A. Iamnitchi, and M. Rippeanu. Individual and Social Behavior in Tagging Systems. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, HT '09, pages 183–192, New York, NY, USA, 2009. ACM.

- [93] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the 10th International Conference on World Wide Web, WWW '01*, pages 285–295, New York, NY, USA, 2001. ACM.
- [94] R. Schenkel, T. Crecelius, M. Kacimi, T. Neumann, J. Parreira, M. Spaniol, and G. Weikum. Social Wisdom for Search and Recommendation. *IEEE Data Engineering Bulletin*, 31(2):40–49, 2008.
- [95] C. Schmitz, M. Grahl, A. Hotho, G. Stumme, C. Cattuto, A. Baldassarri, V. Loreto, and V. D. P. Servedio. Network Properties of Folksonomies. In *Proceedings of the Tagging and Metadata for Social Information Organization workshop held in conjunction with WWW2007*, 2007.
- [96] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. Mining Association Rules in Folksonomies. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Žiberna, editors, *Data Science and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 261–270. Springer Berlin Heidelberg, 2006.
- [97] J. Schrammel, M. Leitner, and M. Tscheligi. Semantically Structured Tag Clouds: An Empirical Evaluation of Clustered Presentation Approaches. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI '09*, pages 2037–2040, New York, NY, USA, 2009. ACM.
- [98] P. Seitlinger and T. Ley. Implicit and Explicit Memory in Social Tagging: Evidence from a Process Dissociation Procedure. In *Proceedings of the European Conference on Cognitive Ergonomics*, 2011.
- [99] P. Seitlinger and T. Ley. Implicit Imitation in Social Tagging: Familiarity and Semantic Reconstruction. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems, CHI '12*, pages 1631–1640, New York, NY, USA, 2012. ACM.
- [100] S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl. tagging, communities, vocabulary, evolution. In *Proceedings of the 20th Anniversary Conference on Computer Supported Cooperative Work, CSCW '06*, pages 181–190, New York, NY, USA, 2006. ACM.
- [101] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering. In *Proceedings of the ACM Conference on Recommender Systems, RecSys '08*, pages 259–266, New York, NY, USA, 2008. ACM.



- [102] B. Sigurbjörnsson and R. van Zwol. Flickr Tag Recommendation Based on Collective Knowledge. In *Proceedings of the 17th International Conference on World Wide Web, WWW '08*, pages 327–336, New York, NY, USA, 2008. ACM.
- [103] H. Simon. On a Class of Skew Distribution Functions. *Biometrika*, 42(3/4):425–440, 1955.
- [104] J. Sinclair and M. Cardew-Hall. The Folksonomy Tag Cloud: When is it Useful? *Journal of Information Science*, 34(1):15–29, 2008.
- [105] L. Steels. Semiotic Dynamics for Embodied Agents. *IEEE Intelligent Systems*, 21(3):32–38, 2006.
- [106] M. Steyvers and J. B. Tenenbaum. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, 29(1):41–78, 2005.
- [107] M. Strohmaier, C. Körner, and R. Kern. Why do Users Tag? Detecting Users' Motivation for Tagging in Social Tagging Systems. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010.
- [108] F. M. Suchanek, M. Vojnovic, and D. Gunawardena. Social Tags: Meaning and Suggestions. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08*, pages 223–232, New York, NY, USA, 2008. ACM.
- [109] M. Tatu, M. Srikanth, and T. D'Silva. RSDC'08: Tag Recommendations using Bookmark Content. In *Proceedings of the ECML PKDD Discovery Challenge*, 2008.
- [110] E. Tisselli. thinkflickrthink: A Case Study on Strategic Tagging. *Communications of the ACM*, 53(8):141–145, 2010.
- [111] C. Trattner, Y.-l. Lin, D. Parra, Z. Yue, W. Real, and P. Brusilovsky. Evaluating Tag-based Information Access in Image Collections. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media, HT '12*, pages 113–122, New York, NY, USA, 2012. ACM.
- [112] K. H. L. Tso-Sutter, L. B. Marinho, and L. Schmidt-Thieme. Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms. In *Proceedings of the 2008 ACM Symposium on Applied Computing, SAC '08*, pages 1995–1999, New York, NY, USA, 2008. ACM.
- [113] D. van Leijenhorst and T. van der Weide. A Formal Derivation of Heaps' Law. *Information Sciences*, 170(2–4):263 – 272, 2005.

- [114] C. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.
- [115] T. Vander Wal. Explaining and Showing Broad and Narrow Folksonomies. Website, February 21 2005. [http://www.personalinfocloud.com/2005/02/explaining\\_and\\_.html](http://www.personalinfocloud.com/2005/02/explaining_and_.html).
- [116] M. Vojnovic, J. Cruise, D. Gunawardena, and P. Marbach. Ranking and Suggesting Popular Items. *IEEE Transactions on Knowledge and Data Engineering*, 21(8):1133–1146, 2009.
- [117] L. von Ahn, M. Blum, N. Hopper, and J. Langford. CAPTCHA: Using Hard AI Problems for Security. In E. Biham, editor, *Advances in Cryptology — EUROCRYPT 2003*, volume 2656 of *Lecture Notes in Computer Science*, pages 646–646. Springer Berlin / Heidelberg, 2003.
- [118] Y. Wang, E. Zhai, C. Cao, Y. Xie, Z. Wang, J. Hu, and Z. Chen. DSpam: Defending Against Spam in Tagging Systems via Users’ Reliability. In *IEEE 16th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 139–146, 2010.
- [119] D. J. Watts and S. H. Strogatz. Collective Dynamics of ‘Small-’World’ Networks. *Nature*, 393(6684):440–442, 1998.
- [120] R. Wetzker, W. Umbrath, and A. Said. A Hybrid Approach to Item Recommendation in Folksonomies. In *Proceedings of the WSDM ’09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR ’09, pages 25–29, New York, NY, USA, 2009. ACM.
- [121] R. Wetzker, C. Zimmermann, and C. Bauckhage. Analyzing Social Bookmarking Systems: A del.icio.us Cookbook. In *Mining Social Data (MSoDa) Workshop Proceedings*, pages 26–30, 2008.
- [122] H. White and B. Griffith. Quality of Indexing in Online Data Bases. *Information Processing & Management*, 23(3):211–224, 1987.
- [123] X. Wu, L. Zhang, and Y. Yu. Exploring Social Annotations for the Semantic Web. In *Proceedings of the 15th International Conference on World Wide Web*, WWW ’06, pages 417–426, New York, NY, USA, 2006. ACM.
- [124] D. Yin, Z. Xue, L. Hong, and B. D. Davison. A Probabilistic Model for Personalized Tag Prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’10, pages 959–968, New York, NY, USA, 2010. ACM.

- [125] N. Zhang, Y. Zhang, and J. Tang. A Tag Recommendation System Based on Contents. In *Proceedings of the ECML/PKDD Discovery Challenge*, pages 285–295, 2009.
- [126] G. Zipf. *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley, 1949.
- [127] P. Zunde and M. Dexter. Indexing Consistency and Quality. *American Documentation*, 20:259–267, 1969.

# Lebenslauf

Dipl.-Inform. Klaas Dellschaft

## Kontaktinformation

**Anschrift** Institut für Web Science and Technologies  
Fachbereich 4: Informatik  
Universität Koblenz-Landau  
Postfach 201 602  
56016 Koblenz

**Telefon** +49 261 287-2444

**e-Mail** [klaasd@uni-koblenz.de](mailto:klaasd@uni-koblenz.de)

## Werdegang

seit 04/2006	Wissenschaftlicher Mitarbeiter am Institut für Web Science and Technologies an der Universität Koblenz-Landau
01/2006 – 03/2006	Wissenschaftlicher Mitarbeiter am Institut für Wissensmedien an der Universität Koblenz-Landau
10/2000 – 12/2005	Studium der Computervisualistik an der Universität Koblenz-Landau. Abschluss: Diplom-Informatiker
04/2003 – 12/2005	Wissenschaftliche Hilfskraft am Institut für Wissensmedien an der Universität Koblenz-Landau
12/2001 – 03/2003	Wissenschaftliche Hilfskraft in der Arbeitsgruppe für Künstliche Intelligenz an der Universität Koblenz-Landau
07/1999 – 06/2000	Zivildienst
1999	Abitur am St.-Ursula-Gymnasium, Düsseldorf

## Lehre

WS11/12	Übung zu Algorithmen und Datenstrukturen, Universität Koblenz-Landau
WS10/11	Übung zu Algorithmen und Datenstrukturen, Universität Koblenz-Landau
WS09/10	Übung zu Algorithmen und Datenstrukturen, Universität Koblenz-Landau
SS09	Übung zu Information Retrieval, Universität Koblenz-Landau
WS08/09	Seminar <i>Interaktive Soziale Medien</i> , Universität Koblenz-Landau
SS08	Projektpraktikum <i>MyTag 2.0</i> , Universität Koblenz-Landau
SS07	Seminar <i>Herkunft und Vertrauenswürdigkeit von Informationen</i> , Universität Koblenz-Landau
SS07	Projektpraktikum <i>MyTag</i> , Universität Koblenz-Landau
WS06/07	Seminar <i>Analyse komplexer Informationssysteme</i> , Universität Koblenz-Landau

## Publikationen

### Konferenzbeiträge

- K. Dellschaft. Das Epistemic Model – Ein Modell zur Erklärung der Dynamik in Tagging Systemen. In *Proceedings der 2. DGI-Konferenz der Deutschen Gesellschaft für Informationswissenschaften und Informationspraxis*, S. 263–278, Düsseldorf, Deutschland, 2012
- K. Dellschaft und S. Staab. Measuring the Influence of Tag Recommendations on the Indexing Quality in Tagging Systems. In: *Proceedings of the 23rd ACM Conference on Hypertext and Hypermedia*, S. 73–82, Milwaukee, USA, 2012
- T. Franz, K. Dellschaft und S. Staab. Unlock your Data: The Case of MyTag. In: *Proceedings of the Future Internet Symposium*, Wien, Österreich, 2008
- K. Dellschaft und S. Staab. An Epistemic Dynamic Model for Tagging Systems. In: *Proceedings of the 19th ACM Conference on Hypertext*

and *Hypermedia*, S. 71–80, Pittsburgh, USA, 2008

- K. Dellschaft und S. Staab. On how to Perform a Gold Standard Based Evaluation of Ontology Learning. In: *Proceedings of the 5th International Semantic Web Conference*, S. 228–241, Athens, USA, 2006

### Buchbeiträge

- K. Dellschaft und S. Staab. Strategies for the Evaluation of Ontology Learning. In: P. Buitelaar und P. Cimiano: *Briding the Gap between Text and Knowledge – Selected Contributions to Ontology Learning and Population from Text*. S. 253–272, IOS Press, Amsterdam, 2008

### Poster und Demos

- K. Dellschaft und S. Staab. On Differences in the Tagging Behavior of Spammers and Regular Users. In: *Proceedings of the 2nd Web Science Conference*, Raleigh, USA, 2010
- K. Dellschaft und S. Staab. Understanding the Dynamics in Tagging Systems. In: *Proceedings of the European Future Technologies Conference*, Prag, Tschechien, 2009
- K. Dellschaft, Q. Ji und G. Qi. CoDR: A Contextual Framework for Diagnosis and Repair. In: *Proceedings of the 9th International Semantic Web Conference*, Washington D. C., USA, 2009
- K. Dellschaft, O. Görlitz und M. Szomszor. Sense Aware Searching and Exploration with MyTag. In: *Proceedings of the 9th International Semantic Web Conference*, Washington D. C., USA, 2009
- M. Braun, K. Dellschaft, T. Franz, D. Heering, P. Jungen, H. Metzler, E. Müller, A. Rostilov und C. Saathoff. Personalized Search and Exploration with MyTag. In: *Proceedings of the 17th International World Wide Web Conference*, S. 1031–1032, Peking, China, 2008
- K. Dellschaft, H. Engelbrecht, J. Monte Barreto, S. Rutenbeck und S. Staab. Cicero: Tracking Design Rationale in Collaborative Ontology Engineering. In: *Proceedings of the 5th European Semantic Web Conference*, S. 782–786, Teneriffa, Spanien, 2008

### Arbeitsberichte

- K. Dellschaft und S. Staab. Unterstützung und Dokumentation kollaborativer Entwurfs- und Entscheidungsprozesse. Arbeitsberichte aus dem Fachbereich Informatik 4/2008, Universität Koblenz-Landau, 2008

## **Auszeichnungen**

- Ted Nelson Newcomer Award, 19th ACM Conference on Hypertext and Hypermedia, 2008

## **Eingeladene Vorträge**

- Vortrag "Kollaboratives Tagging – Wie Schlagwortvorschläge die Erschließungsqualität beeinflussen", Universität Düsseldorf, Düsseldorf, Deutschland
- Vortrag "Eigene Erweiterungen für die SKOS-Spezifikation", Kickoff-Workshop der DINI AG KIM, Mannheim, Deutschland, 2011
- Vortrag "Cicero: Unterstützung von kollaborativen Entwurfs- und Entscheidungsprozessen im Semantic Media Wiki", Corporate Wiki Infotag im Rahmen der XInnovations, Berlin, Deutschland, 2008

## **Review-Tätigkeiten**

### **Journals**

- Journal of Web Semantics, 2011-2012
- International Journal of Human-Computer Studies, 2010

### **Konferenzen**

- 23rd ACM Conference on Hypertext and Social Media (Hypertext), 2012
- 18th International Conference on Knowledge Engineering and Knowledge Management (EKAW), 2012
- 3rd ACM Web Science Conference (WebSci), 2011
- 9th International Semantic Web Conference (ISWC), 2009

### **Sonstige Review-Tätigkeiten**

- 11th International Semantic Web Conference (Poster- und Demotrack), 2012
- 10th International Semantic Web Conference (Poster- und Demotrack), 2011

- 10th International Semantic Web Conference (Poster- und Demotrack), 2010
- 16th International Conference on Knowledge Engineering and Knowledge Management (Poster- und Demotrack), 2010
- 7th Extended Semantic Web Conference (Poster- und Demotrack), 2010
- 6th European Semantic Web Conference (Poster- und Demotrack), 2009
- 5th Workshop on Semantic Web Applications and Perspectives, 2008
- Workshop "Knowledge Reuse and Reengineering over the Semantic Web", 2008
- Workshop "Evaluation of Ontologies and Ontology-based Tools", 2007
- 2nd Workshop on Building and Applying Ontologies for the Semantic Web, 2007