# Impact and Prospect of Social Bookmarks for Bibliographic Information Retrieval

Kazuhiro Seki
Organization of Advanced Science & Technology
Kobe University
1-1 Rokkodai, Nada, Kobe
657-8501, Japan
seki@cs.kobe-u.ac.jp

Huawei Qin
Graduate School of Engineering
Kobe University
1-1 Rokkodai, Nada, Kobe
657-8501, Japan
qin@ai.cs.kobe-u.ac.jp

Kuniaki Uehara
Graduate School of Engineering
Kobe University
1-1 Rokkodai, Nada, Kobe
657-8501, Japan
uehara@kobe-u.ac.jp

## ABSTRACT

This paper presents our ongoing study of the current/future impact of social bookmarks (or social tags) on information retrieval (IR). Our main research question asked in the present work is "*How are social tags compared with conventional, yet reliable manual indexing from the viewpoint of IR performance?*". To answer the question, we look at the biomedical literature and begin with examining basic statistics of social tags from CiteULike in comparison with Medical Subject Headings (MeSH) annotated in the Medline bibliographic database. Then, using the data, we conduct various experiments in an IR setting, which reveals that social tags work complementarily with MeSH and that retrieval performance would improve as the coverage of CiteULike grows.

## Categories and Subject Descriptors

H.3.1 [**Information storage and retrieval**]: Content Analysis and Indexing—*Indexing methods, Thesauruses*; H.3.7 [**Information storage and retrieval**]: Digital Libraries—*Collection, Standards*

## General Terms

Experimentation, Languages, Performance

## Keywords

Subject headings, controlled vocabulary, free keywords, folksonomy

## 1. INTRODUCTION

Recently, social bookmarks are becoming increasingly popular as a means to share, organize, and search public resources on the Web, including web pages, pictures, and videos. These bookmarks are typically shared with others (hence the name "social") and they often contain some descriptions as metadata, such as short keywords. For example, a user may add keywords like "video" or "entertainment" to YouTube, a video sharing website, and these keywords are called social tags.


*JCDL'10,* June 21–25, 2010, Gold Coast, Queensland, Australia.


Social tagging (also known as *folksonomy*) is basically seen as manual indexing of information resources using free keywords, instead of a controlled vocabulary. From the viewpoint of information retrieval (IR), using user-assigned free keywords as index terms inherently suffers from informal vocabularies and, to make matters worse, some social tags may even be inappropriate since tagging can be performed by virtually anyone without any inspection. Nonetheless, the scale of the Web may make social tags useful for IR as we witnessed for other collaborative social media, such as Wikipedia [2]. This brings up an interesting research question that has been asked in the realm of information science [7]; that is, "Is controlled vocabulary better than free keywords?". We will revisit the question in view of the emergence of the folksonomy.

There are numbers of studies related to our work. For example, Heymann et al. [5] and Bischoff et al. [1] studied the characteristics of social tags for web pages. Morrison [6] compared IR performance of folksonomies against several search engines for web search. Zhou et al. [9] and Yang et al. [8] proposed IR models incorporating social tags. The present work is different from these studies in that our focus will be contrasting controlled vocabulary and social tags for bibliographic databases. Another work, more similar to ours, was conducted by Good et al. [3], who compared a variety of statistics of social tags and Medical Subject Headings (MeSH). Our study goes beyond their work and empirically evaluates the current impact and future prospect of social tags for IR on a real-world data set.

## 2. EMPIRICAL ANALYSIS

### 2.1 Overview

To investigate the impact of social tags on IR, our analysis relies on Medline, the world largest bibliographic database in life science, and CiteULike[1], a popular social bookmarking site focusing on academic articles. The former contains subject headings from a controlled vocabulary, called Medical Subject Heading (MeSH)[2], the latter social tags assigned by its users using (uncontrolled) free keywords. Using the data, we first examine the characteristics of social tags and MeSH. Then, we carry out experiments on a standard IR test collection to investigate whether social tags make/will make any difference on retrieval performance.

The Medline data used for this work is the 2010 version of the Medline/Pubmed files, and a snapshot of CiteULike was downloaded on December 17, 2009.

[1] `http://www.citeulike.org`
[2] MeSH terms are manually assigned by experts at the National Library of Medicine.

## 2.2 Size and growth of social tags

While Medline covers only life sciences, CiteULike is not limited to particular domains. For meaningful comparison, we extracted and analyzed only the CiteULike entries that link to Medline entries (i.e., the intersection of CiteULike and Medline). In other words, we restricted our analysis only to the life science domains for CiteULike, too. Also, we eliminated redundant CiteULike tags annotated to the same article for fairer comparison[3]. In addition, since social tags are uncontrolled and contain many writing variants, they were preprocessed (stemmed, lower-cased, and non-alphanumeric symbols removed) for normalization before analysis.

Table 1 presents the descriptive statistics for MeSH terms and social tags form CiteULike. The total number of articles annotated with at least one MeSH term is around 77 (=18,058k/237k) times larger than the case of CiteULike. For the total number of terms/tags, the difference becomes even greater; MeSH terms are annotated around 176 (=178,023k/1,009k) times more than CiteULike tags. These statistics indicate that social tag annotation is much more sparse than MeSH, which is also evidenced by the smaller mean and median of the number of terms/tags per article, called *density* [3]. On the other hand, the total number of distinct tags is around three times larger for CiteULike than for MeSH. This is due to the fact that social tags are uncontrolled and tend to be diverse even after normalization. Note that the statistics reported by Good et al. [3] are generally smaller than what reported here due to the fact that they included CiteULike entries with no social tags in their analysis[4].

**Table 1: Descriptive statistics of MeSH terms and CiteULike tags.**

| | | MeSH | CiteULike |
|---|---|---|---|
| Total # of articles | | 18,058,028 | 237,265 |
| Total # of terms/tags | | 178,023,923 | 1,009,548 |
| Total # of distinct terms/tags | | 25,195 | 80,449 |
| # of terms/tags per article (density) | Max. | 97 | 210 |
| | Mean | 9.86 | 4.25 |
| | Median | 9 | 3 |
| | SD | 5.21 | 4.24 |

It is apparent that social tags (from CiteULike) do not currently match MeSH in either coverage or density. However, MeSH indexing relies on a limited number of experts at NLM, whereas the users of social bookmarking service are voluntary and continuously growing world-wide. Thus, it is conceivable that CiteULike could catch up with Medline in coverage in future. To examine the growth of social tags, Figure 1 shows the number of distinct articles annotated per month in Medline and CiteULike, respectively, on a log-log scale. Although it is risky and unreliable to make a prediction by extrapolation, especially for distant future, a simple regression analysis tells us that the number of social tag annotation per month may reach that of MeSH by the end of 2024—taking a little longer than 10 years from now. Of course, both resources would not keep growing at the current rates and the social tagging system may not remain in the present form, we could at least expect that the coverage of CiteULike would become closer to that of Medline in the next few years.

[3]The same article may be annotated with the same social tag for multiple times by different users, which never happens in Medline.

[4]CiteULike allows entries with no social tag, which are automatically assigned with fictitious "no-tag". We excluded these entries in this study.

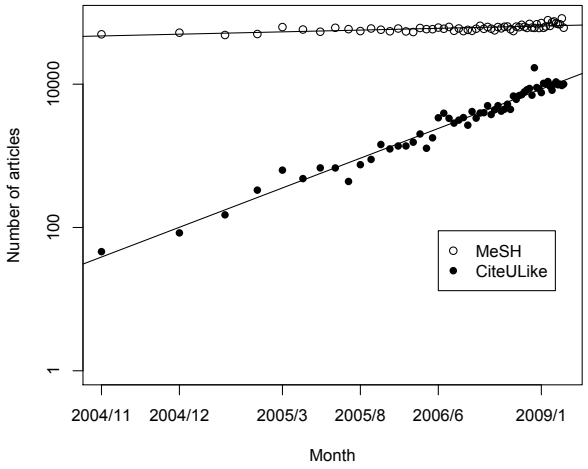


**Figure 1: Number of distinct articles annotated per month.**

## 2.3 Impact of social tags on IR

As discussed in the previous section, the size of CiteULike is currently limited and by no means comparable to MeSH. However, the limited resources may still add some values as additional index terms in users' information seeking process. In fact, it was reported that some social tags not overlapping with MeSH terms were actually good representation of the topical contents of their associated articles [3]. This section examines whether or not social tags could have any impact on the performance of an IR system.

### *2.3.1 Experimental settings*

We used the 2004 Genomics track data set [4] for experiments. The track was held from 2003 to 2007 as a part of the Text Retrieval Conference (TREC) to foster research for IR in biomedicine. The data set is a subset of Medline from 1994 to 2003 and comes with 50 topics representing user information needs. For each topic, pooled results from track participants' submissions were manually judged and labeled as relevant or irrelevant, which can be used to evaluate a given IR system.

For a retrieval model, we simply adopt the Indri search engine[5] out of the box since our focus is not on pursuing the highest retrieval performance on the particular data set but on studying the utility of social tags for IR. Article titles, abstracts, and MeSH terms were used for indexing. Our baseline retrieval performance in mean average precision (MAP) using the resulting retrieval model was 0.278, which is not particularly strong but reasonable, being placed around the 8th position among the 27 participating groups at the Genomics track [4].

To focus on the effect of social tags, the following experiments utilize not the entire Genomics track data set but a subset consisting of the articles referenced from CiteULike; that is, they have at least one social tag annotation Partly because the Genomics track data were produced before the inception of CiteULike, there is not large overlap; the subset consists of 62,035 articles, which means that we can take advantage of only a quarter of the CiteULike data available. Accordingly, we used a subset of relevance judgement data associated with the articles in common between CiteULike and the

[5]`http://www.lemurproject.org`

Genomics track data. This hypothetical setting intends to imitate the case where all the articles included in Medline are annotated with at least one social tag.

### 2.3.2 Results and discussion

*Overall retrieval performance.*

Using the Genomics track data, we first examined the effect of social tags by adding them to the index of our IR system. Table 2 summarizes the results, where "None" did not use either MeSH or social tags (i.e., used only article titles and abstracts), "MeSH" and "CiteULike" used MeSH terms and social tags in addition to "None", respectively, and "Both" used both MeSH and social tags. Surprisingly, using the social tags was better than using MeSH and combining the two further improved the performance. This result indicates that social tags are not alternative to but rather complementary to the controlled vocabulary. Based on the two-sided $t$ test for dependent means, the difference between "None" and "Both" was found statistically significant at the 0.05 significance level ($p$=0.03). We plan to analyze their complementary nature and why social tags were better than MeSH.

It should be noted that significant improvement was obtained despite that this experiment was able to employ only a subset of available social tags due to the small overlap of CiteULike and Genomics track data. Greater improvement may be achieved if more updated test collections are utilized.

**Table 2: Retrieval performance in MAP contrasting the effects of MeSH terms and social tags from CiteULike. Figures in parentheses indicate percent increase/decrease with respect to "None". Asterisk indicates statistical significance at 0.05 level.**

| Index | MAP |
|---|---|
| None | 0.3304 |
| MeSH | 0.3322 (+0.54%) |
| CiteULike | 0.3376 (+2.18%) |
| Both | 0.3477* (+5.24%) |

*Tag quality and retrieval performance.*

Because social tags can be assigned by anyone who wish, in contrast to MeSH terms assigned only by experts, the quality of social tags is a serious concern when used for IR. Using inappropriate tags as index terms would lead to retrieving irrelevant articles not pertinent to user information needs. Conversely, if we selectively use only high quality social tags that well represent topical contents of articles, retrieval performance may increase. To this end, we explored two simple quality measures in the present work, although plenty of others are possible.

**An article-based quality measure.** The first measure is the popularity of articles. The rationale behind is that popular articles are read by more people and thus tend to receive more tags, where there would be a better chance that some quality tags are assigned. Here, we measured the popularity of an article by the number of social tags annotated to it, where duplicate tags were also counted. Then, for only the articles with higher popularity than a threshold, their associated social tags were added as additional index terms. By varying the threshold, Figure 2 plotted the transition of retrieval performance. The leftmost points (threshold=1) correspond to the results shown in Table 2. Note that because "MeSH" and "None" do not use social tags in the first place, their performance is constant irrespective of the threshold.

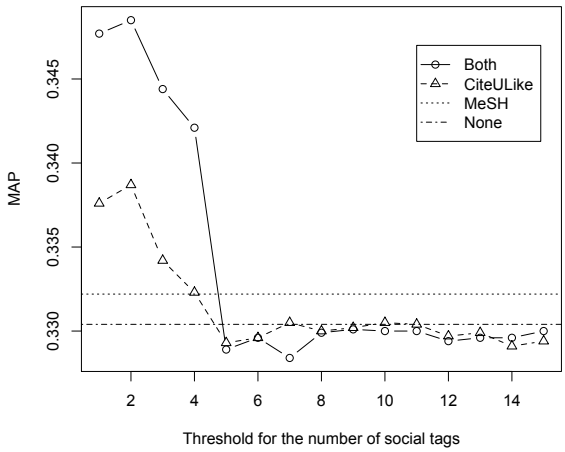


**Figure 2: Transition of retrieval performance when using social tags as additional index assigned to only popular articles measured by the number of tags higher than a threshold.**

When the threshold was increased to 2, MAP moderately increased up to 0.3485 and then rapidly dropped afterward for both "CiteULike" and "Both". This result indicates that 1) social tags exclusively annotated to articles (i.e., the number of tags equals 1) are not generally useful in terms of MAP and removing them does not change—even slightly improve—the retrieval performance and that 2) as opposed to our expectation, the quality of tags measured by the number of tags does not positively relate to the retrieval performance except for the case where it is less than 3.

**A tag-based quality measure.** Whether social tags or MeSH terms, annotated tags/terms are not equally useful for IR; some are more general and thus not very useful for locating documents pertinent to user information needs, whereas the others may be specific and useful in this regard. One way to quantify such usefulness or quality of social tags is to look at inverse document frequency (IDF) commonly used in the IR literature. IDF was developed to measure the specificity of a term to represent a particular document in a given document collection. It is known that there is a "sweet spot" in the range of IDF for choosing good index terms from; it should not be too low (i.e., general terms including function words) or too high (i.e., rare terms including misspellings). IDF is defined as $\log(N/\mathrm{DF}_t)$, where $N$ is the total number of articles in a collection and $\mathrm{DF}_t$ denotes the number of articles containing term $t$. Thus, instead of using all social tags indiscriminately, using only the ones with middle-range IDF may be more effective.

Based on the idea, we chose social tags with higher IDF values than a threshold and added them in the index of our IR system. Figure 3 plotted a transition of MAP with different thresholds, where two different settings were tested. One setting excluded social tags with DF=1, and the other did not. Social tags with DF=1 are those assigned to only one article in the CiteULike database and thus considered to be too rare to be good index terms.

As expected, when social tags with DF=1 were excluded, the retrieval performance was better (but only slightly) than the case where they were included. Also, the trend of the transition was found almost the same; MAP was higher when the threshold was in the middle range, around 7.7 to 9.2. This result confirmed that the quality of social tags as index terms can be in part measured by IDF and has some influence on retrieval performance. However, it

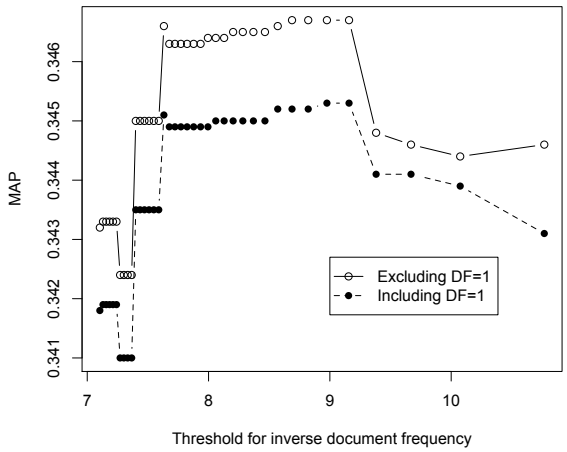


**Figure 3: Transition of retrieval performance when using only social tags with higher quality measured by IDF than threshold.**

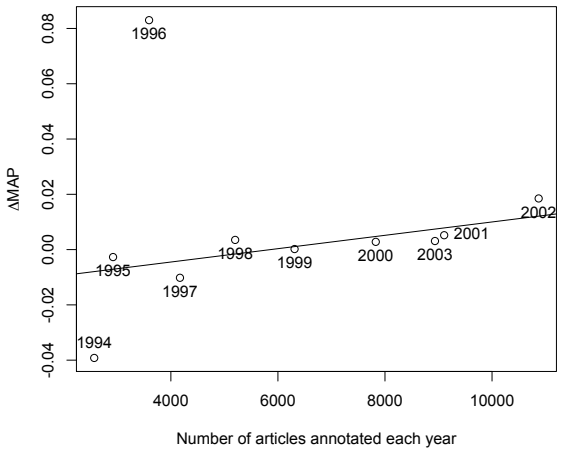


**Figure 4: Relation between the change of retrieval performance and the number of distinct articles annotated each year.**

did not reach the performance that was achieved when all the social tags were utilized (i.e., MAP=0.3477, see Table 2). This is presumably due to the fact that an IDF-like factor is already incorporated in the IR system used for this experiment.

### *Coverage and retrieval performance.*

Lastly, we studied the significance of the coverage of social tagging. We expect that the wider the coverage becomes, the more influence social tags have on the retrieval performance. To examine our hypothesis, we restricted our search by a publication year of articles[6]. For this experiment, article titles and abstracts with/without social tags were utilized for search to investigate the effect of social tags alone. Figure 4 plots for each year a circle with $x$-axis being the number of distinct articles annotated for the year and $y$-axis being the performance increase/decrease, denoted as ΔMAP, as compared with the search results without using social tags.

Although the trend is subtle, we can observe that MAP increases as the number of articles annotated per year increases. Using Pearson's product moment correlation coefficient, there existed a statistically significant association at the 0.05 level ($p$=0.015) when the data point for 1996, which appears to be an outlier, was removed. This result confirms our intuition that the coverage of annotations does influence the retrieval performance. Again, if a newer IR test collection is used for experiments, one may be able to see a greater performance increase since more social tags can be exploited.

## 3. CONCLUSION

In this paper, we analyzed social tags from CiteULike in comparison with MeSH terms from Medline and discussed whether they could add any value as new index terms of a general IR system. Our analysis indicated that although not currently comparable, the coverage of CiteULike might exceeds that of Medline in 14 years. Also, we empirically evaluated the impact of social tags on retrieval performance using the test collection from the TREC Genomics track. The experiments demonstrated that social tags worked complementarily with MeSH and marked a significant improvement and that selectively using quality tags could potentially boost retrieval performance. In addition, our experiment showed that there was a positive correlation between annotation coverage and retrieval performance, suggesting that further improvement is expected simply from the growth of social tags.

Future work would include exploring other ways to measure the quality of tags, such as the number of users who assigned the same tag to the same article. Also, we plan to study the linguistic properties of social tags to better understand their effectiveness for IR.

[6]Although CiteULike was launched in 2004, many articles published before 2004 have been annotated.

## 4. REFERENCES

[1] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. Can all tags be used for search? In *Proc. of the 17th ACM CIKM*, pages 193–202, 2008.

[2] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.

[3] B. Good, J. Tennis, and M. Wilkinson. Social tagging in the life sciences: characterizing a new metadata resource for bioinformatics. *BMC Bioinformatics*, 10(1):313, 2009.

[4] W. Hersh, R. T. Bhuptiraju, L. Ross, A. M. Cohen, and D. F. Kraemer. TREC 2004 genomics track overview. In *Proc. of the 13th Text REtrieval Conference (TREC)*, 2004.

[5] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *Proc. of the 1st International Conference on Web Search and Web Data Mining*, pages 195–206, 2008.

[6] P. Jason Morrison. Tagging and searching: Search retrieval effectiveness of folksonomies on the World Wide Web. *IP&M*, 44(4):1562–1579, 2008.

[7] J. Rowley. The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. *Journal of Information Science*, 20(2):108–118, 1994.

[8] L. Yang, S. Xu, S. Bao, D. Han, Z. Su, and Y. Yu. A study of information retrieval on accumulative social descriptions using the generation features. In *Proc. of the 18th ACM CIKM*, pages 721–730, 2009.

[9] D. Zhou, J. Bian, S. Zheng, H. Zha, and C. L. Giles. Exploring social annotations for information retrieval. In *Proc. of the 17th WWW*, pages 715–724, 2008.