Exploring Social Annotations for Web Document Classification

Michael G. Noll
Hasso-Plattner-Institut,
University of Potsdam
14440 Potsdam, Germany
michael.noll@hpi.uni-potsdam.de

Christoph Meinel
Hasso-Plattner-Institut,
University of Potsdam
14440 Potsdam, Germany
meinel@hpi.uni-potsdam.de

ABSTRACT

Social annotation via so-called collaborative *tagging* describes the process by which many users add metadata in the form of unstructured keywords to shared content. In this paper, we explore and study social annotations and tagging with regard to their usefulness for web document classification by an analysis of large sets of real-world data. We are interested in finding out which kinds of documents are annotated more by end users than others, how users tend to annotate these documents, and in particular how this user-generated folksonomy compares with a top-down taxonomy maintained by classification experts for the same set of documents. We describe what can be deduced from the results for further research and development in the areas of document classification and information retrieval.

Categories and Subject Descriptors

I.7.4 [**Document and Text Processing**]: Electronic Publishing; I.7.1 [**Document and Text Processing**]: Document and Text Editing—*Document Management*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering*; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia—*User issues, Navigation*

General Terms

Experimentation, Human Factors, Measurement

Keywords

classification, folksonomy, semantic web, social annotation, taxonomy

1. INTRODUCTION

Social annotation via so-called collaborative *tagging* describes the process by which many users add metadata in the form of unstructured keywords to shared content. The recent success of web services with such a tagging component like del.icio.us or Flickr has shown the great potential of this simple yet powerful approach to add metadata to documents, and has provided a plethora of user-supplied metadata about web content for everyone to leverage. Unlike traditional categorization systems, the process of tagging is nothing more than annotating documents with a flat, unstructured list of keywords called tags. Users can browse or query documents by tags, and so-called *tag clouds* provide a rudimentary but often sufficient way to find popular and interesting content. Several studies have already analyzed the semantic aspects of tagging and why it is so popular and successful in practice [6, 2, 5, 1]. A common argument is that tagging works because it strikes a balance between the individual and the community: the cost of participation, in particular entering data, is low for the individual, and tagging a document benefits both the individual and the community.

In this paper, we explore and analyze social annotations and tagging with regard to classification of web documents. We are interested in finding out which kinds of documents are annotated more by end users than others, how users tend to annotate these documents, and how this *ad hoc* classification with its free tagging vocabulary differs from classification by experts with a well-defined, controlled vocabulary and category structure.

The rest of this paper is organized as follows. In section 2, we briefly outline the different types and forms of metadata available for describing, classifying and annotating web documents. In section 3, we describe how we obtained real-world data for building the experimental data set used for our analysis. We report and discuss the results of our experiments in section 4, and give a summary of our findings in section 5.

2. WEB DOCUMENTS

2.1 Metadata provided by authors and publishers

The traditional and most common method of adding metadata to web documents is described in the (X)HTML standards[1], which define elements and attributes for specifying metadata in the document source itself. This implies that this kind of metadata is provided by the authors or publishers of online content. For example, authors should use the TITLE element to identify the contents of a document.

[1]http://www.w3.org/MarkUp/

SAC'08 March 16-20, 2008, Fortaleza, Ceará, Brazil

While adding a title to a document is common in practice as we will see, other metadata such as META keywords or META description is often neglected by authors. The purpose of these elements and attributes has been to help users find relevant content. However, search engines like Yahoo or Google often do not trust and therefore discard HTML metadata elements in web documents because those have been abused by spammers in the past [9]. Since search engines do not guarantee to honor this data at all, an incentive for authors to add this information is often missing, thus lowering the amount of available data for proper document classification.

2.2 Classification and categorization by expert editors

The Open Directory[2] is by its own account "the largest, most comprehensive human-edited directory of the Web" and comprises almost five billion web documents in 590,000 categories. It is constructed and maintained by an international community of volunteer editors which evaluate and categorize each web document into one or more predefined categories. Prospective editors have to file an application form which includes a categorization test, and senior Open Directory editors must review and evaluate the application before the new candidate can become an editor. This review process helps to ensure that web documents will properly organized and categorized in the catalogue by all its editors.

In this paper, we consider the Open Directory's categorization of web documents as *expert categorization* (similar to taxonomies) because it is based on a controlled structure of predefined category hierarchies which is used by a "peer-reviewed" group of collaborating human editors with a common goal. We will compare this expert classification with the uncontrolled folksonomy of social annotations at del.icio.us, which is "built" by non-expert end users in a free and unrestricted way and without a common goal.

2.3 Metadata provided by end users

Social bookmarking and tagging services such as del.icio.us, CiteULike and Connotea take a different approach. Here, the recipients and readers of online content supply metadata about web documents in a collaborative fashion. This metadata is not part of the document source but stored at and available from external web services. In the case of del.icio.us, the metadata of a web document is stored as bookmarks of the document's URL with additional tag information. Organizing and sharing bookmarks with the help of tags mitigates some of the problems of traditional, hierarchical bookmarking (for example, where to file a bookmark if it fits to more than one category without filing it twice) and increases findability.

Basically, tagging can be interpreted as a relation

$$R_{tagging} \subseteq D \times U \times T \quad (1)$$

where D is the set of documents, U the set of users and T the set of tags. The act of bookmarking a document with tags by a user creates one or more tuples as described by the relation above. Documents are identified by their URLs and users by their account name in the bookmarking service.

Golder and Huberman [2] and Ames and Naaman [1] analyzed the structure and dynamical aspects of collaborative tagging systems and user motivations for annotation of resources. The evolution of such systems depends on a variety of factors such as the user interface, tagging rights, user incentives, social connectivity, and the personal characteristics of individual users as described in [11, 5].

The social bookmarking service del.icio.us, which we used as information source for user-supplied metadata in this paper, provides a *free-for-all* tagging system (to use the terms of Marlow et al. [5]) in which users can freely annotate any document with as many tags as they want. The del.icio.us interface affords for *suggested-tagging*, i.e. it supports users in tagging documents by recommending tags and displaying a document's most popular tags.

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

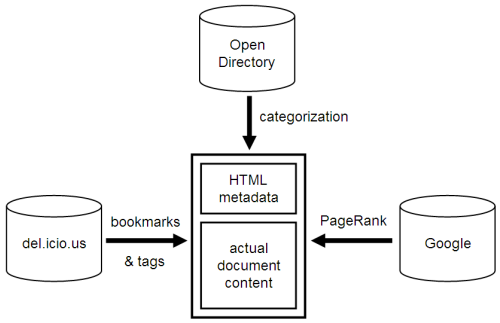

Figure 1: Information sources used for building the experimental data set used in this paper.

3. EXPERIMENTAL DATA SETS

We have created a data set called *DMOZ100k06* by building an initial random sample of 100,000 URLs from the Open Directory, which contained 4,818,944 URLs in over 590,000 categories in December 2006. The initial DMOZ100k06 data set is described in [**anonymized**]. For the work in this paper, we have updated and significantly extended the original data set by retrieving *all* bookmarking and tagging information for documents in the corpus[3] as well as integrating category information from the Open Directory into the data set. For each document in the sample, we retrieved the actual HTML document source from the WWW plus its Open Directory categorization, metadata from the social bookmarking service del.icio.us and from Google as shown in figure 1. We implemented custom software tools for this purpose which relied on the services' official APIs where possible and fell back to alternative techniques for situations where the APIs did not provide the required functionality. The updated data set is available on the author's home page.

4. RESULTS

The total data set was built from an initial random sample of 100,000 documents from which those documents were removed which could not be retrieved from the WWW. Of the remaining 97,574 documents, 18.7% have been bookmarked by a total of 165,192 users, and 17.8% are tagged with a total of 758,242 tag annotations[4] Details are shown in tables

[3]The previous data set contained only a document's so-called *common del.icio.us tags* only, i.e. the up to 25 most popular tags of the document, due to technical restrictions by del.icio.us.

[4]We discarded the special tags `system:unfiled` and

1 and 2.

Total documents	97,574	
Total users	165,192	
Total categories	115,458	unique: 84,663
Total bookmarks	282,529	
Total tags	758,242	unique: 63,594
Bookmarked documents	18,220	18.7%
Tagged documents	17,342	17.8%

Table 1: Overall statistics of the data set.

Looking at bookmarking behavior, a user bookmarked 1.7 documents on average in the data set, and if he decided to annotate the bookmark, he added 3.0 tags on average.

Looking at documents, the average Google PageRank of a document in the data set was about 3 of a maximum of 10. Due to the experimental setup[5], a document was assigned to *at least* one category by the Open Directory expert editors. If a document was bookmarked by end users, its probability to have at least one tag was 95.2%.

statistics per document	mean	std. dev.
Categories	1.18	0.47
Category depth	6.66	1.90
Bookmarks	2.90	71.35
Tags	7.77	190.96
Unique tags	1.90	17.28
PageRank	3.13	1.66

Table 2: Statistics per document in the data set.

4.1 Spatial granularity

We analyzed whether end users tend to bookmark and tag documents higher up or deeply within a website's content hierarchy. URL schemes such as HTTP contain names that can be considered hierarchical, and the components of the hierarchy are separated by a "/" delimiter character[6]. We therefore based the calculation of a URL's depth primarily on the "/" separator so that a top-level URL like

http://www.example.com/

would be assigned a depth of 0 while a URL such as

http://www.example.com/path/file.html

would be assigned a depth of 2, and so on. As shown in table 3, users tend to bookmark and tag top-level URLs rather than those documents located deeply in a website's content hierarchy.

Compared to the average URL depth of documents in the data set, users preferred to bookmark and tag URLs higher up in the hierarchy, in particular the home pages of websites. This is particularly interesting because intuitively, one might think that users would be more likely to bookmark those web documents for quick reference which are harder to find or access than others: documents with deep URLs are often more complicated to navigate to, it takes *per definitionem* longer to type the full URL, etc. Our results suggest that social data available for web document classification tends to gravitate towards the entry or top-level pages of websites so that an examination and processing of deeper documents might require leveraging information from "parent" content, in particular if direct information is not available at all.

URL depth	mean	std. dev.
All documents	1.06	1.74
Bookmarked documents	0.48	1.06
Tagged documents	0.48	1.05

Table 3: URL depth of documents in the data set.

We conducted a second analysis as comparison based on a data set containing 6,459 documents which were featured on the front page of del.icio.us over the course of several months. For this data set, the average URL depth was 2.32 with a standard deviation of 1.77. Annotated documents were located more deeply within a website's hierarchy compared to the findings in the DMOZ100k06 data set. We are still investigating whether this result is the effect of the recommendation algorithm of del.icio.us for featured documents or related to other phenomenons.

4.2 Classification granularity

We studied whether end users tend use tags to classify documents into broad or specific categories. For this, we matched user-supplied tags of a document against its categorization by the expert editors of the Open Directory. We used the Levenshtein distance [4] to vary and relax the matching conditions, so that small variations of tags such as singular-plural ("article" vs. "articles") or different languages ("music" vs. "músika") could be detected to a certain degree of accuracy.

A document in the Open Directory is categorized by one or more category hierarchies such as "arts > crafts > textiles > weaving". We analyzed at which hierarchy depth, or level, matches occurred and normalized the results so that the top category in a hierarchy, e.g. "arts", is represented by 0 and the leaf category by 1, e.g. "weaving". The detailed results are shown in figure 2. In our experiments, the mean normalized category depth for matches is 0.35 for exact matches between tags and categories, and increases when matching conditions are relaxed, e.g. 0.43 for a Levenshtein distance of up to 2 between tags and categories. This means that users tend to prefer broader categories to more specific ones, similar to the findings in [2, 7], even though preprocessing of tags for identifying variations such as singular-plural cases, translations or synonyms can increase the intersection between tags and categories in such a way that it shifts this bias to a more balanced ratio of broad and specific classification. Our results suggest that for document classification in general, tags may help more with broad categorization or clustering of documents rather than finding the specific "needle in the haystack".

4.3 Tag popularity and classification

We studied the popularity of user-supplied tags of a doc-

`imported` in our analysis. The former is a pseudo tag listed by del.icio.us for bookmarks without any user-supplied tags, the latter is automatically added to imported bookmark collections.

[5]All documents in the Open Directory are categorized.

[6]See RFC 3986 "Uniform Resource Identifier (URI): Generic Syntax" available at http://www.ietf.org/rfc/rfc3986.txt.

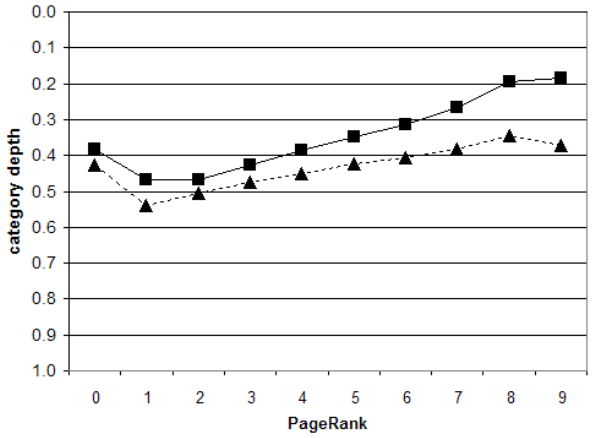

Figure 2: Normalized category depth for matches between tags and categories. A value of 0 denotes a root category ("broad"), a value of 1 a leaf category ("specific"). The solid and dotted lines show exact matches and matches for a Levenshtein distance of up to 2, respectively, between tags and categories.

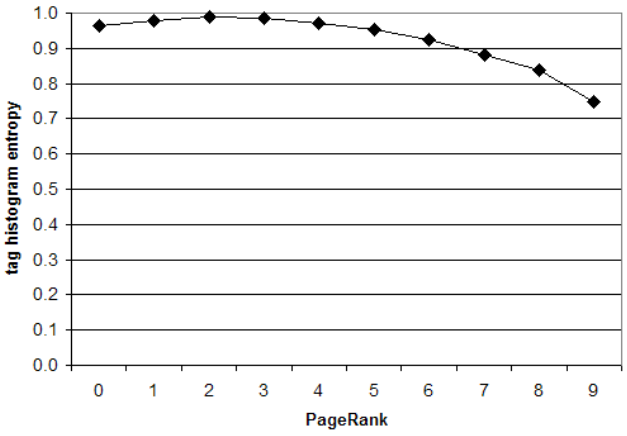

Figure 3: Normalized tag histogram entropy by PageRank. A value of 0 denotes no entropy (full consensus), a value of 1 maximum entropy (no consensus). A decreasing entropy means an increasing consensus.

ument with regard to their likeliness of matching the document's Open Directory categories assigned by expert editors. We describe the popularity of a tag by its tag count, i.e. the number of its annotations per document, and normalize it so that the least popular tag of a document is represented by 0 and the most popular tag by 1. The mean normalized popularity of tags matching a document's categories is 0.71 for exact matches, and decreases when matching conditions are relaxed, e.g. 0.55 for a Levenshtein distance of up to 2 between tags and categories. This means that more popular tags match better than less popular tags as long as no preprocessing is applied to identify variations of tags. It also shows that proper handling of tags can significantly help with extracting information from folksonomies. On the other hand, the shift of matching frequency towards less popular tags when relaxing matching conditions does not *necessarily* mean that non-matching popular tags provide new classification information which is not already represented in a document's categories (otherwise they would match) because the number of less popular tags is generally higher than popular ones.

Our experimental results confirm the intuitive feeling that an analysis of tag popularity and techniques such as thresholding can help with identifying those tags which provide the most relevant classification information. However, also less popular tags deserve a thorough examination as we will show in section 4.5.

4.4 Classification consensus

We studied the "consensus" of end users when adding tags to documents. We were interested in finding out whether and how much end users agree on classification by tagging even though they do not collaborate or share common goals in general. A document's tags and their tag counts can be considered as a "tag histogram", and the entropy of such an histogram of a document d can be computed by

$$E(d) = -\sum_{t_i \in T(d)} p(t_i|d) \log_2 p(t_i|d) \qquad (2)$$

where $T(d)$ is the set of tags with which document d has been annotated. The histogram entropy can be used as an indicator of the consensus of users with regard to classification. We normalize the entropy values so that a full consensus (zero entropy) is represented by 0 and a low consensus (maximum entropy) by 1. The results are shown in figure 3. The mean tag histogram entropy of a document is 0.95, which means a very low consensus. However, the entropy is negatively correlated with the number of users who tagged the document, and the number of tag annotations: the Pearson-r [10] is -0.35 and -0.34, respectively. We also found a negative though weaker correlation with a document's popularity as indicated by its PageRank: here, the Spearman-r is -0.24[7]. Though these results might suggest that higher and higher numbers of users or annotations will lead to higher consensus, a full consensus will most likely never been reached because of different interpretations and perceptions of documents due to varying user characteristics such as educational or cultural background or personal preferences. Because a high consensus could be reached more easily in a homogeneous sub-group of users or tags, we will try to analyze classification consensus more granularly in the future and differentiate between types of documents, tags, and users.

Our experiments show that the consensus of end users with regard to classification via tagging *increases* with a web document's popularity in the Internet and the more it is tagged and bookmarked by human users. It is important to note that a document's popularity is algorithmically estimated by its PageRank through analysis of hyperlinks created by other web document *authors* whereas bookmarks and tags are supplied by the *readers* of web documents, i.e. "average" end users. The two correlations for classification consensus of a document therefore refer to different dimen-

[7]Kendall-τ is -0.30.

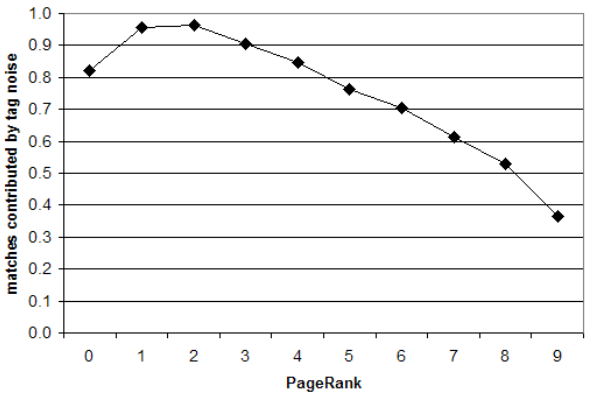

Figure 4: The share of matches with a document's content or metadata, contributed by tag noise. The decrease of matches for tag noise with a document's popularity could be related to the simultaneous increase of tagging "consensus" as shown in figure 3.

sions: the link structure of the World Wide Web and the behavior of users navigating through its contents.

4.5 Tag noise

There is a variety of different kinds of tags [2], and users have differing habits with regard to tagging vocabulary and usage [5]. For example, novice users might focus on annotating documents with tags useful only for themselves whereas more experienced users might have come to know the benefits of others' social annotations and therefore also add such tags that are deemed to be helpful for their local group of friends or the user community at large.

Techniques such as thresholding are often applied to discard rarely used tags because the usage and interpretation of these tags might be too volatile and unpredictable to be treated as a trustable piece of information. On the other hand, this "tag noise", i.e. such tags with which a document has been annotated just once, might as well provide valuable information. We showed in a previous work [7] that user-supplied tags in general provide additional metadata which is not already contained within a document's content or its metadata. With the extended DMOZ100k06 data set we prepared for this paper, we were able to compare a document's content and metadata with tag noise in particular. If we consider tags as an indicator of how end users *perceive* a web document, matching user-supplied tags and a document's content and metadata composed and supplied by its authors is an estimate of how well the intention and perception of a document's authors and its readers intersect.

Generally, the analysis of the updated data set confirms the trends of our previous study [7]: the body of a web document is matched significantly better by user-supplied tags than the metadata provided by the document's authors such as its title, META keywords or META description. Additionally, the more popular a document, the less likely are tags matching the document.

Tag noise in particular constitutes a large percentage of matches as shown in figure 4. The contribution of tag noise decreases with a document's popularity. This observation could be related to a simultaneous *increase* of tagging "consensus" (and also to the number of tag annotations in general) as described in section 4.4 and shown in figure 3: when more users agree on tags for annotating a document, the amount of tag noise decreases because a tag is used by more than one user. We therefore argue that it is worth further research to analyze what kind of and how information can be extracted from tag noise. In particular, our results suggest that tag noise does not only consist of "personal" tags that users add just for themselves like `toread` or `notfunny` but also provides helpful data for information retrieval and classification tasks in general.

5. CONCLUSIONS

In this paper, we measured and analyzed the characteristics of social annotations provided by end users with regard to their usefulness for web document classification. The most important results and findings of our study are summarized below.

Users tend to bookmark and tag top-level web documents rather than pages located deeply within a website's content hierarchy. The majority of user input available to classification tasks will therefore target the entry pages of websites whereas classification of deeper pages might require more direct content analysis (as with traditional machine classification); however, this analysis could be augmented with contextual information derived from user-supplied metadata of the parent pages higher up in the content hierarchy. Additionally, we have shown that tag popularity can indeed help with identifying those tags which provide the most relevant classification information. For example, a deep web document describing a multimedia playback device could be identified as iPod[8] product information if parent documents are primarily tagged with `apple`, `mac`, `itunes` by human users.

Users prefer broad terms rather than specific terms when tagging documents. The information derived from social annotations therefore seems to help more with broad classification of documents than with finding the specific "needle in the haystack". For example, we have shown in a previous work that social annotations can be used for disambiguation of search keywords and queries in the context of web search personalization [8].

We have found that consensus on social annotations increases with a web document's popularity in the Internet and with the number of users who bookmark and tag it. This observation might serve as a starting point to identify trust metrics of social annotations, i.e. how reliable such information is as input for classification tasks. On the other hand, we also observed that tag noise - the opposite of popular tags - provides helpful data for general information retrieval and classification tasks even though it is often believed as consisting mostly of personal, individual annotations.

6. RELATED WORK

The work of Golder and Huberman [2] gives a detailed analysis of social annotations in del.icio.us, studying user activity, tag frequencies, and trends in bookmarking and tagging. A particularly interesting observation was the stabilization of tag proportions for a specific web document after a certain amount of bookmarks. Our analysis of classification consensus in section 4.4 extends these findings and

[8]The iPod is a brand of portable media players made by Apple Inc.

takes web page popularity and the number of users and tag annotations into account. Wu et al. [12] study social annotations in the context of the semantic web. By mapping the entities in folksonomies to a conceptual space, they derive emergent semantics from social annotations, analyze the ambiguity of different tags on "knowledge dimensions" in the conceptual space and apply their results to model semantic search and discovery of web content.

Grosky et al. [3] propose that the semantics of a web page are not only determined by its authors but also how its readers perceive and use the web page. If so, document classification can benefit from metadata derived from social annotations because the latter help to capture this type of contextual information. In a previous work, we conducted a quantitative and qualitative analysis of metadata and information provided by the authors and publishers of web documents and compared it with metadata supplied via social annotations by readers of the same content [7]. We found that popular web documents - measured by their Google PageRank - are bookmarked and tagged much more frequently than less popular documents, and that distributions of bookmarks and tags for documents show power law curves, similar to the findings of [2] for users. Our experiments suggest that tags provide additional information about a web document, which is not directly contained within its content. An example of putting the results of this paper and [7, 3] into practice is [8] where we describe how to design and implement a new approach to web search personalization based on social annotations.

7. REFERENCES

[1] M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. In *Proceedings of CHI '07*, 2007.

[2] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[3] W. I. Grosky, D. V. Sreenath, and F. Fotouhi. Emergent semantics and the multimedia semantic web. *SIGMOD Rec.*, 31(4):54–58, 2002.

[4] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, February 1966.

[5] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proceedings of HT '06*, pages 31–40, 2006.

[6] A. Mathes. Folksonomies - cooperative classification and communication through shared metadata. Technical report, UIC, 2004.

[7] M. G. Noll and C. Meinel. Authors vs. readers: A comparative study of document metadata and content in the www. In *Proceedings of 7th Int'l ACM Symposium on Document Engineering '07*, pages 177–186, Canada, 2007.

[8] M. G. Noll and C. Meinel. Web search personalization via social bookmarking and tagging. In *Proceedings of 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*, Busan, South Korea, 2007.

[9] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proc. of WWW '06*, pages 83–92, Scotland, 2006.

[10] J. A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, Canada, 1995.

[11] 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 CSCW '06*, pages 181–190, 2006.

[12] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 417–426, 2006.