IADIS International Journal on Computer Science and Information Systems
Vol. 1, No. 2, pp. 132-143
ISSN: 1646-3692

FOLKSONOMIES VERSUS AUTOMATIC KEYWORD EXTRACTION: AN EMPIRICAL STUDY

Hend S. Al-Khalifa and Hugh C. Davis *Learning Technology Group- ECS- Southampton University, Southampton, SO17 1BJ, 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 contain higher semantic value than keywords extracted by machines. The experiment has been carried-out in two ways: subjectively, by asking a human indexer to evaluate the quality of the generated keywords from both systems; and automatically, by measuring the percentage of overlap between the folksonomy set and machine generated keywords set.
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.

KEYWORDS

Folksonomy, Keyword Extraction, Tags, Semantics.

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 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

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

words they feel best describe what it is about; these words or tags are popularly known as folksonomies.

We believe that most folksonomy words have a higher 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 describe what a document is about, or do they go beyond the functionality of index keywords? What is the relationship between folksonomy tags and keywords assigned by an expert such as a librarian? Where are folksonomies positioned in the spectrum from professionally assigned keywords to context-based machine extracted keywords?

To answer these questions, our paper is organized as follows: In section 2 the background and related work will be reviewed. 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. BACKGROUND AND RELATED WORK

Scholarly work encompassing our paper theme comes in two themes. The keyword extraction theme, with its various techniques will be outlined, and in the other theme, related work discussing folksonomies compared to other indexing mechanisms will be highlighted.

Keywords extraction -as a field of Information Retrieval (IR)- is an approach to formally study document text to obtain "*cognitive content hidden behind the surface*" (Hunyadi, 2001). Keyword extraction tools vary in complexity and techniques. Simple term extraction is based on term frequency (*tf*) while complex ones use statistical techniques e.g. (Matsuo and Ishizuka, 2004), or linguistic techniques 'Natural Language Processing (NLP)' e.g. (Sado et al., 2004) supported by domain specific ontologies e.g. (Hulth et al., 2001). There are a wide variety of applications that use automatic keyword extraction, among these are document summarization and news finding e.g. (Martinez-Fernandez et al., 2004). Keyword analyzer services[4] used by most Search Engine Optimization (SEO) companies are another type of keyword extraction application using *tf*. Most complex keyword extraction techniques require corpus training in a specific domain for example Kea[5] - a keyphrase extraction algorithm- (Witten et al., 1999).

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 of a document except stop words. After all the keywords that have been extracted from the document need to be filtered; since not all words can be adequate for indexing. The filtering can be done using vector space model or more specifically the latent semantic analysis (Landauer et al., 1998; Martinez-Fernandez et al., 2004).

From our previous discussion we find that most indexing methods are based on *tf*, which ignores the semantics of the document content. This is because *tf* technique is based on the occurrences of terms in a document by assigning a weight to indicate its importance. Most indexing techniques rely on statistical methods or on the documents term distribution

[4] Example: http://www.searchengineworld.com/cgi-bin/kwda.cgi
[5] http://www.nzdl.org/Kea/

tendency. Statistical methods lack precision and they fail in extracting the semantic indexes to represent the main concepts of a document (Kang and Lee, 2005). This problem might be partially solved by using people assigned keywords or tags (i.e. folksonomies) on bookmarking systems like del.icio.us.

Kipp (2006) has examined the differences and similarities between 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[6] and connotea[7], which are specialized for academic articles. Her 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) 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 did not compare folksonomies to keywords extracted automatically using context-based extraction methods. This extra evaluation method will be significant in measuring the relationship between automatic machine indexing mechanisms led by a major search engine like Yahoo compared to human indexing mechanisms, and whether folksonomies have greater semantic value than automatically extracted keywords.

3. EXPERIMENT SETUP AND TEST DATA

There are plenty of keyword extraction techniques in IR literature, most of which are either experimental or proprietary; so they do not have a corresponding freely available product that can be used. Therefore we were limited by what exists in this field such as, SEO keyword analyzer tools, Kea, an open source tool released under the GNU General Public License, and Yahoo API term extractor[8]. Of these the Yahoo API was the preferred choice.

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:

[6] http://www.citeulike.org
[7] http://www.connotea.org
[8] Yahoo API term extractor service was launched on May 2005

- The technique used by Yahoo's API to extract terms is context-based as described in (Kraft et al., 2005), 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, the experiment was conducted in four phases: in the first phase we exposed a sample of both folksonomy and Yahoo keywords sets to a human indexer to evaluate -in general- which set holds greater semantic value than the other. In the second phase, we used another modified instrument from (Kipp, 2006) to further explore what semantic value did the folksonomy tags and Yahoo keywords gave us. In the third phase, we measured, 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 final 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 set and the Yahoo set to measure the degree of overlap. Thus, the analysis of the experiment can be thought of as being in two forms: term comparison (phase 1 and 2) and descriptive statistics (phase 3 and 4).

The rest of this paper will talk about the comparison system framework used for evaluating phase 3 and 4, the data set and the different phases of the experiment along with the accomplished results.

3.1 The Comparison System Framework

Our system was composed of three distinct components: Term Extractor, Folksonomy Extractor and the Comparison Tool as shown in Figure 1. *Term Extractor* is composed 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 website from HTML tags the result is passed to Yahoo TE to generate the appropriate keywords.

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 (1):

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

The above equation can be also expressed using set theory as (2):

$$P = \frac{F_s \cap K_s}{F_s \cup K_s} \times 100 \qquad (2)$$

[9] http://sourceforge.net/projects/jtidy
[10] http://developer.yahoo.net/search/content/V1/termExtraction.html

Where:

P Percentage of overlap
N Number of overlapped keywords
F_s Size of folksonomy set
K_s Size of keyword set

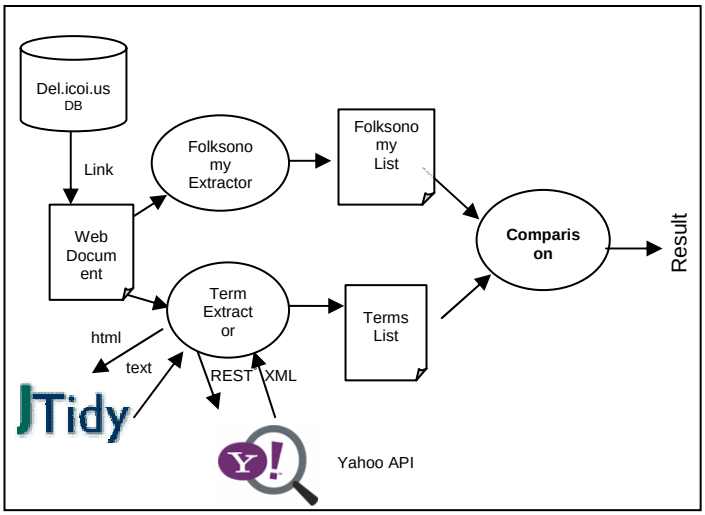

Figure 1. The Comparison System Framework

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 spanning various topics from the popular tags page, as shown in Table I.

Table I: 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

Web resources were manually selected 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.

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:

1. Most websites that use Google Adsense (an advertisement tool by Google) 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.
2. Yahoo TE is limited to produce only twenty terms, which may consist of one or more words to 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 AND DISCUSSION

4.1 Phase 1

The role of phase one is to determine whether or not folksonomies carry more semantics 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 data set, the results in Table 2 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.

Table 2. The mode value of Likert score for Folksonomy (F) set and Yahoo TE (K) set

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

4.2 Phase 2

The role of phase two 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, 2006). 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 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 evaluation the results are plotted as a bar graph, where the Blue bars denote the Yahoo keywords frequency and the Purple bars denote the Folksonomy keywords frequency.

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

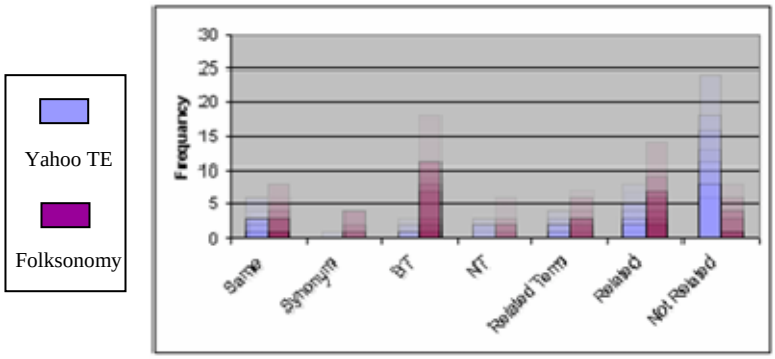

Figure 2. The similarity comparison results of the 10 web resource are layered on top of each other shaping a ghost effect.

The figure also 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. Finally, the figure shows that in all similarity categories, the folksonomy set outperforms the Yahoo keyword set.

4.2.1 More in depth analysis of Phase 2

In this section a detailed analysis of both the Yahoo keywords set and the folksonomy set falling in the 'not related' and 'related' categories will be discussed.

A) Unrelated tags

To explore more the nature of tags falling in the 'not related' category, a further inspection was carried out to analyze the type of tags and keywords found in this category.

Folksonomy tags falling in the 'not related' category tend to be either time management tags e.g. todo, toread, toblog, etc., or expression tags e.g. cool, and sometimes they might be unknown/uncommon abbreviations.

Time management tags, as Kipp said, suggest that the users want to be reminded of the bookmarked resource, but have not yet decided what to do with it. These kinds of tags do not appear in any controlled vocabulary or thesaurus; they are made up for the user's own needs and do not have any value to anyone except the individual who created it. Therefore, time management tags, from the thesis perspective, can not be used in the process of semantic annotation.

Another common type of unrelated tags is the use of expression tags e.g. 'cool', 'awesome', etc. These reflect what the users think of the bookmarked resource. The expression tags suggest that the bookmarked web resource might be useful.

On the other hand, Yahoo keywords falling in the 'not related' category do not follow a recognized pattern as folksonomy tags do. Most keywords seem to be words that have occurred frequently in the text or in the URL of a web resource or the position of the word and its style (e.g. heading or sub-title) might be one reason for extracting it. The algorithms that

Yahoo TE uses to extract keywords from web sites are obscure, thus we can not elaborate more in analyzing the extracted keywords.

B) Related tags

This category represents relationships that are ambiguous or difficult to place into the previous 1-5 similarity categories. These tags often occur when there is a relationship between the tag or keyword and its field of study, or/and a relationship between two fields of study (Kipp, 2006)

An example of the first mentioned relationship would be of a web resource talking about open source software which has tags such as 'code' or 'download'. These two tags do not appear explicitly in the dmoz.org directory listing nor in the title of the web resource; however, they describe the field of 'open source' software where someone can download and play with the code.

Another example of a relationship between two fields of study is a web resource about an open source office application called 'NeoOffice' for Mac operating system. This web resource is tagged with tags such as 'Microsoft' and 'OpenOffice' to denote the relationship between the 'Mac OS' and 'Microsoft' and between 'NeoOffice' and 'OpenOffice' application.

4.3 Phase 3

As mentioned in the experiment setup, the role of phase (three and four) was to find the percentage of overlap between folksonomy set and keywords generated by Yahoo TE. This extra evaluation phase was needed to see where folksonomies are positioned in the spectrum from professionally assigned keywords to context-based machine extracted keywords, and to measure the scope of this overlap. The overlap can be interpreted using set theory (Stoll, 1979).

We considered the folksonomy set of tags as set *F*, keywords set from Yahoo TE as set *K* and keywords set 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:

1. No overlap e.g. $F \neq K$ or $F \cap K = \varnothing$ (i.e. empty set).
2. 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. $F = K$).

After observing the results of 100 websites as shown in Figure 3 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 neither complete overlap nor no overlap at all. Finally, the most frequent percentage of overlap (i.e. mode) was 12.5%.

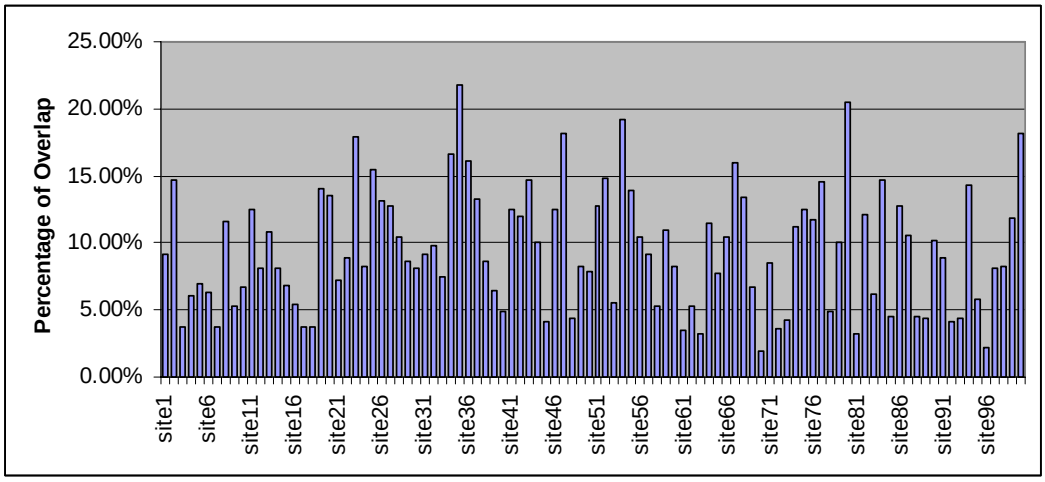

Figure 3. Distribution of the percentage of overlap for 100 websites

4.4 Phase 4

The role of phase four 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 data set and to check them against folksonomy and Yahoo TE sets. The assignment of keywords was done using the following guidelines:

1. The use of controlled vocabularies of terms for describing the subject of a website, such as DMOZ[11] (the Open Directory Project) and Yahoo directory.
2. The source code of each website was checked to see if it contains any keywords provided by the website creator.
3. The position (i.e. in titles) and emphasis (such as bold) of words in a website were considered.
4. The indexer also was asked to read the content of the website and generate as many keywords as possible.

After the end of this process the set of produced keywords for each website was compared once 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 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.

[11] http://dmoz.org/

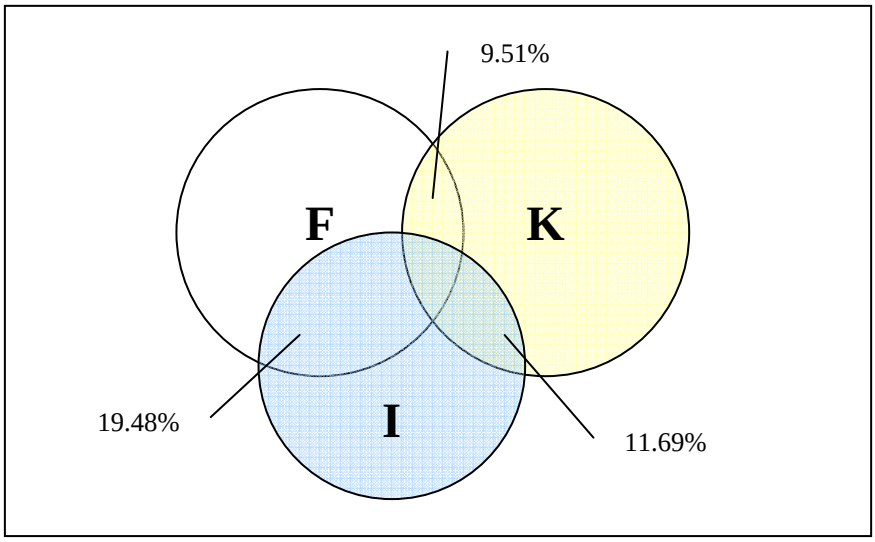

Figure 4. The mean of the percentage of overlap between Folksonomy (F), Yahoo TE (K) and the human indexer (I) set

5. DISSCUSSION

Unsurprisingly, some folksonomy tags did fall in the 'Narrower Term' and 'synonym' relationships; however, these relationships were less common than the Broader Term, 'Same' and 'Related Term' categories. This might be due to the low number of specialized people who use the del.icio.us bookmarking service, or it might be due to the varied backgrounds of the del.icio.us users.

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.

Finally, the results were vary encouraging, and illustrate 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. This can justify the potential use of folksonomies in the process of semantic annotation.

6. CONCLUSION AND FUTURE WORK

After completing the four 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 showing that folksonomies can be semantically richer than keywords extracted using a major search engine extraction service like Yahoo TE.

The rational of this work was built on the motivation of investigating whether folksonomies can be used for automatically annotating learning resources; see (Al-Khalifa and Davis, 2006). 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.

REFERENCES

Al-Khalifa, H. S. and Davis, H. C. (2006). 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*. IEEE Computer Society, Dubai, UAE [Accepted]

Hulth, A., et al. (2001). Automatic Keyword Extraction Using Domain Knowledge. *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*.

Hunyadi, L. (2001). Keyword extraction: aims and ways today and tomorrow. *In: Proceedings of the Keyword Project: Unlocking Content through Computational Linguistics*.

Kang, B.-Y. and S.-J. Lee (2005). "Document indexing: a concept-based approach to term weight estimation." *Information Processing and Management: an International Journal* 41(5): 1065 - 1080.

Kipp, M.E. (2006). Exploring the context of user, creator and intermediate tagging. *in IA Summit 2006*. Vancouver, Canada.

Kraft, R., et al. (2005). Y!Q: Contextual Search at the Point of Inspiration. *The ACM Conference on Information and Knowledge Management (CIKM'05)*, Bremen, Germany.

Landauer, T. K., et al. (1998). "Introduction to Latent Semantic Analysis." *Discourse Processes* 25: 259-284.

Martinez-Fernandez, J. L., et al. (2004). "Automatic Keyword Extraction for News Finder." *LNCS* 3094.

Matsuo, Y. and M. Ishizuka (2004). "Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information." *International Journal on Artificial Intelligence Tools* 13(1): 157-169.

Sado, W. N., et al. (2004). A linguistic and statistical approach for extracting knowledge from documents. *Proceedings of the 15th International Workshop on Database and Expert Systems Applications (DEXA'04)*, IEEE Computer Society.

Stoll, R. R. (1979). *Set Theory and Logic*. Mineola, N.Y., Dover Publications.

Witten, I., et al. (1999). KEA: Practical Automatic Keyphrase Extraction. *In Proceedings of ACM DL'99*.