DESIDOC Journal of Library & Information Technology, Vol. 38, No. 1, January 2018, pp. 11-15, DOI : 10.14429/djlit.38.1.10969

Social Semantics and Similarities from User-generated Keywords to Information Retrieval: A Case Study of Social Tags in Marine Science

Praveenkumar Vaidya[*] and N.S. Harinarayana[#]

[*]*Tolani Maritime Institute, Induri, Pune - 410 507, India*
[#]*Department of Studies in Library and Information Science, University of Mysore, Mysuru - 570 006, India*
[*]*E-mail: vaidyapraveen@gmail.com*

ABSTRACT

Of late, social tagging has become popular trend in information organisation. In context of digital resources the tags assigned by users also play vital role in information retrieval. For information discovery the 'terms' used to retrieve the results also depend upon the 'relevancy' or 'weightage' of the keywords. This study investigates 'relevancy ranking' of terms used in the full text of the resource. The common words present in both full text of the article and social tags were considered for the study by employing TF-IDF statistical technique and Jaccard similarity test. The results show that it is possible to assign 'weight' to keywords for better results and also determine the significant tags assigned by the user. The Jaccard similarity coefficient test adopted to understand the word similarity between full text words of an article and marine social tags. This work reveals the social tags can enrich metadata for information retrieval.

Keywords: Web 2.0; Folksonomies; Social tags; TF-IDF; Jaccard similarity; Coverage ratio; Information retrieval; Knowledge organisation

1. INTRODUCTION

As a popular component of Web 2.0 technology, social tagging is a new phenomenon in the organisation, management, and discovery of digital resources. In tagging system any one can participate in the process and assign the keywords they prefer. These social tags assigned by majority of participants for resources are treated as valuable vocabulary to organise and share resources within the community[1]. The concept of social tagging is adopted by many information communities and presently vast amount of online information is operated by social tags as information discovery tool. The semantic value of social tags is also crucial to measure the performance of information retrieval.

It is vital for researchers to understand the semantic and similarity value of these social tags. This study examines the semantic and similarity information associated with social tags and the terms used in the full text of the research articles. It is assumed that there is a semantic and similarity relationship between social tags and terms in the full text article. The analysis of tags exhibit the semantic similarity in information retrieval. The social tags, author keywords, subject headings and few important terms from the article will also enhance the information retrieval by providing additional access points.

This study makes an effort to understand the 'relevance ranking' with the help of term frequency-inverse document frequency (TF-IDF) and 'similarity' by employing Jaccard similarity coefficient. TF-IDF is a common mathematical method of weighing texts for information retrieval and automatic indexing[2]. In this context of information retrieval the term 'relevance ranking' suggests 'statistically significant search results'. It is believed that the application of statistical analysis against texts has greater information retrieval advantage over Boolean search[3].

The Jaccard index is used for comparing the similarity and diversity of sample sets, which measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets[4]. Hence, with the help of TF-IDF and Jaccard index methods this work will investigate the semantic and similarity analysis of user-generated keywords of marine science full text research articles.

2. LITERATURE REVIEW

Conventionally, information retrieval (IR) performance is determined in terms of speed, precision and recall and these measures can be extended to web IR systems[5-6]. The main purpose of information retrieval system is to provide users the information they require in efficient and effective way. The researchers are interested to find whether the social tags have any such influence in improving the retrieval performance with respect to the traditional methods.

It is widely accepted feature that social tagging is not only the established method to organise and index the user generated content but also employ them for retrieval when information is requested[7-8].

Received : 14 December 2016, Revised : 07 August 2017
Accepted : 09 November 2017, Online published : 01 January 2018

2.1 Semantic Analysis of Social Tags Through TF-IDF

Various studies focuses on the semantic aspects of social tagging. Peters and Stock[9-10] detailed the number of inconsistencies in social tagging. According to them unstructured nature of social tags create restriction, rather than enhancing the results, in information retrieval and they proposed a criteria for relevance ranking of tagged documents with the help of TF-IDF.

Yi[11] devised a conceptual frame work for information retrieval for folksonomy on the basis of Library of Congress Subject Headings (LCSH). He proposed TF-IDF algorithm to find the similarities between folksonomies and subject headings. Yi[12] again investigated the ways of predicting relevant subject headings from resources from social tags assigned to the resources. The prediction of subject headings was measured by TF-IDF. Zubiaga[13], *et al.* performed a tag based resource classification study by adopting TF-IDF weighting scheme. The researchers found that the tag distribution in social tags can help to determine the relevance of tags and hence useful retrieval. Studies show that TF-IDF is one of the popular statistical tool used to find out the relevance tags to enhance search results in information retrieval.

2.2 Similarity Analysis Through Jaccard Coefficient Method

Jaccard similarity analysis is one of the popular method to determine the similarity coefficient between two set of datasets. Heyman and Garcia-Molino[14] in their comparison work of tags with controlled vocabularies by adopting Jaccard similarity method, observed that many of the keywords designated by the tags and controlled vocabularies are similar or the same but the usage of keywords by annotators are different.

Jaccard index method is being used in determining the similarity between tags and controlled vocabularies[15-17]. These researchers found that it was essential to use the Jaccard similarity index to compare social tags and subject terms. The results reveal that the social tags and controlled vocabularies are quite distinct lexically and semantically, reflecting the different viewpoint and processes between them. It was found the lexical overlap between the two sets of data was marginal. However, despite limitations, the social tags have the potential to become a complementary source to expand and enrich the controlled vocabulary system.

Yi[12] investigated the ways to predict relevant subject headings for resources from social tags assigned to the resources. The predication of subject headings was measured by Jaccard similarity method. This study demonstrated the application of similarity technique to predict correct Library of Congress Subject Headings.

These studies indicate the importance of TF-IDF and Jaccard similarity in context of information retrieval, knowledge organisation and metadata enrichment.

3. RESEARCH OBJECTIVES

In this study, an attempt is made to answer following research objectives.

(a) Is it possible to assign weights to search results and arrange them statistically using TF-IDF?

(b) Is it possible to weigh terms and provide relevance ranking without knowing the semantic meaning of a word in the index by using TF-IDF?

(c) To what extent Social tags comprise similar vocabulary from the article?

(d) To what extent readers assign similar words that the author uses in the article?

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

4. SCOPE AND LIMITATION OF STUDY

In this study, the full text articles were chosen from marine science subject. The researchers identified the marine science journal from where the articles were identified then collected the respective social tags from CiteULike. The CiteULike is a social web service where users can save and share citations to academic papers. However, the care was taken that full text of the article was available for analysis. In this context, we considered three full text articles from the open access journal *Ocean Science* where users have assigned minimum ten tags to these articles.

Table 1. Glimpse of the data collected

Doc No.	Word count	Tag count A	Wordlist B	Common words A&B
1	8082	31	1268	27
2	8748	22	1154	16
3	7216	24	1051	14

5. METHODOLOGY

To determine social semantic and similarity values of the text the methodology adopted is both term frequency-inverse document frequency (TF-IDF) and Jaccard similarity test. In this perspective, the data was collected as described below.

The tags extracted were transferred to MS Excel sheet and arranged them as required for the study. Table 1 gives the nature of data collected for this study. The full text articles chosen for the study were converted to wordlist by removing the stop words (the, an, a) used in the article. For example, in Table 1, the document 1 consists of 8082 words. Instead of using just title, keywords and abstract, the full text mining creates extra difficulties and noise. Hence after processing, the terms were reduced to 1268. The acknowledgement part, authors' name and emails, references were omitted from the full text to analyse only the academic part of the scientific work[18].

The social tags provided by users were also split into single word to make more precise in composition. For example 'Geostrophic-turbulence' was split into two separate words like 'geostrophic' and 'turbulence'[19]. It is quite obvious that by splitting the words the chance of finding common words will increase. The word list also gives an advantage of finding the frequency of the words appeared in the document. The wordlist of the article were compared to find common terms among the corpus created. The common words were considered in setting up the calculation as these words are contributed by both author and users. The tags annotated by user and the terms used by author in the article will provide the researcher to find the 'significant terms' for information retrieval. Taking such an approach-the application of statistical analysis against texts- does have its information retrieval advantages over straight Boolean logic[3].

Table 2. TF-IDF for common words for doc 1 and doc 2

Terms	TF 1	IDF 1	TF-IDF1	TF 2	IDF 2	TF-IDF 2
Water	0.006682	1.405465	0.009391	0.0053727	1.405465	0.0075511
Ocean	0.006805	1.405465	0.009565	0.0017147	1.405465	0.0024099
Interannual	0.001114	1.405465	0.001565	0.0008002	1.405465	0.0011246

6. ANALYSIS AND INTERPRETATION

6.1 TF-IDF for Sematic Analysis

TF-IDF is a conventional term weighting method as well as a fundamental way of documenting similarity based on mathematical Boolean operations. This is regularly applied method in the field of information retrieval. In simple words, TF of a term in a document is the number of occurrences of the term in the document. TF is usually normalised over the size of a document to provide standardised measure regarding document size. DF is the number of documents in which a term occurs in a corpus of documents. IDF which measures how important the term is. While computing TF all terms are considered equally important. However it is known that certain terms, such as 'is', 'of', and 'that', may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing IDF (t) = log_e(Total number of documents / Number of documents with term t in it)

The basic assumption of DF in TF-IDF is that as a term appears in fewer distinct documents, the value of the term in DF increases so that it holds more weight in TF-IDF.

Therefore TF and IDF are calculated as follows for this research work.

$$TF(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

$$IDF(t) = 1 + \log_e \frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}$$

For the three documents selected for this work the statistical analysis divulges very significant results.

In document 1, we could find 27 common terms provided by both user and author of the article. The word 'salinity' appeared 128 times followed by 'data' with 74 times in the article. In document 2, 'layer' and 'deep' were top two terms whereas 'sea' and 'straight' were most preferred term in document 3. In whole of these three documents we could find three common terms (water, ocean, and interannual) in doc 1 and doc 2 and then two words (terms, circulation) in doc 2 and doc 3. The present study uses these words for further analysis.

TF and IDF for doc 1 and doc 2 for terms 'water', 'ocean', and 'interannual' are as shown in Table 2. For doc 3 the results will be 0, in absence of any common terms between doc. 1 and 3. Similarly TF and IDF for doc 2 and doc 3 for words 'terms' and 'circulation'.

For doc 1 the results will be 0 in absence of any common words between doc 3 and doc 1. It is now observed that if there is any query like 'water circulation' then all three documents will be in search for the retrieval of the results. Hence the resultant TF-IDF for this query in all documents would be as mentioned in Table 4.

Table 3. TF-IDF for common words for doc 2 and doc 3

Circulation	Terms
0.001029	0.000114
1.405465	1.405465
0.001446	0.000161
0.001247	0.000693
1.405465	1.405465
0.001753	0.000974

Table 4. TF-IDF for search term water circulation

	Doc 1	Doc 2	Doc 3
Water	0.009391	0.007551	0
Circulation	0	0.001446	0.001753

With TF-IDF analysis another important observation is manifested in this research work. In document 1 the word 'water' appears 54 times and another word 'variability' appears 41 times. But in list of relevance words 'variability' finds 4th place and 'water' much below in the list. Hence the word 'variability' is more in weight than 'water' because 'water' appears in two documents and is considered as less in relevance. This statistical analysis definitely helps to enhance the information retrieval results.

We can also surely distinguish between all words in the respective documents in terms of their relevance. Given such a list, it is possible to take first three terms from each document and call them most significant subject tags. TF-IDF also works rightly in the context of enhancing and enriching metadata of the resources to facilitate retrieval with speed and precision.

For document 1 the list of significant words is presented in graphical format to understand it better. In this given order weightage distribution of the tags is obtained.

Table 5. Significant terms identified for doc 1, doc 2, and doc 3

Doc 1	TF*IDF	Doc 2	TF*IDF	Doc 3	TF*IDF
Salinity	0.033237	Layer	0.020871	Sea	0.015996
Data	0.019215	Deep	0.013194	Strait	0.006689
Time	0.014541	Convection	0.012954	Gibraltar	0.006107

6.2 Similarity Analysis by Jaccard Similarity Method

The Jaccard similarity coefficient is used to measure the similarity between the frequent sets of tags and the terms employed in the article[15] and is calculated according to following equation

$$J(T,W)=\frac{T\cap W}{T\cup W}$$

where *T* represents social tags assigned by users and *W* indicates the words used by the author in the fulltext article. The Jaccard similarity coefficient measures the similarity between finite sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets. The Jaccard similarity coefficient[20] measures the share properties of both social tags and words whereas all of the objects are represented by 0 and 1, respectively. It is also true that if Jaccard value is close to 0, it means that they are not similar at all.

With respect to Table 1, it indicates that in doc 1 the common words are 27 whereas 31 tags are assigned by users and 1268 unique words are listed in the full text of the article chosen for study. Subsequently for document 2 and document 3 the data is reflected in the table. By replacing the values in the above equation Jaccard coefficient value is determined for these three articles individually.

Table 6 indicates the Jaccard value calculated for all three articles together is 0.0163 which is less than one. This work has taken consideration of all unique words in the three articles, as well as social tags extracted for these same articles. By observing the Jaccard coefficient value from the table, it is obvious that there are very less similar words in the corpus of social tags and unique words of the article.

Table 6. Jaccard coefficient value for three articles

Articles	Tags T	Unique words W	Common words in T & W	Jaccard coefficient
1	31	1268	27	0.021226415
2	22	1154	16	0.013793103
3	24	1051	14	0.013195099
ALL 3	77	3473	57	0.016318351

6.2.1 Jaccard Distance

We can also calculate the Jaccard dissimilarity for these sample sets, as we determined similarity coefficient. The 'Jaccard distance' measures the dissimilarity between sample sets, which is complementary to the Jaccard similarity coefficient and is obtained by subtracting the Jaccard similarity coefficient from 1 or by dividing the difference between the sizes of the union and the intersection of the two sets of the size of the union.

$$d\ j(T,W)=1-J\,(T,W)=\frac{|T\cup W|-|T\cap W|}{|T\cup W|}$$

Substituting the values to the above equation the Jaccard dissimilarity is found to be 0.9837 which is very high with reference to the above full text articles chosen for this study. The dissimilarity index also indicates a very small percentage of similar words in the sets of words selected. The poor similarity proves that social tags are free in nature and have no explicitly defined relationship or hierarchy between the terms.

6.2.2 Coverage Ratio

The corpora of social tags and full text terms in our study have different sizes, so we also calculated the coverage ratio. The coverage ratio is defined as the fraction of the common annotations for an article covered by its full text terms and tags respectively. Examining the coverage ratio can help determine whether full text terms could be substituted for social tags or vice versa.

$$\text{Coverage Ratio of unique terms of the article}=\frac{T\cap W}{W}=0.0164$$

$$\text{Coverage ratio of Tags}=\frac{T\cap W}{T}=0.7402$$

In this case, it may make sense to suggest existing tags to users, because they are more likely to contain appropriate terms for annotations than the full text terms.

7. DISCUSSION AND CONCLUSIONS

The study of TF-IDF and Jaccard similarity test are important in context of information retrieval process in library and information science. This research work prominently tried to highlight the significance of social tags in enriching metadata with emphasis on retrieval. These are conventional tools to test the similarities between two sets of words extracted.

The research objective (RO) (a) is adequately answered in proving the possibility of assigning weight to arrange them statistically with the help of Table 2-4 by using TF-IDF. It is important to note that the terms are converted into statistical values and calculated them with the amount of frequency appeared in the document. The meaning of the word has no role to play in providing the search results. Again the statistical values are regrouped into words to make them meaningful text. This is also the response to RO (b), where in Table 5 provides list of significant words in terms of relevance or weightage.

RO (c) is answered sufficiently in Table 6 by finding the results from Jaccard similarity calculation. This similarity result has clearly proved that the social tags are less in similar to the words used by author in their research articles. The user driven social tags may be useful to some extent but may not replace other varieties of structured vocabularies for information retrieval. In reply to RO (d) the researcher has used Jaccard distance and coverage ratio to demonstrate dissimilarity of words and also significant words respectively between user driven tags and full text terms of author. This also indicates the implication of social tags for metadata enrichment and may prove potential complementary source for information retrieval and in knowledge organisation.

For future work, these document values can be used to derive a vector. These set of documents in collection will be viewed as a set of vectors in a vector space and similarity between two documents can be determined. The cosine similarity measure can be worked out for the same set of values.

REFERENCES

1. Mathes, A. Folksonomies - cooperative classification and communication through shared metadata. 2004. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html (Accessed on 15

January, 2016).

2. Hjørland, Birger. Term frequency-inverse document frequency. 2006. http://www.iva.dk/bh/Core%20 Concepts%20in%20LIS/articles%20a-z/weighting. htm(Accessed on 15 January, 2016).
3. Morgan, E.L. TF-IDF in libraries: Part II of III (For programmers) http://infomotions.com/blog/2009/04/tfidf-in-libraries-part-ii-of-iii-for-programmers/(Accessed on 15 January, 2016).
4. Term frequency-inverse document frequency. Wikipedia. https://en.wikipedia.org/wiki/Tf%E2%80%93idf (accessed on 15 January, 2016).
5. Kobayashi, M. & Takeda, K. Information retrieval on the web. *ACM Computing Surveys (CSUR)*, 2000, **32**(2), 144–173.
6. Manning, Christopher D. Prabhakar Raghavan & HinrichSchütze, Introduction to information retrieval. Cambridge University Press. 2008. http://nlp.stanford.edu/IR-book/html/htmledition/contents-1.html(Accessed on 15 January, 2016).
7. Hotho, A.; Jaschke, R.; Schmitz, C. & Stumme, G. Information retrieval in folksonomies: Search and ranking. Lecture Notes in Computer Science, Volume 4011. Chapter 31, 411-426. Springer Berlin, 2006.
8. Golder, S.A. & Huberman, B.A. Usage patterns of collaborative tagging systems. 2006. *J. Info. Sci.*, **32**(2), 198-208.
9. Peters, I. & Stock, W.G. Folksonomy and information retrieval. *In* Annual Meeting of the American Society for Information Science and Technology, Stuttgart, Germany. 2007, **45**, p33.
10. Peters, I. & Stock, W.G. Power tags in information retrieval. *Library Hi Tech*, 2010, **28**(1), 81-93.
11. Yi, K. A conceptual framework for improving information retrieval in folksonomy using Library of Congress subject headings. *In* Proceedings of American Society Information Science and Technology, 2008. 45, pp. 1–6. http://onlinelibrary.wiley.com/doi/10.1002/meet.2008.1450450368/pdf (Accessed on 15 January 2016)
12. Yi, K. A semantic similarity approach to predicting library of congress subject headings for social tags. *J. Am. Society Info. Sci.*, 2010, **61**(8), 1658-1672.
13. Zubiaga, A.; Martinez, R. & Fresno, V. Analyzing tag distributions in folksonomies for resource classification, *In* Proceedings of the 5th International Conference on Knowledge Science, Engineering and Management, 2011, KSEM ‘11.
14. Heyman, C & Garcia-Molino, L. Contrasting controlled vocabulary and tagging: Do experts choose the right names to label the wrong things? *In* WSDM 09 Barcelona, Spain. 2009.
15. Lu, C.; Park, J.R. & Hu, X. User tags versus expert-assigned subject terms: A comparison of library thing tags and library of congress subject headings. *J. Info. Sci.*, 2011, **36**(6), 763-779.
16. Lee, T. Social tagging is no substitute for controlled indexing: A comparison of medical subject headings and CiteULike tags assigned to 231, 388 papers. *J. Am. Society Info. Sci. Technol.*, 2012, **63**(9), 1747-1759
17. Wu, D.; He, D.; Qiu, J.; Lin, R. & Liu, Y. Comparing social tags with subject headings on annotating books: A study comparing the information science domain in English and Chinese. *J. Info.Sci.*, 2013, **39**(2), 169-187.
18. Janssens, Frizo; Leta, Jacqueline; Glanzel, Wolfgang & De Moor, Bart. Towards mapping library and information science. *Info. Proces. Manag.*, 2006, **42**(6), 1614-1642 doi: 10.1016/j.ipm.2006.03.025
19. Chen, Yi-Chen, Analysis of social tagging and book cataloging: A case study. 2008. www.hkla.org/events/2008/conf/**yi**.ppt (Accessed on 13 December, 2015).
20. Niwattanakul, Suphakit; Singthongchai, Jatsada; Naenudorn, Ekkachai; & Wanapu, Supachanun. Using of Jaccard coeffcient for keywords similarity. *In* Proceedings of the International Multi Conference of Engineers and Computer Scientists, 2013, Hong Kong. IMECS 2013, 13-15 March, 2013, **1**.

CONTRIBUTORS

Mr Praveenkumar Vaidya, received Masters in Library and Information Science from Karnataka University, Dharwad. He is a research student at the Department of Studies in Library and Information Science, University of Mysore, Mysuru. Presently, working as Librarian at Tolani Maritime Institute, Pune, Maharashtra. His research area include Folksonomies, Social tagging, Metadata and Information retrieval.

Contribution in the current study, he conceptualised the article, research questions and carried out the data collection, literature review, analysis and data interpretation. He also contributed in drafting and revising the manuscript.

Dr N.S. Harinarayana received PhD in Library and Information Science from University of Mysore, Mysuru, is working as an Associate Professor at the Department of Studies in Library and Information Science, University of Mysore, Mysuru. He teaches the courses Library automation, Information retrieval, and library classification and content organisation. His research area include, metadata and information retrieval, automation and networking and scientometrics.

Contribution in the current study, he suggested, drafted, revised and improvised the contents of the manuscript to put it in better shape for final approval.