EXPLORING THE VALUE OF FOLKSONOMIES FOR CREATING SEMANTIC METADATA

Hend S. Al-Khalifa and Hugh C. Davis
Learning Societies Lab
School of Electronics and Computer Science (ECS)
University of Southampton, UK
{hsak04r/hcd}@ecs.soton.ac.uk

ABSTRACT

Finding good keywords to describe resources is an on-going problem: typically we select such words manually from a thesaurus of terms, or they are created using automatic keyword extraction techniques. Folksonomies are an increasingly well populated source of unstructured tags describing web resources. This paper explores the value of the folksonomy tags as potential source of keyword metadata by examining the relationship between folksonomies, community produced annotations, and keywords extracted by machines. The experiment has been carried-out in two ways: subjectively, by asking two human indexers 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 results of this experiment show that the folksonomy tags agree more closely with the human generated keywords than those automatically generated. The results also showed that the trained indexers preferred the semantics of folksonomy tags compared to keywords extracted automatically.

These results can be considered as evidence for the strong relationship of folksonomies to the human indexer's mindset, demonstrating that folksonomies used in the del.icio.us bookmarking service are a potential source for generating semantic metadata to annotate web resources.

KEYWORDS

Metadata, Web Technologies, Folksonomy, Keyword Extraction, Tags, Social Bookmarking Services, Collaborative Tagging

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 words they feel best describe what it is about; these words or tags are popularly known as folksonomies and the process as collaborative tagging.

We believe that most folksonomy words are more related to a professional indexer's mindset than keywords extracted using generic or proprietary automatic keyword extraction techniques.

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 indexer? Where are folksonomies positioned in the spectrum from professionally assigned keywords to context-based machine extracted keywords?

In order to find out if folksonomies can improve on automatically extracted keywords, it is significant to examine the relationship between them, and between them and professional human indexer keywords.

To study these relationships, our paper is organized as follows: we begin with an overview of folksonomies and social bookmarking services, followed by a review of related work concerning folksonomies and keyword extraction techniques. We then discuss the experimental setup and the data selection, along with the four experiments we have carried out to examine the degree of the relationship. Finally, the results of these experiments, as well as a case study, conclusions and future work are discussed.

FOLKSONOMY AND SOCIAL BOOKMARKING SERVICES

The growing popularity of folksonomies and social bookmarking services has changed how people interact with the Web. Many people have used social bookmarking services to bookmark web resources they feel most interesting to them, and folksonomies were used in these services to represent knowledge about the bookmarked resource.

Folksonomies

The word folksonomy is a blend of the two words 'Folks' and 'Taxonomy'. It was first coined by the information architect Thomas Vander Wal in August of 2004. Folksonomy as Thomas (Vander Wal, 2006) defines is: "… *the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval. The tagging is done in a social environment (shared and open to others). The act of tagging is done by the person consuming the information*."

From a categorization perspective, folksonomy and taxonomy can be placed at the two opposite ends of categorization spectrum. The major difference between folksonomies and taxonomies are discussed thoroughly in (Shirky, 2005) and (Quintarelli, 2005).

Taxonomy is a top-down approach. It is a simple kind of ontology that provides hierarchical and domain specific vocabulary which describes the elements of a domain and their hierarchal relationship. Moreover, they are created by domain experts and librarians, and require an authoritative source.

In contrast, folksonomy is a bottom-up approach. It does not hold a specific vocabulary nor does it have an explicit hierarchy. It is the result of peoples' own vocabulary, thus, it has no limit (it is open ended), and tags are not stable nor comprehensive. Most importantly, folksonomies are generated by people who have spent their time exploring and interacting with the tagged resource (Wikipedia, 2006).

Social Bookmarking Service

Social bookmarking services are server-side web applications; where people can use these services to save their favorite links for later retrieval. Each bookmarked URL is accompanied by a line of text describing it and a set of tags (aka folksonomies) assigned by people who bookmarked the resource (as shown in Figure 1).

Figure 1: Excerpt from the del.icio.us service showing the tags (Blogs, internet, ... ,cool) for the URL of the article by Jonathan J. Harris, the last bookmarker (pacoc, 3mins ago) and the number of people who bookmarked this URL (1494 other people)

A plethora of bookmarking services such as Furl [4], Spurl [5] and del.icio.us exists; however, del.icio.us is considered one of the largest social bookmarking services on the Web. Since its introduction in December 2003, it has gained popularity over time and there have been more than 90,000 registered users using the service and over a million unique tagged bookmarks (Sieck, 2005; Menchen, 2005). Visitors and users of the del.icio.us service can browse the bookmarked URLs by user, by keywords (i.e. tags or folksonomies) or by a combination of both techniques. By browsing others' bookmarks, people can learn how other people tag their resources thus increasing their awareness of the different usage of the tags. In addition, any user can create an inbox for other users' bookmarks, by subscribing to another user's del.icio.us pages. Also, users can subscribe to RSS feeds for a particular tag, group of tags or other users.

RELATED WORK

In this section we review related work discussing state-of-the-art folksonomy research and various techniques in the keyword extraction domain.

State-of-the-art Folksonomy Research

There is a lot of recent research dealing with folksonomies, among these are overviews of social bookmarking tools with special emphasis on folksonomies as provided by Hammond et al., (2005) and other research papers that discuss the strengths and weaknesses of folksonomies (Mathes, 2004) (Kroski, 2006) (Quintarelli, 2005) (Guy & Tonkin, 2006).

Another genre of research has experimented with folksonomy systems. For instance, (Mika, 2005) has carried out a study to construct a community-based ontology using del.icio.us as a data source. He created two lightweight ontologies out of folksonomies; one is the actor-concept (user-concept) ontology and the other is the concept-instance ontology. The goal of his

experiment was to show that ontologies can be built using the context of the community in which they are created (i.e. the del.icio.us community). By the same token, Tom Guber is working on a system called 'TagOntology' to build ontologies out of folksonomies, and in his paper entitled "*Ontology of Folksonomy: A Mash-up of Apples and Oranges*" (Gruber, 2005) he cast some light on some design considerations needed to be taken into account when constructing ontologies from tags. In addition, (Ohmukai et al., 2005) proposed a social bookmark system, called 'socialware', using several representations of personal network and metadata to construct a community-based ontology. The personal network was constructed using Friend-Of-A-Friend (FOAF), Rich Site Summary (RSS), and simple Resource Description Framework Schema (RDFS), while folksonomies were used as the metadata. Their system allows users to browse friends' bookmarks on their personal networks, and map their own tag onto more than one tag from different friends, so that they are linked by the user. This technique will allow for efficient recommendation for tags because it is derived from personal interest and trust. They also used their social bookmark system 'socialware' to design an RDF-based metadata framework to support open and distributed models.

(Golder & Huberman, 2006), from HP Labs, have analyzed the structure of collaborative tagging (folksonomies) to discover the regularities in user activity, tag frequencies, the kind of tags used and bursts of popularity in bookmarked URLs in the del.icio.us system. They also developed a dynamic model that predicts the stable patterns in collaborative tagging and relates them to shared knowledge. Their results show that a significant amount of tagging is done for personal use rather than public benefit. However, even if the information is tagged for personal use other users can benefit from it. They also state that del.icio.us, for most users, functions as a recommendation system even without explicitly providing recommendation.

In MIT labs, an experiment was carried out by (Liu et al., 2006) to generate a taste fabric of social networks. Folksonomies were used in the experiment to weave the taste fabric. Their idea was based on philosophical and sociological theories of taste and identity to weave a semantic fabric of taste. They mined 100,000 social network profiles, segmented them into interest categories and then normalized the folksonomies in the segments and mapped them into a formal ontology of identity and interest descriptors. Their work has inspired us with the idea of using folksonomies in the process of semantic annotation.

(Hotho et al., 2006) have presented a new search algorithm for folksonomies, called 'FolkRank', which exploits the structure of the folksonomy. Their proposed algorithm is used to support the retrieval of resources in the del.icio.us social bookmarking services by ranking the popularity of tags. They demonstrated their findings on a large-scale dataset (around 250k bookmarked resources) and showed that their algorithm yielded a set of related users and resources for a given tag. Therefore, 'FolkRank' can be used to generate recommendations within a folksonomy system.

(Versa, 2006a) has presented a study in which the linguistic properties of folksonomies demonstrated that users engaged in resource tagging are performing classification according to principles similar to formal taxonomies. To prove his findings, Versa analyzed the kinds of classification observed in user tags using the non-taxonomic categories proposed by the linguist Anna Wierzbicka. He then compared users' patterns to those observed for two well known sources of classification schemes on the Internet: the open directory project (DMOZ) and the Yahoo directory. His findings showed that there is a clear difference between folksonomy tags and the two classification schemes. Tags are drawn from most categories while DMOZ and YAHOO were biased only towards one category (namely functional category). In another paper by the same author, entitled "Concept Modeling by the Messes: Folksonomy Structure and

Interoperability", (Versa, 2006b) has used folksonomies to model concepts in a domain. He used a method, based on the linguistic properties of the tags, to extract structural properties of free form user tags to construct ontology. The resultant ontology is a simple conceptual domain model built from automatically mediated collaboration; this ontology has been used to facilitate interoperability between applications dependent tag sets.

Finally, 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 such as Yahoo compared to human indexing mechanisms.

From the previous discussion the reader can observe that most research on folksonomies is either user-centric e.g. (Mika, 2005) and (Ohmukai et. al, 2005) or tag-centric e.g. (Gruber, 2005), (Versa, 2006a,b), (Liu et. al, 2006) and (Hotho et. al, 2006). Little research has been conducted on examining the relationship between folksonomies and other indexing systems.

KEYWORD EXTRACTION- A BRIEF LITERATURE REVIEW

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, Fontaine & Fontaine, 2004) supported by domain specific ontologies e.g. (Hulth, Karlgren, Jonsson, Boström & Asker, 2001). There are a wide variety of applications that use automatic keyword extraction; among these are document summarization and news finding e.g. (Martínez-Fernández et al., 2004). Keyword analyzer services [8] used by most Search Engine Optimization (SEO) companies are another type of keyword extraction application using term

frequency. Most complex keyword extraction techniques require corpus training in a specific domain for example Kea [9] - 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 have been extracted, the document needs to be filtered; since not all words can be adequate for indexing. The filtering can be done using the vector space model or more specifically by 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 term frequency, which ignores the semantics of the document content. This is because the term frequency technique is based on the occurrences of terms in a document assigning a weight to indicate its importance. Most indexing techniques rely on statistical methods or on the documents term distribution tendency. Statistical methods lack precision and they fail in extracting the semantic indexes to represent the main concepts of a document (Kang & Lee, 2005). This problem might be partially solved by using manually assigned keywords or tags (i.e. folksonomies) in bookmarking systems like del.icio.us.

EXPERIMENT SETUP AND TEST DATA

There are plenty of keyword extraction techniques in the *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 to 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 [10]. 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). The decision to use Yahoo API was made for the following reasons:

- 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.
- Also, 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.

The experiment was conducted in four phases: in the first phase we exposed a sample of both folksonomy and Yahoo keywords sets to two trained-human indexers who, given a generic classification, evaluated which set held greater semantic value than the other. In the second phase, we used another modified instrument from (Kipp, 2006) to further explore the semantic value of folksonomy tags and the Yahoo keywords. 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, one of the human indexers 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 TE 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.

The Comparison System Framework

We constructed a system to automatically compare the overlap between the folksonomy, Yahoo TE and human indexer keywords and generate the desired statistics. The system consisted of three distinct components: the Term Extractor, the Folksonomy Extractor and the Comparison Tool as shown in Figure 2. The *Term Extractor* consists of two main components: JTidy [11], an open source Java-based tool to clean up HTML documents and *Yahoo Term Extractor* (TE) [12], a web service that provides "*a list of significant words or phrases extracted from a larger content*". After removing HTML tags from a website, the result is passed to Yahoo TE to generate the appropriate keywords.

The *Folksonomy Extractor* that we developed is designed to fetch the keywords (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 folksonomy list 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):

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

Where:

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

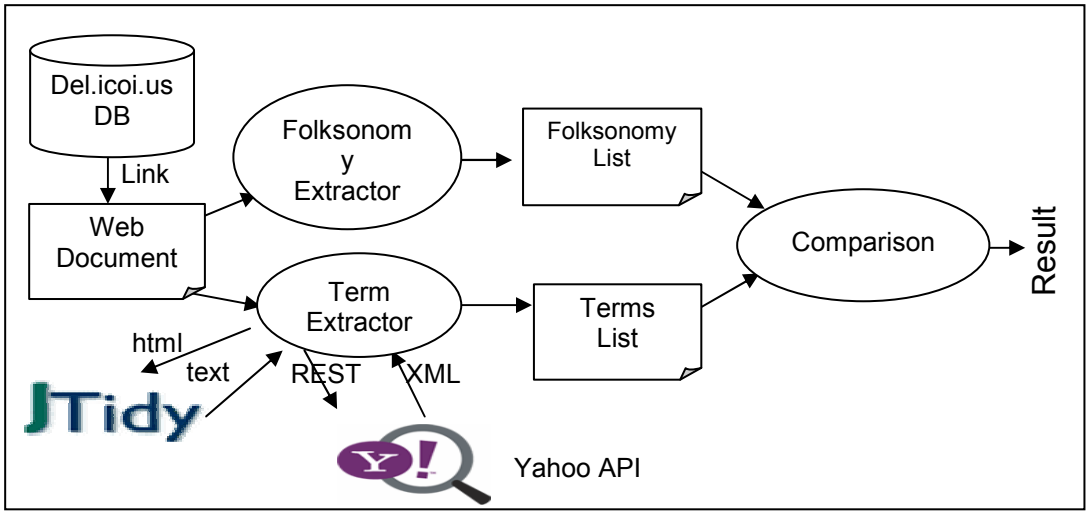

Figure 2: The Comparison System Framework

Data Selection

The test data used in this experiment was randomly collected from the del.icio.us social bookmarking service. One hundred bookmarked websites spanning various topics from the popular tags webpage were selected [13], as shown in Table 1.

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

The selected web resources were chosen based on the following heuristics:

- Bookmarked sites that are of a multimedia nature such as audio, video, flash, Word/PDF documents, etc. were avoided, as the Yahoo term extraction service only extracts terms from textual information. By the same token, whole Blog sites were avoided because they usually hold a diversity of topics; 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 participating taggers; this was necessary to ensure there were enough tags describing the website.

Other General Heuristics

Some other 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.

RESULTS

Phase 1

The role of phase one is to determine whether or not folksonomies carry more semantic value than keywords extracted using Yahoo TE. In this phase the phrase 'semantic value' means that the tag or keyword used to describe a web resource is relevant to its gist, i.e. the tag or keyword contributes to the description of the resource meaning.

Thus, given the sets of keywords from Yahoo TE and del.icio.us; the two indexers were asked to blindly [14] evaluate each keyword from both sets. The indexers were provided by a five-category table to classify the keywords from both sets. The table has the following values: "Strongly relevant" encoded 5, "Relevant" encoded 4, "Undecided" encoded 3, "Irrelevant" encoded 2 and "Strongly irrelevant" encoded 1.

After evaluating 10 websites from our data set, an inter-rater reliability test was conducted for each evaluated web resource to measure the evaluation agreement between the two indexers. This step is essential to measure the consistency among the two indexers.

The inter-rater agreement reliability test that we used to measure the consistency of classifying keywords into categories without any ordering (i.e. nominal data), was the Kappa (k) coefficient, the widely accepted measurement developed by (Cohen, 1960). The value of the resulting Kappa coefficient indicates the degree of agreement between the two raters. For interpreting the meaning of the resulting Kappa value we used (Landis & Koch, 1977) interpretation, where $0 \leq k < 0.2$ means slight agreement, $0.2 \leq k < 0.4$ means fair agreement, $0.4 \leq k < 0.6$ means moderate agreement, $0.6 \leq k < 0.8$ means substantial agreement, and $0.8 \leq k < 1.0$ means almost perfect agreement.

Table 2 shows the overall average degree of agreement between the two indexers for the 10 evaluated web resources. The obtained Kappa value for both sets falls in the fair level of agreement, which is considered satisfactory for the purpose of this experiment. However, the results show that agreement between the indexers about the folksonomy set is slightly lower (0.2005) than their agreement about the Yahoo TE set (0.2162); the difference is statistically significant at $p < 0.001$. The lower kappa value for the folksonomy set was due to a slight disagreement in evaluating one of the websites in that set, which affected the results accordingly.

Table 2: Average Inter-Rater agreement for the ten evaluated web resources in phase 1

	Average Inter-Rater Agreement [Kappa-coefficient value]
Folksonomy	0.2005
Yahoo TE	0.2162

The values summarized in Table 3 show the average mode value for each evaluated website from both indexers. For all values except for site 2, 5 and 8, the results for the folksonomy set was higher or equal to Yahoo TE values. By further inspecting the three cases (2, 5 and 8), the authors have found that what affected the average mode value in these three cases in the folksonomy set was, the amount of general tags used to describe these web resources compared to the same Yahoo TE set for these resources, which extracted more specific keywords (i.e. same or narrower term).

The results also show that the folksonomy and Yahoo TE sets scored an equal mode value (4 = relevant) for all sites. The values for the Yahoo TE varied considerably compared to the folksonomy values but the most frequent value in Yahoo TE was still (4) which appeared 3 times compared to 7 times in the folksonomy set.

Moreover, the results show that the folksonomy set has a higher mean and lower standard deviation i.e. 4.15(0.24), this indicates a low variance in the views of the two indexers towards classifying folksonomy tags, compared to the values for Yahoo TE, i.e. 3.55(1.01), which indicates a high variance in the views of the two indexers.
These results indicate that the folksonomy tags are more relevant to the human indexer's conception than Yahoo TE keywords. Furthermore, the difference between the two means was statistically significant at $p < 0.001$.

Table 3: The average mode values for each website in both Folksonomy (F) and Yahoo TE (K) set along with the mean, mode and standard deviation for all 10 evaluated websites

Site	F	K
1	4.5	4
2	4	4.5
3	4	3
4	4	2.5
5	4	4.5
6	4.5	3
7	4	1.5
8	4	4.5
9	4	4
10	4.5	4
Mean	4.15	3.55
SD.	0.24	1.01
Mode	4	4

The results of this phase give us the big picture of the semantic relationships held in the folksonomy and Yahoo TE keywords compared to the two indexers views. To better understand the semantics of each classified keyword in the folksonomy and Yahoo TE sets, an in depth analysis is carried out in phase 2.

Phase 2

The role of phase two was to inspect in more detail the semantic categories of the folksonomy set 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 two indexers were provided with another categorization. The new categorization values were adopted from (Kipp, 2006). Kipp built her scale instrument based on the different relationships in a thesaurus as an indication of closeness of match, into the following categories:

- Same - the descriptors and tags or keywords are the same or almost the same (e.g. plurals, spelling variations and acronyms); encoded 7,
- Synonym - the descriptors and tags or keywords are synonyms; encoded 6,
- Broader Term (BT) - the keywords or tags are broader terms of the descriptors; encoded 5,
- Narrower Term (NT) - the keywords or tags are narrower terms of the descriptors, encoded 4,
- Related Term - the keywords or tags are related terms of the descriptors; encoded 3,
- Related - there is a relationship (conceptual, etc) but it is not obvious to which category it belongs to; encoded 2,
- 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; encoded 1.

The two indexers applied the modified categorization scale to a sample of 10 bookmarked websites that were chosen from the experiment corpus.

After evaluating the 10 bookmarked websites, an inter-rater reliability test was conducted to evaluate the agreement between the two indexers for their evaluation of each web resource.

Table 4 shows the degree of agreement between the two indexers. The agreement between the two indexers gave us a fair level of agreement with almost equal scores for the folksonomy set (0.2257) and the Yahoo TE set (0.2241). The difference between the two means was statistically significant at $p < 0.001$.

Table 4: Average Inter-Rater agreement for the ten evaluated web resources in phase 2

	Average Inter-Rater Agreement [Kappa-coefficient value]
Folksonomy	0.2257
Yahoo TE	0.2241

The values summarized in Table 5 show the average mode value for each evaluated website from both indexers. Notice this time for all values, except for site 3, the results for the folksonomy set was higher than Yahoo TE values. By further inspecting site 3, the authors have found that what caused the drop down of the average mode value in this site was the number of tags assigned to this website, i.e. 18 tags compared to 28 keywords from Yahoo TE, and also the class of the tags used to describe the website, which fall more in the related category.

The results also show that the folksonomy set scored a higher mode value (5) compared to Yahoo TE (2). However, the results show that the folksonomy set has a higher mean and higher standard deviation i.e. 4.45(1.28), which indicates a high variance in the views of the two indexers towards classifying folksonomy tags, compared to the values for Yahoo TE, i.e. 2(0.71), which indicates a lower variance in the views of the two indexers, the difference between the two means was statistically significant at $p < 0.001$.

The resultant statistical analysis of this phase stressed the finding of the previous phase and gave us more insight in how folksonomies are considered semantically richer than Yahoo TE keywords.

Table 5: The average mode values for each website in both Folksonomy (F) and Yahoo TE (K) set along with the mean, mode and standard deviation for all 10 evaluated websites

Site	F	K
1	5	1.5
2	5	1
3	1.5	2
4	5	2.5
5	5	2
6	3.5	2
7	5	3
8	6	2
9	3.5	3
10	5	1
Mean	4.45	2
SD.	1.28	0.71
Mode	5	2

Furthermore, to visualize the results of this phase, a two-column bar graph was generated for each evaluated web resource to reflect the result of each category, i.e. the Blue bars denote the Yahoo keywords frequency and the Purple bars denote the folksonomy tags frequency.

Figure 3 shows the accumulated bar-graph obtained by juxtaposing each individual bar graph of the 10 evaluated web resources, for both indexers, in a layered fashion so that a general conclusion can be drawn easily.

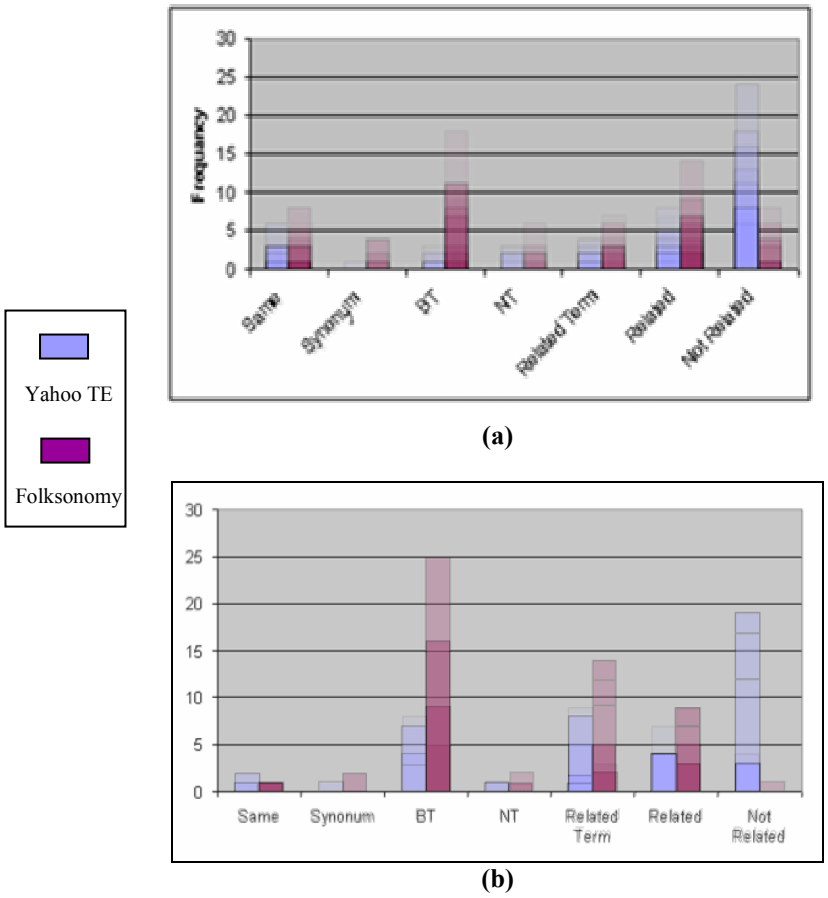

Figure 3: A visualization of the categorization results for the 10 web resources layered on top of each other resulting in a ghost effect, (a) corresponds to the results of the first indexer (b) corresponds to the results of the second indexer.

Comparing the two figures shows us that there is almost a good agreement between indexer (a) and indexer (b) in the assignment of Yahoo TE keywords in the 'not-related' category. However, this agreement starts to fluctuate, in order of magnitudes between the two indexers, in the similarity categories (i.e. Same, Synonym, BT, NT, Related Term and Related).

For instance, in Figure 3.(a), the folksonomy tags are accumulating more around the 'Broader Term' and 'Related' category, while in Figure.(b), the folksonomy tags are accumulating more around the 'Broader Term' and 'Related Term' 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 the folksonomy set outperforms the Yahoo keyword set.

Finally, we believe that the variance between the two indexers categorization was due to either the different interpretation of the meaning of the categories or the use of single category with different frequencies, as in the case of indexer (b), thus a further marginal homogeneity analysis using the Stuart and Maxwell test to identify the sources of variability will be considered for future work.

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 is discussed.

A) Unrelated tags
To explore in greater depth 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', self-reference tags and sometimes 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 them.

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

Self-reference tags include any tags that have to do with the user's own interest. Examples are dates, e.g. 'January', 'monthly' and 'night', names, e.g 'tojack' and/or own reference, e.g. 'mylink', 'mysite' and 'myblog'. These tags usually appear once or twice among all the tags in a given bookmarked web resource.

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; alternatively the position of the word and its style (e.g. heading or sub-title) might be the reason for extracting it. The algorithms that Yahoo TE uses to extract keywords from web sites are obscure which affects further analyses of the extracted keywords.

B) Related tags
This category represents relationships that are ambiguous or difficult to place into the previous 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. Furthermore, in a web resource that gives examples about FreeBSD, a particular version of the UNIX operating system, del.icio.us users' have tagged the web resource with related tags such as: tutorial, tips, and how-to, these tags were not explicitly mentioned in the web resource; however, they contributed to the description of the web resource by giving it a new contextual dimension.

Another example of a relationship between two fields of study is a web resource about an open source office application called 'NeoOffice' for the 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' applications.

Phase 3

As mentioned in the experiment setup, the role of phases three and four was to find the percentage of overlap between the folksonomy set and the keywords generated by Yahoo TE. In this phase and the next one, folksonomy tags, Yahoo TE keywords and the indexer keywords are treated as abstract entities which do not hold any semantic value. This assumption will help us 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 measurement used in our comparison framework was 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. $F \cap K$
3. 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. $F = K$).

The collected data set (described in a previous subsection on Data Selection) was dispatched to our comparison framework to measure the percentage of overlap between folksonomy tags and Yahoo TE keywords.

After observing the results of 100 websites 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, and the most frequent percentage of overlap (i.e. mode) was 12.5%.

Figure 4 shows a histogram of the frequency of the results which graphically summarizes and displays the distribution of the percentage of the overlaps using short intervals (2.5 percentages wide). Notice that most of the overlap values (14 values) fall in the interval between 7.5 and 8.75, while the least of the overlap values fall at the ends of the histogram. The shape of the histogram forms the beginning of a normal curve, thus, we believe that with more evaluated websites the histogram will ends up being an approximate normal curve, which can be used as a tool to estimate proportion of overlaps with appropriate margins of errors.

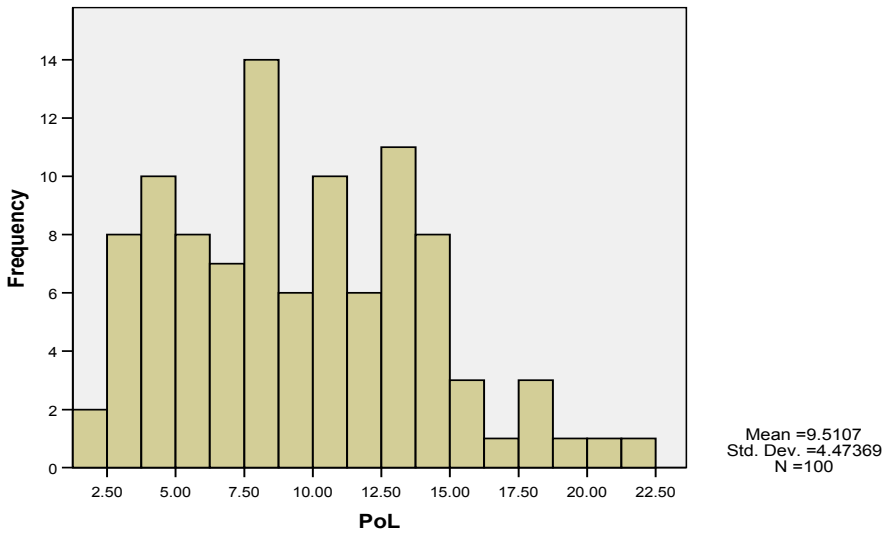

Figure 4: Histogram of the Percentage of Overlap (PoL) for 100 websites

Finally, the results of this phase showed us that folksonomy tags can not be replaced by automatically extracted keywords, even if there was a marginal overlap between the two sets. However, to inspect in more depth the position of folksonomies in the spectrum from professionally assigned keywords to context-based machine extracted keywords, phase 4 is carried out to envision the place of folksonomy tags in this spectrum.

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 most closely 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 [15] (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 using our comparison framework, once with the keywords from the Yahoo TE set and another with the folksonomy set. This step is essential to see whether folksonomies produced the same results as if a human indexer was doing the process.

The results show (see Figure 5) that there is 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%. Furthermore, the experiment showed one case where there is a complete overlap (inclusion) between the folksonomy set and the indexer set.

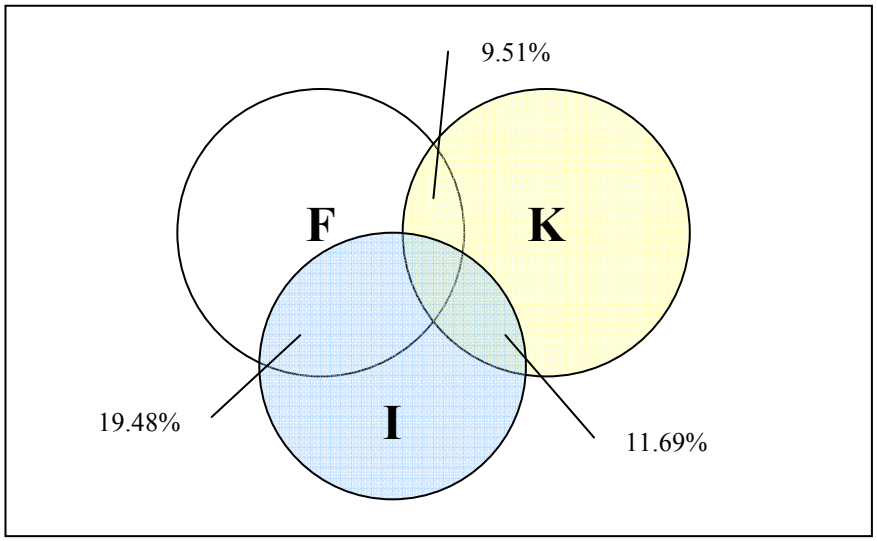

Figure 5: A Venn diagram that shows Folksonomy (F), Yahoo TE (K) and the human indexer (I) sets as three distinct circles and highlights the percentage of the overlap between the three sets

The results of this phase showed us that folksonomy tags are more oriented toward the professional indexer keywords. Therefore, this finding positioned the folksonomy tags nearer to the indexer keywords in the spectrum from professionally assigned keywords to context-based machine extracted keywords.

DISCUSSION

After completing the four phases of this experiment, a number of observations were made. As a first impression, phase 1 was used to evaluate the relevance of the folksonomy tags and Yahoo TE keywords to the human conception. Thus, the results of this phase indicate 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 2, 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.

In phase 3 and 4, the folksonomy tags showed a greater tendency to overlap with the professional indexer produced keywords than with the Yahoo TE keywords. Thus, in phase 3, the average overlap between the folksonomy set and Yahoo keywords was 9.51%, which implies that there was only a minor intersection between the two sets, and that folksonomy tags cannot be replaced completely with keywords generated by machine (in this case Yahoo

TE). This finding also opens the door for other potential research directions, for instance in the field of language technology and semantics, which is out of the scope of this experiment.

In phase 4, 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 about the better 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 indexers. 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.

THE FOLKSANNOTATION TOOL – A CASE STUDY

To emphasis the usefulness of the results obtained from this experiment; i.e. the rich semantics of folksonomies, a working example illustrating the value of the findings is demonstrated using a prototypical tool called FolksAnnotation (Al-Khalifa & Davis, 2006). This tool uses folksonomy tags to semantically annotate web resources with educational semantics from the del.icio.us bookmarking service, guided by appropriate domain ontologies.

Figure 6 shows the system architecture of the implemented FolksAnnotation tool; the detail of the implementation of the tool has been previously reported in (Al-Khalifa & Davis, 2006); however, a brief description of the tool is presented here.

The tool consists of two processes: 1) tags extraction/normalization pipeline and 2) semantic annotation pipeline.

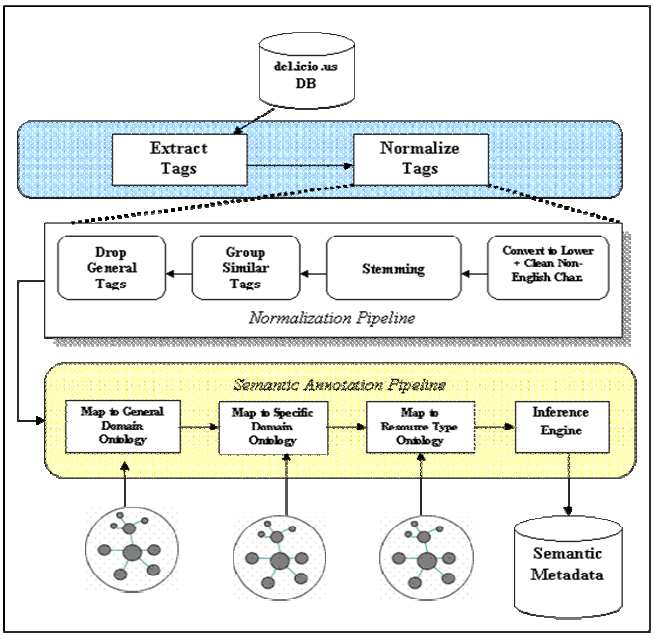

Figure 6: System Architecture of the 'FolksAnnotation' Tool

The normalization process is responsible of cleaning and pruning tags. The process starts by fetching all tags assigned to a web resource bookmarked in the del.icio.us bookmarking service then passes these tags to the normalization pipeline which does the following. First, tags are converted to lower case so that string manipulation (e.g. comparison) can be applied to them easily. Secondly, non-English characters are dropped; this step is to ensure that only English tags are present when doing the semantic annotation process. Thirdly, tags are stemmed (e.g. convert plural to singular) using the Porter stemmer [16] then similar tags are grouped (e.g. inclusion of substrings). Finally, general concept tags in our domain of interest are eliminated. The process of normalization is done automatically and it is potentially useful to clean up the noise in people's tags.

The semantic annotation process is the backbone process that generates semantic metadata using pre-defined ontologies. The process attempts to match folksonomy terms (after normalizing them) from the bookmarked resource against terms in the ontology (which it uses as a controlled vocabulary) and only selects those terms that appear in the ontology.

After assigning semantic descriptors to the web resource, the inference engine is responsible for associating pedagogical semantics (i.e. difficulty level and instructional level) to the annotated web resource. These two values are generated from a set of reasoning rules when enough information is available in the basic semantic descriptors.

One of the evaluation procedures we have carried out on this tool was to compare the number of folksonomy tags attached to our ontologies concepts against the Yahoo TE keywords attachment for the same web resource. For the purpose of this evaluation a set of 30 web resources were randomly selected from the del.icio.us bookmarking service, and for each web resource a two sets of keywords (namely, folksonomy tags and Yahoo TE keywords) were prepared to be passed through the semantic annotation pipeline. The results of this experiment showed that the number of attached keywords from the folksonomy set is much higher than the Yahoo TE set with mean and standard deviation of 14.17(8.25) and 4.24(2.47), respectively. The difference between the means was statistically significant at $p < 0.001$. The results demonstrate that folksonomy tags are more useful in generating semantic metadata than context-based keywords.

CONCLUSION AND FUTURE WORK

In this paper we have described four experiments to explore the value of folksonomies in creating semantic metadata. The first and second experiments evaluated the relevance of the folksonomy tags and Yahoo TE generated keywords to the human conception. The evaluation was performed by two trained indexers using an evaluation scale based on the different relationships in a thesaurus as an indication of the closeness of match. The third and fourth experiments were conducted to find the percentage of overlap between the folksonomy tags, keywords generated by Yahoo TE and the human indexer keywords.

The results of phases one and two show that the two human indexers have both agreed on the richer semantics of the folksonomy tags compared to Yahoo TE, with $p < 0.001$. The results of

phase three showed that the average overlap between the folksonomy set and Yahoo keywords was 9.51%, and the results of phase four showed that the folksonomy set was more correlated to the human indexer set with a mean of 19.48%, while Yahoo TE set scored a mean of 11.69%.

It is clear from the results of this experiment that the folksonomy tags agree more closely with the human generated keywords than those automatically generated. The results also showed that the trained indexers preferred the semantics of folksonomy tags compared to keywords extracted by Yahoo TE. These results were very encouraging, and illustrated the power of folksonomies. We have demonstrated that folksonomies have an added new contextual dimension that is not present in automatic keywords extracted by machines.

This experiment was a first step towards future evaluation techniques on which we are planning to embark. These techniques will measure the semantic value of folksonomies based on knowledge engineering principles and methods, such as Formal Concept Analysis (FCA) and frame-based systems (Stuckenschmidt, 2004). In such techniques concept hierarchies (or 'concepts lattices') are used to define a given term. By using this approach, the intended meaning of a term is addressed instead of finding the exact syntactic match.

So to conclude, folksonomies are very popular and a potential rich source for metadata. The rational of this work was based on the motivation of investigating whether folksonomies could be used to automatically annotate web resources. The findings of this experiment was used to justify the use of folksonomies in the process of generating semantic metadata for annotating learning resources; see (Al-Khalifa & Davis, 2006).

REFERENCES

Al-Khalifa, H. S., & 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.

Cohen, J. (1960). A coefficient of agreement of nominal data. *Educational and Psychological Measurements*. 20(1), 37-46.

Golder S. & Huberman B.A. (2006). "Usage Patterns of Collaborative Tagging Systems." *Journal of Information Science*, 32(2). 198-208.

Gruber, T. (2005). Ontology of Folksonomy: A Mash-up of Apples and Oranges. AIS SIGSEMIS Bulletin, 2(3&4).

Guy, M. , & Tonkin E. (2006). Folksonomies: Tidying up Tags? D-Lib Magazine, V 12(1).

Hammond, T., Hannay T., Lund B. & J. Scott. (2005) Social Bookmarking Tools (I): A General Review. D-Lib Magazine, 11(4).

Hotho, A., Jäschke R., Schmitz C. & Stumme G. (2006). Information Retrieval in Folksonomies: Search and Ranking. in Proceedings of the 3rd European Semantic Web Conference (ESWC2006). Budva, Montenegro: LNCS, Springer.

Hulth A, Karlgren J., Jonsson A., Boström H., & Asker L. (2001). Automatic Keyword Extraction Using Domain Knowledge. *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*. : LNCS, Vol.2004, pp. 472 - 482. Springer.

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

Kang, B.-Y., & Lee S.-J. (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., Maghoul F., Chang C.C., & Kumar K. (2005). Y!Q: Contextual Search at the Point of Inspiration. *The ACM Conference on Information and Knowledge Management (CIKM'05)*, Bremen, Germany.

Kroski, E. (2006). The Hive Mind: Folksonomies and User-Based Tagging. Retrieved January 14, 2006. from http://infotangle.blogsome.com/category/folksonomies

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). "Introduction to Latent Semantic Analysis." *Discourse Processes* **25**: 259-284.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Liu, H., Maes P., & Davenport G. (2006). Unraveling the taste fabric of social networks. International Journal on Semantic Web and Information Systems, 2(1): p. 42-71.

Martínez-Fernández J. L., García-Serrano A., Martínez P. & Villena J. (2004). "Automatic Keyword Extraction for News Finder." Adaptive Multimedia Retrieval, Lecture Notes in Computer Science, Vol. 3094, pp. 99-119, Springer.

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

Mathes, A. (2004). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Accessed on February 28, 2006. Available on line http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Menchen, E. (2005). Feedback, Motivation and Collectivity in a Social Bookmarking System. in Kairosnews Computers and Writing Online Conference. Retrieved February 3, 2007 from http://kairosnews.org/node/4338.

Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. in In Proceedings of the Fourth International Semantic Web Conference (ISWC 2005). Galway, Ireland: Lecture Notes in Computer Science.

Ohmukai, I., Hamasaki M., & Takeda. H. (2005). A Proposal of Community-based Folksonomy with RDF Metadata. in The 4th International Semantic Web Conference (ISWC2005). Galway, Ireland.

Quintarelli, E. (2005). Folksonomies: power to the people. in ISKO Italy-UniMIB meeting. 2005. Milan, Itlay. Retrieved February 3, 2007 from http://www.infospaces.it/docs/Folksonomies%20-%20Power%20to%20the%20People.doc.

Sado, W.N., Fontaine, D. & Fontaine, P. (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.

Shirky, C. (2005). Ontology is Overrated: Categories, Links, and Tags. Retrieved March 27, 2006. from http://shirky.com/writings/ontology_overrated.html

Sieck, S. (2005). connotea and citeulike: "folksonomies" emerge within scholarly communities, E. Insight, Editor. Retrieved February 3, 2007 from http://www.connotea.org/EPS-Connotea.pdf.

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

Stuckenschmidt H. & Harmelen F.V. (2004). *Information Sharing on the Semantic Web*. Berlin: Springer.

Vander Wal, T. (2006). Folksonomy definition and Wikipedia. Retrieved April 29, 2006. from http://www.vanderwal.net/random/category.php?cat=153

Versa, C. (2006a). The language of Folksonomies: What tags reveal about user classification. NLDB 2006. LNCS 3999, pp. 58-69.

Versa, C. (2006b). Concept modeling by the messes: Folksonomy structure and interoperability. ER 2006. LNCS 4215, pp. 325-338.

Witten I., Paynter G., Frank E., Gutwin C. & Nevill-Manning C. (1999). KEA: Practical Automatic Keyphrase Extraction. *In Proceedings of ACM DL'99*.

Wikipedia. (2006). Folksonomy. Retrieved March 26, 2006. from http://en.wikipedia.org/wiki/Folksonomy

ENDNOTES

[1] http://www.flickr.com
[2] http://del.icio.us
[3] http://www.furl.net
[4] http://www.furl.net/
[5] http://www.spurl.net/
[6] http://www.citeulike.org
[7] http://www.connotea.org
[8] Example: http://www.searchengineworld.com/cgi-bin/kwda.cgi
[9] http://www.nzdl.org/Kea/
[10] Yahoo API term extractor service was launched on May 2005
[11] http://sourceforge.net/projects/jtidy
[12] http://developer.yahoo.net/search/content/V1/termExtraction.html
[13] http://del.icio.us/tag/, Data was collected between 24/2 and 27/2 2006
[14] By blindly, we mean that both indexers do not know which keyword list belongs to which set (i.e. folksonomy or Yahoo TE).
[15] http://dmoz.org/
[16] http://www.tartarus.org/~martin/PorterStemmer/index-old.html

Authors' Bibliography

Hend S. Al-Khalifa is a PhD candidate in her final year of study in the Learning Societies Lab research group within the School of Electronics and Computer Science (ECS) at the University of Southampton, UK. She received her M.Sc degree in Information Systems (2001) from King Saud University, Riyadh, KSA. Her publications are in the fields of web technologies, computers for people with visual impairments and applications of e-learning.

Hugh C. Davis is the Head of the Learning Societies Lab within the School of Electronics and Computer Science (ECS) at the University of Southampton, UK. He is also the University Director of Education with responsibility for

eLearning strategy. He has been involved in hypertext research since the late 1980's and has interests in the applications of hypertext for learning, open hypertext systems and architectures for adaptation and personalization. He has extensive publications in these fields, and experience of starting a spin-off company with a hypertext product. His recent research interests revolve around social hypertext, web and grid service frameworks for eLearning and he has a particular focus on the e-assessment domain. He has led many projects focusing on both the technology and application of e-learning.