Retrieval Effectiveness of Tagging Systems

Isabella Peters, Laura Schumann, Jens Terliesner, Wolfgang G. Stock
Heinrich-Heine-University
Department of Information Science
Universitätsstraße1, 40225 Düsseldorf, Germany
{isabella.peters; laura.schumann; jens.terliesner}@uni-duesseldorf.de

ABSTRACT
Social tagging is a widespread activity for indexing user-generated content on Web services. This paper summarizes research on folksonomies and their retrieval effectiveness. A TREC-like retrieval test was conducted with tags and resources from the social bookmarking system delicious, which resulted in recall and precision values for tag-only searches. Moreover, several experimental tag-based databases (i.e., power tags, Luhn-tags) have been tested regarding their retrieval effectiveness. Test results show that folksonomies work best with short queries although recall values are high and precision values are low. Here, a search function "power tags only" greatly enhances precision values.

Keywords
Information retrieval, folksonomies, tagging systems, recall, precision, F-measure, power tags, Luhn-tags.

INTRODUCTION

Folksonomies and social tagging may now be called established methods to organize and index user-generated content like photos or bookmarks. Although tags primarily serve individual needs in managing personal collections of digital resources on the Web (Peters, 2009) and act as "hooks" for retrieving resources which have once been found from one person, in sum they help all users of the Web service accessing saved resources. Therefore, the study presented in this poster aimed at finding out how effective tags can be used for retrieving relevant resources in tagging systems.

The paper is structured as follows: first we will situate our study in the context of the retrieval research on folksonomies and we will motivate our approach. Then we will explain the basic terminology and ideas behind the conducted research. After that we give details about the methods used. The paper ends with the presentation and discussion of results as well as implications for future work.

ASIST 2011, October 9-13, 2011, New Orleans, LA, USA.

RELATED WORK

Retrieval research in context of folksonomies mainly looks at overlaps between tag vocabulary and controlled vocabulary (Kipp, 2011), relevance ranking based on folksonomic structures (Hotho et al., 2006) or tags as search functionality in combination with search in bibliographic metadata and controlled keywords (Kipp & Campbell, 2010). Closest related to our study is the work of Lu and Kipp (2010) who also carried out a retrieval test with folksonomies. They found that tags as additional search aids to other metadata improve mean average precision by nearly 5%. Moreover, searching with tags retrieves more relevant resources at the top of the result list.

Motivation

The review of related work reveals that some efforts are made in researching and testing folksonomies' effects in information retrieval but many questions are still open. First of all there is the lack of standardized and high quality test collections which enable structured testing and comparison of test results. Although tagging is widely accepted among users, only a small portion of Web content has been tagged yet. This portion becomes smaller again when it is required that each Web resource should be tagged from at least x people to reflect collective intelligence of the users. On the other hand studies mostly explore recall of folksonomies or the tags' strength for discriminating resources, but precision of search hits is often not part of the agenda as it depends on relevance judgments for retrieved resources. Or, tags are evaluated in combination with other metadata which does not reveal what are the core benefits of tags in retrieval.

We aimed at creating a high quality test collection subject to TREC standards which serves as test bed for our and future folksonomy retrieval tests. Moreover, we focus on tags-only searches to find out which kinds of tags work best in retrieval. Our considerations result in two research questions leading our study: (1) How do tags effect retrieval in folksonomies?, (2) Do particular portions of tags (power tags, Luhn-tags) enhance precision of search results?

BASIC TERMINOLOGY AND ASSUMPTIONS

This section describes the terminology used in the study.

Folksonomy vs. Docsonomy

The folksonomy of a Web service F_{web} can be defined as a tuple $F_{web} := (U, T, R, Y)$ where U, T, R are finite sets of the elements user names U, tags T and resource identifiers R, and Y is a ternary relation between them, i.e., $Y \subseteq U \times T \times R$ whose elements are called tagging actions (Hotho et

al., 2006). In broad folksonomies the multiple assignment of a single tag to a particular resource is allowed. In contrast to this, narrow folksonomies only allow the addition of new tags to the resource (Vander Wal, 2005). The major difference between both types of folksonomies is that in broad folksonomies tag frequency distributions on resource level can be observed. Folksonomies for a single resource are called docsonomies D_{web} as they comprise only the tags assigned to this particular resource of the Web service. That is why D_{web} is defined as multiset $D_{web} := (T, R, Z)_b$ where $Z \subseteq T \times R$ and $Z := \{(t, r) \in T \times R | (t, r) \in Z\}, b \in \mathbb{N}^+$. D_{web} becomes a docsonomy for a particular resource (D_r) by substituting Z with Z_r where $r \in R$. The differentiation between F_{web} and D_{web} and the notion of D_{web} as multiset is important for our discussion of tag distributions.

Tag Frequency Distributions of Docsonomies

Tag frequency distributions on resource level reflect how often a particular tag was assigned to the resource. To be able to include tag frequencies docsonomies are defined as multisets as every tag of the tag set may be assigned more than once. Frequency distributions of docsonomies appear in different forms (Stock, 2006), most often as power law-like curves with a characteristic long tail of seldom used tags (Egghe, 2005). But there are also shapes which show a "long trunk" of tags of similar high frequency. Depending on which form of distribution is given for a docsonomy we see that different and different numbers of tags are of similar quality. Popular tags may be weak in discriminating resources from each other but may also reflect collective intelligence of users for allocating semantics to a resource. Thus, we separate tags of docsonomies into groups and test their potential for effective information retrieval.

Power Tags vs. Luhn-Tags

The goal is to separate tags into groups of power tags and Luhn-tags depending on the type of frequency distribution of the resource. Power tags are the most popular tags for the resource (Peters & Stock, 2010). In power law-like distributions their number can be low (e.g., 2-3), in long trunk-distributions more tags are marked as power tags (e.g., 5-15). Our approach separates power tags near the first turning point of the curve. Luhn-tags reflect considerations of Luhn (1958) who observed that both frequently and little used words in texts have low semantic quality. Often used words are known as stop words in text statistics, seldom words may contain typographical errors. Luhn stated that moderately often used words are best for describing texts and show high semantic quality. Therefore, retrieval systems should disregard low- and high-frequent words as they do not deliver the most relevant documents. Luhn-tags begin where power tags have been cut and end where the long tail of very seldom tags begins. We developed an algorithm for automatically separating three disjunctive groups of tags: power tags, Luhn-tags, long tail-tags. The algorithm locates cutting points, where the quotient of the frequencies of neighboring tags is maximal and cuts at first maximum for power tags and at second maximum for Luhn-tags.

Here, we first hypothesize that power tags enhance precision of search results because they are very valuable for resource indexing (Lux, Granitzer, & Kern, 2007). The second hypothesis regards Luhn's idea that popular tags have low quality and should not be used for retrieval. Following this notion Luhn-tags with moderate frequency will retrieve more precise results than power tags.

METHOD

This section gives details about methods used in the study.

Retrieval Test Preparation and Design

We followed the recommendations of TREC (Harman, 1993) and created a Cranfield-like retrieval scenario (Cleverdon, 1978) in a laboratory setting. According to the TREC conventions we had to develop a test collection, topics which represent information needs and relevance judgments for all resources of the test collection. The test collection contains 1,989 resources collected from delicious.com in October 2010. The resources consist of the docsonomies of the delicious-bookmarks. The tags form the indexed databases for the retrieval test runs. Each collected resource is tagged at least from 30 users and with either "folksonomy", "folksonomies", or "seo" (search engine optimization). The request for at least 30 users per resource ensures a great variety in the tag vocabulary and stable tag distributions (Halpin, Robu, & Shepherd, 2007). The topical restrictions guarantee that docsonomies belong to a specific domain which is important for evaluators to judge topical relevance of the resources. 55 information needs and search tasks have been chosen as topics. These search tasks vary in their complexity as the following examples illustrate: (1) simple lookup ("find a thesaurus"), (2) complex lookup ("find articles which present social bookmarking tools"), and (3) exploratory ("find articles which advise the combination of folksonomies and controlled vocabularies for indexing").

As retrieval effectiveness of tagging systems is reflected by recall and precision values relevance judgments for each of the 1,989 resources and for each information need had to be found. 24 students and 9 people from staff of the department acted as assessors and decided which resource is (ir)relevant for the information need. Assessments are binary and are conducted manually. Assessors have access to resources (i.e., websites) to be able to judge relevance properly. Tags are hidden for relevance assessments. To avoid contradictory relevance decisions every resource is judged from two different assessors. If both assessors agree in their decision the relevance judgment is directly saved. If they disagree, both have to discuss if this resource is relevant for this information need. This procedure results in 109,395 relevance judgments. After relevance evaluation 25 students and 2 people from staff created queries for each topic, resulting in 730 queries. Boolean operators and brackets are allowed because they can be used in delicious. The minimum number of query terms is 1, the maximum

number is 13 and the average query length is 3 terms. Moreover, experts from staff built queries for each search task with a maximum number of 5 search terms to simulate real-world-users who use 2 terms and less for Web search (Spink et al., 2001).

Databases

The test databases consist only of the tags of the 1,989 collected resources. But, as we aim at finding out which tags and which linguistic processing of tags have which effect on the retrieval effectiveness we construct twelve databases grouped into three sets: (1) four databases of original delicious-tags as downloaded, (2) four databases of unified tags in lower case, special characters removed, and (3) four databases like second set but stemmed with Porter stemmer (http://snowball.tartarus.org). The databases within the sets persist of different numbers of unique tags which are searchable in the retrieval test and which sum up in twelve retrieval runs: (a) all tags, (b) restricted to Luhn-tags, (c) restricted to Luhn- and power tags, and (d) restricted to power tags. Our retrieval algorithm works with an exact match-approach so query terms are equally processed as the tags in database sets. For searches in the first set query terms are kept as entered from the searchers, for the second set query terms are also unified and in the third set search terms are also stemmed. Brackets and Boolean operators are kept as entered.

Retrieval Test

The retrieval test follows a repeated measures design where always the same queries are searched in each of the twelve databases. Queries only retrieve results when query terms and tags match exactly. Matching involves character-wise comparison of tags and queries as well as Boolean logic. If a tag database does not contain one of several query terms (combined via AND) recall is 0. Due to our variety of search topics we can differentiate between information needs answerable with a one-word-query and which have to be represented with at least two query terms. All simple lookup-tasks only request one-word-queries (e.g., “find information about Twitter” via *twitter*). The two other search task categories request at least two words in the query to retrieve relevant results (e.g., “find articles on bookmarking at school” via *bookmarking AND school*). Search runs only receive unranked result lists. Thus, we only use the simple retrieval measures recall, precision and F-measure (harmonic mean) to compare effectiveness of databases. Because our test design allows construction of several queries for each topic we had to modify our understanding of recall and precision. We use average precision/ recall that is the mean of precision and recall values for *x* information needs gained from *x* queries of a single searcher.

RESULTS

Figures 1 to 3 visualize the results of our retrieval test and display average recall values, average precision values and values for F-measure for conducted search runs. The results refer to the simple lookup-search tasks requesting one query term. Here, average precision and recall values are given for retrieval results of five different queries. Figure 1 tells that retrieval with original delicious tags works quite well in sense of recall (0.91) but poor for precision (0.28), F-measure=0.43. The restriction to power tags for search yields the best value for F-measure (0.64) meaning recall and precision are well-balanced at a fairly good quality. Moreover, with power tags option precision is twice as

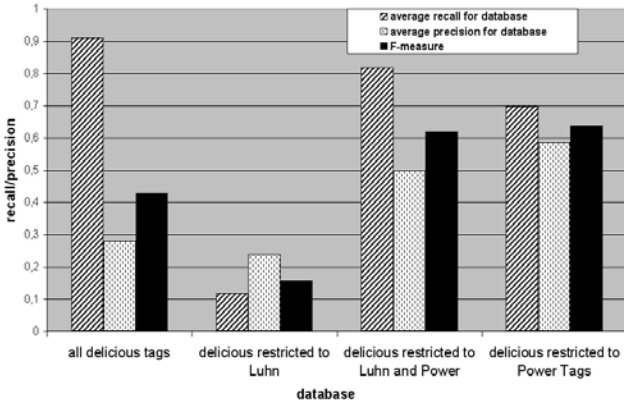

Figure 1. Average recall values, average precision values and F-measure for 5 one-word-information needs and queries – original delicious tags.

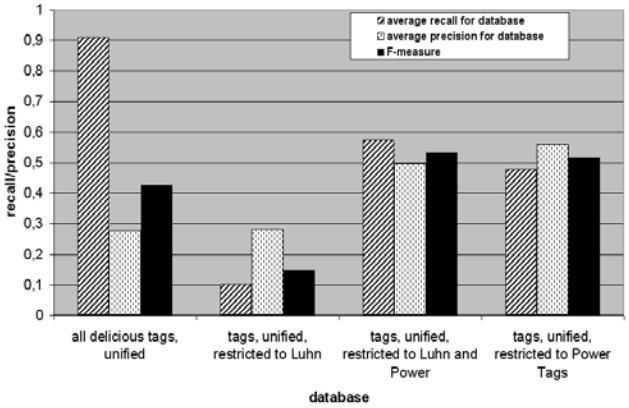

Figure 2. Average recall values, average precision values and F-measure for 5 one-word-information needs and queries – unified tags.

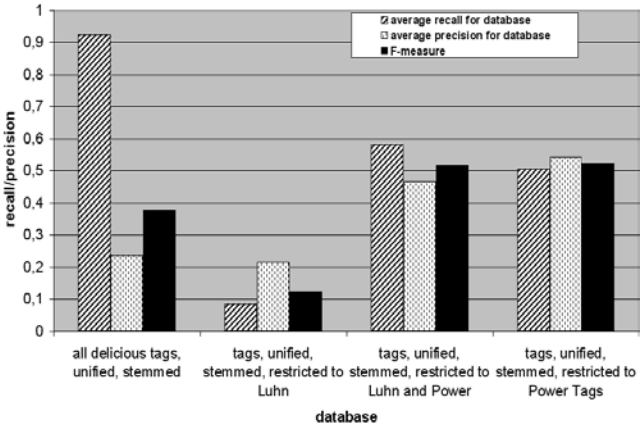

Figure 3. Average recall values, average precision values and F-measure for 5 one-word-information needs and queries – unified and stemmed tags.

good (0.58) as precision for search results of all delicious tags. The restriction to Luhn-tags shows a disappointing low F-value of 0.15. Figures 2 and 3 show similar results. The unification and stemming of tags do not enhance precision of results. The F-measure for search runs in unified power tags and in unified stemmed power tags is 0.53 each (after rounding). Precision is slightly better in unified, but not stemmed, power tags (0.56 vs. 0.54). These results confirm our first hypothesis that power tags enhance retrieval effectiveness at least for one-word queries, but our second hypothesis (Luhn-tags enhance precision) is wrong.

We also tested queries with a length of 1 to 5 terms (average=2.6 terms) created for 55 search tasks from two expert searchers of our staff. Averaging their average precision/recall values result in mean average precision values for the databases. The results of these search runs are quite different than the results before. Best perform searches in unified delicious tags with mean average precision of 0.67 and mean average recall of 0.56 (F=0.61). The restriction to power tags results in F-values between 0.17 and 0.21, the restriction to Luhn-tags gives F-values between 0.03 and 0.04 for all runs.

CONCLUSION AND FUTURE WORK

The retrieval test gives surprising insights in tags as means for information retrieval. We assumed that tags-only searches in high-frequent power tags retrieve most relevant search hits resulting in enhanced precision values. This could be confirmed for one-word-queries and information needs. Complex queries with up to 5 query terms drastically loose in precision, both in searches in power tags and in Luhn-tags, while searches in original delicious-tags perform quite well. This is due to our exact match approach which only finds results when all query terms are available in the tag database and which favors large tag databases. The notion of Luhn that moderately used words are retrieving more relevant results could not be confirmed for tags, neither for original tags nor for unified or stemmed tags. The combination of Luhn- and power tags and along with it the extension of searchable tags privileges recall by reducing precision of search results of one-word-queries. The unification and stemming of tags, which aimed at reducing language and spelling variety, have almost no positive effects on precision. The inverse relation between recall and precision is also not clearly observable in this retrieval results. Though, an increase in precision reduces recall, but recall drops not as deep as known in Web search.

Future work comprises deeper investigation of behavior of tags in information retrieval. We will conduct more search runs with varying query length to verify the results of this study. Adjustment of the tag separation algorithm may also enhance precision values of short queries. The comparison of searches in the full text of the resources and in their tags is also part of our agenda.

ACKNOWLEDGMENTS

The research is funded by the DFG (STO 764/4-1). We thank students and staff for their valuable contribution.

REFERENCES

Cleverdon, C. W. (1978). User evaluation of information retrieval systems. In D. W. King (Ed.), *Key Papers in Design and Evaluation of Retrieval Systems* (pp. 154-165). New York: Knowledge Industry.

Egghe, L. (2005). *Power Laws in the Information Production Process: Lotkaian Informetrics.* Amsterdam: Elsevier Academic Press.

Harman, D. (1993). Overview of the first text retrieval conference (TREC-1). In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, USA* (pp. 36-47). New York: ACM.

Halpin, H., Robu, V., & Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th Conference on World Wide Web* (pp. 211-220). New York: ACM.

Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. *Lecture Notes in Computer Science*, *4011*, 411-426.

Kipp, M.E.I., & Campbell, D.G. (2010). Searching with tags: Do tags help users find things? *Knowledge Organization*, *37*(4). 239-255.

Lu, K., & Kipp, M.E.I. (2010). Can collaborative tagging improve retrieval effectiveness? An experimental study. In *Proceedings of the 73th Annual Meeting of the American Society for Information Science and Technology, Pittsburgh, Pennsylvania, USA*.

Luhn, H.P. (1958). The automatic creation of literature abstracts. *IBM Journal*, 2(2), 159-165.

Lux, M., Granitzer, M., & Kern, R. (2007). Aspects of broad folksonomies. In *Proceedings of 18th International Conference on Database and Expert Systems Applications, Regensburg, Germany*.

Peters, I. (2009). *Folksonomies: Indexing and Retrieval in Web 2.0*. Berlin: De Gruyter Saur.

Peters, I., & Stock, W.G. (2010). Power tags in information retrieval. *Library Hi Tech*, *28*(1), 81-93.

Spink, A., Wolfram, D., Jansen, B.J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, *52*(3), 228-230.

Stock, W.G. (2006). On relevance distributions. *Journal of the American Society for Information Science and Technology*, *57*(8), 1126–1129.

Vander Wal, T. (2005). Explaining and showing broad and narrow folksonomies. Retrieved June 30, 2011 from http://www.vanderwal.net/random/entrysel.php?blog=1635.