JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 7(1): 29-38, 2019
https://doi.org/10.1633/JISTaP.2019.7.1.3

Comparison of User-generated Tags with Subject Descriptors, Author Keywords, and Title Terms of Scholarly Journal Articles: A Case Study of Marine Science

Praveenkumar Vaidya*
Department of Studies in Library and Information Science,
University of Mysore, Mysuru, India
Tolani Maritime Institute, Pune, India
E-mail: praveenv@tmi.tolani.edu

N. S. Harinarayana
Department of Studies in Library and Information Science,
University of Mysore, Mysuru, India
E-mail: ns.harinarayana@gmail.com

ABSTRACT
Information retrieval is the challenge of the Web 2.0 world. The experiment of knowledge organisation in the context of abundant information available from various sources proves a major hurdle in obtaining information retrieval with greater precision and recall. The fast-changing landscape of information organisation through social networking sites at a personal level creates a world of opportunities for data scientists and also library professionals to assimilate the social data with expert created data. Thus, folksonomies or social tags play a vital role in information organisation and retrieval. The comparison of these user-created tags with expert-created index terms, author keywords and title words, will throw light on the differentiation between these sets of data. Such comparative studies show revelation of a new set of terms to enhance subject access and reflect the extent of similarity between user-generated tags and other set of terms. The CiteULike tags extracted from 5,150 scholarly journal articles in marine science were compared with corresponding Aquatic Science and Fisheries Abstracts descriptors, author keywords, and title terms. The Jaccard similarity coefficient method was employed to compare the social tags with the above mentioned wordsets, and results proved the presence of user-generated keywords in Aquatic Science and Fisheries Abstracts descriptors, author keywords, and title words. While using information retrieval techniques like stemmer and lemmatization, the results were found to enhance keywords to subject access.

Keywords: Web 2.0, social tagging, information retrieval, Jaccard similarity, subject descriptors

Open Access

Accepted date: February 28, 2019
Received date: July 27, 2018

***Corresponding Author:** Praveenkumar Vaidya
Librarian
Tolani Maritime Institute, Induri, Talegaon, Pune 410507, India
praveenv@tmi.tolani.edu

1. INTRODUCTION

Information retrieval in the context of information overload is the challenge for library and information architects. The adversity in recalling relevant information with precision is exacerbated when substantial information afforded by the Internet is available in abundance. In order to organize such profusely accessible information, library professionals have designed many hierarchical classification systems or subject related controlled vocabularies. The shift in this order arose due to the impact of advancement in Web 2.0 (Anfinnsen, Ghinea, & de Cesare, 2011) applications wherein many social networking platforms enabled users to organize their personal information resources in the form of social tags or folksonomies. Hence, the folksonomies are user created metadata (Furner, 2010; Guy & Tonkin, 2006; Wal, 2004) for web resources and are used extensively for content categorization and retrieval in the age of Web 2.0. Unlike a controlled vocabulary which is designed by top-down ways, a folksonomy is constructed from bottom-up by user-centred ways to organize personal information resources.

Mathes (2004) indicates about three groups which are predominantly involved in providing keywords to resources which are also used for effective retrieval: the authors, users, and subject experts. But generally, the keywords provided by subject experts, known as controlled vocabulary, are a popular dataset. The hierarchical structure of subject-specific taxonomies is prevalent in knowledge organisation but with some limitation (Golder & Huberman, 2006; Kipp, 2006). In the case of author-assigned keywords, authors are normally asked to choose a few keywords which describe the content of their own article (Névéol, Doğan, & Lu, 2010), but which may not be sufficient to greater precision and recall. Furthermore, user-generated keywords or collaborative tags have the ability to facilitate both retrieval and discovery. Folksonomies can be navigated through tags, resources, and users for any user query within a single user-centric environment for effective retrieval system. Hence, tags can also be a useful dataset for content categorization and knowledge organisation (Peters et al., 2011; Rafferty, 2017; Stan & Maret, 2017). In scholarly journal articles or any other source of the document the 'title'grabs the attention of the user at first sight. Therefore for any researcher the 'title'plays an important role that provides aboutness and contents of the document. Hence, the title terms are also a dominant source of metadata in information retrieval (Davarpanah & Iranshahi, 2005; Voorbij, 1998).

As 'social tags'represent a tagger's conceptual understanding or categorization of a resource from a personal point of view, hence researchers consider social tagging as related to sense making (Hotho, Jäschke, Schmitz, & Stumme, 2006). The 'subject descriptors'or index terms, which are also descriptive metadata like social tags, come from highly structured controlled vocabularies. The 'author keywords'consist of conceptual and content categorization from the author's perspective, and add important value to resources. Similarly, title words accurately describe the contents of the manuscript, hence are presented as significant metadata. Given their conceptually shared purpose of social tags, subject taxonomies, author keywords, and title words, it makes sense to investigate whether social tags can complement subject descriptors, author keywords, and title words. Essentially, the purpose of this work is to understand whether social tags can also emerge as alternative access points to subject access despite the presence of subject descriptors, author keywords, and title words.

All these above-mentioned datasets have some limitations in precise retrieval and hence need to be studied for useful application. The combination of folksonomy, controlled vocabulary systems, author-supplied keywords, and title terms is an effective way to make up for the shortcomings of all these metadata for effective information retrieval.

2. LITERATURE REVIEW

There are many studies where comparative works are done to understand the significance of the datasets. Such studies demonstrated the emergence of additional useful terms for search and information retrieval which also enhance the process of knowledge discovery.

Several studies are found where comparison of datasets is conducted between social tags and subject descriptors. In their study, C. Lu, Park, and Hu (2010) examined the "difference and connections between social tags and expert-assigned subject terms and further explored the feasibility and obstacles of implementing social tagging in library systems. The results show the possible use of social tags to improve the accessibility of library collections."In another study, Wu, He, Qiu, Lin, and Liu (2013) believe that tagging has the potential to become a complementary resource for expanding and enriching controlled vocabulary systems. They also propose that "the help of future technology to regulate and promote features related to controlled vocabulary in social tags would greatly improve people's organizational and access capabilities within information resources."Hence, there was an attempt to enhance information retrieval using social tags in addition to subject vocabularies.

But this comparison work also involves author keywords and title words in addition to subject descriptors to compare them with social tags. In one of the early studies on comparing user, creator, and intermediary tagging, Kipp (2006) examined these three set of words and found the presence of many user terms which were related to the author and controlled vocabularies. A few terms were also found which were not available in controlled vocabularies and it was concluded that user tags can provide additional access points to discover information. In other studies by Kipp (2011a, 2011b), similar datasets were compared and analysed by using descriptive statistics method, informetric measures, and thesaural term comparison. The results showed the presence of additional access terms in tags and it was recommended to take advantage of these terms over traditional systems.

Similarly, Lu and Kipp (2014) and Syn and Spring (2010) conducted studies to evaluate whether user tags can represent resources as author keywords do and are used to categorize resources as keywords. The cosine similarity test was conducted to measure the similarity value. The results showed that author provided keywords were more consistent in describing the content of the resources. But, the user-assigned tags showed more variation in describing the content of the resources. In the same study, the researchers also conducted a comparative study of both the title and abstract terms of papers. In case of comparison of tags with title keywords, it was observed that the title of papers seems to be the main source for users to assign tags and therefore tags and title keywords represent the content of the paper.

In another early study Voorbij (1998) compared title keywords with subject descriptors to demonstrate that the subject descriptors retrieve more precise and far more successful results than by searching through title keywords. The study concludes that many relevant records cannot be retrieved by title keywords because of the wide diversity of ways to express the topic. In a similar study, Ansari (2005) tried matching between assigned descriptors and title keywords of medical theses. The results show that the keywords in the title comprise genuine information value and it was recommended that such words should be taken into consideration while introducing them into the indexing descriptors. The other study by Engelson (2013) worked to determine the correlation between title keywords and Library of Congress Subject Headings (LCSH) terms, and found that books with a popular content level designator had high-level matches.

Strader (2009) examined the overlap between author-assigned keywords with LCSH terms. It was observed that both keywords and controlled vocabularies complement one another and the ability to provide unique access points for the majority of the searches was demonstrated. But both LCSH and keywords provide significant numbers of unique terms that may increase the discoverability of resources.

The above studies suggest that the comparison of user assigned tags with author keywords, title words, and subject descriptors will result in new access points to information discovery and retrieval.

This study stands apart due to comparison work undertaken with different datasets and methodology as well. The CiteULike tags have been compared with subject descriptors, author keywords, and title words also. Even though the above review shows such works, they differ in the methodology adopted for this work. In some other works, where the same methodology is adopted, they differ in the datasets considered for this work.

3. RESEARCH QUESTIONS

In this study, an attempt is made to address the following research questions:

A. Is there any similarity between CiteULike tags with Aquatic Science and Fisheries Abstracts (ASFA) subject descriptors, author-assigned keywords, and title terms of marine science literature?

B. Do social tags enhance the effectiveness of keywords to subject access better than controlled descriptive terms, author-assigned keywords, and title terms?

The findings of this research work will exhibit the importance of social tags for information retrieval and knowledge organisation.

4. SCOPE AND METHODOLOGY OF THE STUDY

Essentially, for this research work the user-generated tags were primary data which were extracted from the social bookmarking site CiteULike. CiteULike is a popular social web service where users can save and share citations from scholarly journal articles. With its great compatibility with subject databases and publishers, it can capture bibliographic data of research articles. This also provides an opportunity to users to annotate personal keywords (tags) to the articles for repeat access. Not only are these tags personally useful, but also are to other researchers of the same field. If a profile is created with subject interests, users can join them and idea

exchange can be facilitated to access the reference articles of other researchers at one place and understand the research carried out by peers. CiteULike also allows users to import/export the citation details in many formats. The tags created by many such users can be useful for research work. As CiteULike is popular among researchers it attracts listings of many articles and a good number of social tags also. Hence, CiteULike has an edge over other available reference management tools. For this research work, marine science scholarly journal articles were chosen due to the dynamic nature of the subject. Globally, between 2010 and 2014 more than 370,000 manuscripts were published and more than 2 million articles were cited in marine science. The research and development expenditure of countries with high gross domestic product show high ocean science performance in terms of publications and citations (United Nations Educational, Scientific and Cultural Organization, 2017).

Marine science journal titles were collected from the list of ASFA. Consequently, the researcher identified and gathered 5,150 articles from the ASFA journal list published during 1954 to 2015, in which 1,405 articles belonged to publication year 1954 to 2000 and the remaining 3,745 were published during 2001 to 2015. The collected journal articles were searched in CiteULike to collect the tags, which resulted in 42,369 tags from 356 marine science journals. Similarly, these articles were also searched in the ASFA database to collect the corresponding subject headings and author keywords. WebCorp, an online tool, was used to convert the selected titles into a wordlist. All these datasets were transposed to Excel (Microsoft, Redmond, WA, USA) to manipulate the data. For these 5,150 articles the corresponding 49,478 subject headings, 10,752 author-assigned keywords, and 8,019 title terms were accumulated (Table 1). The research did require unique words, hence all these datasets were tested for duplication and the overlapped words were removed. For stemmer and lemmatization, it was also necessary to convert datasets to single words. Hence, during this process of converting multi-words to single words, the researcher could uncover 6,391 unique CiteULike tags, 5,695 ASFA words, 6,391 author keywords, and 7,213 title words.

The extracted CiteULike tags were transposed to Microsoft Excel sheets and the tags were preprocessed by removing the trashy tags (Thomas, Caudle, & Schmitz, 2010). These chosen articles were searched in the ASFA database and the corresponding controlled vocabularies were collected which were also transposed to Microsoft Excel sheets. Simultaneously, these 5,150 articles were explored with their DOI and author keywords were mined. Additionally, from these selected scholarly articles, title terms were also created. The user-generated CiteULike tags were compared with ASFA controlled vocabularies, author keywords, and title words to recognize to what extent these user tags resemble and enhance the keywords to subject access.

Table 1. Summary of all distinct dataset of each type

Datasets	Total words	Distinct words
CiteULike tags	42,369	9,015
ASFA descriptors	49,478	10,106
Title words	8,019	8,019
Author keywords	10,752	6,545

4.1. Preprocessing of Words

Preprocessing of CiteULike tags, author keywords, and title words is a very significant process for effective comparison work. Generally, the social tags were likely to consist of unpredictable and inconsistent words assigned by different users, unlike the controlled vocabularies which are organised and hierarchical in nature. The tags were assigned with several variations of singular, plural, hyphen, underscore, words with numerals or just numerals, acronyms/abbreviations, compound words, and also foreign language words. The tags have a serious deficit of synonym control and lack of precision and recall, which creates a challenge for retrieval effectiveness. Such inconsistencies are natural because the social tags are user-oriented, collaborative, democratic, cheap, dynamic, and distributed. CiteULike prevents users from assigning two words for any resource. Hence, users assign tags interpolated with 'underscore'or 'hyphen'between two concepts. With such limitations, these CiteULike tags must be normalized to compare them with ASFA descriptors, author keywords, and title words. Furthermore, these tags were converted into more meaningful words by removing hyphens, underscores, numerals, and foreign words.

This research work also includes the process of stemming and lemmatization of CiteULike tags, ASFA descriptors, author keywords, and title words. Hence, these datasets were converted to single words. Such single words were stemmed/lemmatized (http://text-processing.com/demo/stem/) to reduce variants (Mohammad, 2018). The stemmer is a function used in many text retrieval systems and search engines. A stemming algorithm is used to reduce words to their stems or roots after removing their prefixes and suffixes. For example, 'activatio n,''active,''activities,'and 'activity'were stemmed to 'activ.'With stemmer, these four words will turn into a single word, which enhances the efficiency of comparison (Lee & Schleyer, 2010,

2012; Syn & Spring, 2013). On the other hand, lemmatization works on morphological analysis of words and tries to remove inflectional endings thereby returning words to their dictionary form. For example, the word form of 'studies'and 'studying'was lemmatized to 'study'in both cases (Risueño, 2018).

4.2. Comparison of Words by Jaccard Similarity Coefficient

A similarity coefficient represents the similarity between two set of keywords, two documents, two queries, or one document and one query. A similarity coefficient is a function which computes the degree of similarity between a pair of text objects (Heymann & Garcia-Molina, 2009; Niwattanakul, Singthongchai, Naenudorn, & Wanapu, 2013; Thada & Jaglan, 2013).

The Jaccard Similarity coefficient is used to measure the similarity between the frequent sets of tags and the terms employed in the dataset (C. Lu et al., 2010). The Jaccard similarity index is a statistical tool used to compare similarity and diversity of sample datasets and is defined as the size of the intersection divided by the size of the union of the sample dataset. For example, A is the tag dataset comprising distinct tags for articles and B is the term dataset, comprising distinct terms for articles. The Jaccard similarity, ranging from 0 to 1, suggests the amount of overlap between the two data sets (C. Lu, Zhang, & He, 2016). This can be represented in the following formula:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

(if A and B are both empty, we define J (A, B) = 1 where $0 \leq J(A,B) \leq 1$)

4.3. Comparison of CiteULike Tags with ASFA Descriptors, Author Keywords, and Title Words

In this work, the CiteULike tags were compared with

- ASFA descriptors
- author keywords and
- title words

Furthermore, the comparison is tested in four different formats of following word structures. All CiteULike tags were compared with ASFA descriptors, author keywords, and title words in the following manner:

- Preprocessed CiteULike tags were compared with preprocessed ASFA descriptors, preprocessed author keywords, and preprocessed title words.
- Lemmatized CiteULike tags were compared with lemmatized ASFA descriptors, lemmatized author keywords, and lemmatized title words.
- Stemmed CiteULike tags were compared with ASFA stemmed descriptors, stemmed author keywords, and stemmed title words.
- CiteULike single tags were compared with ASFA single descriptors, single author keywords, and single title words.

The similarity or overlap between these sets of words was measured by adopting Jaccard's coefficient method, which is the commonly used method in such studies that also helps to answer the research questions considered for this study.

5. ANALYSIS AND INTERPRETATION

It is interesting to know the results of the comparison study conducted for this work. It will be tested to what extent the CiteULike tags will overlap with subject descriptors, author keywords, and title words. This comparison is tabulated in the form of tables for the benefit of understanding in detail. For this comparison work, Microsoft Excel functions were used in an extensive manner. As the extracted data runs into thousands of rows, Microsoft Excel was used for data manipulation. The following tables reveal the outcome and analysis of comparative study between different datasets. The results showed that Jaccard similarity was enhanced when the CiteULike tags, ASFA descriptors, author keywords, and title words were employed with stemmer, lemmatizer, and single words. But information retrieval depends on precision and recall. It was interesting to notice that when the terms were stemmed and lemmatized the Jaccard similarity index was consistently improved more than the results of preprocessed and single words to indicate the presence of more common terms in the datasets. It was also observed that the non-similar terms also play a vital role in such comparative studies.

5.1. Comparison Between CiteULike Tags with ASFA Descriptors

Table 2 shows a glimpse of comparison of terms between CiteULike tags with ASFA descriptors. Table2 illustrates the similarity measurement between CiteULike tags and ASFA descriptors. The results show that the Jaccard coefficient is just 9.17% when compared with preprocessed words of both datasets, which is minimal in the context of parameters of comparison. However, the results show maximum similarity when these words were stemmed (30.73%). But it was also observed that when these words were either lemmatized or

stemmed the similarity seems to go higher and they are in close proximity to each other (25.87% and 30.73%), but the similarity is reduced to 22.73% when compared with single tags of CiteULike and ASFA descriptors, which indicates the importance of stemmer or lemmatized words or even single words for retrieval. Further, whenever there is a high rate of similarity the effectiveness of retrieval also increases, but may hamper precision.

It was also observed that the Jaccard index, when compared with preprocessed CiteULike tags and ASFA descriptors, is 9.17% whereas when the tags and descriptors are converted to single words, the Jaccard similarity index rose to 22.73%. This describes the importance of splitting tags or descriptors into single words to find the common words in order to enhance retrieval efficiency. Due to the splitting of words, the retrieval precision may be affected but recall will be enhanced.

Additionally, this comparison between CiteULike tags and ASFA shows that users do not really have any knowledge of subject taxonomies. The tags were assigned to the sources which were convenient for users to recall and retrieve when needed. Due to this, there may be just 9.17% of common words or similar words, which is very low. However, controlled vocabularies play a vital role for precise information retrieval and hence cannot be neglected (Heymann & Garcia-Molina, 2009; Lee & Schleyer, 2010, 2012; C. Lu et al., 2010; C. Lu et al., 2016; Wu et al., 2013). It was also noticed that some important words which were present in tags but did not find a place in taxonomies may help to enhance the taxonomical dataset, which in turn may help in retrieval precision. For example, 'accretionary wedge'was listed in tags but did not find a place in ASFA descriptors. Similarly, the term 'nutrient starvation'was recorded in user tags, but in ASFA it was registered as 'nutrient deficiency'and 'nutrient depletion.'Therefore, the terms available in tags can also be used as 'Related Term'in controlled vocabulary entries.

5.2. Comparison of CiteULike Tags with Author Keywords

Author keywords are an integral part of the articles and these keywords were compared with user-generated CiteULike tags. Table 3 demonstrates the comparison of CiteULike tags with author keywords. As mentioned earlier, author keywords characterize the content of the research work published in any document. The author keywords are always considered as an important feature of information retrieval.

In this case, the CiteULike tags were compared with author keywords and interesting results were found. It was witnessed that when preprocessed CiteULike tags were compared with author keywords, the overlap was relatively higher or almost double (19.21%) than for the ASFA descriptors, as suggested in Table 2 (9.17%). It infers that users were probably influenced by keywords provided by authors. Hence the overlap was 19.21% between CiteULike tags and author keywords.

And it was also noticed that when these tags and keywords were converted into single words, the Jaccard index of the overlap was 39.61%, which is considerably high. Subsequently, when these same words were stemmed the overlap rose to 44.47%. Besides this, even when compared with lemmatized, stemmed, and single words, the overlap results show high in the case of stemmed words. It can also be understood that when the words are stemmed the overlapped result was the highest (44.47%) among these three entities. Conversely, it was also true that the author keywords and social tags did not match to a large extent (80.79%) and differ in the context of assigning. Similarly, the comparison with CiteULike tags and author keywords also indicated that the author keywords were not matched by around 59.44% when compared as single words. This also indicates that there is a distinct difference between the context of the user and author of the article (Kipp, 2006, 2007). Both author and user think in a diverse direction while assigning keywords to the article.

Table 2. Comparison of CiteULike tags with ASFA words

Datasets	Words before preprocess	Lemmatized words	Porter stemmer words	Single words
CiteULike tags	9,015	6,391	6,391	6,391
ASFA words	10,106	5,695	5,695	5,695
Common words	1,606	2,484	2,841	2,238
Jaccard index	0.0917 (9.17%)	0.2587 (25.87%)	0.3073 (30.73%)	0.2273 (22.73%)

ASFA, Aquatic Science and Fisheries Abstracts.

Table 3. Comparison of CiteULike tags with author keywords

Datasets	Words before preprocess	Lemmatized words	Porter stemmer words	Single words
CiteULike tags	9,015	6,391	6,391	6,391
Author keywords	6,545	4,261	4,261	4,261
Common words	2,507	3,022	3,279	3,074
Jaccard index	0.1921 (19.21%)	0.3961 (39.61%)	0.4447 (44.47%)	0.4056 (40.56%)

As discussed above, more similarity in tags indicates that users also tend to derive the tags from the author keywords. This can be seen by looking into long multi-word keywords. Author keywords like 'altricial versus precocial development,''taxonomic and functional approaches,'and 'western and central pacific fisheries commission'appeared in the dataset of CiteULike tags. These author keywords were mentioned as tags by the users.

5.3. Comparison of CiteULike Tags with Title Terms

Title words play an important role in information retrieval as controlled vocabularies and author keywords. In this section, CiteULike tags were compared with title terms and analysed for their overlap in the context of social tags, as title words are also one of the important datasets for retrieval. Table 4 explains the comparison of these two datasets and analysis is explained for better understanding.

By observing Table 4, there is 38.93% of overlap with CiteULike tags when these words were lemmatized, while in single words the overlap is 36.29%. In Tables 2 and 3, the rate of overlap was more in the case of stemmed words (30.73% and 44.47%), while in Table 4 the similarity result shows 22.7%, which is quite less than in Table 2 and 3. However, it can be noted that social tags were derived from both title terms (36.29%) and author keywords (40.56%) significantly. The reverse is also true for social tags where users not only rely on author and title keywords but they also prefer to provide tags, whichever was convenient for them and for their personal retrieval.

Table 4 also indicates the common terms between CiteULike tags and title terms. The presence of common terms was 2,986 when comparison was done with preprocessed words, which signifies the user was influenced by the title of the article to assign tags. The similar terms were more when CiteULike tags were compared with ASFA descriptors and author keywords. When lemmatized tags and title words were compared the Jaccard ratio was found to be 38.93%, which was more than the comparative result of stemmed words (22.7%) and also preprocessed words (21.26%). Hence it can be also derived that the processed words yield poor Jaccard values, in comparison with lemmatized, stemmed, or single words.

5.4. Comparison of the Jaccard Index of All Datasets

It is essential to analyse the Jaccard index of all these compared datasets. With reference to Table 5, the Jaccard index of all these datasets suggests that the tags and keywords throw consistent results when they were either lemmatized or stemmed or in the form of single words. While the words or tags before preprocessing narrowly result in any significant overlaps. Hence there is a need to determine the retrieval richness, if words were in single, lemmatized, or stemmed format.

This Jaccard index of the words before preprocessing indicates a very low overlap for datasets produced by users, experts, and authors because the titles to scholarly articles were also provided by authors. This does mean that the terms assigned by users, experts, and authors were very different, and even though a few terms are very popular among users, they are not used by experts and vice versa (C. Lu et al., 2010).

Table 5 shows that social tags were in higher agreement with author keywords and title words while describing the content with the controlled vocabularies. The Jaccard index for author keywords (19.21%) and title words (21.26%) was almost the same compared to controlled vocabularies (9.17%). This indicates less overlap in controlled vocabularies in respect to comparison with author keywords and title words. Fig. 1 provides a graphical representation of the same.

With the usage of more techniques like lemmatization and stemmer the researcher had tried to reduce the lexical variation and compared them to enhance the overlap which is visible from the above tables and figure. The Jaccard index for lemmatized words, stemmed words, and single terms stands higher than preprocessed words. This indicates that after preprocessing the tags the rate of similarity or overlap will increase considerably and provide rich dividends in retrieval but may also affect precision.

Table 4. Comparison of CiteULike tags with title words

Datasets	Words before preprocess	Lemmatized words	Porter stemmer words	Single words
CiteULike tags	9,015	6,391	6,391	6,391
Title words	8,019	7,213	7,213	7,213
Common words	2,986	3,812	2,517	3,622
Jaccard index	0.2126 (21.26%)	0.3893 (38.93%)	0.2270 (22.70%)	0.3629 (36.29%)

Table 5. Jaccard index of compared dataset

Comparison of CiteULike tags with	Words before preprocess	Lemmatized words	Porter stemmer words	Single words
ASFA words	0.0917	0.2587	0.3073	0.2273
Author keywords	0.1921	0.3961	0.4447	0.4056
Title words	0.2126	0.3893	0.2270	0.3629

ASFA, Aquatic Science and Fisheries Abstracts.

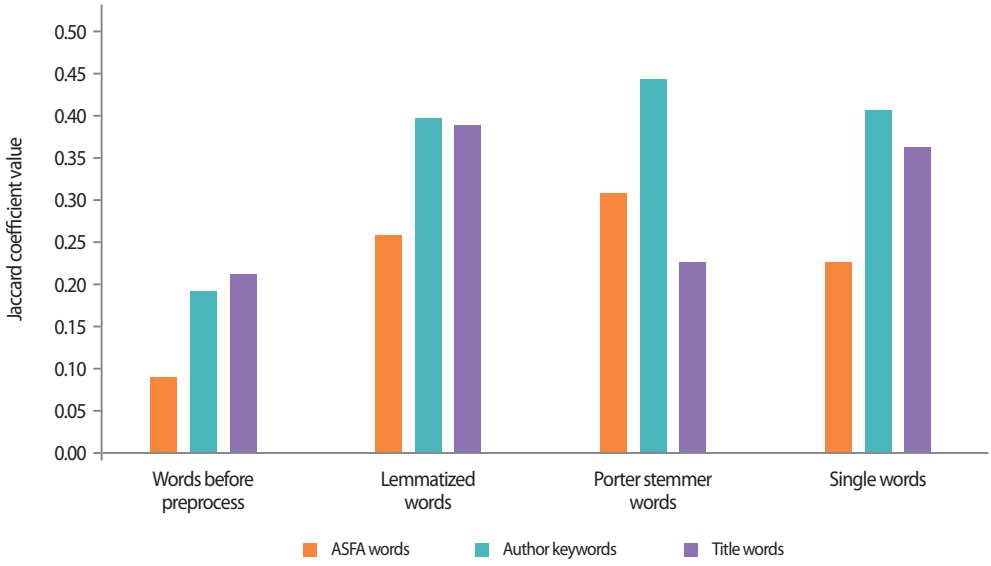

Fig. 1. Jaccard index of comparison of CiteULike tags with other datasets. ASFA, Aquatic Science and Fisheries Abstracts.

6. DISCUSSION AND CONCLUSION

The emergence of Web 2.0 technology has provided an enormous opportunity for users to access their resources by assigning tags to their sources, and typically 'social tagging'allows users to participate and interact with professionals. However, social tags have limitations because of their more uncontrolled and inconsistent nature, and scepticism exists about the value of these tags. In this context, this research work tried to address some of the reservations of social tags and an attempt has been made to throw light on the effectiveness of social tags for the retrieval process and in what way the tags may enhance subject access.

In this research work, the researchers have presented an exhaustive assessment of the association between ASFA (controlled vocabularies), author keywords, and title terms in comparison with CiteULike tags, particularly in the domain of marine science by using Jaccard similarity coefficient method. In context to the research questions of this study, the comparison task was conducted between ASFA descriptors with CiteULike tags (Table 2) to determine the presence of similar or overlap words among them. The result shows a minimal existence (9.17%) of similar words was found when compared before preprocessing. However, the similarity compliance was enhanced when these words were subjected to text processing techniques like lemmatization and stemmer to reduce lexical variation. As a result, the similarities were increased up to 30.73%. These results adequately answer the research question A, which emphasize the presence of common or overlap terms between these datasets. Hence, the users and experts share common terms even though lexical overlap between the corpora is very negligible.

The author keywords and title terms were considered to indicate the content of the article published and their respective CiteULike tags may comprise of tags significantly related to the article. The comparison work showed also in Table 3 and 4 that suggested the rise in overlap or similarity (19.21% and 21.26%) against CiteULike tags.

In an attempt of comparison between CiteULike tags and ASFA descriptors, author keywords, and title words, it was implied that the user was mostly influenced by either author keywords or title words before assigning the tags to resources. The comparison of tags with author keywords and title words resulted in good similarity ratios, which attributes the importance of author keywords and title words. This work clearly illustrated the gain in retrieval when the datasets were compared in single, lemmatized, or stemmed format. The implication of this study can be summarized that information retrieval can be enhanced when multiple words were split into single words but this may affect the precision. Future study could involve finding the appropriate techniques for precision retrieval.

However, it is interesting to know whether social tags can enhance 'subject access'in comparison with the subject

descriptors, author keywords, or title words. This comparison work emphasizes that there is an overlap of terms among subject descriptors, author keywords, and title words. But it can be very well presumed that the non-overlap tags also convey 'subject access'value in the ASFA database, as these CiteULike tags are subject specific to marine science. For example, when the preprocessed CiteULike tags in the form of single words were compared with ASFA single terms the Jaccard index was found to be 22.73%. The non-overlapped terms in the dataset, which is also known as 'Jaccard distance,'was found to be 77.27%. This can be further elaborated, as the presence of 4,153 non-overlapping terms in the dataset has also 'subject access'value. These words may be absent in ASFA yet may throw search results related to the subject, but may hamper precision. This explanation reflects research objective B considered for this study, which specifies the enhancement of keywords to subject access.

Overall, the introduction of social tags in the Web 2.0 context is an opportunity for libraries to enhance their access to resources. Many studies have concluded that their overlap is relatively low but still are very different in their nature and cannot be neglected, and also similarly cannot be considered as an alternative schema for a controlled vocabularies system. However, the user generated tags have a potential to become a complementary source to enhance and enrich a controlled vocabulary system which has the presence of multiple semantic relationships between them.

With the help of semantic technology the integration of social tags and controlled vocabularies can be achieved. With this combination there is a possibility to improve the access and organisation of information resources.

REFERENCES

Anfinnsen, S., Ghinea, G., & de Cesare, S. (2011). Web 2.0 and folksonomies in a library context. *International Journal of Information Management*, 31(1), 63-70.

Ansari, M. (2005). Matching between assigned descriptors and title keywords in medical theses. *Library Review*, 54(7), 410-414.

Davarpanah, M. R., & Iranshahi, M. (2005). A comparison of assigned descriptors and title keywords of dissertations in the Iranian dissertation database. *Library Review*, 54(6), 375-384.

Engelson, L. (2013). Correlations between title keywords and LCSH terms and their implication for fast-track cataloging. *Cataloging & Classification Quarterly*, 51(6), 697-727.

Furner, J. (2010). Folksonomies. In M. J. Bates, & M. N. Maack (Eds.), *Encyclopedia of library and information sciences* (3rd ed.). New York: Taylor and Francis.

Golder, S. A, & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2), 198-208.

Guy, M., & Tonkin, E. (2006). Folksonomies: Tidying up tags? *D-Lib Magazine*, 12(1). Retrieved July 30, 2018 from http://www.dlib.org/dlib/january06/guy/01guy.html.

Heymann, P., & Garcia-Molina, H. (2009). *Contrasting controlled vocabulary and tagging: Do experts choose the right names to label the wrong things?* Paper presented at the Second ACM International Conference on Web Search and Data Mining (WSDM), Barcelona, Spain.

Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In Y. Sure & J. Domingue (Eds.), *The semantic web: Research and applications* (pp. 411-426). Berlin/Heidelberg: Springer.

Kipp, M. E. I. (2006). *Exploring the context of user, creator and intermediary tagging*. Retrieved July 30, 2018 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.172.9783&rep=rep1&type=pdf.

Kipp, M. E. I. (2007). *Tagging practices on research oriented social bookmarking sites*. Retrieved July 30, 2018 from http://hdl.handle.net/10150/105837.

Kipp, M. E. I. (2011a). Tagging of biomedical articles on CiteULike: A comparison of user, author and professional indexing. *Knowledge Organization*, 38(3), 245-261.

Kipp, M. E. I. (2011b). User, author and professional indexing in context: An exploration of tagging practices on CiteULike. *Canadian Journal of Library and Information Science*, 35(1), 17-48.

Lee, D. H., & Schleyer, T. (2010). *A comparison of meSH terms and CiteULike social tags as metadata for the same items*. Paper presented at the 1st ACM International Health Informatics Symposium, Arlington, VA, USA.

Lee, D.H., & Schleyer, T. (2012). Social tagging is no substitute for controlled indexing: A comparison of Medical Subject Headings and CiteULike tags assigned to 231,388 papers. *Journal of the American Society for Information Science and Technology*, 63(9), 1747-1757.

Lu, C., Park, J., & Hu, X. (2010). User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings. *Journal of Information Science*, 36(6), 763-779.

Lu, C., Zhang, C., & He, D. (2016). Comparative analysis of book tags: A cross-lingual perspective. *The Electronic*

Library, 34(4), 666-682.

Lu, K., & Kipp, M. E. I. (2014). Understanding the retrieval effectiveness of collaborative tags and author keywords in different retrieval environments: An experimental study on medical collections. *Journal of the Association for Information Science and Technology*, 65(3), 483-500.

Mathes, A. (2004). *Folksonomies: Cooperative classification and communication through shared metadata*. Retrieved July 30, 2018 from http://www.bibsonomy.org/bibtex/245ae9616f7c7e480384d43cb2f6aec4d/jil.

Mohammad, F. (2018). Is preprocessing of text really worth your time for online comment classification? *ArXiv:1806.02908*. Retrieved July 30, 2018 from http://arxiv.org/abs/1806.02908.

Névéol, A., Doğan, R. I., & Lu, Z. (2010). Author keywords in biomedical journal articles. *AMIA Annual Symposium Proceedings*, 2010, 537-541.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013). Using of Jaccard coefficient for keywords similarity. In *Proceedings of the International MultiConference of Engineers and Computer Scientists, March 13-15, 2013*. Hong Kong.

Peters, I., Kipp, M. E. I., Heck, T., Gwizdka, J., Lu, K., Neal, D., & Spiteri, L. (2011). Social tagging & folksonomies: Indexing, retrieving... and beyond? *Proceedings of the 74th Annual Meeting of the American Society for Information Science and Technology*, 48(1), 1-4.

Rafferty, P. M. (2017). *ISKO Encyclopedia of knowledge organization: Tagging*. Retrieved July 30, 2018 from http://www.isko.org/cyclo/tagging.

Risueño, T. (2018). *What is the difference between stemming and lemmatization?* Retrieved August 29, 2018 from https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/.

Stan, J., & Maret, P. (2017). Social bookmarking or tagging. In R. Alhajj, & J. Rokne (Eds.), *Encyclopedia of social network analysis and mining*. New York: Springer. https://doi.org/10.1007/978-1-4614-7163-9_91-1.

Strader, C. R. (2009). Author-assigned keywords versus Library of Congress Subject Headings: Implications for the cataloging of electronic theses and dissertations. *Library Resources & Technical Services*, 53(4), 243-251.

Syn, S. Y., & Spring, M. B. (2010). Tags as keywords: Comparison of the relative quality of tags and keywords. *Proceedings of the American Society for Information Science and Technology*, 46(1), 1-19.

Syn, S. Y., & Spring, M. B. (2013). Finding subject terms for classificatory metadata from user-generated social tags. *Journal of the American Society for Information Science and Technology*, 64(5), 964-980.

Thada, V., & Jaglan, V. (2013). Comparison of Jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. *International Journal of Innovations in Engineering and Technology*, 2(4), 202-205.

Thomas, M., Caudle, D. M, & Schmitz, C. (2010). Trashy tags: Problematic tags in LibraryThing. *New Library World*, 111(5-6), 223-235.

United Nations Educational, Scientific and Cultural Organization. (2017). *The current status of ocean science around the world*. Paris: United Nations Educational, Scientific and Cultural Organization.

Voorbij, H. (1998). Title keywords and subject descriptors: A comparison of subject search entries of books in the humanities and social sciences. *Journal of Documentation*, 54(4), 466-476.

Wal, T. V. (2004). *You down with folksonomy?* Retrieved July 30, 2018 from http://www.vanderwal.net/random/entrysel.php?blog=1529.

Wu, D., He, D., Qiu, J., Lin, R., & Liu, Y. (2013). Comparing social tags with subject headings on annotating books: A study comparing the information science domain in English and Chinese. *Journal of Information Science*, 39(2), 169-187.