2009 Ninth International Conference on Intelligent Systems Design and Applications

Improved Search in Tag-Based Systems

Ruba Awawdeh
School of Computing and Mathematics
University of Ulster, Newtownabbey
BT37 0QB, Northern Ireland.
awawdeh-r@email.ulster.ac.uk

Terry Anderson
School of Computing and Mathematics
University of Ulster, Newtownabbey
BT37 0QB, Northern Ireland.
tj.anderson@ulster.ac.uk

Abstract

Social bookmarking systems are used by millions of web users to tag, save and share items. User-defined tags, however, are so variable in quality that searching on these tags alone is ineffective. One way to improve search in bookmarking systems is by adding more metadata to the user-defined tags to enhance tag quality. Such an approach would add value by incorporating information about the content of the resource while retaining the original user-defined tag. Tags automatically extracted from the resource could be the main source for tag enhancement.

This paper describes how users' tags can be enhanced with metadata in the form of additional tags automatically extracted from the original document. An evaluation study shows how the enhanced tag set improved user searching in comparison to using only user-defined tags.

1. Introduction

Social bookmarking systems are a major part of the interactive, collaborative trend in web software known as Web 2.0. They allow users to save a set of document links, and add one or more freely chosen tags to each one. While the set of tagged links remains easily accessible to the creator in his/her account, they are normally publicly available (unless explicitly declared private). This form of informal collective indexing allows users to search for all (or at least the most recently added) documents tagged with a tag of interest to them. Del.icio.us [22], which started in 2003 and is one of the largest social bookmarking sites, has more than five million users and 175 million bookmarked URLs tagged by different users [23]. Because users are free to add any tags they wish, these tags are arguably not true metadata, often reflecting idiosyncratic views or planned uses [5,1]. Moreover, as there is no controlled vocabulary or even widely accepted set of guidelines, tag based systems, or 'folksonomies', exhibit limitations in searching and re-finding items [16]. However, a major benefit of folksonomies is that users have evaluated the resources and presumably consider the ones they tag as being of particular value. To others with like interests, this can lead to the discovery of related items that may not be reported by conventional search engines. This is rather like being, say, a bee-keeper, and being able to browse other bee-keepers' bookshelves – there's a good chance of finding interesting materials, even if some of it is only tangentially related to bee-keeping.

The paper is structured as follows: in the next section the research motivation is given. Section 3 provides a brief literature review related to research in the area and section 4 describes the approach used in the prototype to enhance user tags with terms automatically extracted from the original document. The additional terms are therefore based on document content and largely avoid the idiosyncratic and ambiguous terms too often evident in user-generated tags. A brief description of ETS (Enhanced Tag Search engine) is given as well. Section 5 reports on the experiment used to evaluate ETS. Finally, future work is presented.

2. Motivation

Since around 2002/2003, collaborative tagging systems have multiplied, including sites such as Digg, StumbleUpon, Del.icio.us, Flickr and Connotea. While the freedom users have in creating their own tags is a major reason behind their success, the ambiguity in tags, including non-standard abbreviations, misspellings, polysemy (one word with multiple meanings, e.g. bank: a financial institution, an earthen slope, an aircraft manoeuver)

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.170

and synonymy (multiple words with the same meaning, e.g. symbol, character) reduces the value of tags in searching a folksonomy [11]. Also, the number of tags a user attaches to an item is highly variable. In an analysis of 60,000 tagged items on delicious we found that the number of tags per item ranged from none to 19, though the modal number was two. It is therefore unsurprising that searching based on user tags is often disappointing. From this, it is logical to argue that user-defined tags need to be enhanced in some way to improve their value in searching a folksonomy-based site.

3. Related Work

Current research on social tagging and folksonomies can be grouped into 3 broad approaches:

- Tags and their role in indexing
- Tagging and user behavior
- The nature of tagging systems [18].

Tags and their role in indexing

Quintarelli [16] studied tagging as a potential tool for information retrieval despite the fact that tags are not restricted to specified vocabularies. Guy and Tonkin [11] advocate tags as supplements to formal classification systems, not as wholesale replacements. The purpose is to improve tag quality for later reuse as searchable keywords. This approach, in which a formal classification is enhanced with tags is largely theoretical, but crucially retains the major drawback of all formal systems, that it is inherently costly in time and dependent on significant human expertise.

Tagging and user behavior

Guy and Tonkin [11] declare that tags could be improved in two ways: "educating users to add 'better' tags and improving the systems to allow 'better' tags to be added". Golder and Huberman [9] analysed collaborative tagging and found that appreciating the nature of the tags, e.g. content-descriptive or personal plans, can help understand the differences between tags. They studied tag frequency, user performance and trends in bookmarking on sample data from del.icio.us, and found that the majority of tags are for personal use but that most of them are also helpful to other users. Examples of purely personal tags include 'to-read' or 'interesting-paper', but these kinds of tags do not reflect the document content.

Michlmayr [15] also studied tag properties, concluding that while it is appropriate to filter items using tags, search based purely on user-defined tags is of limited success.

The nature of tagging systems

Bao et. al. [3,19] utilized social tagging to improve web searches by using tagging based on 2 ideas: (1) similarity ranking, that is "the estimated similarity between a query and a web page", and (2) static ranking, which means "the amount of annotations assigned to a page which indicates the page popularity". They found that tags can provide a multi-faceted summary of the web page and its quality. Byde [7] describes a tag recommendation approach for bookmarking systems. His method involves suggesting tags which may be derived from two sources, (1) the existing set of user-defined tags, and (2) content-based tags based on the cosine similarity matrix.

Brooks and Montanez [6] compared the effectiveness of two approaches to clustering and naming blogs. Direct use of user-defined tags was the basis of the first approach, while the second involved automatic extraction of words from the blog articles where the three words with the top TF/IDF (term frequency – inverse document frequency) were extracted. Their results illustrate that user-defined tags did help in clustering the blogs, although they were poor for describing the overall subject matter of an article, a task much more successfully achieved using automatically extracted words.

In an enhanced web search prototype by Yanbe et al. [20] they develop a method for ranking search results based on social bookmarking data such as users' tags, time stamp, page popularity and users' comments, combined with weightings from a PageRank algorithm. They report experiment results that support the effectiveness of their method for enhancing the search systems.

Al-Khalifa [2] has performed a comparison between user-defined tags and automatically extracted words from the Yahoo API term extraction tool. While her experiment demonstrates the value of these tags when used as annotations, she does not demonstrate how they can be used as searchable keywords [2]. Her results were particularly influenced by the scientific nature of the information in her dataset. The users were professionals who strongly tended to use well recognized scientific terms as tags. The terms extracted by the Yahoo API service were less satisfactory for this dataset in that it did not reliably return scientific terms.

4. Prototype Development

In this research we aim to improve tag-based search by enhancing the user-defined tag set with additional terms which accurately extend the

description of the web page. This was attempted by going back to the original documents and extracting further descriptive terms. Clearly, there is no standard method for automatically extracting the 'best' descriptive terms from a document, not least because the criteria for identifying the best terms may be very difficult to determine.

Using 73,000 records from the Del.icio.us RSS feed as a basis, we first retrieved each html document and applied 3 alternative techniques for term extraction:

- Yahoo Term extraction API, which is a widely used but proprietary service.
- Selecting keywords and description meta tags if present in the html file.
- Selecting terms based on the highest word frequency, with extra weighting for words in the title of the document.

Each html document contains various pieces of information that are not needed for the extraction process; therefore an initial cleaning process was carried out on the files. This included (1) stripping html tags, and (2) removing unwanted characters and stop words. Then for each document the url, title, user-defined tags, automatically extracted tags, and keywords and description were stored in a record in a MySQL database, forming the dataset for our prototype search engine. User-defined tags and the most frequent keywords were stemmed using the Porter Stemming algorithm, returning the roots of the words. Table 1 gives an example of one record in the database.

The Enhanced Tag Set Search engine (ETS) is designed to allow comparison of tag-based searching on our del.icio.us data set. As shown in Table 1, ETS stores 4 different tag sets, ie:

- Tag-Set1: User tags alone
- Tag-Set2: Tags from Yahoo API
- Tag-Set3: Keyword and description meta tags in html file
- Tag-Set4: Tags based on term frequency in document

The search algorithm was instrumented to employ, at any one time, one of a combination of tag sets so that evaluation experiments could help determine which tag set returned the set of results deemed most helpful by the participants. The search algorithm selects urls that share similar tags based on tag co-occurrence. Table 2 very briefly illustrates the concept of co-occurrence among records [1], and how related documents can be retrieved that do not explicitly contain any of the tags used as part of the query.

Table 1. Example database record

Link	**www.webteacher.com/javascript/index.html**
Title	**Java script tutorial for the total non-programmer**
User-Tags	**Java, Good.site, learn**
Yahoo API	**JavaScript tutorial, programming language, object oriented programming, non-programmer tutorial**
Frequency	**JavaScript, tutori, JavaScript tutori, program language**
Keywords in Meta tags	**Javascript, programming language, javascript tutorial, non-programmer, object oriented programming**
Description in Meta tags	**This java script tutorial for non-programmers step-by-step fundamendals of the javascript programming language.**

Table 2. Co-occurrence example

1	Url-1	Belfast	Northern-Ireland	Cold-weather
2	Url-2	Giants-Causeway	Belfast	Tourist-places
3	Url-3	Carrick-a-Rede	Giants-Causeway	Rope-bridge

In Table 2, the tag 'Belfast' occurs in both record 1 and record 2, while 'Giants-causeway' occurs in both record 2 and record 3. Because of the co-occurrence of 'Belfast' and 'Giants-causeway' in record 2, record 3 which is tagged with 'Giants-causeway' would also be returned in a search using the tag 'Belfast' even though record 3 is not tagged with 'Belfast'.

The location and frequency of a query term in the tags being searched is of central importance. If a query term is found in the title and more than once within the database of tags then it is given a higher rank as it is deemed to be more relevant to the user's query.

Table 4. Paired t-test results for user preferences

	Phase 1	Phase 2	T	df	2-tailed p	Std. Error Diff
	Average	Average				
Groups						
A	3.39	3.73	-2.52	17	0.0221	0.137
B	3.36	3.73	-3.82	15	0.0017	0.098
C	3.39	3.48	-0.91	15	0.3753	0.103
D	2.99	3.35	-2.99	19	0.0078	0.118

The top 10 Yahoo API terms extracted were stored in the database for each record. The decision to take the top 10 tags each time was based on an inspection of the tags returned by the Yahoo system and a desire to ensure, as far as possible, that all of the most useful terms had been retained in the database. In many cases a smaller number of tags would have been perfectly appropriate, but it was felt that the remaining tags, even if they were of less clear relevance, were unlikely to degrade the quality of the results returned. Similarly, the number of terms in Tag-Set 3 was limited to 10. In practice the average number of tags was around 15.

5. Evaluation

To compare the search engine's effectiveness when using user-defined tags (Tag-Set1) with enhanced tag sets (Tag-Set2, 3 and 4) we devised a lab-based experiment. It was carried out in 2 phases using 69 participants, followed by a questionnaire. Participants were divided into 4 groups (A, B, C and D), where each group used Tag-Set1 and one of the enhanced tag sets.

We use term 'phase1' for use of Tag-Set1 and 'phase 2' for use of one of the other tag sets, but the order in which students completed theses phases was randomly reversed to cancel out any learning effect. Groups were assigned to different data-sets as shown in Table 3.

Table 3. Groups and data-sets

A	Tag-set1 Tag-set2 (Yahoo API)
B	Tag-set1 Tag-set3 (keywords and description)
C	Tag-set1 Tag-set4 (term frequency)
D	Tag-set1 Tag-set2 + Tag-set3 + Tag-set4

In each phase the participants were asked to use 10 different query phrases. These were supplied to them using our knowledge of the tags in the dataset. In the second phase, the participants used the same 10 search phrases. In each phase participants were asked to review reasonably carefully at least 10 of the search results and to score on a 1 – 5 scale their impression of how relevant the results were to the initial query.

5.1 Results

A detailed analysis of the relevance ratings reveals that users perceived real differences in the relevance of the results supplied by ETS using the different tag-sets, and that invariably the enhanced tag sets outperformed, or at least did as well as, Tag-Set1, i.e. the user tags alone.

For groups A, B and D, the average rating on phase 2 is higher than for phase 1 at the 5% level of statistical significance, as shown by the paired t-test results in Table 4. These figures indicate an improvement of the order of 10% between phase 1 and phase 2. Only in the case of group C (term frequency) are the average responses so close that there is no statistical difference. The same information is shown in Figure 1.

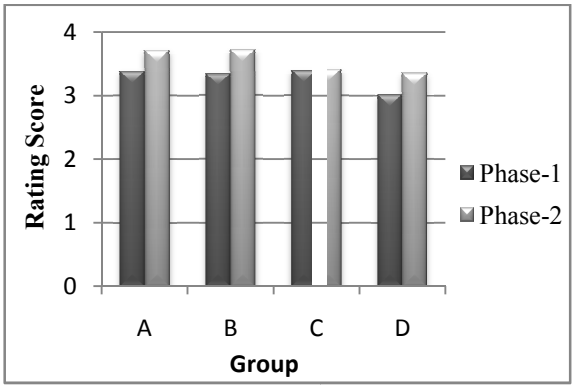

Figure 1. Overall groups average rating

After the experiment, users were asked which of the two experimental phases had provided the more helpful search results. Figure 2 shows the clear majority of participants preferred the tool they had used in phase 2. 50 of the 69 participants opted for phase 2 as providing better results for their search queries.

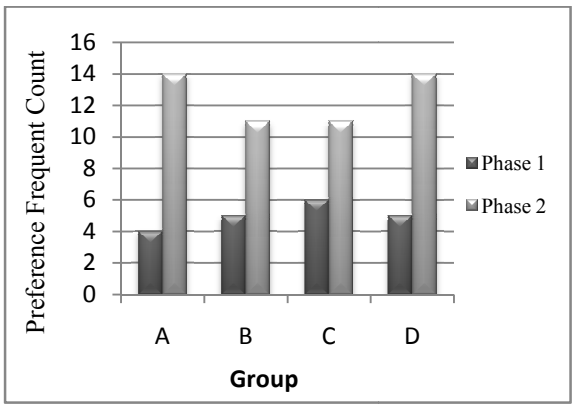

Figure 2. User preference for ETS using different tag sets

6. Future Work

The Yahoo terms and the keyword metadata shows improvement in the search results, therefore a combination of both tag sets could result in significant improvement in tag-based searching. While the simple term frequency method was the least effective method for improving the search results, it could be used in combination with Tag-Set 2 and 3 to weight the extracted keywords and store only the most valuable terms. Coping with tag ambiguity remains a challenge and further research needs to be carried out to remove it as far as possible. Using the GN algorithm [8] could perhaps help us to partially remove ambiguities from tags, which could improve the quality of the search results. Yeung et al. [20] developed a method based on the GN algorithm for tag disambiguation. Employing this method could help us understand the meaning of the ambiguous tags through grouping web documents into clusters. Documents within each cluster share the same context, therefore the attached tags will share similar meaning and can help resolve out the meaning of ambiguous tags. Revision of ETS is in progress, as we hope to further improve the relevance of results returned and incorporate a tag recommendation system with a tag-cloud visualization feature [17].

References

[1] Alag, S. *Collective Intelligence in Action,* Manning, USA, October-2008.

[2] Al-Khalifa, H. & Davis, H. "Folksonomies versus Automatic Keyword Extraction: An Empirical Study", IADIS Web Applications and Research, 2006, Available from: http://eprints.ecs.soton.ac.uk/14292/.

[3] Bao, S., Wu, X., Fei, B., Xu, G., Su, Z. & Yu, Y. "Optimizing Web Search Using Social Annotations", International World Wide Web Conference Committee (IW3C2), Canada, 2007, pp. 501-510.

[4] Begelman, G. "Automated Tag Clustering: Improving search and exploration in the tag space", 2006, Available from: www.pui.ch/phred/automated_tag_clustering/automated_tag_clustering.pdf.

[5] Berendt, B. & Hanser, C. "Tags are not Metadata, - but Just More Content – to Some People", 2007, Available from: http://www.icwsm.org/papers/2--Berendt-Hanser.pdf.

[6] Brooks, C. & Montanez, N. "Improved annotation of the blogspere via autotagging and hierarchical clustering", In WWW '06: Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland, 2006, pp. 625-632.

[7] Byde, A., Wam, H. & Cayzer, S. "Personalized Tag Recommendations via Tagging and Content – Based Similarity Metrics", international Conference on Weblogs and Social Media, 2007.

[8] Girvan, M. & Newman, M. "Finding and evaluating community structure in networks", 2004. Available from: http://www.uvm.edu/~pdodds/research/papers/others/2004/newman2004b.pdf.

[9] Golder, S. & Huberman, B. "The structure of collaborative tagging systems", HP labs, 2006, Available from: http://www.citeulike.org/user/zelig/article/305755.

[10] Gruber, T. "Ontology of Folksonomy: a mash-up of apples and oranges", 2006, Available from: www.tomgruber.com.

[11] Guy, M. & Tonkin, E. "Folksonomies: Tidying up Tags", D-Lib Magazine, 2006, Available from: http://www.dlib.org/dlib/january06/guy/01guy.html.

[12] Hayes, C. & Avesani, P. "Using Tags and Clustering to Identify Topic-Relevant Blogs", 2007, Available from: http://www.icwsm.org/papers/2--Hayes-Avesani.pdf.

[13] Hotho, A., Jäschke, R., Schmitz, C. & Stumme, G. "Information Retrieval in Folksonomies: Search and Ranking", The Semantic Web: Research and Applications, Springer Berlin, Heidelberg, 2006, pp. 411-426.

[14] Mathes, A. "Folksonomies - Cooperative classification and communication through shared metadata", 2004, Available from: www.adammathes.com.

[15] Michlmayr, E. "A Case Study on Emergent Semantics in Communities", International Semantic Web Comference (ISWC), 2002.

[16] Quintarelli, E. "Folksonomies: power to the people", 2005, Available from: http://www.iskoi.org/doc/folksonomies.htm.

[17] Sinclar, J. "The Folksonomy tag cloud: when is it useful?", Vol. 34, No. 1, 2008, pp. 15-29.

[18] Trant, J. "Studying Social Tagging and Folksonomy: A Review and Framework", Vol. 10, 2009, Available from: http://dst.sir.arizona.edu/2595/.

[19] Xu, S., Bao, S., Fei, B., Su, Z. & Yu, Y. "Exploring folksonomy for personalized search", ACM, 2008, pp. 155-162.

[20] Yanbe, Y., Jatowt, A., Nakamura, S. & Tanaka, K. "Can social bookmarking enhance search in the web?", JCDL '07, 2007, pp. 107-116.

[21] Yeung, C., Gibbins, N. & Shadbolt, N. "Tag Meaning Disambiguation through Analysis of Tripartite Structure of Folksonomies", The 2007 IEEE/WIC/ACM international Conference on intelligent Agent Technology, 2007, pp.3-6.

[22] www.delicious.

[23] www.ebizmba.com/articles/social-bookmarking.