Measuring the Semantic Value of Folksonomies

Hend S. Al-Khalifa and Hugh C. Davis
Learning Technology Research Group, ECS, The University of Southampton, Southampton, UK
{hsak04r/hcd}@ecs.soton.ac.uk

Abstract

Semantic Metadata, which describes the meaning of documents, can be produced either manually or else semi-automatically using information extraction techniques. Manual techniques are expensive if they rely on skilled cataloguers, but a possible alternative is to make use of community produced annotations such as those collected in folksonomies. This paper reports on an experiment that we carried out to validate the assumption that folksonomies carry more semantic value than keywords extracted by machines. The experiment has been carried-out in two ways: automatically, by measuring the percentage of overlap between the folksonomy set and machine generated keywords set; and subjectively, by asking a human indexer to evaluate the quality of the generated keywords from both systems.

The result of the experiment can be considered as evidence for the rich semantics of folksonomies, demonstrating that folksonomies used in the del.icio.us bookmarking service can be used in the process of generating semantic metadata to annotate web resources.

1. Introduction

Nowadays, contemporary web applications such as Flickr[1] and del.icio.us[2] rely extensively on folksonomies. Folksonomies, as a widely accepted neologism and one of Web 2.0 signatures, can be thought of as keywords that describe what a document is about.

Since people started using the del.icio.us service in late 2003, many resources have been bookmarked and tagged collaboratively. Using the service, people usually tag a resource with words they feel best describes what it is about; these words or tags are popularly know as folksonomies.

Folksonomies can be thought of as keywords that describe what a document is about. We believe that most folksonomy words hold more semantic value than keywords extracted using generic or proprietary automatic keyword extraction techniques ('semantics' here means that a word can be a synonym or a generalization of a concept, etc.).

The main questions this experiment tries to answer are: do folksonomies only represent a set of keywords that describes what a document is about? Or do they go beyond the functionality of index keywords? How about the relation between the folksonomy tags and a librarian or an expert assigned keywords? Where are folksonomies positioned in the spectrum from professionally assigned keywords to context-based machine extracted keywords?

Therefore to answer these questions, our paper is organized as follows: In section 2, related work will be overviewed. In section 3, the experiment setup and the data selection will be discussed along with the four experiments we have carried out to assess our claim. Finally, the results of these experiments, as well as conclusions and future work will be discussed in sections 4, 5 and 6 respectively.

2. Related work

To the best of our knowledge, there was only one related work that has explored the area of folksonomies compared to other indexing mechanisms. Kipp [1] has examined the differences and similarities between the user keywords (folksonomies), the author and the intermediary (such as librarians) assigned keywords. She used a sample of journal articles tagged in the social bookmarking sites citeulike[3] and connotea[4], which are specialized for academic articles. Her

[1] http://www.flickr.com/, is a photo sharing website and web services suite, and an online community platform.
[2] http://del.icio.us, is a social bookmarking web service for storing, sharing, and discovering web bookmarks.

[3] http:// citeulike.org
[4] http://connotea.org

1-4244-0674-9/06/$20.00 ©2006 IEEE.

selection of articles was restricted to a set of journals known to include author assigned keywords and to journals indexed in Information Service for Physics, Electronics, and Computing (INSPEC[5]) database, so that each article selected would have three sets of keywords assigned by three different classes of metadata creators. Her methods of analyses were based on concept clustering via the INSPEC thesaurus, and descriptive statistics. She used these two methods to examine differences in context and term usage between the three classes of metadata creators. Kipp's findings showed that many users' terms were found to be related to the author and intermediary terms, but were not part of the formal thesauri used by the intermediaries; this was due to the use of broad terms which were not included in the thesaurus or to the use of newer terminology. Kipp then concluded her paper by saying that "*User tagging, with its lower apparent cost of production, could provide the additional access points with less cost, but only if user tagging provides a similar or better search context*."

Apparently, the method that Kipp used lacks comparing folksonomies to keywords extracted automatically using context-based extraction methods. This extra evaluation method will be significant to carry out to measure the relation between automatic machine indexing mechanisms lead by a major search engine like Yahoo compared to human indexing mechanisms, and whether is it possible to replace folksonomies with automatically extracted keywords.

3. Experiment setup and test data

There are plenty of keyword extraction techniques in information retrieval literature. Most of which are either experimental or proprietary that do not have a corresponding freely available product that can be used. Therefore, we are limited by what exists in this field such as, Search Engine Optimization (SEO) services[6] keyword analyzer tools, Kea[7] [2], an open source tool released under the GNU General Public License, and Yahoo Term Extractor[8].

Kea requires an extensive training in a specific domain of interest to come up with reasonable results; SEO tools on the other hand, were biased; they look for the appearance of popular search terms in a webpage when extracting keywords, besides the extraction technique they are using is very basic (e.g. they use word frequency/count). Therefore, the decision to use Yahoo API was made because the technique used by Yahoo's API to extract terms is context-based as described in [3], which means that it can generate results based on the context of a document; this will lift the burden of training the system to extract the appropriate keywords, and

Based on that, the experiment was conducted in four phases: the first phase was to measure, for a corpus of web literature stored in the del.icio.us bookmarking service, the overlap between the folksonomy set and Yahoo extracted keyword set. In the second phase, a human indexer was asked to generate a set of keywords for a sample of websites from our corpus and compare the generated set to the folksonomy and the Yahoo sets to measure the degree of overlap. The third phase was to expose a sample of the two sets (folksonomy and Yahoo keywords) to the indexer to evaluate in general which set holds greater semantic value than the other. The final phase was to use another modified instrument from Kipp [1] to further explore what semantic value did the tags and keywords gave us. Thus, the analysis of the experiment can be thought of as being in two forms: descriptive statistics (phase 1 and 2) and term comparison (phase 3 and 4).

The rest of this paper will talk about the comparison system framework, the data set and the different phases of the experiment along with the accomplished results.

3.1. The Comparison system framework

The system consisted of three distinct components: the Term Extractor, the Folksonomy Extractor and the Comparison Tool, as shown in Figure 1.

The *Term Extractor* consists of two main components which are: JTidy[9], an open source Java-based tool to clean up HTML documents and *Yahoo Term Extractor* (TE)[10], a web service that provides "*a list of significant words or phrases extracted from a larger content*". After cleaning up a webpage of HTML tags the result was passed to Yahoo TE to generate the appropriate keywords.

The Folksonomy Extractor that we developed was designed to fetch tags list for a particular website from the del.icio.us service and then normalize the tags, see [5].

The *Comparison Tool* role was to compare the list of folksonomy tags to Yahoo's keywords; by counting the number of overlapped keywords between the two sets.

[5] A database which provides an intermediary assigned controlled vocabulary for searchers.
[6] Example: http://www.searchengineworld.com/cgi-bin/kwda.cgi
[7] http://www.nzdl.org/Kea/
[8] http://developer.yahoo.com/search/content/V1/termExtraction.html
[9] http://sourceforge.net/projects/jtidy
[10] http://developer.yahoo.net/search/content/V1/termExtraction.html

The tool then calculates the percentage of overlap between the two sets using the following equation (1):

$$P = N/(F+K)-N \quad (1)$$

Where:
P: Percentage of overlap; N: Number of overlapped keywords, F_s: Size of folksonomy set; K_s: Size of keyword set.

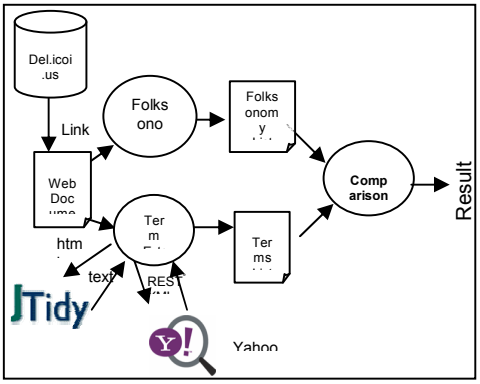

Figure 1. The Comparison System Framework.

3.2. Data selection

The test data used in the experiment was collected from the del.icio.us social bookmarking service. One hundred bookmarked websites[11] spanning various topics from the popular tags webpage were manually selected, as shown in Table 1.
The selection was based on the following heuristics:

- Bookmarked sites that are of a multimedia nature such as audio, video, flash, Word/PDF documents, etc. were avoided due to the limitation of Yahoo term extraction service (i.e. it only extracts terms from textual information). By the same token, whole Blog sites were avoided because they usually hold a diversity of topics. So, we tried to look for web pages with a single theme (e.g. a specific post in a Blog).
- We only choose bookmarked sites with 100+ participants; this was necessary to ensure there were enough tags describing the website.

[11] Data was collected between 24/Feb and 27/Feb 2006

Table 1. Topics covered in the experiment data set

Topic	Number of Web Sites
Software	11
Open source	14
Education	6
Programming	18
Sciences	8
Linux	10
References	13
Development	20
Total	100

4. Results

As mentioned in the experiment setup, the role of phases one and two was to find the percentage of overlap between folksonomy set and keywords extracted by Yahoo TE. The overlap can be interpreted using set theory. While the role of phase three and four was to manually scrutinize the list of folksonomy tags and Yahoo keywords to determine which list is semantically richer.

4.1. Results of phases 1, 2 and 3

The role of phase 1 was to measure the overlap between folksonomy set and Yahoo TE set. The results show that there is a partial overlap ($F \cap K$) between folksonomy set and keywords extracted using Yahoo TE with an overlap mean of 9.51%. Also the results show both maximum and minimum possible overlap with values equal to 21.82% and 1.96% respectively. This indicates that there is never likely to be complete overlap or no overlap at all between the two sets.

The role of phase two was to determine automatically the correlation between the folksonomy set and the human keyword assignment, and also between Yahoo TE keywords and the human assignment. This step is necessary to see which technique is highly related to a cataloguing (indexation) output. The results of this phase show that there is a partial overlap between the two sets and the indexer set, but this time with higher scores. The folksonomy set was more correlated to the indexer set with a mean of 19.48% and a standard deviation of 5.64%, while Yahoo TE set scored a mean of 11.69% with a standard deviation of 7.06%.

The role of phase three was to determine manually, from a general perspective, whether or not folksonomies carry more semantic value than keywords extracted using Yahoo TE. Thus, given the sets of

keywords from Yahoo TE and del.icio.us; the indexer was asked to evaluate each keyword from both sets compared to the main theme of a given web resource. The indexer was given a 5-point Likert scale that has the following values: "Strongly relevant"= 5, "Relevant"= 4, "Undecided"= 3, "Irrelevant"= 2 and "Strongly irrelevant"= 1.

After evaluating 10 websites from our data set, the results show that the folksonomy set scored a higher mode in Likert scale with a value of 4; which means that the folksonomy tags are more relevant to the human indexer conception. While, the Yahoo keywords set scored a mode of 1; which means keywords extracted using the Yahoo TE do not agree with the human conception.

For further in depth analysis of the first three phases results, the reader is referred to [4].

4.2. Results of phase 4

The role of phase four was to inspect in more detail the semantics of the folksonomy set (tags) and the Yahoo keywords set compared to the web resource hierarchical listing in the dmoz.org directory and to its title keywords (afterwards, these will be called descriptors). Thus, the indexer was provided with another 7-point Likert scale. The new Likert scale values were adopted from Kipp [1]. Kipp built her Likert scale instrument based on the different relationships in a thesaurus as an indication of closeness of match, into the following categories:

1. Same - the descriptors and tags or keywords are the same or almost the same (e.g. plurals, spelling variations and acronyms)
2. Synonym - the descriptors and tags or keywords are synonyms
3. Broader Term (BT) - the keywords or tags are broader terms of the descriptors
4. Narrower Term (NT) - the keywords or tags are narrower terms of the descriptors
5. Related Term - the keywords or tags are related terms of the descriptors
6. Related - there is a relationship (conceptual, etc) but it is not obvious to which category it belongs to
7. Not Related - the keywords and tags have no apparent relationship to the descriptors, also used if the descriptors are not represented at all in the keyword and tag lists.

The indexer applied the modified Likert scale on a sample of 10 bookmarked websites that were chosen from the experiment corpus. She first evaluated the folksonomy keywords based on the Likert scale then she evaluated the Yahoo extracted keywords based on the same scale. For each evaluated web resource, a two-column bar graph was generated to reflect the result of each category, i.e. the Blue bars denote the Yahoo keywords frequency and the Purple bars denote the Folksonomy keywords frequency.

Figure 2 shows a generated graph from the accumulated 10 bar graphs of the evaluated web resources, juxtaposed in a layered fashion, so that a general conclusion can be drawn easily.

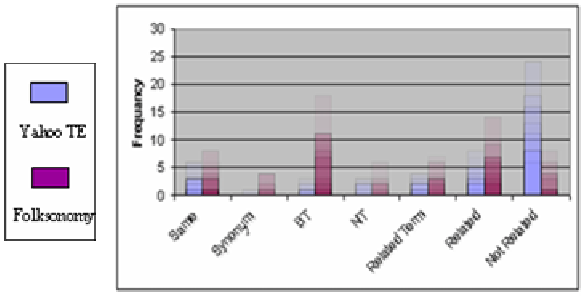

Figure 2. The similarity comparison results of the 10 web resource are layered on top of each other shaping a ghost effect.

The figure shows that the folksonomy tags are accumulating more around the 'Broader Term' and 'related' category, while the Yahoo keywords are accumulating more around the 'not related' category. The figure also shows that most of the folksonomy tags fall in the similarity categories compared to a small portion which falls in the 'not related' category. In contrast, most of the Yahoo keywords fall in the 'not related' category compared to a small portion distributed in the similarity categories. Also, the figure shows that in all similarity categories (i.e. Same, Synonym, BT, NT, Related Term and Related), the folksonomy set outperforms the Yahoo keyword set.

5. Discussion

After completing the four phases of this experiment, a number of observations were made. As a first impression, phase 3 was used to evaluate the relevance of the folksonomy tags and Yahoo TE keywords to the human conception. Thus, the results of this phase indicates a significant tendency of the folksonomy tags towards depicting what a human indexer might think of when describing what a web resource is about compared to Yahoo TE keywords.

Another interesting observation was found in phase 4, where some folksonomy tags fall in the 'Narrower

Term' and 'synonym' categories. These categories were less common than the 'Broader Term', 'Same' and 'Related Term' categories, which implies, from our point of view, that this might be due to the low number of specialized people who uses the del.icio.us bookmarking service, or it might be due to the varied backgrounds of the del.icio.us users.

While in phase 1 and 2, the folksonomy tags have showed more statistical significance than Yahoo TE keywords. In phase 1, the average overlap between the folksonomy set and Yahoo keywords was 9.51%, which implies that even if there was a minor intersection between the two sets, the folksonomy tags can not be replaced completely with machine generated keywords, in this case Yahoo TE. This finding also opens the door for other potential research directions, for instance in the filed of language technology and semantics, which is out of this experiment scope.

In phase 2, the results showed that the folksonomy set was more correlated to the indexer set with a mean of 19.48%, while Yahoo TE set scored a mean of 11.69%. This finding also emphasis our claim that there is a good correlation between folksonomies and professional indexing compared to the correlation between professional indexing and context-based machine extracted keywords.

Finally, it is worth mentioning that the results from this experiment have not been evaluated against a large corpus, especially where this concerns the sample size used by the indexer. This was due to the high effort needed for manual indexing. However, to get a fair judgment we have attempted to choose varied websites topics spanning multiple domains as shown in Table 1. We also think that the estimated sample size for each stage of the experiment was proportional to the amount of time and effort needed for the evaluation.

6. Conclusion and future work

It is clear from the results of this experiment that the folksonomy tags agree more closely with the human generated keywords than the automatically generated ones. The results also showed that the professional indexer has valued the semantic value of folksonomy tags compared to keywords extracted by Yahoo TE, when manually evaluating the experiment web resources. These results were vary encouraging, and illustrated the power of folksonomies. Folksonomies showed that they have added new contextual dimension that is not present in either automatic keywords extracted by machines or manually assigned keywords by an indexer.

Hopefully, the purpose of this experiment was satisfied by showing that folksonomies can be semantically richer than keywords extracted using a major search engine extraction service like Yahoo TE. This can justify the potential use of folksonomies in the process of semantic annotation.

So to conclude, the rational of this work was based on the motivation of investigating whether folksonomies can be used for automatically annotating web resources; as folksonomies are very popular and a potential rich source for metadata. Thus, the findings of this experiment can be used to justify the use of folksonomies in the process of generating semantic metadata for annotating learning resources; see [5].

7. References

[1] Kipp, M.E. Exploring the context of user, creator and intermediate tagging. in IA Summit 2006. Vancouver, Canada.

[2] Witten, I., G. Paynter, E. Frank, C. Gutwin, and C. Nevill-Manning. KEA: Practical Automatic Keyphrase Extraction. in In Proceedings of ACM DL'99. 1999.

[3] Kraft, R., F. Maghoul, and C.C. Chang. Y!Q: Contextual Search at the Point of Inspiration. in The ACM Conference on Information and Knowledge Management (CIKM'05). 2005. Bremen, Germany.

[4] Al-Khalifa, H.S. and H.C. Davis. Folksonomies versus Automatic Keyword Extraction: An Empirical Study. In Proceedings of IADIS Web Applications and Research 2006.

[5] Al-Khalifa, H. S. and Davis, H. C. FolksAnnotation: A Semantic Metadata Tool for Annotating Learning Resources Using Folksonomies and Domain Ontologies. *Proceedings of the Second International Conference on Innovations in Information Technology*. 2006, Dubai, UAE [In Press]