Project no. 34721

TAGora

Semiotic Dynamics in Online Social Communities

http://www.tagora-project.eu

Sixth Framework Programme (FP6)

Future and Emerging Technologies of the Information Society Technologies (IST-FET Priority)

D4.6: Report describing the results of the control experiments performed

Period covered: from 01/06/2008 to 31/08/2009
Start date of project: June 1st, 2006
Due date of deliverable: October 15th, 2009
Distribution: Public

Date of preparation: 15/09/2009
Duration: 39 months
Actual submission date: September 15th, 2009
Status: Final

Project coordinator: Vittorio Loreto
Project coordinator organisation name: PHYS-SAPIENZA
Lead contractor for these deliverables: PHYS-SAPIENZA

2009 © Copyright lies with the respective authors and their institutions

Contents

1 "Phenotypes / Limited Forms" & Ikoru 4

1.1 The installation 4

1.2 Goal of analysis 5

1.3 Preliminary analytical results 5

1.3.1 The dataset 5

1.3.2 Distribution pattern of users, tags and resources 5

1.3.3 Features of the three-mode network 7

1.3.4 Preliminary conclusions 8

1.4 Outlook 8

2 Semantic Similarity and Recommendations 10

2.1 Measuring the Semantic Similarity of Tags 10

2.1.1 Comparison of measures for tag similarity 10

2.1.2 Focus Group at Dagstuhl seminar. 13

2.2 Tag Recommendations 14

2.2.1 Collaborative Filtering 14

2.2.2 Content-Based Recommendations 15

2.3 User Recommendations 15

2.3.1 Tag-based user recommendation 16

2.3.2 Graph-based user recommendation 16

3 Analysing and Influencing User Behavior 18

3.1 Analyzing User Behavior 18

3.2 Spam Detection 19

3.2.1 Direct Computation of User Features 20

3.2.2 Model-based Computation of User Features 21

List of Figures

1.1 Visitors interacting with the “Phenotypes / Limited Forms” installation 4
1.2 TAS distribution of the Armin Linke dataset 6
1.3 TAS distribution Delicious Data (Hotho et al., 2006b) 6

3.1 Delicious: Growth of the number of distinct tags before removing spam postings. . . 22
3.2 Delicious: Tag frequencies for the *Phone* stream with and without the spam postings. 23
3.3 Bibsonomy: Tag frequencies for the *Software* stream with and without spam postings. 23
3.4 Scatter plot of the values for the *Sum of Distances* and *Tendency Ratio* features. . . 26
3.5 Scatter plot of the values for the *Sum of Distances* and *Average of Distances* features. 27
3.6 Scatter plot of the values for the *Sum of Distances* and *Standard Deviation of Distances* features. 27
3.7 Scatter plot of the values for the *Sum of Distances* and *Tag-count Ratio* features. . . 28
3.8 ROC curve for the training data set. 29
3.9 ROC curve for the test data set. 30

2009 © Copyright lies with the respective authors and their institutions

Chapter 1

“Phenotypes / Limited Forms” & Ikoru

“Phenotypes / Limited Forms” is an art installation by artist and photographer Armin Linke[1] in collaboration with Peter Hanappe of SONY-CSL. The installation was exhibited at several locations in Europe and beyond. Ikoru, the tag-based navigation application for images and music developed at SONY-CSL, was an important component of this work of art.

In this chapter we will discuss the preliminary analysis of the tagging data that was collected using the installation. The data analysis was conducted using technologies developed within the scope of the TAGora project.

1.1 The installation

Armin Linke’s “Phenotypes / Limited Forms” installation, shown by the photo in figure 1.1, has been on display in 2008 at the Bienal de São Paulo in Brazil and the “Selective Knowledge” exhibition at the Institute for Contemporary Art and Thought (ITYS) in Athens, Greece and in 2009 at the “YOU_ser” exhibition at the Zentrum für Kunst und Medien (ZKM) in Karlsruhe, Germany and the “Concrete & Samples” exhibition in the Museum of Contemporary Art in Siegen, Germany.

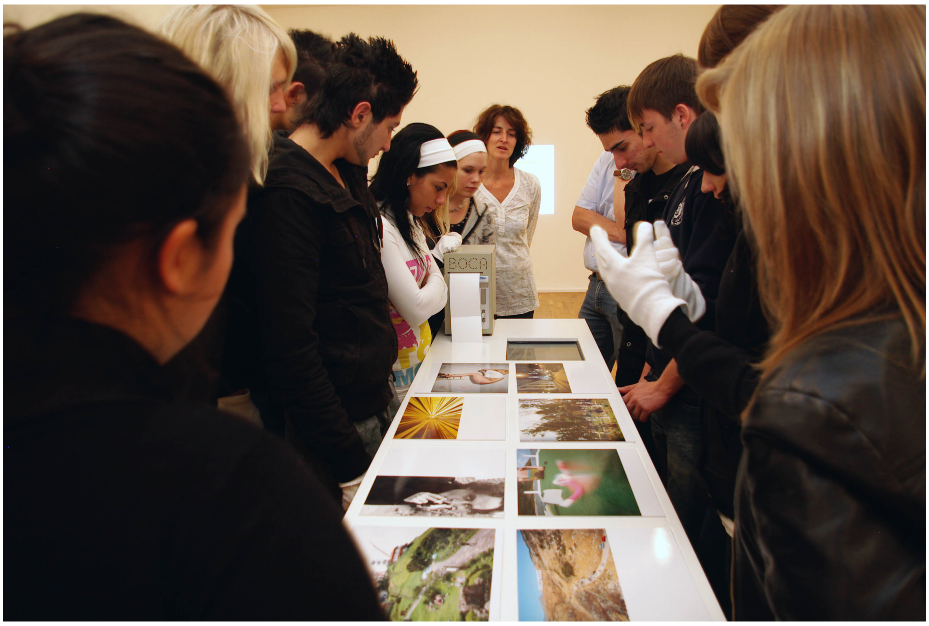

Figure 1.1: Visitors interacting with the “Phenotypes / Limited Forms” installation

[1] http://www.arminlinke.com

The installation consisted of shelves displaying a selection (1000) of Linke's photos and a working table in the middle. On this table visitors of the exhibition had the opportunity to create their own photo album out of the displayed photos and invent a title for it. The albums were then printed as books and given to the visitors. The titles they gave to these photo albums were considered as tags, associated to specific books, and thus to all the images in there. These assignments were then fed into the Ikoru system to be stored in the Ikoru database and made accessible through the Ikoru web interface. In this way, museum visitors were engaged in tagging in a physical space, and the exhibition served as a physical extension of the otherwise virtual Ikoru interface.

The main artistic aspect of this exhibition was not only the opportunity to display Like's photos but also to experience their diversity in meaning. The tags collected in this exhibition should therefore be as diverse as possible, a goal which is opposite to most common (online) tagging systems where a clear description of the tagged resource is intended.

1.2 Goal of analysis

In this document we will discuss the analysis of the data that was gathered through the "Phenotypes / Limited Forms" installation, which we will refer to as the Armin Linke dataset. This analysis was conducted using technologies developed within the TAGora project.

The album creation process can be considered as a one-directional interface from the exposition participant to the Ikoru system. In this configuration there exists no direct feedback from Ikoru back to either the artist or the visitors of the exposition. The goal of the data analysis is therefore to close the communication loop and provide feedback from the virtual tagging-application of Ikoru to the real world tagging-application.

The secondary goal is to find evidence that the imposed conditions on the tagging process truly supports the collection of many different, meaningful tags.

1.3 Preliminary analytical results

In this section the preliminary results of the analysis of the Armin Linke dataset will be presented. First an initial description of the dataset is given and of the conditions imposed on the tagging process are discussed. Then follows a brief interpretation of the TAS distribution in comparison to other online tagging systems such as Delicious (formerly known as Del.icio.us). Hereafter a closer look is taken on the network structure of the dataset. Finally the analysis concludes with preliminary conclusions and an outlook.

1.3.1 The dataset

The dataset collected from the exhibition consists of around 190 000 tag assignments (TAS) and splits up into 24000 different users, 17000 tags and 2400 different images. In each exhibition only 1000 images are show which is always a subset of the 2 400 within the dataset.Because the exhibition visitors stayed anonymous, each user in the dataset is only associated to a single album which he or she created. The photo albums consist of seven to eight images. It is also worth mentioning that exhibition visitors were presented with an instructions list, advising them to first think of a topic before creating their actual photo album.

1.3.2 Distribution pattern of users, tags and resources

Figure 1.2 illustrates the TAS distribution of the Armin Linke dataset. Both axis are in logarithmic scale to improve readability. On the X-axis the number of TAS occurrences for each user, tag and image is shown. On the Y-axis the corresponding percentage of users, tags and images with the same number of TAS occurrences, in comparison to all users, tags or images, is displayed.

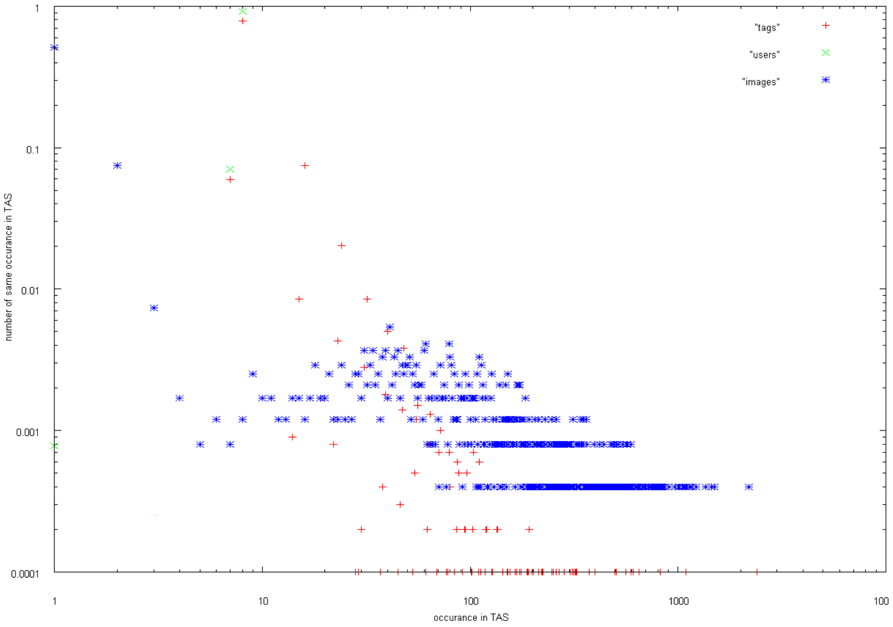

Figure 1.2: TAS distribution of the Armin Linke dataset

As a reference for comparison, figure 1.3 shows the TAS distribution calculated from data of the Delicious bookmarking service collected by Hotho et al (Hotho et al., 2006b). Examining the plot in figure 1.2 reveals significant differences in the distribution of user, tags and images in the Armin Linke dataset in comparison to the plot for Delicious in figure 1.3.

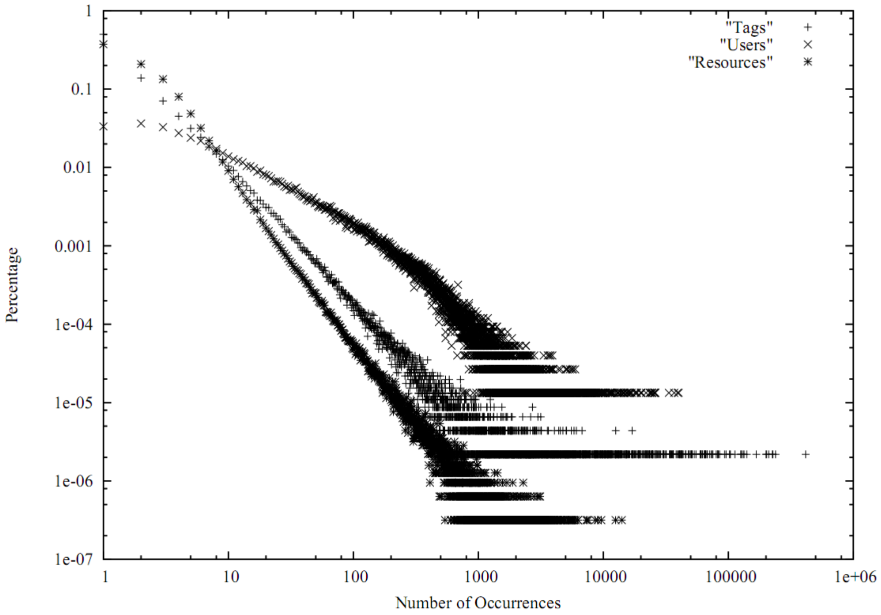

Figure 1.3: TAS distribution Delicious Data (Hotho et al., 2006b)

If we focus on the plot of the users in figure 1.2 and take into account that users stayed anonymous, so each user has tagged only one album, and that an album consists of 7 to 8 images, it becomes clear that the users plot corresponds to the distribution of album size. It also shows that most users actually created albums with 8 images but also many users settled with 7 image albums.

Looking at the distribution of tags in figure 1.2, the complete lack of low TAS count (1-6) tags becomes clear. This can be explained by taking batch tagging into account, meaning that visitors tagged always a set of images with one tag. Together with the album size limitation of 7-8 images it is clear that no tag with a TAS count lower than 7 can exist. Apart from this small difference the tag distribution follows the long tail distribution typically seen in tagging systems.

The distribution of images in figure 1.2, in contrast to users and tags, shows a slightly different distribution which is not logarithmic in nature. We can see that images with a TAS count between 5 and 50 have an unusual low proportion in comparison to images with higher or lower TAS counts. The explanation for this can be derived from the fact that the set of resources (images) is in comparison to other tagging systems rather small, and very selected since every exhibition location had only a selection of one thousand images. It follows that dependant on how long each set of images was presented also the number of usages for certain images changes. Therefore the distribution of image usage is heavily altered by the selection which images are actually shown in the exhibition.

1.3.3 Features of the three-mode network

To gain a deeper understanding of the differences of the Armin Linke dataset to usual tagging systems such as Delicious we will now take a closer look at the three-mode network structure of the folksonomy. This network is based on the folksonomy data where the users, tags and images form the vertices within in the network and the TAS the three-mode edges .The key values to get a better understanding of the network structure are the Characteristic path length, the Cliquishness and the Connectivity.

The Characteristic path length is the average shortest path from any node in the network to reach any other node. As shown in table 1.3.3 this value is in the Armin Linke dataset with 3.5 comparable with the one measured in Delicious. It is a prove that both networks have the small world property, meaning that the whole network can be traversed in very few steps.

The Cliquishness as well as the Connectivity are derivations of the clustering coefficient as it is defined by Watts (Watts, 1999) for normal non-hyper graphs. There the clustering coefficient is the proportion of edges that actually occur in the graph in comparison to all possible edges, which resembles at the same time the probability that two vertices are connected. This feature combines two aspects which are not equal in three-mode data as in the case of tagging datasets. One aspect is how many of the possible edges around a node actually do occur, and the other is whether the neighbourhood of a given vertex approaches a clique. According to this differentiation, Cattuto et al. (Cattuto et al., 2007c) extended the concept of the clustering coefficient to Cliquishness and Connectivity.

Cliquishness is the proportion of TAS which actually occurs in the dataset for a certain vertex in relation to all which are possible. For an image node it is therefore the number of TAS where this image occurs in divided by the product of all users and tags which have ever used or been assigned to this image. As shown in table 1.3.3 both datasets have comparable values for Cliquishness which indicates a dense network structure, since the neighbourhood of vertices include nearly all possible neighbourhoods. In case of the Armin Linke dataset this value is actually misleading as it becomes recognisable if the Cliquishness is only calculated for users, images and tags separately. Since every user has created only one album it becomes reasonable that the Cliquishness for users is 1.0 since there is only one tag assigned to all the images of the album. This again is also the reason why the Cliquishness for tags is with 0.92 comparable to the users. In contrast to this the neighbourhood of images is not as dense. For images we measured a Cliquishness value of 0.55. These values show that every user vertex defines an own clique around it and the tags are the major vertexes to interconnect between them.

The Connectivity is the proportion of nodes which stay connected even if the vertex in question would be removed from the network. For an image it would be the proportion of tag-user pairs where the image occurs in, which would still be possible with a replacement image if the image in question would be removed. In Delicious the Connectivity is measured as 0.85 which indicates that most of all connections would remain present if the vertex in question would be removed. In contrast to this table 1.3.3 shows that for the Armin Linke dataset the Connectivity has only a value of 0.14 meaning that the removal of a single vertex from the network has significant impact on the overall network structure. Looking closer at the Connectivity only for users or tags proves this small value with a user-Connectivity or 0.18 and a tag-Connectivity or 0.0 . In contrast to these values the image-Connectivity has a value of about 0.55 which indicates a higher amount of interconnection formed through the images. As stated above these values again prove that the tags are the major type of vertexes connecting the different user-formed cliques.

Characteristic path length	mean path length	Cliquishness	Connectivity
Armin Linke	3.5	0.95	0.14
Delicious	3.6	0.85	0.85

Table 1.1: Values of different network parameters for the Armin Linke and the Delicious dataset

1.3.4 Preliminary conclusions

The previous sections drew the following picture about the Armin Linke dataset. The constrains imposed on the underlying tagging system, such as batch tagging, anonymous users and limited set of resources had significant influence on the development of the folksonomy. Most changes occurred in the distribution of the users and the resources, but minor changes are also visible in the distribution of tags. In comparison to the common tagging system, for which the Delicious system was chosen as a representative example, the users form tiny cliques together with the assigned tag and the images used in the album the user used. These cliques are strongly interconnected through single tags and in a minor way through the multiple images as described by the Cliquishness and Connectivity measurements. The conclusion of these observations is that in contrast to usual tagging systems were a high interconnection among different tagged resources exists, the Armin Linke dataset is separable in many small topics which are interconnected lightly through tags. This can be interpreted as that the tagging process itself is less influenced by the popularity of the tags, in contrast to common tagging systems, and therefore supports a rather diverse usage of tags.

An investigative look into the actual tags, which are used within the Armin Linke dataset provides hints for the following conclusions. Most of the most popular tags are actually dates, and description of the location or event of the exhibition itself. Also tags in different languages, namely German, French and Greek, are dominant within the dataset and have often the same meaning once translated into English. A manual clustering and disambiguation of the most popular tags indicates that the set of tags can be significantly reduced and it is still to be determined how this densification would influence the network structure.

1.4 Outlook

This analysis should be seen as a first step in the process to find an appropriate feedback for the users or to understand the true impact of the imposed constrains on the tagging process as present in the “Phenotypes / Limited Forms” installation.

The conclusions drawn out of this first look is that the current setup indeed supports a rather diverse tagging creation process with the drawback that the resulting network between users, tags and images is weakly interconnected. It is still to be examined if this is only the result of a slowed down growing process of the interconnections with the folksonomy, caused by external factors such as different languages, or a true result of the constrains.

The next steps in finding an appropriate feedback for the visitors of the exhibition lies in our opinion in the definition of a measure to determine the true diversity of an image based on the assigned tags. A first step to achieve this would be to automatically translate and maybe even disambiguate (perhaps using the web service developed by the team at the University of Southampton) the tags.

Chapter 2

Semantic Similarity and Recommendations

2.1 Measuring the Semantic Similarity of Tags

2.1.1 Comparison of measures for tag similarity

The structure of folksonomies differs fundamentally from that of e.g., natural text or web resources, and sets new challenges for the fields of knowledge discovery and ontology learning. Central to these tasks are the concepts of similarity and relatedness. In this task, we have focussed on similarity and relatedness of tags, because they carry the semantic information within a folksonomy, and provide thus the link to ontologies. Additionally, this focus allows for an evaluation with well-established measures of similarity in existing lexical databases.

Budanitsky and Hirst pointed out that similarity can be considered as a special case of relatedness (Budanitsky and Hirst, 2006). As both similarity and relatedness are semantic notions, one way of defining them for a folksonomy is to map the tags to a thesaurus or lexicon like Roget's thesaurus[1] or WordNet (Fellbaum, 1998), and to measure the relatedness there by means of well-known metrics. The other option is to define measures of relatedness directly on the network structure of the folksonomy. One important reason for using measures grounded in the folksonomy, instead of mapping tags to a thesaurus, is the observation that the vocabulary of folksonomies includes many community-specific terms which did not make it yet into any lexical resource. Measures of tag relatedness in a folksonomy can be defined in several ways. Most of these definitions use statistical information about different types of *co-occurrence* between tags, resources and users. Other approaches adopt the *distributional hypothesis* (Firth, 1957; Harris, 1968), which states that words found in similar contexts tend to be semantically similar.

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

In most studies, the selected measures of relatedness seem to have been chosen in a rather ad-hoc fashion. We believe that a deeper insight into the semantic properties of relatedness measures is an important prerequisite for the design of ontology learning procedures that are capable of harvesting the emergent semantics of a folksonomy.

In (Cattuto et al., 2008b) (see also (Cattuto et al., 2008a)), we analysed five measures of tag relatedness: the *co-occurrence count*, *three distributional measures* which use the cosine similar-

[1] http://www.gutenberg.org/etext/22

ity (Salton, 1989) in the vector spaces spanned by users, tags, and resources, respectively, and *FolkRank* (Hotho et al., 2006a), a graph-based measure that is an adaptation of PageRank (Page et al., 1998) to folksonomies. Our analysis is based on data from a large-scale snapshot of the popular social bookmarking system Delicious.[2] To provide a semantic grounding of our folksonomy-based measures, we map the tags of Delicious to synsets of WordNet and use the semantic relations of WordNet to infer corresponding semantic relations in the folksonomy. In WordNet, we measure the similarity by using both the taxonomic path length and a similarity measure by Jiang and Conrath (Jiang and Conrath, 1997) that has been validated through user studies and applications (Budanitsky and Hirst, 2006). The use of taxonomic path lengths, in particular, allows us to inspect the edge composition of paths leading from one tag to the corresponding related tags, and such a characterization proves to be especially insightful.

The contribution of our work (Cattuto et al., 2008b) is twofold: First, it introduces a systematic methodology for characterizing measures of tag relatedness in a folksonomy. Several measures have been proposed and applied, but given the fluid and open-ended nature of social bookmarking systems, it is hard to characterize – from the semantic point of view – what kind of relations they establish. As these relations constitute an important building block for extracting formalized knowledge, a deeper understanding of tag relatedness is needed. In this paper, we grounded several measures of tag relatedness by mapping the tags of the folksonomy to synsets in WordNet, where we used well-established measures of semantic distance to characterize the investigated measures of tag relatedness. As a result, we showed that distributional measures, which capture the context of a given tag in terms of resources, users, or other co-occurring tags, establish – in a statistical sense – *paradigmatic* relations between tags in a folksonomy. Strikingly, our analysis shows that the behavior of the most accurate measure of similarity (in terms of semantic distance of the indicated tags) can be matched by a computationally lighter measure (tag context similarity) which only uses co-occurrence with the popular tags of the folksonomy. In general, we showed that a semantic characterization of similarity measures computed on a folksonomy is possible and insightful in terms of the type of relations that can be extracted. We showed that despite a large degree of variability in the tags indicated by different similarity measures, it is possible to connotate *how* the indicated tags are related to the original one.

The second contribution addresses the question of emergent semantics: Our results indicate clearly that, given an appropriate measure, globally meaningful tag relations can be harvested from an aggregated and uncontrolled folksonomy vocabulary. Specifically, we showed that the measures based on tag and resource context are capable of identifying tags belonging to a common semantic concept. Admittedly, in their current status, none of the measures we studied can be seen as *the* way to instant ontology creation. However, we believe that further analysis of these and other measures, as well as research on how to combine them, will help to close the gap towards the Semantic Web.

Our results can be taken as indicators that the choice of an appropriate relatedness measure is able to yield valuable input for learning semantic term relationships from folksonomies. In an application context, the semantic characterization we provided can be used to guide the choice of a relatedness measure as a function of the task at hand. We will close by briefly discussing which of the relatedness measures we investigated is best for ...

- *... synonym discovery.* The tag or resource context similarities are clearly the first measures to choose when one would like to discover synonyms. As shown in this work, these measure delivers not only spelling variants, but also terms that belong to the same WordNet synset. This kind of information could be applied to suggest concepts in tagging system or to support users by cleaning up the tag cloud.

- *... concept hierarchy.* Both FolkRank and co-occurrence relatedness seemed to yield more

[2] http://del.icio.us/

general tags in our analyses. This is why we think that these measures provide valuable input for algorithms to extract taxonomic relationships between tags.

- *. . . tag recommendations.* The applicability of both FolkRank and co-occurrence for tag recommendations was demonstrated in Ref. (Jäschke et al., 2007). Both measures allow for recommendations by straightforward modifications. Our evaluation in Ref. (Jäschke et al., 2007) showed that FolkRank delivered superior and more personalized results than co-occurrence. On the other hand, similar tags and spelling variants as frequently provided by the context similarity are less accepted by the user in recommendations.

- *. . . query expansion.* Our analysis suggests that resource or tag context similarity could be used to discover synonyms and – together with some string edit distance – spelling variants of the tags in a user query. The original tag query could be expanded by using the tags obtained by these measures.

- *. . . discovery of multi-word lexemes.* Depending on the allowed tag delimiters, it can happen that multi-word lexemes end up as several tags. Our experiment indicates that FolkRank is best to discover these cases. For the tag *open*, for instance, it is the only of the three algorithms which has *source* within the ten most related tags and vice versa.

The work along this line was continued in (Markines et al., 2009), where we described an evaluation framework to compare various general folksonomy-based similarity measures. The main contributions of this paper are:

- A general and extensive foundation for the formulation of similarity measures in folksonomies, spanning critical design dimensions such as the symmetry between users, resources, and tags; aggregation schemes; exploitation of collaborative filtering; and information-theoretic issues. Some of the measures considered have been introduced and investigated before, but no systematic study including all dimensions of a folksonomy and all measures exists to date about their application to social similarity.

- An experimental assessment of the effectiveness of several similarity measures for both tags and resources. For the former, we measure effectiveness by comparison with user-created tag relations. For both tags and resources, as a second step we gauge the similarity measures against reliable grounding measures validated by user studies on large and open reference data sets. This evaluation addresses several key limitations of traditional user based assessments.

- An analysis of the empirical evaluation results in the context of their scalability, in particular their viability for practical implementations in existing social bookmarking systems. A clear trade-off between effectiveness and efficiency is demonstrated and discussed.

In summary, we have discussed a general and extensive foundation for the formulation of similarity measures in folksonomies, spanning critical design dimensions such as the symmetry between users, resources, and tags; aggregation schemes; exploitation of collaborative filtering; and information-theoretic issues. Experiments with resource and tag similarity alike have pointed to folksonomy-based mutual information measures as the best at extracting semantic associations from social annotation data.

The question of scalability has highlighted a critical trade-off between accuracy and computational complexity. Some social aggregation methods achieve good accuracy in a non-scalable way. On the other hand, measures based on collaborative aggregation of annotations achieve competitive quality while minimizing computation time thanks to incremental updates. This leads to the best performance/cost trade-off; we underscore the key role of scalability for the practical viability of similarity computations in existing social bookmarking systems.

Other similarity measures that we have not yet explored include matrix-normalized mutual information with binary projection aggregation and the integration of collaborative filtering with distributional aggregation.

The similarity measures analyzed in this paper can readily be employed to support many Social and Semantic Web applications, such as tag clustering for ontology construction and learning, query expansion, and recommendation. Our group has begun the use of these similarity measures in Web navigation during knowledge exploration (Krause et al., 2008b). Another straightforward application of the socially induced resource similarity would be to enrich, e.g., results of a search engine with semantically similar resources. This would resemble an annotation-based version of Google's "`related`" operator, exemplifying a possible synergy between traditional and socio-semantic Web technologies.

2.1.2 Focus Group at Dagstuhl seminar.

At the seminar on social online communities, which was organized by the Tagora team at the Leibniz Center for Informatics Schluss Dagstuhl in 2008, a group of researchers continued this kind of analysis in a 'hacking session' (Benz et al., 2008). Understanding of annotation properties is crucial for constructing accurate and efficient navigation and browsing mechanisms, including content recommendations (favorites), ranked retrieval of relevant items for user queries, or user assistance in annotating new contents (tag recommendation). For this reason, the discussion in the focus group was centered around the semantic grounding of tagging. Our objective was to exploit state-of-the-art Information Retrieval methods for finding associations and dependencies between tags, capturing and representing differences in tagging behavior and vocabulary of various folksonomies, with the overall aim to better understand tags and the tagging process. To this end, we analyzed the semantic content of tags in the Flickr and Delicious folksonomies. We observed the following interesting findings:

While many of the frequently occurring Delicious tags also appear in Flickr, applying the tag context similarity measures at a global scale does not give exciting insights. However, comparison of an individual's co-occurrence network could be used to some extent to measure whether ambiguous terms are used with the same sense. Such measures are noisy and do not provide stable results. Improvement might be made by filtering the tags so morphological variations and synonyms are merged.

We performed the analysis of tag context similarity in the narrow folksonomy of Flickr and confirmed the result obtained for Delicious in a previous work. We find that tags in Flickr are obviously oriented towards their visual meaning, whereas in Delicious they are biased more towards their technical meaning. Moreover, we restricted the analysis of tag context to those users belonging to the same group of interest and found no particular variations in tag similarities with respect to the unrestricted set.

We embedded in a three dimensional space the representation of the tag-tag space with the cosine similarity metric by means of the software OntoGen. We were able to navigate in such space and find regions of high similarity density, where the cosine similarity distance between tag pairs is higher than the average.

Finally, by constructing a *directed* tag-tag co-occurrence network, in which nodes represent tags and links connect two adjacent tags inside a post from left to right, we showed that tag order in posts has a relevant semantic value.

2.2 Tag Recommendations

Collaborative tagging systems allow users to assign keywords – so called “tags” – to resources. Tags are used for navigation, finding resources and serendipitous browsing and thus provide an immediate benefit for users. These systems usually include tag recommendation mechanisms easing the process of finding good tags for a resource, but also consolidating the tag vocabulary across users. In practice, however, only very basic recommendation strategies are applied.

There are two typical approaches to the recommendation problem: content-based approaches and collaborative filtering approaches (Burke, 2002). While the former rely solely on the content of the documents, the latter take into account the behavior of similar users. Social bookmarking systems are an ideal scenario for the collaborative filtering approach, as the similarity of users can be measured by comparing their tagging behavior. Nevertheless the so-called *cold start problem* also occurs in social bookmarking systems: When a resource is tagged for the first time by some user, all other users – and in particular those who are similar to him – do not yield any recommendation about which tags to use. Therefore, content-based recommendations also have their use in social bookmarking systems.

2.2.1 Collaborative Filtering

In (Jäschke et al., 2008b), we have evaluated and compared several recommendation algorithms on large-scale real life datasets: an adaptation of user-based collaborative filtering, a graph-based recommender built on top of the FolkRank algorithm, and simple methods based on counting tag occurrences.

Most recommender systems are typically used to call users’ attentions to new objects they do not know yet and have not rated already in the past. This is often due to the fact that there is no repeat-buying in domains like books, movies, music etc. in which these systems typically operate. In social bookmarking systems, on the contrary, re-occurring tags are an essential feature for structuring the knowledge of a user or a group of users, and have to be considered by a tag recommender.

This means that the fact that a tag already has been used to annotate a resource does not exclude the possibility of recommending the same tag for a different resource of the same user. Overall, recommending tags can serve various purposes, such as: increasing the chances of getting a resource annotated, reminding a user what a resource is about and consolidating the vocabulary across the users.

Recommender systems (RS) in general recommend interesting or personalized information objects to users based on explicit or implicit ratings. Usually RS predict ratings of objects or suggest a list of new objects that the user hopefully will like the most. The task of a tag recommender system is to recommend in a folksonomy $\mathbb{F} = (U, T, R, Y)$, for a given user $u \in U$ and a given resource $r \in R$ with $tags(u, r) = \emptyset$, a set $\tilde{T}(u, r) \subseteq T$ of tags. In many cases, $\tilde{T}(u, r)$ is computed by first generating a ranking on the set of tags according to some quality or relevance criterion, from which then the top n elements are selected.

Notice that the notion of tag relevance in folksonomies can assume different perspectives, i. e., a tag can be judged relevant to a given resource according to the society point of view, through the opinion of experts in the domain or based on the personal profile of an individual user. For all the evaluated algorithms, we concentrate here on measuring the individual notion of tag relevance, i. e., the degree of likeliness of a user for a certain set of tags, given a new or untagged resource.

The results of our empirical evaluation showed that the graph-based approach of FolkRank is able to provide tag recommendations which are significantly better than those of approaches based on tag counts and even better than those of state-of-the-art recommender systems like Collaborative Filtering. The tradeoff is, that computation of FolkRank recommendations is cost-intensive so that one might prefer less expensive methods to recommend tags in a social bookmarking system.

The *most popular tags ρ–mix* approach proposed in this work has proven to be considered as a solution for this problem. It provides results which can almost reach the grade of FolkRank but which are extremely cheap to generate. Especially the possibility to use index structures (which databases of social bookmarking services typically provide anyway) makes this approach a good choice for online recommendations. Finally, despite its simplicity and non-personalized aspect, the *most popular tags* achieved reasonable precision and recall on the small datasets (last.fm and BibSonomy).

2.2.2 Content-Based Recommendations

In (Illig et al., 2009 (to appear), we studied different content-based recommenders, and compared them on a real-world dataset – a crawl of the delicious bookmarking system. The main contribution is a comparison of state of the art recommenders, the adaption of classifiers to this problem and a demonstration that content based recommenders are able to generalize and to make predictions for new web pages. A more detailed discussion of the findings can be found in the bachelor thesis (Illig, 2008).

We evaluated the effectiveness of multiple text classification methods and variants applied to a scenario that is compatible with the common text classification evaluation practice of disjoint training and test scenarios but still represents a realistic and pure cold start tag recommender evaluation scenario.

Some algorithms have been slightly modified in various ways to make use of tag assignment frequencies by multiple users. Improvements by these extensions have been detected for the case of a TAS weighted 30-Nearest-Neighbors algorithm. Nevertheless, we found that an one-vs-one SVM variant on length normalized document feature vectors is the most effective of all evaluated classifiers. Concluding, we could show that folksonomy tag assignments can be learned by application of machine learning techniques to address the cold start problem of collaborative recommender systems.

2.3 User Recommendations

The ultimate reason that boosted the emergence of folksonomies is the search for information. While either navigating through the web or looking for references to cite in a paper, users encounter lots of informations and find suitable to keep track of their research efforts by annotating the interesting resources in a folksonomy (Delicious, CiteULike, Bibsonomy, etc.). In this sense, tags are an added value to be exploited in a second moment during the process of information retrieval. It is now well established that tags carry semantics and that this semantics can be disclosed thanks to the uncorrelated tagging activity of users, who use them wisely for their own advantage. Therefore, given the undiscussed usefulness of tags, one may use them to draw possible similarities between either resources or users. It is logical that resources or users, who have similar corresponding tag clouds, are similar themselves. This similarity may be used to collect suggestions about unknown interesting resources, which is one of the major strengths of folksonomies. In order to suggest to users those unknown resources that might be of their utility, one could set up an environment proposing resources annotated by others, according to their similarity with respect to a previously selected known resource. This procedure, although logically irreproachable, has its drawback in the enormous number of resources annotated, which makes the respective tag cloud comparison a formidable non-scalable task. Despite the fact that other (e.g. graph-based) approaches to item recommendation in folksonomies exist, another elegant way to suggest resources is to use the similarity between users in an indirect way. The number of users is in fact usually two order of magnitudes less than the number of resources in a folksonomy, and active users are even much less than that. Given a well quantified similarity measure between users, which we shall deal with in the next section, one can suggest to single users a ranked list of most similar users. After that, the user may browse the resource list of the most similar users, look at their resources, and pick the interesting resources she could have missed. If the similar users own a large number of resources, personalized ranking schemes can greatly alleviate this task by sorting the resources by relevance for the target user. First simple approaches again based on tags (like counting the number of relevant tags for each post, whereby a tag is considered as relevant if it appears in the tag cloud of the target user) seemed to yield a helpful resource ranking. Additionally, the user may add those most similar users in a personal user list and be informed by the system whenever those watched users annotate new resources. In this way, user recommendation serves not only the purpose of facilitating the discovery of interesting content, but also supports the user in finding relevant communities within the folksonomy. In addition to these improvements of the navigation and interaction experience of users, the overall procedure also provides a way to control whether the chosen similarity and ranking measures deliver senseful informations. This user recommendation system along with personalized ranking algorithms has been implemented in Bibsonomy (see section 1.1.2 of Deliverable 2.5). Unfortunately, we do not have up to now a sufficient number of data to explore the introduced control-feedback mechanism. In summary, we have implemented three tag-based and one graph-based approaches of computing user similarity, which are explained in the following two subsections.

2.3.1 Tag-based user recommendation

In this section we revise the tag-based methods we implemented to quantify the similarity measure between users. Say we have to compare the similarity between user A and user B with tag clouds T_A and T_B respectively. Each tag has of course its own weight in the tag cloud, according to the number of posts into which it appears. The three implemented methods are:

Jaccard Given two sets, i.e. tag-clouds T_A and T_B in our case, the Jaccard similarity is defined as $J(T_A, T_B) = |T_A \cap T_B|/|T_A \cup T_B|$. It is a symmetric quantity. Tag multiplicity has to be taken into account: if tag t occurs x_A times in tag-cloud T_A and x_B times in tag-cloud T_B, then $\min(x_A, x_B)$ is its contribution for the intersection cardinality and $\max(x_A, x_B)$ for the union.

Cosine In this procedure, we build two multi dimensional vectors V_A and V_B each component of which contains the number of occurrences of the corresponding tag in the tag-cloud. After that, the cosine similarity is defined with the usual euclidean scalar product as $C(T_A, T_B) = (V_A \cdot V_B)/(|V_A||V_B|)$.

TF/IDF This method is a variation of the cosine similarity method. The difference is in the construction of the vectors V_A and V_B, whose components now contain a sort of TF/IDF score for the corresponding tag. To be precise, the TF term is the frequency of that tag inside user tag-cloud and the IDF term is the base 2 logarithm of the reciprocal value of that tag global frequency. This measure favors the similarity between users who share uncommon tags.

2.3.2 Graph-based user recommendation

Complementing the three tag-based metrics described avobe, the BibSonomy users can also test a graph-based approach of computing user similarity. The main difference hereby lies in the fact that this method takes into account the complete graph structure of users, tags and resources instead of being based solely on tag clouds.

FolkRank This user similarity measure is based on the FolkRank (Hotho et al., 2006a) calculated for the user A. FolkRank is a differential and biased PageRank-like procedure and ranks tags/users/resources according to the topology of the folksonomy seen as a tripartite network. In this particular context, only the user rank is used. To summarize this method, the more two users are similar, the more nodes they share of the folksonomy tripartite network.

Chapter 3

Analysing and Influencing User Behavior

3.1 Analyzing User Behavior

Continuing our work (Krause et al., 2008a) of the second year of the Tagora project, we have, in (Krause et al., 2008c) (see also (Jäschke et al., 2008a)), discussed the realization of "search communities" within search engines by building an anonymized folksonomy similar to the del.icio.us social bookmarking system from search engine logdata. As logdata contain queries, clicks and session IDs, the classical dimensions of a folksonomy can be reflected: Queries or query terms represent tags, session IDs correspond to users, and the URLs clicked by users can be considered as the resources they tagged with the query terms. Search engine users can then browse this data along the well known folksonomy dimensions of tags, users, and resources.

A search engine folksonomy, which we will call *logsonomy* in the sequel, brings a variety of features to search engines. Partly discussed in blogs (Smith, 2005) one can picture users adding additional tags to their pages to have them higher ranked. Temporal aspects can be introduced by incorporating a fourth dimension and showing popular tags, users or resources at a certain time. Finally, search engine users may interact with each other, commenting and copying search results of each other.

Logsonomies open a wide field of exploration. What kind of semantics can we extract from logsonomies? Is the serendipitous discovery of information also possible in logsonomies? How does the structure of logsonomies differ from folksonomies? In this paper, we address these questions by analyzing the topological properties of two logsonomy datasets and comparing our findings to a social bookmarking system. In previous work (Cattuto et al., 2007b), we have shown that folksonomies exhibit specific network characteristics (e.g. small world properties, power laws, and long tail degree distributions). These characteristics help to explain why people are fascinated from this structure: A small world leads to short ways between users, resources and tags, which allows for finding interesting resources by browsing the system randomly. High clustering coefficients show dense neighbourhoods which are tracked by the formation of communities around different topics. Finally, cooccurrence graphs show the building of user enabled shared semantics.

By looking at a logsonomy graph's components we find that logsonomies collapse in more disconnected components than folksonomies do. In contrast, small world properties considering the shortest path length and the clustering coefficient, compared to random graphs and del.icio.us, can be confirmed, and finally, the strength of each node expresses similar tagging semantics as folksonomies do. Most of the differences in topological structure can be explained by the differences in user behaviour and the creation of metadata in both systems. Overall, we think that our findings strenghten the idea that clickdata can enable social information retrieval and serve as a basis for further analysis.

We have analyzed the graph structure of logsonomies to find similarities and dissimilarities to the existing folksonomy del.icio.us. We found similar user, resource and tag distributions, whereby the split query datasets are closer to the original folksonomy than the complete query datasets. We could show that both graph structures have small world properties in that they exhibit relatively short shortest path length and high clustering coefficients. Finally, the analysis of the strength in the tag-tag–co-occurrence network revealed very similar properties between folksonomies and logsonomies with split queries.

In general, the differences between the folksonomy and logsonomy model did not effect the graph structure of the logsonomies. Minor differences are triggered by the session IDs which do not have the same thematic overlap as user IDs have. Also, full queries show less inherent semantics than the splitted datasets do. In future work, a more thourough analysis of these differences will be interesting.

Overall, the results support our vision to merge the search engine and folksonomy worlds into one system. While some search engines already allow to store and browse search results, they do not provide folksonomy-alike navigation or the possibility to add or change tags. From a practical point of view, the following considerations are further arguments for a logsonomy implementation and its combination with a folksonomy system:

- Users could enrich visited URLs with their own tags (besides the automatically added words from the query) and the search engine could use these tags to consider such URLs for later queries — also from other users. Thus, those tags could improve the quality of the search engine.
- The popularity of folksonomy systems could increase the customer loyalty for a search engine. The community-feeling known from folksonomies could pass over to search engines.
- Search engines typically have the problem of finding new, unlinked web pages. Assumed, users store new pages in the folksonomy, the search engine could direct its crawlers better to new pages. Additionally, those URLs would have been already annotated by the user's tags — even without crawling the pages it would be possible to present them in result sets.
- As described in (Röttgers, 2007), folksonomies can assist in finding trends in society. Many social bookmarking users can be viewed as trend setters or early adopters of innovative ideas — their data is valuable for improving a search engine's topicality.
- Bookmarked URLs of the user may include pages, the search engine can not reach (intranet, password-protected pages, etc.). These pages can then be integrated into personalized search results.

However, privacy issues are very important when talking about search engine logs. They provide details of a user's life and often allow to identify the user himself (Adar, 2007). Certainly, this issue needs attention when implementing a logsonomy system.

3.2 Spam Detection

Web spam detection is a well known challenge for search engines. Spammers add specific information to their web sites that solely serve to increase the ranking and not the quality or content of a page. They thereby increase the traffic to their web sites – be it for commercial or political interests or to disrupt the service provided. Ranking algorithms need to detect those pages using elaborate techniques.

Not only search engines need to fight with malicious web content. Social bookmarking systems also have become an attractive place for posting web spam. These systems allow users to annotate and share bookmarks. Within the last few years, a large community of users who add, share and work with the content of these systems has evolved. Delicious[1] is a popular example, but also other systems targeting more specific communities such as the scholarly world, exist (Connotea[2], CiteULike[3], BibSonomy[4]).

Spammers (mis)use the popularity and the high PageRank of social bookmarking systems for their purposes. All they need is an account; then they can freely post entries which bookmark the target spam web site. In recent months, different spamming techniques have been developed to frequently show up on popular sites, recent post sites or as highly ranked posts after the search for a specific tag. For instance, spammers request several accounts and publish the same post several times. Besides appearing on the recent post page, the bookmark may show up on the popular page, since "many" users have considered the bookmark. Another technique is to add diverse tags to the bookmark or use popular tags.

In order to retain the original benefits of social bookmarking systems, techniques need to be developed which prevent spammers from publishing in these systems, or at least from having their malicious posts published. The problem can be considered as a binary classification task. Based on different features that describe a user and his posts, a model is built from training data to classify unknown examples (on a post or user level) either as ("spam" or "non-spam"). As we consider "social" systems in which users interact with each other and one incentive to use the system is to see and be seen, an exclusion of non-spammers from publishing is a severe error which might prevent the user from further participation. Similar to other spam detection settings, this problem needs to be taken into consideration when classifying users.

The adaptation of classification algorithms to this task consists of two major steps. The first one is to select features for describing the users. The second step is the selection of an appropriate classifier for the problem.

In the following, we will present two different approaches for computing the features which describe a user. The first approach in Section 3.2.1 uses features that are directly computed from the profile and the tagging activities of a user. The second approach in Section 3.2.2, then uses the Epistemic Model from (Dellschaft and Staab, 2008) to compute features which compare the actual tagging behaviour of a user with the predicted behaviour of a non-spam user as it is modelled by the Epistemic Model.

3.2.1 Direct Computation of User Features

In (Krause et al., 2008d), we introduced a set of initial features that can be used for spam classificiation. These features are evaluated with well-known classifiers (SVM, Naive Bayes, J48 and logistic regression) against a simple baseline of representing a user by the usage of tags.

The paper introduced a variety of features to fight spam in social bookmarking systems. The features were evaluated with well-known machine learning methods. Combining all features shows promising results exceeding the AUC and F1 measure of the selected baseline. Considering the different feature groups, cooccurrence features show the best ROC curves.

Our results support the claim of (Heymann et al., 2007), that the problem can be solved with classical machine learning techniques – although not perfectly. The difference to web spam classification are the features applied: on the one hand, more information (e. g., email, tags) is given, on the other hand spammers reveal their identity by using a similar vocabulary and resources. This is why cooccurrence features tackle the problem very well.

Several issues considering our approach need to be discussed. First of all, a switch from the user level to the post level is an interesting next step to consider. This would also facilitate the

[1] http://del.icio.us

[2] http://www.connotea.org

[3] http://www.citeulike.org

[4] http://www.bibsonomy.org

handling of borderline cases, as users, though some of their posts were flagged as spam, can still participate. A consideration of a multiclass classification introducing classes in between "spam" and "non spam" or a ranking of classified instances may also help to identify those borderline users a moderator needs to manually classify. A further issue regards the evaluation method chosen. In future work, we want to consider more than one chronological separated training/test set. This may also help to reduce the ratio between training and test data. The large ratio between spam and non-spam users could be reduced by identifying spammers which have created several user accounts and therefore are counted several times. Finally, the feature groups presented have been intuitively chosen – they may be extended in different ways. We also think of adding more features such as topological information, clustering coefficients and tag similarity in posts.

Overall, our contribution represents a first step towards the elimination of spam in social bookmarking systems using machine learning approaches. Currently, we are constructing a spam detection framework to flexibly combine features and learning algorithms. Besides the practical need to eliminate spam, we intend to use this platform to develop and evaluate further social spam detection mechanisms.

3.2.2 Model-based Computation of User Features

During the quantitative evaluation of the Epistemic Model, it became obvious that the simulations better reproduce co-occurrence streams if spam postings were removed from them. This is an indicator that the Epistemic Model is a model of an average non-spammer and that spammers deviate from the modelled behaviour.

In the following, we will describe our initial experiments how we can use the Epistemic Model for detecting spammers in social tagging systems. We will start with a short summary of the quantitative evaluation. We will show how removing spammers influences the tag growth and the tag frequencies in co-occurrence streams. From these observations we will derive features which compare the tagging behaviour of users in real co-occurrence streams with the simulated tagging behaviour. Finally, we will present an initial evaluation of the model-based spam detection features and give an outlook over future work.

Quantitative Evaluation of the Epistemic Model

In the following, we will describe the quantitative evaluation of the Epistemic Model which extends the evaluation already available in (Dellschaft and Staab, 2008). The objective of the evaluation was to quantitatively measure the distance between simulated tag frequencies and the real frequencies in selected streams from Delicious and Bibsonomy. Furthermore, the results should be compared to the respective results of the Yule-Simon Model with Memory (Cattuto et al., 2007a).

For the evaluation, we also did some extensions to the model which allowed for simulating complete postings instead of single tag assignments. The simulation of postings is a simple extension in which we draw a given posting length from a pre-defined distribution. This distribution might e. g. be taken from real tagging systems or co-occurrence streams. When simulating a posting, we ensure that within such a posting, every tag can occur at most once.

During the evaluation it became obvious that the collected Delicious data set contains a larger amount of spam postings. Especially one kind of spam leads to significant fingerprints in the tag growth and tag frequencies: There exist several postings in the Delicious data set which contain more than 50 tag assignments. Single postings may even contain more than 4000 tag assignments. These very large postings lead to visible patterns in the tag growth graphs (see Fig. 3.1).

Tab. 3.1 shows the statistics of the used Delicious co-occurrence streams after removing all postings which contain more than 50 tag assignments.[5] In Fig. 3.2 it is exemplary shown for one

[5] A manual evaluation showed that approximately 80% of users with such a posting are spammers and 90% of

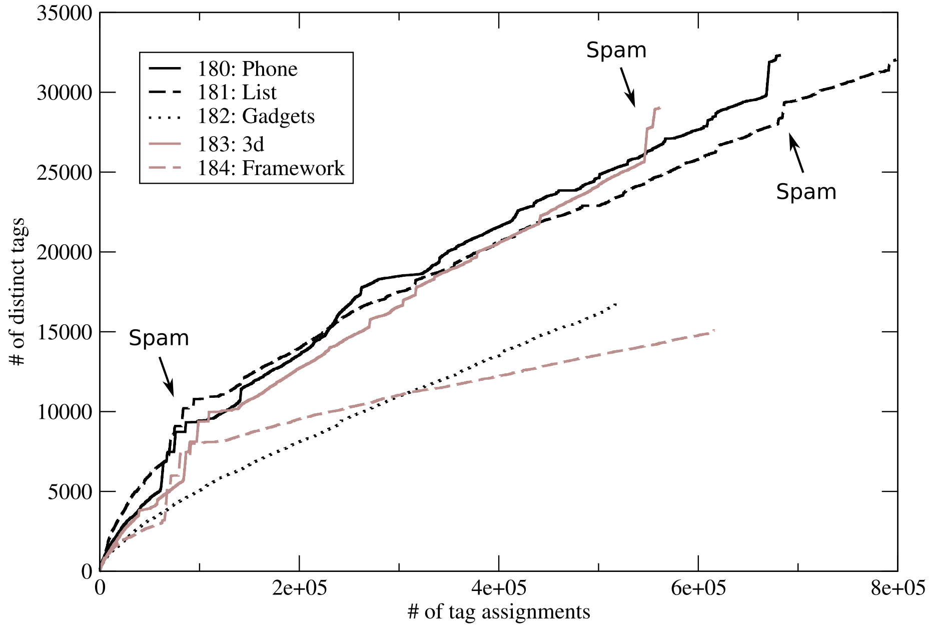

Figure 3.1: Delicious: Growth of the number of distinct tags before removing spam postings.

#	Tag	TAS	Spam Postings	Spam TAS
180	phone	561,451	448	121,244
181	list	657,654	903	140,628
182	gadgets	482,577	416	38,165
183	3d	480,610	226	80,919
184	framework	557,490	195	59,089

Table 3.1: Delicious: Statistics of the streams after removing spam.

Delicious co-occurrence stream, how removing spam postings changes the tag frequencies. Similar effects can be shown for the Bibsonomy data set (see Fig. 3.3).

For comparing the simulated and the real tag frequencies, we used the Smirnov test as it is described in (Conover, 1999). In general, the Smirnov test should be prefered over the χ^2 test if the random samples consist of ordinal data because the Smirnov test avoids unnecessary binning of data. It measures the maximal distance D between the cumulative distribution functions of the simulated and the real tag frequencies. The cumulative distribution function $P(x)$ of a random variable X gives the probability that the random variable takes on a value greater than or equal to x:

$$P(x) = Pr(X \geq x) \tag{3.1}$$

We used this distance measure for determining the best fit between the real tag frequencies and the simulated tag frequencies. For this purpose, we used Powell's direction set method as it is described in (Press et al., 1992). This is an algorithm for finding the minimum of a given function in dependency on multiple variables.

In Tab. 3.2 and 3.3 we show the best results achieved for the Epistemic Model and the Yule-Simon Model together with the parameter values for which this best result was achieved. The results show that the Epistemic Model is significantly better in reproducing the tag frequencies of real co-occurrence streams compared to the Yule-Simon Model. The same tendency could be observed

postings with more than 50 tag assignments are spam.

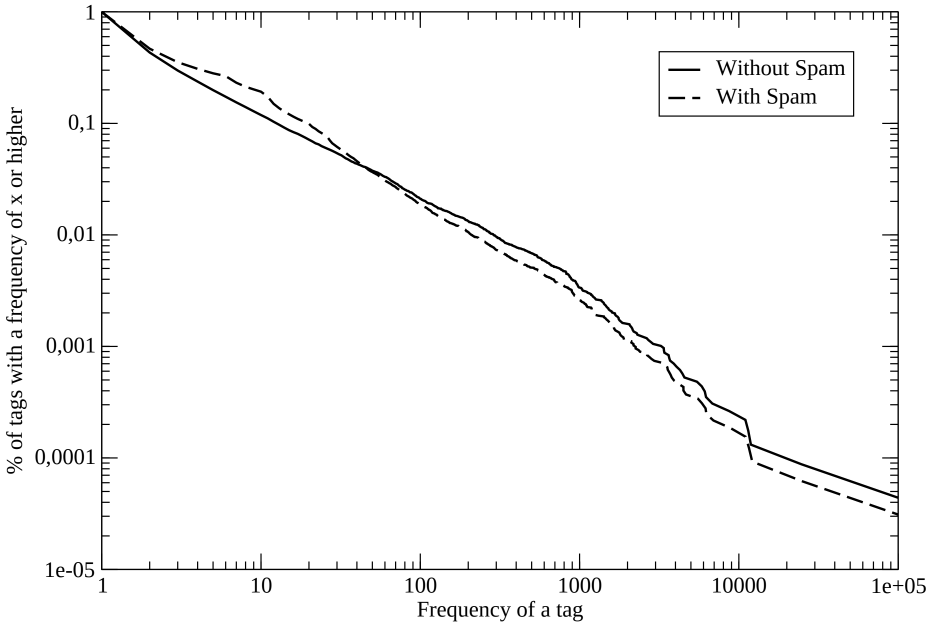

Figure 3.2: Delicious: Tag frequencies for the *Phone* stream with and without the spam postings.

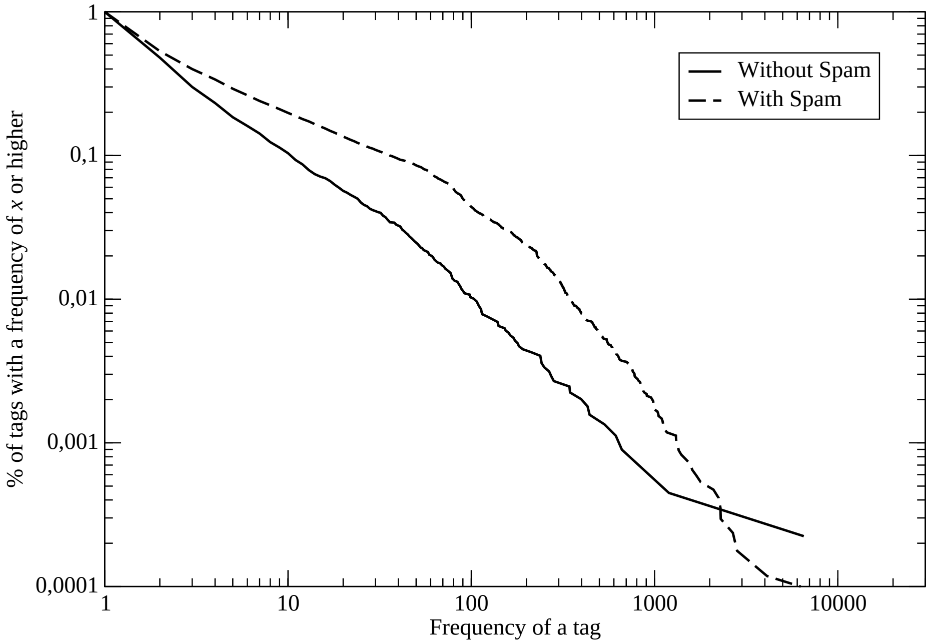

Figure 3.3: Bibsonomy: Tag frequencies for the *Software* stream with and without spam postings.

Stream	D	Parameters
Phone	0.0151	$I = 0.743; n = 1200; h = 8000$
List	0.025175	$I = 0.807; n = 1200; h = 8000$
Gadgets	0.037357	$I = 0.915; n = 665; h = 5000$
3d	0.026381	$I = 0.711; n = 2020; h = 14000$
Framework	0.024396	$I = 0.838; n = 1260; h = 8000$

Table 3.2: Parameter combinations for the Epistemic Model that resulted in the best reproduction of the tag frequencies.

Stream	D	Parameters
Phone	0.171231	$p = 0.046; \tau = 152; n_0 = 100$
List	0.176563	$p = 0.042; \tau = 149; n_0 = 100$
Gadgets	0.166466	$p = 0.036; \tau = 58; n_0 = 100$
3d	0.170942	$p = 0.042; \tau = 152; n_0 = 100$
Framework	0.150115	$p = 0.021; \tau = 157; n_0 = 100$

Table 3.3: Parameter combinations for the Yule-Simon Model with Memory that resulted in the best reproduction of the tag frequencies.

on streams from Bibsonomy.

Model-based Features

Based on the observations of the quantitative evaluation of the Epistemic Model, we designed several features for detecting spam in the data sets. These features compare for each user u the functions $g_u(x)$ and $g'_u(x)$ which give how often u uses tags which occur at least x times in the real or simulated co-occurrence streams. The values of the functions are computed as follows:

- **Extracting Co-occurrence Streams** In a first step, we extracted the co-occurrence streams of all tags t which occur at least 100 times in the training data set of the RSDC spam challenge.[6] This ensures, that the extracted co-occurrence streams have a certain length. All in all, we extracted 10,874 co-occurrence streams for the training data set and 10,874 co-occurrence streams for the test data set of the spam challenge.

- **Simulating Co-occurrence Streams** In a second step, we used the Epistemic Model for simulating each co-occurrence stream. During the simulation, we took the order of postings and posting lengths from the real stream. For example, when simulating the first posting of the *ajax* stream, we took from the real stream the information about who did the first posting and how many tag assignments are contained in the posting. The Epistemic Model was then used for simulating which tags the user assigned in his posting.

- **Computing Tag Frequencies of a User** In a third step, we computed the functions $f_{u,t}(x)$ and $f'_{u,t}(x)$ which represent for each user u how often he used in the co-occurrence stream of tag t a tag which occurs at least x times in that stream. $f_{u,t}(x)$ represents the information from the real streams and $f'_{u,t}(x)$ the information from its simulated counterpart. These functions were then further aggregated to the user specific functions $g_u(x) = f_{u,t_1}(x) + \cdots + f_{u,t_i}(x)$ and $g'_u(x) = f'_{u,t_1}(x) + \cdots + f'_{u,t_i}(x)$.

[6] http://www.kde.cs.uni-kassel.de/ws/rsdc08/

During the quantitative evaluation of the Epistemic Model we observed that removing spammers from a co-occurrence stream leads to a decreased probability of observing tags with specific frequencies. For example, in case of the *phone* stream of Delicious (see Fig. 3.2) it affected tags with a frequency between 5 and 30 while in case of the *software* stream of Bibsonomy (see Fig. 3.3) it affected tags with a frequency between 3 and 2000. Based on these observations, the following 5 features were computed for each user in the extracted co-occurrence streams:

1. **Sum of Distances** For this feature, we calculated the distance between the real and the simulated frequencies for all $5 \leq x \leq 200$ and summed them up. The feature value is computed as follows: $\sum g_u(x) - g'_u(x)$. Based on the observations from the evaluation of the Epistemic Model, we expect that many spammers will have feature values > 0 while for non-spammers we expect values ≤ 0. This prediction can be confirmed by looking at the distribution of feature values in the training and test data which is amongst others shown in Fig. 3.4 and 3.5.

2. **Tendency Ratio** This feature counts how often the real frequencies $g_u(x)$ are lower than the simulated frequencies $g'_u(x)$ and how often they are higher. The feature value is then computed as follows: $\frac{\#(g_u(x) < g'_u(x))}{\#(g_u(x) \leq g'_u(x)) + \#(g_u(x) > g'_u(x))}$. Like all other features, the region of interest is restricted to $5 \leq x \leq 200$ because here the most significant distances can be observed. We expect that spammers to have a lower tendency ratio than non-spammers because for spammers $g_u(x)$ will less often be lower than $g'_u(x)$. This prediction can be confirmed by looking at the distribution of feature values shown in Fig. 3.4.

3. **Average of Distances** This feature is very similar to the *Sum of Distances* feature but instead of summing the single distances, it computes the average value of distances for $5 \leq x \leq 200$. For non-spammers we expect an average value of the distances ≤ 0 and for spammers > 0. This prediction can be confirmed by looking at the distribution of feature values shown in Fig. 3.5.

4. **Standard Deviation of Distances** This feature computes the standard deviation of the single distances for $5 \leq x \leq 200$. No specific prediction about the distribution of feature values can be made based on the observations in the evaluation of the Epistemic Model. But if spammers and non-spammers have different kinds of distributions with regard to the distances, it can be expected that this will also show up in different standard deviations. This can also be confirmed by looking at the distribution of feature values shown in Fig. 3.6.

5. **Tag-count Ratio** For this feature, we compare the size of the vocabulary of a user in the real co-occurrence streams and the simulated co-occurrence streams, i. e. how many different tags he used in the overall stream. The feature is computed as follows: $\frac{Realvocabularysize}{Predictedvocabularysize}$. If the Epistemic Model correctly captures the dynamics in tagging systems, we expect that at least for non-spammers the values of this feature will be around 1. For spammers, we would expect to have values > 1 because it is part of their strategy to assign many different tags to a resource so that it is returned for many different search queries in the tagging system. Thus, also their vocabulary size will be much larger than of a non-spammer. This prediction can only be partially confirmed by the distribution of feature values shown in Fig. 3.7. It shows that quite most of the spammers have a tag-count ratio < 1 and that also the majority of non-spammers have a ratio which is < 0.5. It thus seems that the Epistemic Model has weaknesses in correctly predicting the vocabulary sizes of non-spammers in co-occurrence streams.

	Spammers	Non-Spammers
Training	10476	562
Test	7034	171

Table 3.4: Number of spammers and non-spammers in the used training and test data set.

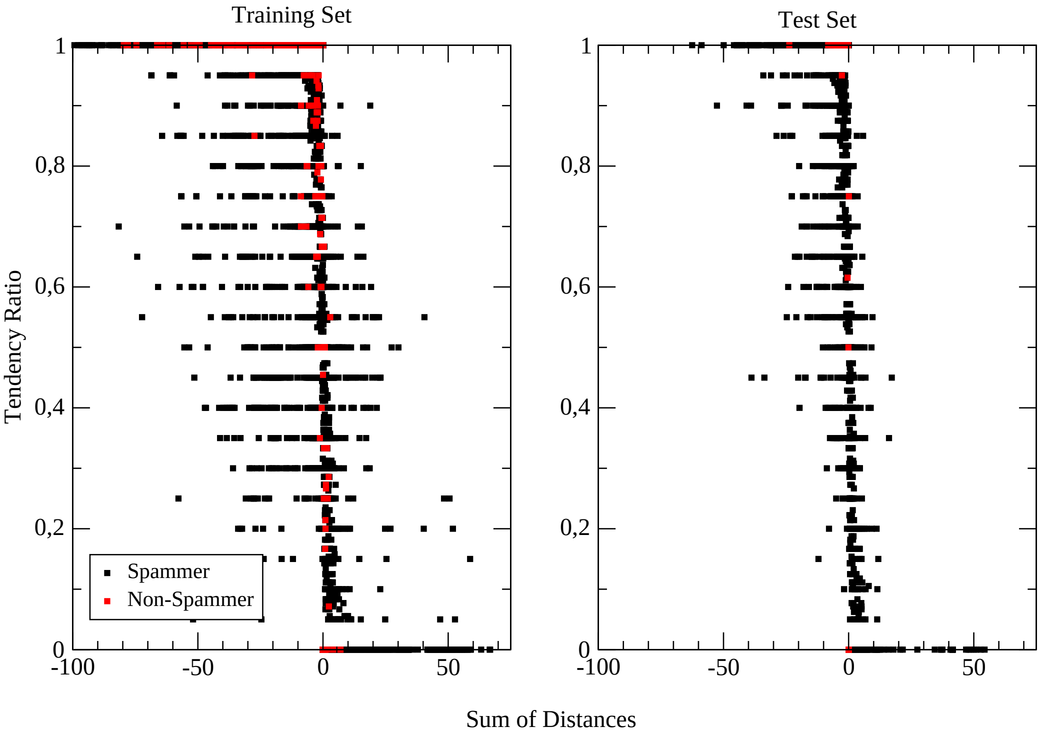

Figure 3.4: Scatter plot of the values for the *Sum of Distances* and *Tendency Ratio* features.

Evaluation Setup

For the evaluation of the model-based spam detection features, we first restricted the training data set of the spam challenge to users which appeared in the 3 month period between January 1st 2008 and March 31st 2008. By restricting the training data set, we ensured that the users from the training data set are approximately of the same age as the users in the test data set which covers the 3 month period from April 1st 2008 to June 30th 2008. This is necessary because the feature values are partially dependent on the age of a user (i. e. how long he already posts new resources to the tagging system). Tab. 3.4 shows the number of spammers and non-spammers in the used training and test data set.

Subsequently, we computed all feature values for the users contained in the co-occurrence streams of the training and test data set. In Fig. 3.4–3.7, exemplary feature values for the users in the training and test data are shown. Already these scatter plots show that the features are able to separate spammers from the set of non-spammers. But it also becomes obvious that a larger portion of the spammers have feature values which are very similar to that of non-spammers.

Evaluation Results

For the evaluation, we trained a SVM classifier on the previously described training data and all of the model-based feature values. Subsequently, we predicted the spammers and non-spammers by applying the trained SVM on the test data. For evaluating the approach we used the ROC curves and the AUC value as proposed in (Fawcett, 2006). These measures were also used in the spam

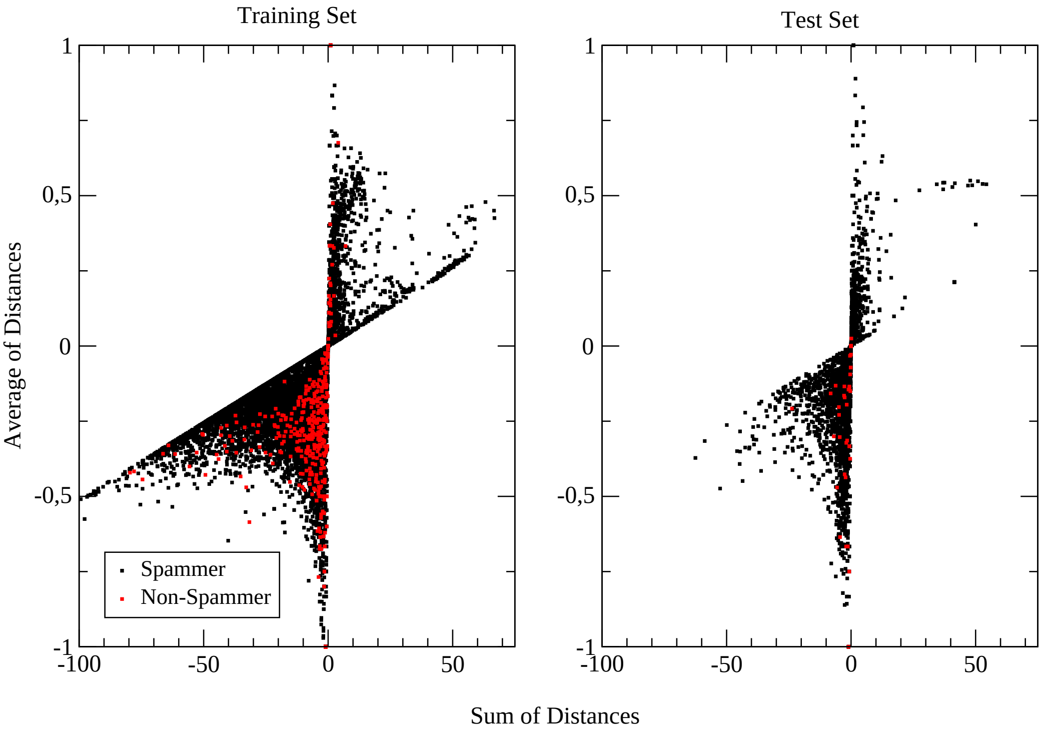

Figure 3.5: Scatter plot of the values for the *Sum of Distances* and *Average of Distances* features.

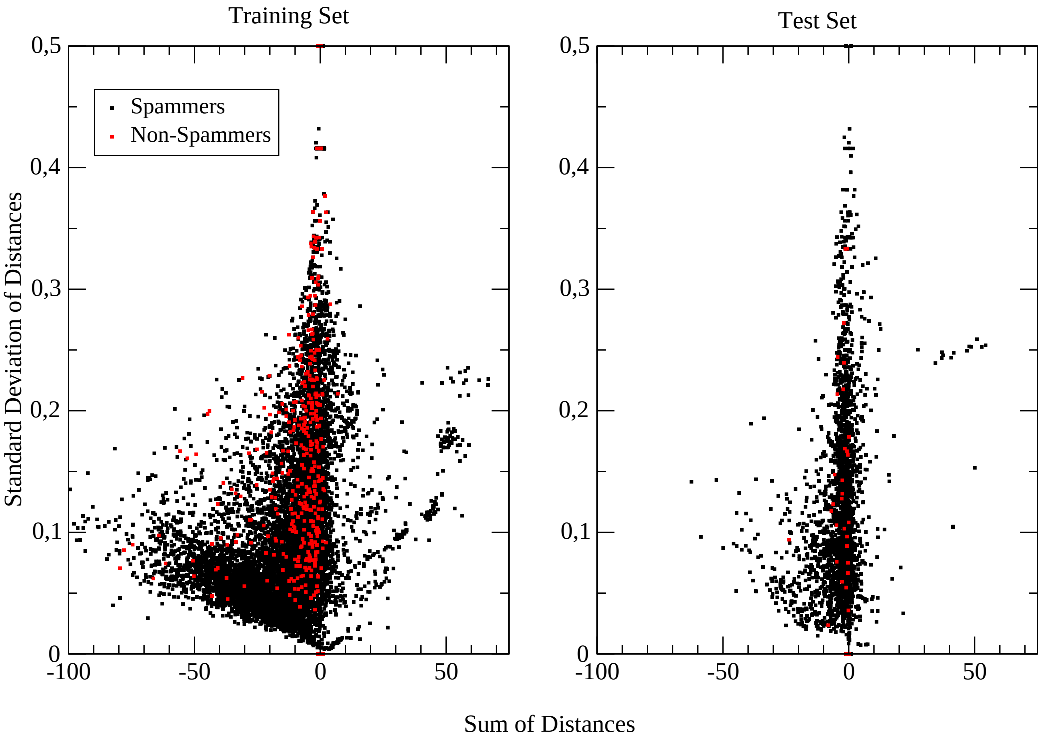

Figure 3.6: Scatter plot of the values for the *Sum of Distances* and *Standard Deviation of Distances* features.

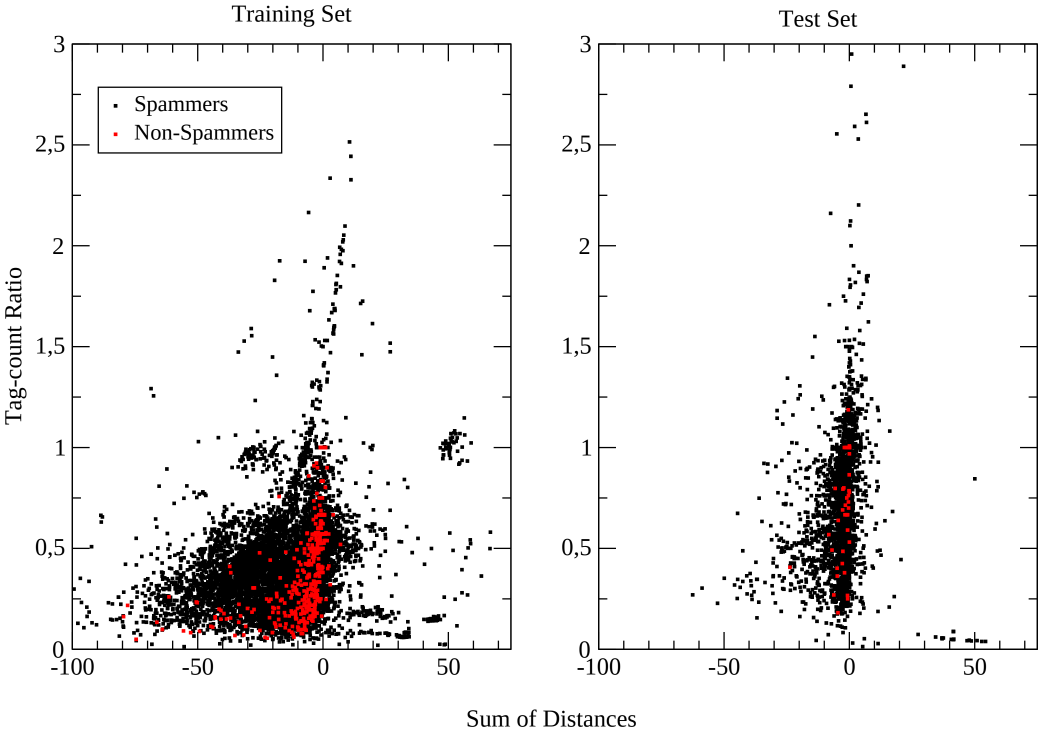

Figure 3.7: Scatter plot of the values for the *Sum of Distances* and *Tag-count Ratio* features.

challenge.

Furthermore, we use the weighted F_β-measure for finding the optimal parameter values during training the SVM classifier and for comparing the results. The F_β-measure is defined as follows:

$$F_\beta = \frac{(1+\beta^2) * Precision * Recall}{\beta^2 * Precision + Recall} \tag{3.2}$$

We use the weighted F_β-measure because we want to express that it is more important for us to correctly detect the non-spammers and leave them in the system than to remove almost all spammers from the system. Thus, we assign a higher cost to misclassifying non-spammers. In our evaluation, we use $\beta = 5$. Furthermore, we use the non-spammers as the positive class during calculating the F_β-measure. This ensures that the F_β-measure orders the different system states according to our intuition:

1. A spam filter which marks all users as spammers should get the lowest possible value of the evaluation measure because than the filtered system will be empty and completely useless for a user of the system. If we use the non-spammers as the positive class, the precision and recall will be 0 in this case and thus also F_β will be 0.

2. A spam filter which marks all users as non-spammers should get a value which is dependent on the ratio between spammers and non-spammers in the system. For example, if the system originally contains a very low number of spammers than already the unfiltered system will be perceived as quite good by the user. But if the system originally contains a very high number of spammers, than the unfiltered system will be perceived worse but still better than the system which is completely empty (i. e. if all users would have been marked as spammers and removed from the system).

3. A spam filter which correctly marks all spammers as spammers and all non-spammers as non-spammers should of course get the highest possible value from the evaluation measure.

	F_5	AUC	TP	FP	TN	FN	uncategorised
Training Data	0.6535	0.78	347	3750	5948	83	910
Test Data	0.4309	0.72	42	1793	1340	3	4027

Table 3.5: Evaluation results of the model-based filter for the training and test data of the spam challenge. For calculating the F_5-measure we treat uncategorised users as non-spammers.

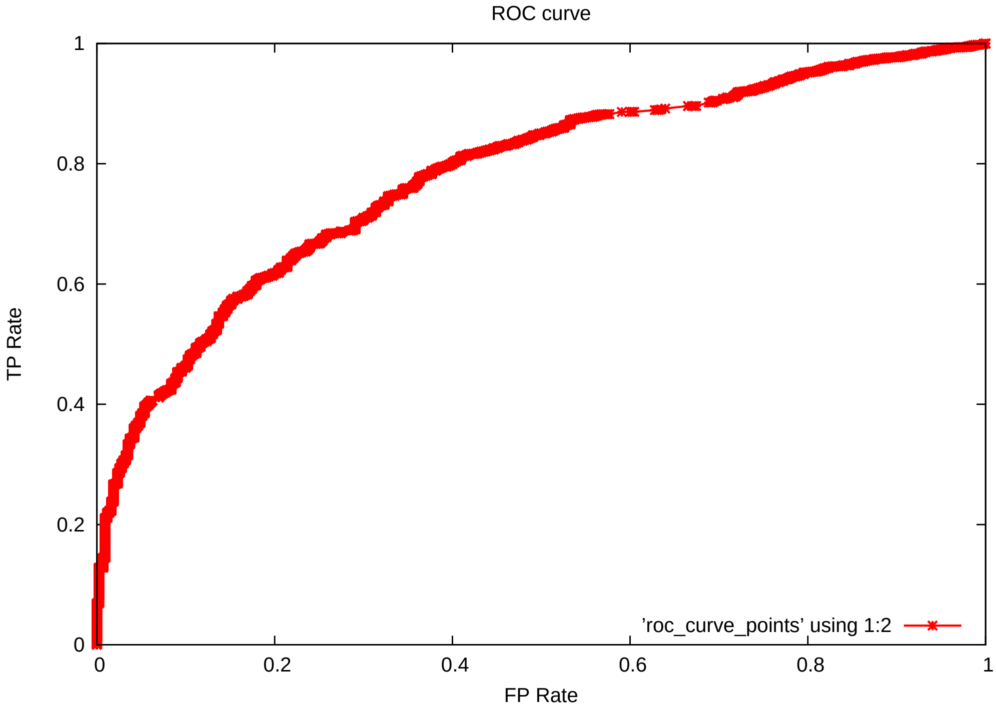

Figure 3.8: ROC curve for the training data set.

Only if we define the non-spammers as the positive class during calculating the F_β-measure, we will get this intuitive ordering of system states by the measure. For example, in case of the used training data set the F_5 measure would give a value of 0.5824 for the unfiltered system and for the test data set a value of 0.3872. As shown in Tab. 3.5, the model-based filter achieves for the training and the test data set a higher value and is thus better than not filtering at all. Furthermore, as we can see on the ROC curves in Fig. 3.8 and 3.9 the filter is also better than a filter which randomly guesses the class of a user in which case the ROC curve would be close to the diagonal line $y = x$ (see (Fawcett, 2006)).

Future Work

When comparing the evaluation results of the model-based features with the results of the winner of the spam challenge who achieved an AUC value of 0.9796 (see (Gkanogiannis and Kalamboukis, 2008)) it becomes obvious that the initial version of the model-based spam detection has to be further improved until it reaches equally good results. In this section, we will describe possible directions of future work.

One major disadvantage of the model-based features is that for the test set many users are not contained in the extracted co-occurrence streams and thus remain uncategorised. In a first step we will try to find ways how to reduce the number of uncategorised users. This will dramatically decrease the number of false positive categorizations because we no longer need to put the uncategorised users into the non-spammer category.

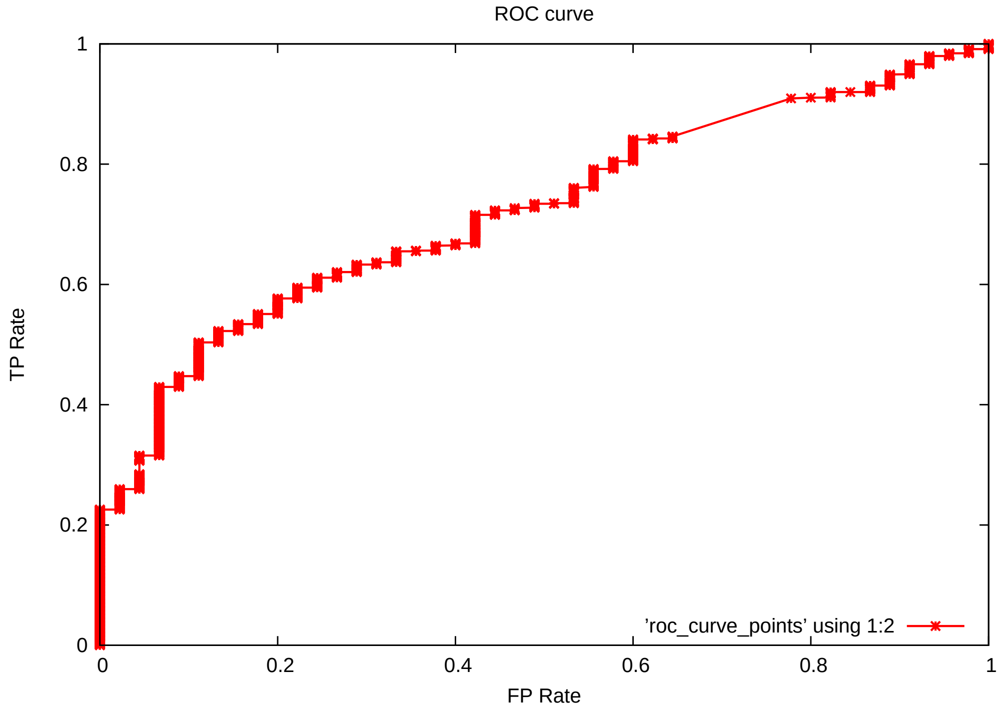

Figure 3.9: ROC curve for the test data set.

Another problem which becomes obvious by looking at the evaluation results is that there are larger variations in the distribution of the feature values during different periods of the system. We tried to counteract this problem by restricting the training set to a period of 3 months like the test set. But it seems that this is only partially successful. We have to take a closer look at the reasons for these larger variations. This may lead to interesting feedback for improving the Epistemic Model. Furthermore, we have to think about features which are more invariant over the different 3 month periods which can be extracted from the complete training set (for our evaluation, we only used one of the available 3 month periods in the training data).

All in all, it can be said that the model-based spam detection, as presented here, is in its current state only a proof of concept. Further improvements are required in order to close the gap to the current state of the art. In the future, model-based spam detection might show its benefits over the current state of the art especially in the following fields: (1) It requires less manually labelled training data in order to reach a good detection rate, and (2) it better deals with newly introduced topics in tagging systems because it is independent of the actual meaning of a tag. Most of the current approaches depend on the meaning of tags because they learn which tags are more often used by spammers or respectively non-spammers. Such features are less robust with regard to newly introduced topics and to adapted tag usage of spammers.

Bibliography

Eytan Adar. User 4XXXXX9: Anonymizing query logs. In *Query Logs Workshop at WWW2006*, 2007.

Dominik Benz, Marko Grobelnik, Andreas Hotho, Robert Jäschke, Dunja Mladenic, Vito D. P. Servedio, Sergej Sizov, and Martin Szomszor. Analyzing Tag Semantics Across Collaborative Tagging Systems. In Harith Alani, Steffen Staab, and Gerd Stumme, editors, *Proceedings of the Dagstuhl Seminar on Social Web Communities*, 2008. URL http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=08391.

Alexander Budanitsky and Graeme Hirst. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–47, 2006. URL http://dblp.uni-trier.de/db/journals/coling/coling32.html#BudanitskyH06.

Robin Burke. Hybrid Recommender Systems, Survey and Experiments. *User Modeling and User Adapted Interaction*, 12(4):331–370, 2002.

Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic Dynamics and Collaborative Tagging. *Proceedings of the National Academy of Sciences (PNAS)*, 104:1461–1464, 2007a.

Ciro Cattuto, Christoph Schmitz, Andrea Baldassarri, Vito D. P. Servedio, Vittorio Loreto, , Andreas Hotho, Miranda Grahl, and Gerd Stumme. Network Properties of Folksonomies. *AI Communications Journal, Special Issue on "Network Analysis in Natural Sciences and Engineering"*, 20 (4):245–262, 2007b. ISSN 0921-7126. URL http://www.kde.cs.uni-kassel.de/stumme/papers/2007/cattuto2007network.pdf.

Ciro Cattuto, Christoph Schmitz, Andrea Baldassarri, Vito D. P. Servedio, Vittorio Loreto, Andreas Hotho, Miranda Grahl, and Gerd Stumme. Network Properties of Folksonomies. *AI Communications Journal, Special Issue on Network Analysis in Natural Sciences and Engineering*, 20(4): 245–262, 2007c. ISSN 0921-7126. URL http://www.kde.cs.uni-kassel.de/stumme/papers/2007/cattuto2007network.pdf.

Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic Analysis of Tag Similarity Measures in Collaborative Tagging Systems. In *Proceedings of the 3rd Workshop on Ontology Learning and Population (OLP3)*, pages 39–43, Patras, Greece, July 2008a. ISBN 978-960-89282-6-8. URL http://olp.dfki.de/olp3/. ISBN 978-960-89282-6-8.

Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In Amit P. Sheth, Steffen Staab, Mike Dean, Massimo Paolucci, Diana Maynard, Timothy W. Finin, and Krishnaprasad Thirunarayan, editors, *The Semantic Web – ISWC 2008, Proc.Intl. Semantic Web Conference 2008*, volume 5318 of *LNCS*, pages 615–631, Heidelberg, 2008b. Springer. URL http://dx.doi.org/10.1007/978-3-540-88564-1_39.

Daniel Chandler. *Semiotics: The Basics*. Taylor & Francis, second edition, 2007.

William J. Conover. *Practical Nonparameteric Statistics*. John Wiley, 3rd edition, 1999.

Ferdinand de Saussure. *Cours de linguistique générale*. v.C. Bally and A. Sechehaye (eds.), Paris/Lausanne, 1916. English translation: Course in General Linguistics. London: Peter Owen, 1960.

Klaas Dellschaft and Steffen Staab. An Epistemic Dynamic Model for Tagging Systems. In *HYPERTEXT 2008, Proceedings of the 19th ACM Conference on Hypertext and Hypermedia*, 2008.

Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

J. R. Firth. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32, 1957.

Anestis Gkanogiannis and Theodore 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.

Z. S. Harris. *Mathematical Structures of Language*. Wiley, New York, 1968.

Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges. *IEEE Internet Computing*, 11(6): 36–45, 2007. ISSN 1089-7801. doi: http://dx.doi.org/10.1109/MIC.2007.125. URL http://portal.acm.org/citation.cfm?id=1304062.1304547&coll=GUIDE&dl=.

Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information Retrieval in Folksonomies: Search and Ranking. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 411–426, Heidelberg, 2006a. Springer. URL http://.kde.cs.uni-kassel.de/hotho.

Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information Retrieval in Folksonomies: Search and Ranking. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 411–426, Heidelberg, June 2006b. Springer. URL http://www.kde.cs.uni-kassel.de/stumme/papers/2006/hotho2006information.pdf.

Jens Illig. Machine Learnability Analysis of Textclassifications in a Social Bookmarking Folksonomy. Bachelor thesis, University of Kassel, Kassel, 2008.

Jens Illig, Andreas Hotho, Robert Jäschke, and Gerd Stumme. A Comparison of content-based Tag Recommendations in Folksonomy Systems. In *Postproceedings of the International Conference on Knowledge Processing in Practice (KPP 2007)*. Springer, 2009 (to appear).

Robert Jäschke, Leandro Balby Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. Tag Recommendations in Folksonomies. In *Proc. PKDD 2007*, volume 4702 of *Lecture Notes in Computer Science*, pages 506–514, Berlin, Heidelberg, 2007. Springer. ISBN 978-3-540-74975-2. URL http://dx.doi.org/10.1007/978-3-540-74976-9_52.

Robert Jäschke, Beate Krause, Andreas Hotho, and Gerd Stumme. Logsonomy – A Search Engine Folksonomy. In *Proceedings of the Second International Conference on Weblogs and Social Media(ICWSM 2008)*. AAAI Press, 2008a. URL http://www.kde.cs.uni-kassel.de/hotho/pub/2008/Krause2008logsonomy_short.pdf.

Robert Jäschke, Leandro Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme. Tag Recommendations in Social Bookmarking Systems. *AI Communications*, 21(4):231–247, 2008b. ISSN 0921-7126. doi: 10.3233/AIC-2008-0438. URL http://dx.doi.org/10.3233/AIC-2008-0438.

Jay J. Jiang and David W. Conrath. Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics (ROCLING)*. Taiwan, 1997.

Beate Krause, Andreas Hotho, and Gerd Stumme. A Comparison of Social Bookmarking with Traditional Search. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *30th European Conference on IR Research, ECIR 2008*, volume 4956 of *Lecture Notes in Computer Science*, pages 101–113, Glasgow, UK, April 2008a. Springer. ISBN 978-3-540-78645-0.

Beate Krause, Robert Jäschke, Andreas Hotho, and Gerd Stumme. Logsonomy - social information retrieval with logdata. In *Hypertext*, pages 157–166, 2008b.

Beate Krause, Robert Jäschke, Andreas Hotho, and Gerd Stumme. Logsonomy - Social Information Retrieval with Logdata. In *HT '08: Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, pages 157–166, New York, NY, USA, 2008c. ACM. ISBN 978-1-59593-985-2. doi: http://doi.acm.org/10.1145/1379092.1379123. URL http://portal.acm.org/citation.cfm?id=1379092.1379123&coll=ACM&dl=ACM&type=series&idx=SERIES399&part=series&WantType=Journals&title=Proceedings%20of%20the%20nineteenth%20ACM%20conference%20on%20Hypertext%20and%20hypermedia.

Beate Krause, Christoph Schmitz, Andreas Hotho, and Gerd Stumme. The Anti-Social Tagger - Detecting Spam in Social Bookmarking Systems. In *AIRWeb '08: Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 61–68, New York, NY, USA, 2008d. ACM. ISBN 978-1-60558-159-0. doi: http://doi.acm.org/10.1145/1451983.1451998. URL http://airweb.cse.lehigh.edu/2008/submissions/krause_2008_anti_social_tagger.pdf.

Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme. Evaluating Similarity Measures for Emergent Semantics of Social Tagging. In *18th International World Wide Web Conference*, pages 641–641, April 2009. URL http://www2009.eprints.org/65/.

L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. In *WWW'98*, pages 161–172, Brisbane, Australia, 1998. URL http://dbpubs.stanford.edu:8090/pub/1999-66.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian 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.

Janko Röttgers. Am Ende der Flegeljahre — Das Web 2.0 wird erwachsen. *c't 25/2007*, page 148, 2007.

Gerard Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

Gene Smith. Search tagging, 2005. http://atomiq.org/archives/2005/05/search_tagging.html.

D. J. Watts. *Small worlds : the dynamics of networks between order and randomness*. 1999.