# Folksonomies versus Automatic Keyword Extraction: An Empirical Study

Hend S. Al-Khalifa and Hugh C. Davis

Learning Technology Research Group, ECS, University of Southampton, Southampton, SO17 1BJ, 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 a 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 such as del.icio.us can be used in the process of generating semantic metadata to annotate web resources.

## 1 Introduction

Nowadays, contemporary web applications such as Flickr[1], del.icio.us[2] and Furl[3] rely extensively on folksonomies. Folksonomies, as a widely accepted neologism, 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 folksonomy.

We believe that most folksonomy words hold more semantics 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

[1] http://www.flickr.com/
[2] http://del.icio.us
[3] http://www.furl.net/

concept, etc). The value of this experiment is to prove that folksonomies carry more semantic value than keywords extracted by machines.

This paper is organized as follows: In section 2, a literature review of some of the different kinds of keyword extraction techniques will be overviewed. In section 3, the experiment setup and the data selection will be discussed along with the three experiments we have carried out to assess our hypothesis. Finally, the results of these experiments, as well as some conclusions and future work are discussed in sections 4, 5 and 6 respectively.

## 2 Literature Review

Keyword extraction -as a field of Information Retrieval (IR)- is an approach to formally study document text to obtain "*cognitive content hidden behind the surface*" [1]. Keyword extraction tools vary in complexity and techniques. Simple term extraction is based on term frequency (*tf*) while complex ones use statistical e.g. [2], or linguistic techniques such as Natural Language Processing (NLP) [3] supported by domain specific ontologies e.g. [4].

There are a wide variety of applications that use automatic keyword extraction and among these are document summarization and news finding e.g. [5]. The keyword analyzer service[4] used by most Search Engine Optimization (SEO) companies is another type of keyword extraction application using term frequency. Most complex keyword extraction techniques require corpus training in a specific domain for example Kea[5], a keyphrase extraction algorithm, [6].

On the other hand, search engines use one kind of keyword extraction called indexing, where the full search is constructed by extracting all the words in a document except stop words. After all the keywords have been extracted from the document they need to be filtered, since not all words are good for indexing. The filtering can be done using the vector space model or more specifically, latent semantic analysis [5, 7].

Most indexing techniques rely on statistical methods or on the documents term distribution tendency and are typically based on term frequency, which ignores the semantics of the document content. This is because term frequency techniques are based on the frequency of occurrences of terms in a document.

Statistical methods suffer from limitations that diminish the precision of the extracted indexes and they also fail in extracting semantic indexes to represent the main concepts of a document [8]. This problem might be partially solved by using people assigned keywords or tags (i.e. folksonomies) on bookmarking systems like del.icio.us.

[4] Example: http://www.searchengineworld.com/cgi-bin/kwda.cgi

[5] http://www.nzdl.org/Kea/

## 3 Experiment Setup and Test Data

There are plenty of keyword extraction techniques in IR literature. Most of which are either experimental or proprietary that do not have a corresponding freely available product that can be used. So we were limited by what exists in this field such as, SEO keyword analyzer tools, Kea, an open source keyphrase extraction tool released under the GNU General Public License, and Yahoo API term extractor[6].

Kea requires an extensive training in a specific domain of interest to come out with reasonable results. SEO tools on the other hand, were biased (i.e. they look for the appearance of popular search terms in a webpage when extracting keywords), besides the IR techniques they are using are very basic (e.g. word frequency/count). Therefore, the decision to use Yahoo API was made for the following reasons:

- Yahoo is one of the leading search engines on the Internet, and the index size is relevantly close to its competitor[7] (i.e. Google).
- The technique used by Yahoo's API to extract terms is context-based as described in [9], which means 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.
- Finally, Yahoo's recent policy of providing web developers with a variety of API's encouraged us to test the quality of their term extraction service.

Based on that, our experiment was conducted in three phases: the first phase was to measure, for a corpus of web literature, the overlap between the folksonomy set and Yahoo generated 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 final phase was to expose a sample of the two sets (folksonomy and Yahoo keywords) to the indexer to evaluate which set holds greater semantic value than the other.

The test data used in the experiment was collected from the social bookmarking service del.icio.us. Del.icio.us is one of the largest bookmark services on the internet with more than 60,000 users and over one million unique tagged bookmarked websites[8].

### 3.1 Comparison System Framework

Our system consisted of three distinct components: the Term Extractor, the Folksonomy Extractor and the Comparison Tool as shown in Fig.1. The *Term Extractor* consists of two main components which are: JTidy[9], an open source Java-

6 Yahoo API term extractor service was lunched on May 2005

7 A study by Matthew Cheney and Mike Perry from UIUC, http://vburton.ncsa.uiuc.edu/indexsize.html

8 From del.icio.us

9 http://sourceforge.net/projects/jtidy

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 is passed to Yahoo TE to generate the appropriate keywords.

*The Folksonomy Extractor* that we developed is designed to fetch keywords (aka tags) list for a particular website from del.icio.us and then clean-up the list by pruning and grouping tags. Finally, the *Comparison Tool* role is to compare the list of folksonomy to Yahoo's keywords; by counting the number of overlapped keywords between the two sets. The tool then calculates the percentage of overlap between the two sets using the following equation:

$$P = \frac{N}{(\mathrm{Fs} + \mathrm{Ks}) - \mathrm{N}} \times 100 \quad (1)$$

Where:

| | |
|---|---|
| P | Percentage of overlap |
| N | Number of overlapped keywords |
| Fs | Size of folksonomy set |
| Ks | Size of keyword set |

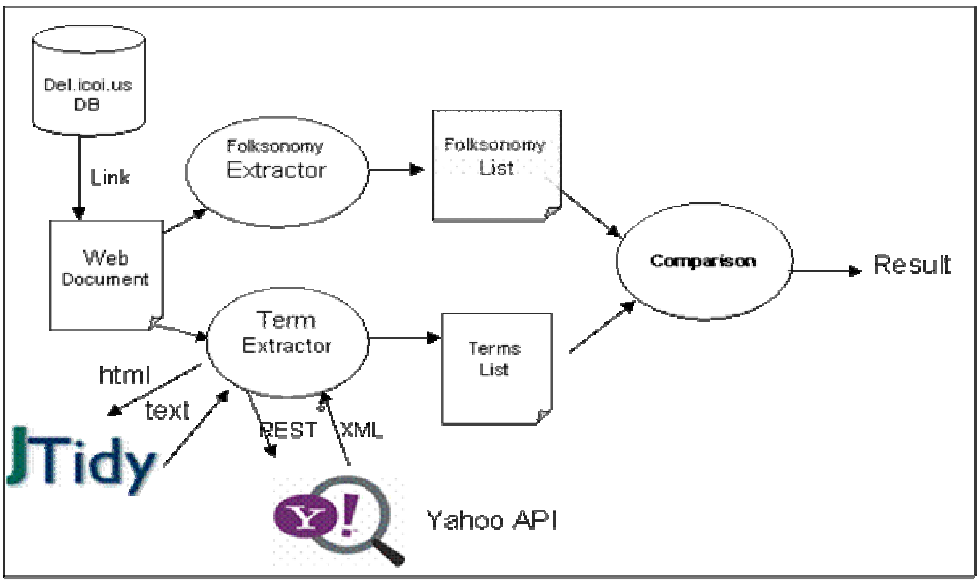


**Fig. 1.** The Comparison System Framework

10 http://developer.yahoo.net/search/content/V1/termExtraction.html

º Representational State Transfer (REST) is a software architectural style for distributed hypermedia systems like the World Wide Web.

### 3.2 Data Selection

One hundred bookmarked websites[11] spanning various topics (*software, open source, education, programming, sciences, Linux, references and development*) were manually selected from del.icio.us bookmarking service based on the following heuristics:

- The topics selected represent the most popular tags at the time of the experiment[12].
- 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 or more participants; this was necessary to ensure there were more than 10 tags describing the website.

### 3.3 Other General Heuristics

Some heuristics were used during the experiment lifecycle, to improve the quality of the extraction results which are listed as follows:

- Google Adsense (an advertisement tool by Google), where present, often affected the results of the terms returned by Yahoo extractor. Therefore, in some cases we were forced to manually enter (i.e. copy and paste) the text of a website and place it in a web form that invokes the Yahoo TE service.
- Since Yahoo TE is limited only to twenty terms that represent the best candidate for a website (as mentioned on the service website); these terms were split out into single words so that they might match del.icio.us style single word tags.

## 4 Results

As mentioned in the experiment setup, the role of phases one and two is to find the percentage of overlap between folksonomy list and keywords generated by Yahoo TE. The overlap can be interpreted using set theory [10].

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

[12] To see a tag cloud for the most popular tags please visit http://del.icio.us/tag/

**o** Representational State Transfer (REST) is a software architectural style for distributed hypermedia systems like the World Wide Web.

We considered the folksonomy list of tags as set F, keywords list from Yahoo TE as set K and keywords list from the indexer as set I, hence:

F = {the set of all tags generated by people for a given URL in del.icio.us}
K = {the set of all automatically extracted keywords for a given URL}
I = {the set of all keywords provided by the indexer}

Using set theory the degree of overlap was described using the following categories:

- No overlap e.g. $F \neq K$ or $F \cap K = \varnothing$ (i.e. empty set).
- Partial overlap (this is know as the intersection) e.g. $F \cap K$
- Complete overlap (also know as containment or inclusion). This can be satisfied if the number of overlapped keywords equals to the folksonomy set (i.e. $F \subset K$) or if the number of overlapped keywords equals to the Yahoo keyword set (i.e. $K \subset F$) or if the number of overlapped keywords equals both folksonomy and keyword set (i.e. $F = K$).

### 4.1 Phase 1

After observing the results of 100 websites as shown in Fig. 2, we can detect that there is a partial overlap ($F \cap K$) between folksonomies and keywords extracted using Yahoo TE. The results show that the mean of the overlap was 9.51% with a standard deviation of 4.47% which indicates a moderate deviation from the sample mean. Also the results show both the maximum and the 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. Finally, the most frequent percentage of overlap (i.e. mode) was 12.5%.

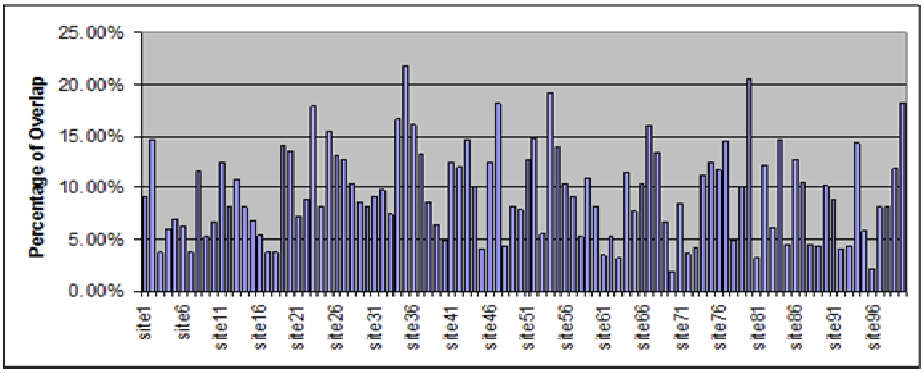


**Fig. 2.** Distribution of the percentage of overlap for 100 websites

### 4.2 Phase 2

The role of phase two is to check the correlation between folksonomy and 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.

Therefore, tools from library and information science were used to index a sample of 20 websites taken from our corpus and to check them against folksonomy and Yahoo TE sets. The assignment of keywords was done using the following guidelines:

- The use of controlled vocabularies of terms for describing the subject of a website, such as DMOZ[13] (the Open Directory Project) and Yahoo directory.
- The source code of each website was checked to see if it contains any keywords provided by the website creator.
- The position in titles and emphasis (such as bold) of words in a website were considered.
- The indexer was also asked to read a website and generate as many keywords as possible.

After the end of this process the produced keywords for each website was compared with the keywords from the Yahoo TE set and also with the folksonomy set. This step is essential for us to see whether folksonomies produced the same results as if a human indexer was doing the process.

The results show (Fig.3) that there is a partial overlap between the two sets and the indexer set, but this time with higher scores. For instance, 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%.

Furthermore, the experiment showed one case where there is a complete overlap (inclusion) between the folksonomy set and the indexer set. This supports our assumptions about the semantic value of folksonomies.

[13] http://dmoz.org/

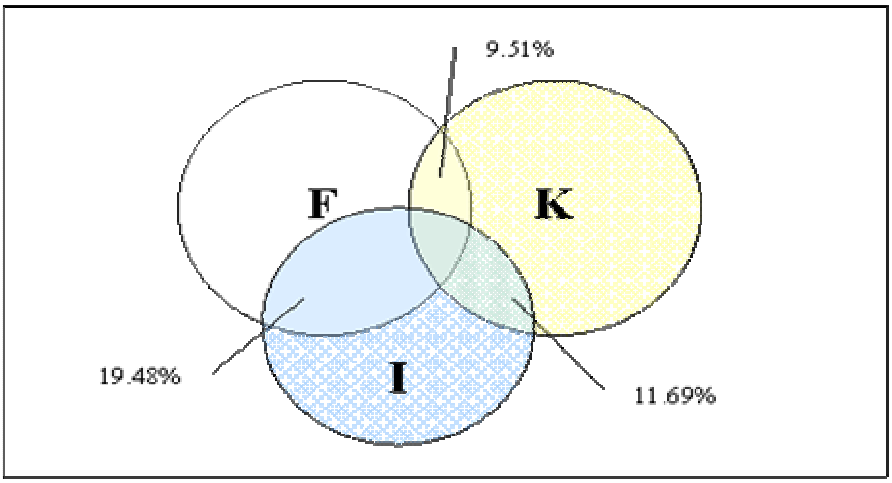


**Fig. 3.** : A Venn diagram showing the mean of the percentage of overlap between Folksonomy (F), Yahoo TE (K) and the human indexer (I) set

### 4.3 Phase 3

The role of phase three is to determine 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. 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 corpus, the results in Table 1 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 keyword set scored a mode of 1; which means keywords extracted using the Yahoo TE do not agree with the human conception.

We note that there are a small number of "perfect" matches in the results of the automatic keyword extraction (sites 4 and 8). This was due to the small number of keywords generated by Yahoo TE for these sites; there were very few keywords but what was there was indeed perfect. On the other hand, the folksonomy set never reached the perfection condition (i.e. 5); this was due to the indexer's unfamiliarity with some of the websites domain.

| Site | F | K |
|---|---|---|
| 1 | 4 | 1 |
| 2 | 4 | 1 |
| 3 | 4 | 1 |
| 4 | 4 | 5 |
| 5 | 4 | 4 |
| 6 | 4 | 1 |
| 7 | 4 | 1 |
| 8 | 4 | 5 |
| 9 | 4 | 4 |
| 10 | 4 | 1 |
| | | |
| Mode | 4 | 1 |

**Table 1.** The mode value of Likert score for Folksonomy (F) set and Yahoo TE (K) set

## 5 Discussion

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 discussed in section 3.2). 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.

Finally, the results are vary encouraging and do show the power of folksonomies. This can prove the potential use of folksonomies in the process of semantic annotation.

## 6 Conclusion

After completing the three phases of the experiment it is clear from the results that the folksonomy tags agree more closely with the human generated keywords than the automatically generated ones.

In addition, the purpose of this experiment was satisfied by proving that folksonomies can be semantically richer than the keywords extracted using a major search engine service like Yahoo TE. The experiment also showed the percentage of overlap between folksonomies and automatically extracted keywords for a given website.

As a result, the findings of this experiment will be used to justify the use of folksonomies in the process of generating semantic metadata to annotate web resources.

## 7 Future Work

This experiment is part of PhD research to build a document-level semantic metadata annotation tool using folksonomies and domain ontologies. The experiment justified the rich semantic value of folksonomies. Therefore, the logical next step in the PhD lifecycle is to implement the document-level semantic metadata annotation tool to annotate web resources bookmarked using del.icio.us bookmarking service.

## References

1. Hunyadi, L. Keyword extraction: aims and ways today and tomorrow. in In: Proceedings of the Keyword Project: Unlocking Content through Computational Linguistics. 2001.
2. Matsuo, Y. and M. Ishizuka, Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information. International Journal on Artificial Intelligence Tools, 2004. 13(1): p. 157-169.
3. Sado, W.N., D. Fontaine, and P. Fontaine. A linguistic and statistical approach for extracting knowledge from documents. in Proceedings of the 15th International Workshop on Database and Expert Systems Applications (DEXA'04). 2004: IEEE Computer Society.
4. Hulth, A., J. Karlgren, A. Jonsson, H. Boström, and L. Asker. Automatic Keyword Extraction Using Domain Knowledge. in Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing. 2001.
5. Martinez-Fernandez, J.L., A. García-Serrano, P. Martínez, and J. Villena, Automatic Keyword Extraction for News Finder. LNCS, 2004. 3094.
6. 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.
7. Landauer, T.K., P.W. Foltz, and D. Laham, Introduction to Latent Semantic Analysis. Discourse Processes, 1998. 25: p. 259-284.
8. Kang, B.-Y. and S.-J. Lee, Document indexing: a concept-based approach to term weight estimation. Information Processing and Management: an International Journal, 2005. 41(5): p. 1065 - 1080.
9. 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.
10. Stoll, R.R., Set Theory and Logic. 1979, Mineola, N.Y.: Dover Publications.