Research Journal of Applied Sciences, Engineering and Technology 7(24): 5172-5178, 2014
DOI:10.19026/rjaset.7.914
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: February 08, 2014 Accepted: April 09, 2014 Published: June 25, 2014

Research Article

Measuring the Relevancy between Tags and Citation in Social Web

[1]Saif Ur Rehman, [2]Ghani-ur-Rehman, [1]Aftab Ali Haider, [1]Tanveer Afzal and [3]Kamran Aziz
[1]Department of Computer Science, Muhammad Ali Jinnah University, Islamabad, Pakistan
[2]Khushal Khan Khattak University, Karak, Pakistan
[3]Abasyn University, Islamabad, Pakistan

Abstract: With the advent of web, massive information is available to the internet users. One can acquire information from this according to his or her own field of interest; for example we can have large amount of information on bioinformatics available on the web, computer researcher community can found any type of published data at any period of time with just a single click on the Google or any other well renewed web search engines. Filtering the most relevant information from a large dump of online information is considered a challenging task, which is gaining popularity in the web research community. Now, various scientific tools and techniques have been introduced which enable the users to extract the relevant and required information. The accuracy of the information extracted is an interrogative mark. In research community the citation is very common term. Citations are used to extract the historic information relevant to some particular topic. But the citation of a specific research article requires enough time to cite a paper. However, in today's social bookmarking period of the concept of tags is gaining popularity. Tags are assigned to papers or some topic by reviewers or its readers in a short period of time. In this study, we worked to find the relevance among the citations of a research paper to the tags assigned to these papers. Furthermore, we have obtained the titles of the cited papers and then perform comprehensive analysis on how much the tags are involved in titling the upcoming research articles. This will be helpful to argue that a tag can be used to assess the future diffusion of a research paper. For this we have provided our own framework. For validation of our framework, we have used the CiteULike, a very renewed social bookmarking web site articles to evaluate the performance of our proposed framework.

Keywords: Citation, co-relation, knowledge discovery, social bookmarking, statistical analysis, web

INTRODUCTION

It is the age of information technology. In information technology, we cannot live alone without internet. Internet is the biggest source of huge amount of useful information. This information contains: text data, audio, video, pdf documents, images etc. Among these sources of information, the Research publications are the source of knowledge for the scholars or researchers working in various domain including science, arts etc. The ideas and thoughts discussed in these papers are used as a basic knowledge for further research. A theory may progress from biology into computer science e.g., genetic algorithms or vice versa e.g. genes ontology. This transfer of knowledge is termed as knowledge diffusion. It is challenging task to track this diffusion of knowledge. References to a research publication are called citations. Citations have been used as a means to study the diffusion of knowledge. Patent to paper citation based knowledge diffusion has been studied (Carpenter *et al*., 1980). They found that basic sciences are major source of knowledge diffusion as compared to engineering and applied sciences. Citations to a research paper take time before these are available to research community. Sometimes a citation to a research paper may not have strong relevancy to it. Tags can be used as an alternative to citations to trace the knowledge diffusion (Saeed *et al*., 2008a).

Web 2.0 has revolutionized the web by giving a user more flexibility. Social bookmarking sites like bibsonomy and cite Ulike has emerged that gave users privilege to review and tag the available information. Empowerment of user to tag a paper has prospects to optimize the tracing of knowledge diffusion. These tags describe a paper and this description can be used to identify future diffusion of knowledge (Wu *et al*., 2006).

In the recent past, social data mining has started to gain interest in academia and the practitioner world alike. Social book marking systems have been very successful in attracting and retaining users. This success initially originated from members' ability to centrally store bookmarks on the web (Wetzker *et al*., 2008).

Corresponding Author: Saif Ur Rehman, Department of Computer Science, Muhammad Ali Jinnah University, Islamabad, Pakistan

With the coming of age of these services, however, the user's perceived value shifted toward the underlying social effects, such as trend indication, advanced web search or recommendation functionality (Wetzker *et al*., 2008). For researchers, these services are an invaluable source of information, since they provide a vast amount of user-generated annotations (tags) and reflect the interests of millions of users. The social-aspect of these systems derives from the fact that resources (in general web pages) are tagged by the community and not by the creator of content alone. This characteristic, called collaborative tagging, provides relevant metadata (Paul *et al*., 2008) and is expected to boost the semantic quality of labels (James, 2004).

Tags when compared to citations are assigned by the serious reviewers of an article. These tags are much like keywords and are essence of a paper. A strong semantic relationship exists between the tags and a research paper. In the previous literature (Saeed *et al*., 2008b), researchers have established a co-relationship between the citations and tags. This study was based on papers of WWW'06 where author verified tag citation correlation and recurrence of tags terms in the titles of the citations. In order to further strengthen and investigate our previous research findings on a small dataset we have adopted a diversified dataset. This dataset consists of nearly 3000 research papers of cite Ulike, a very popular social bookmarking site. In cite Ulike, the users create their own libraries and add the research articles of their own research domain. We have selected 15 such users from various domain of computer science including software engineering, Business Intelligence, Decision Support System, Data Mining, Databases, Data Warehousing, Theory of computation and Artificial Intelligence and web related topics etc. We have performed extensive experiments on this dataset and from these experiments we have concluded that there is some co-relation between and citations. Furthermore, we have founded that tags of a research article have worth in the title of future citations of these papers.

Social bookmarking is the beginning of Web 2.0. Social bookmarking is highly dynamic, real time scenario that includes sentiments of the reviewer as well. In literature; some useful studies have been conducted to identify some relationship between the tags and their citations (Saeed *et al*., 2008b). The authors have obtained the tags and citation count from the CiteULike and perform the co-relation analysis between the tags count and citation count. They have obtained tags as well as citation count manually and the number of paper for analysis was less than one thousands. They have performed the analysis on small dataset manually and concluded that there exists some positive relationship between tags count and citation count of specific research articles.

An improved search model based on social bookmarking was proposed for ranking and filtering the web pages (Yanbe *et al*., 2007). User sentiments were given worth to rank and filter pages. This model was supposed to give a new direction to link based ranking algorithm. Research on tagging has gain momentum and is thought to be much realistic approach for analytical studies based on purely user's sentiments and reviews. Recently, very useful taxonomy was suggested to classify tagging to facilitate users (Marlow *et al*., 2006). Tagging was realized as tripartite ontology model (Peter, 2007). Tripartite ontology used to find diffusion of knowledge and to rank topics (Hotho *et al*., 2006).

A recommender system was suggested that was based on past interaction of user. This system enables users to suggest some meaningful tags. Social bookmarking describes collective behavior of public that is visible to all. Social bookmarking sites allow sharing of resources and community can benefit from each other (Pierpaolo *et al*., 2007). Tags were used to gather information about the user and further based on this information they developed user models to classify the users in various groups (Carmagnola *et al*., 2007). The knowledge diffusion has been analyzed using geospatial model. They used terror activities, avian influenza and news stories to trace the knowledge diffusion across the globe (Chen *et al*., 2007).

Tags and bookmarking have been presented a new perspective by using it to find diffusion of knowledge. In this study, they argued that tags have potential to replace citations. Citations had been previously used to trace the knowledge diffusion (Saeed *et al*., 2008b). Some studies have been worked on the knowledge diffusion using patents and published papers. This research revealed that basic sciences are major source of knowledge diffusion as compared to applied sciences. As discussed, tagging systems are real time scenario that includes sentiments of the users; it is possible to spontaneously capture the knowledge diffusion using tags (Carpenter *et al*., 1980). Bookmarking based ranking was successfully carried out with citation and co-author network based ranking. They found encouraging results that supported there thought to replace citations with tags (Saeed *et al*., 2008b, 2010).

The focus of this research study is to find the relationship between the total tags that a research article has received and the total citation that the same research article obtained. Furthermore, we also have found that there exists a strong relationship between the tags of the research article and the articles that has cited the tagged articles in the future. To validate our work, we mainly divided this study into two sections. Firstly, we have obtained the tags count of each article and then obtained the citation code of each article respectively from Google scholar. We have performed the analysis to identify a positive relationship between tags count and citation count that a specific article has received

from the readers and the cited articles respectively. Second objective of this research study is to analyze the sentiment analysis of the tags in the title of the future coming articles of the cited papers. For this purpose we have obtained the cited articles titles from the Google scholar.

MATERIALS AND METHODS

CiteULike is one of the popular social bookmarking sites where millions of research papers are available for review. Users have the option to tag and bookmark these papers. The objective of this study is to find a co-relation between tags and citations. If a strong co-relation between tags and citations is found, it will be possible to take tags as an alternative to citations. Secondly, it is possible to find a semantic relationship between the titles of the citations and tags assigned to a paper. If these tags are semantically matched with the titles of citations, we can infer that tags are an alternative to citation.

In this study we have selected more than 3000 research papers randomly from CiteULike. The core modules or parts of this framework have been described in depth hereafter:

The CiteULike has allocated a special IDs to each article that have added to someone own personnel collection of articles. In this setup, first the articles IDs have been extracted from the CiteULike web site. We have developed a module called Paper IDs Extraction (). This module has extracted the papers' ID automatically from the CiteULike web resource and has updated the database with these article IDs.

Papers' details extraction: In the next step, we have used the articles' ID obtained in the previous step and pass each ID to the module called article Details Extraction (). This method has browsed the CiteULike against the article IDs passed to it and has retried all the relevant details about each articles including:

- Title of each article
- Authors names
- Tags
- Total number of tags that have been allocated by different users to this particular article

This has updated our database with these details of each article.

Citation acquisition: This is the core module in this setup. After having the details of the article, we have used it for obtaining the citation details. For this we have passed the article titles from the article Citation Details () method. This method has browsed the google scholar site. Based on the exactly match, we have updated our database with the total number of citation against each article as well as the title of the maximum 50 articles. Thus at the end of this method execution we have obtained the following detail against the article retrieved in the Papers' Details Extraction step including:

- Total citation count
- Article titles of the cited articles

Comparing tags versus citation count: In this step, we have used the module compare Tags Citation Count () count. This method has compared the total number of tags with the total number of citation obtained in the previous steps. The details of these results have been given in the Experimental Details. Based on this comparison, we have come up with the result that there exists some kind of positive relationship between total number of tags and the total citation that a particular article has received.

Comparing tags in cited papers' title: In this step, we have just used the articles' details obtained in the previous two steps. We have compared the tags of each article with the titles of their respective cited papers. The module that is used for this objective is called compare Tags Citations (). We have obtained the results for this purpose based on:

- Full title of the paper with full tag including stop words
- Comparing the tags and title of the articles after removing the stop words in both. This also has depicted that there tags that a paper has received from various users has also been used in the future papers that have cited the citing paper. Further details are given in the Experimental Section

Methodology of our research is given in Fig. 1. For each selected paper, we browsed Google Scholar to collect the count and titles of the citations to these papers. This information was again stored in local database. We analyzed the metadata by finding co-relation between the tags and citations of the papers. Secondly, we semantically matched the tags to the titles of the citations. Using fusion charts the results found were visualized. The framework of our research is shown in above Fig. 1. Moreover; the above discussed modules have been summarized in the following Fig. 2 and 3.

In order to carry out this study, we used MS Visual Studio, Net Framework 3.5 and C-Sharp with ASP.Net. Microsoft SQL Server 2005 was used as a back end database server. We have used the fusion charts for graphs and charts for the experimental results. The pseudo code is given in Fig. 2. We have collected the data from the CiteULike. In Cite ULike, users create

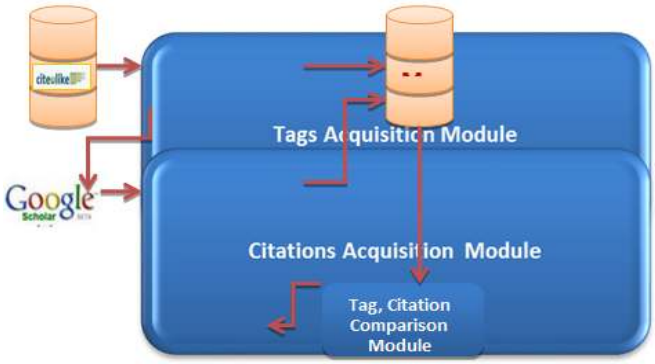

Fig. 1: Architecture of tool

Step:1 Selection of Research Papers in diversified areas through random Article ID generation for citeUlike
Step:2 Storing IDs in local database for further analysis
Step:3 Iterating Article ID to collect metadata of research papers from citeUlike e.g. Author name, Title, and Tags
Step:4 Storing metadata in local database
Step:5 For each title of a paper, collecting titles of citations from Google Scholar with max bound of citations equal to 100.
Step:6 Storing Titles of citations in local database
Step:7 For each paper, comparing tags and titles of citations
Step:8 Results are stored in local database
Step:9 Visualizing the facts found using fusion charts

Fig. 2: Architecture modules in steps

```
//comparing tags with titles of the citations for a paper
Void  compare tags, citations
Begin
For inti = 0; i< = total_ no._of _Tags;
   Fetch tags from db();
 For int j = 0; j< = total_no._of_Titles_for_citations
     Fetch titles_for_citationsfromdb();
     // Compare tags and Citations
     Compare tag with titles_for_citations
     // Store results in local database
     Save result in db();
     // Iterate till end
     Loop next
     End
     // Use fusion charts to display results
Visualize results using fusion charts
End
```

Fig. 3: Pseudo code of our approach

their own libraries and add the articles of their own research interest. The other users who have access these libraries can read these articles and then can tags to these articles. We have collected 3000 research papers of almost 15 CiteULike users of various domain of computer science including: software engineering, Business Intelligence, Decision Support System, Data Mining, Databases, Data Warehousing, Theory of computation and Artificial Intelligence.

```
//This program is used to retrieve the titles of citations
//from the Google Scholar web site when the Title of
//paper is passed to it

Void reterive_title of citations()
Begin
For inti = 0 I < = Title_ of_ Papers
Fetch_titles from db();

   //open the Google Scholar  and pass title to it to
  //get the paper citations not more than 100

//To get the paper title
Get_paper_title();

// Save in local db()
Save_titles_for_citationsto_DB();

Loop Next
 // Iterate till execution of loop
End
```

Fig. 4: Pseudo code for the fetching titles of the citations

Similarly, we stored the titles of citations in our local database. Pseudo code for it is given in Fig. 4.

We have taken the equal number of articles from these users' libraries. After getting the CiteULike article IDs of these papers. We have passed these IDs to our program to get the article titles, their authors' details and relevant tags. The routine that we have developed for this purpose is shown in Fig. 5. Furthermore, based on these article titles, we have used the routine given in Fig. 4 to get their total number of citations from the Google scholar. Then following the link given in the Cited by word in the Google scholar, we have obtained the 15-20 top most citation of the article. Thereafter, we have matched the tags of the paper with its cited papers' title and we have concluded the results as shown in the Table 1 and 2. All of these results clearly shows that there is strong relationship between the papers' tags and their cited by papers' titles.

```
//This program is used to retrieve the article details from
the citeulike web site when the article ID is passed to it

Void reterive_artilcle_details()
Begin
For inti = 0 I < = total_no_of_articles
Fetch_article_ID from db();

  //open the citeulike and pass the Id to it to
  //get the paper details

//To get the paper title
Get_paper_title();

// To get the papter authors details
Get_paper_authors();

//To get the papers' tags details
Get_paper_tags();

Save_article_details_to_DB();

Loop Next
 // get next article Id and perform the steps
 //again until loop satisfied

End
```

Fig. 5: Pseudo code for the fetching article details

Table 1: Results of the tags used in the cited papers' titles (with stop words included in tags and title of the study)

No. of papers	Matched (%)
698	64
482	52
350	45
293	35
371	21

Table 2: Results of the tags used in the cited papers' titles (after the deletion of stop words from tags and title of the study)

No. of papers	Matched (%)
698	92
482	87
350	78
293	71
371	47

RESULTS AND DISCUSSION

All of these experiments have been performed on MS Windows 7, with 4GB RAM core i3 processor and 200 GB hard drive. We have used the dataset from the CiteULike, a very popular social bookmarking web site, which allow the users to build their own libraries and add the research papers of their interest. We have collected about three thousand research article from the libraries of more than 15 CiteULike users. These users were from different domains including Decision Support System, Web, Software Engineering, Business Intelligence, Bioinformatics etc., we have collected almost 150 to 200 of research articles from each of the user's library. These papers have been published in different renowned international journals and conferences.

We have extracted the tags and titles of citations for these papers. As we earlier discussed, the objective of this research is to find a correlation between tags count and citation count of particular research articles. This study has proved positive relationship between tags and citation count. The results as indicated in the Fig. 6 below clearly depicts that the articles that have higher number of citation also has higher number of tags obtained from various users.

We semantically matched tags with the titles of citations. We have shown this relationship using fusion charts and result is given in Table 1. Further, we have performed an analysis that how much tags have been used in the cited papers' titles. These results revealed that there are 880 papers that have strong semantic matching with the titles of the citations. There are 350 research papers where tags adequately matched with the titles of the citations. Remaining papers were below threshold and were not semantically matched. We found papers that have no citations. This was mainly due to the fact that these were either latest papers or these papers were a review papers. However, in our study these papers were included and were part of that portion of papers that were not semantically matched.

The equation that we have used to obtain the results that are shown in the above Table 1 and 2 is given by:

$$Matched\ \%ge\ = \frac{N}{M} \times 100 \quad (1)$$

In Eq. (1), N represents the number of words in the title of the paper and the M denotes the total tags that an article has received. We have removed all the stop words from the tags as well as title of the cited papers and then we have done our calculations. The vocabularies of stop words have been obtained from (Stopwords, 2014).

The experimental results are shown in the above Table 1. These results are obtained when the tags and the title of the papers included the stop words. The results depicts that among the 3000 research articles; there are 698 research articles that have found tags matched with about 64% of the cited papers' titles and there are 371 articles that have found just 21% match of the tags in their future referring articles. There were some articles that have either no tags and also there were some articles that have tags but have very few % of their in the cited article titles participation.

Furthermore, we have also performed extensive experiments on the test dataset after removing the stop words from the tags of the articles as well as from the titles of the cited papers. These stop words have been obtained from the popular stop words vocabulary set (Stopwords, 2014). The results obtained from these experiments have been shown in the above Table 2. We have observed that results were perfectly enhanced. For example, as shown in the above Table 2 the result of 698 papers have been raised from 64 to 92%. The other results have also significantly improved.

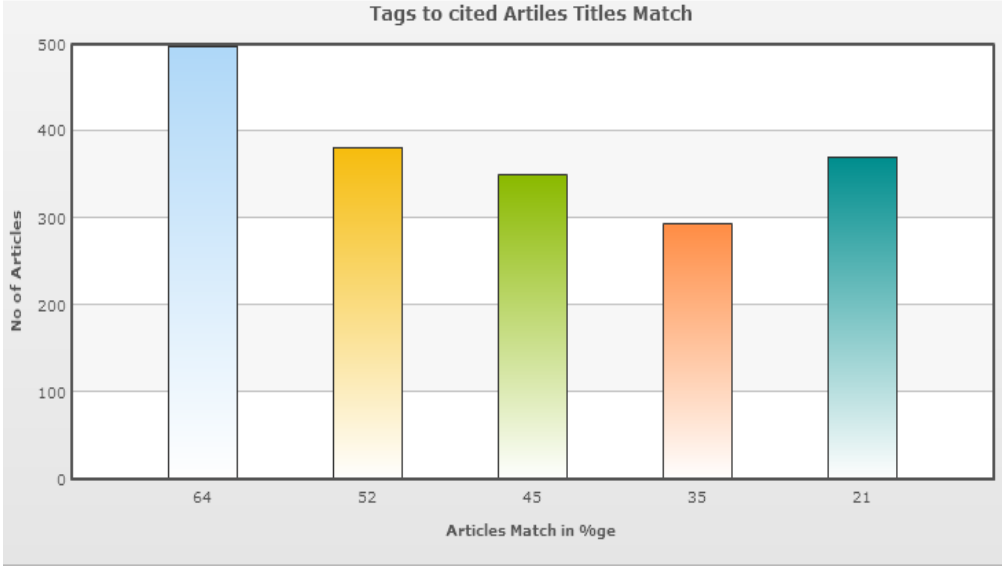

Fig. 6: Results of the tags matching in the cited articles titles

In addition to comparing the number of counts of tags with the total number of citation that a paper has received, we also have compared the diffusion of tags in the cited papers' title. This has proved that the tag a citing paper has received is also used in the future cited article of this study. This comparison has been performed in two dimensions. On one side we have performed the comparison of all tags with the full title of the articles including the stop words (the results of this process is shown in the Fig. 6 given above). Furthermore, we have removed all the stop words from the tags as well as from the title of the articles and the results obtained from this setup are given in the Fig. 6 above.

CONCLUSION AND RECOMMENDATIONS

Web is a huge source of information. It is hectic to reach the exact information. Knowledge is scattered and requires linkage. Citations have been source of liking the knowledge. Due to its inefficiency, efforts had been made to find alternative that can trace and link the knowledge. Social bookmarking has emerged as a powerful tool to include the reviews and views of users. Keywords and tags assigned by a user to a research paper can be used to trace the flow of knowledge in different domains. Tags are real time scenario that are highly dynamic and include the sentiments of a user. We made an effort to find semantic relationship and correlation between the tags and citations. We found that there are more that 70% papers that have tags or keywords in the titles of future citations. This trend is quite encouraging and favorable to opt to replace citations as a means to link knowledge transfer with tags. Tags appear in no time for a research paper and there are no formalities involved in assigning a tag to a research paper. However, there exists risk of tag assignments by non-serious users. Such vague tags may affect the objective of this research. In order to ensure that tags are assigned by the serious users, we selected users from citeULike. This site is used by those who are serious users and intend to use it for research purposes.

This research work has also opened some of our future directions. We will focus on research papers available in DBLP++ dataset to verify the validity and accuracy of this framework. We will find tags and citations for these research papers from CiteULike and Google Scholar. The results from the experiments on this dataset will further strength our idea of using the research article's tags as alternative of the citation for the ranking of the papers.

REFERENCES

Carmagnola, F., F. Cena, O. Cortassa, C. Gena and I. Torre, 2007. Towards a tag-based user model: How can user model benefit from tags? In: Conati, C., K. McCoy and G. Paliouras (Eds.), Proceedings of User Modeling 2007. Lecture Notes in Computer Science. Corfu, Greece, 4511: 445-449.

Carpenter, M.P., M. Cooper and F. Narin, 1980. Linkage between basic research literature and patents. Res. Manage., 23: 30-35.

Chen, C., W. Zhu, B. Tomaszewski and A. MacEachren, 2007. Tracing conceptual and geospatial diffusion of knowledge. In: Schuler, D. (Ed.), Online Communities and Social Comput. Proceedings of HCII LNCS 4564, Springer-Verlag, Berlin, Heidelberg, pp: 265-274.

Hotho, A., J. Robert, C. Schmitz and G. Stumme, 2006. Trend detection in folksonomies. Lect. Notes Comput. Sc., 4306: 56-70.

James, S., 2004. The Wisdom of Crowds. Doubleday, May 2004.

Marlow, C., M. Naaman, M. Boyd and M. Davis, 2006. HT06, tagging paper, taxonomy, flickr, academic article, to read. Proceeding of the 17th Conference on Hypertext and Hypermedia. in HT, Odense, Denmark, pp: 31-40.

Paul, H., G. Koutrika and H.G. Molin, 2008. Can socialbookmarking improve web search? Proceedings of the International Conference on Web Search and Data Mining (WSDM, 2008), pp: 195-206.

Peter, M., 2007. Ontologies are us: A unified model of social networks and semantics. J. Web Semant. Sci. Serv. Agents, 5(1).

Pierpaolo, B., D. Gendarmi, F. Lanubile and G. Semeraro, 2007. Recommending smart tags in a social bookmarking system. Proceedings of Bridging the Gep between Semantic Web and Web.

Saeed, A., M.T. Afzal, A. Latif, A. Stocker and K. Tochtermann, 2008a. Does tagging indicate knowledge diffusion? An exploratory case study. Proceeding of 3rd International Conferences on Convergence and Hybrid Information Technology, pp: 605-610.

Saeed, A., M.T. Afzal, A. Latif and K. Tochtermann, 2008b. Citation rank prediction based on bookmark counts: Exploratory case study of www06 papers. Proceedings of the 12th IEEE International Multitopic Conference, pp: 392-397.

Saeed, A., M.T. Afzal, A. Latif and K. Tochtermann, 2010. Disseminating knowledge through tags: Recommending tags for scienctific resources. J. IT Asia, 3(2010): 25-36.

Stopwords, 2014. Retreieved form: http://www.ranks.nl/resources/stopwords.html.

Wetzker, R., C. Zimmermann and C. Bauckhage, 2008. Analyzing social bookmarking systems: A del.icio.us cookbook. Proceedings of the ECAI Mining Social Data Workshop, pp: 26-30.

Wu, H., M. Zubair and K. Maly, 2006. Harvesting socail knowledge from folksonomies. Proceeding of ACM 17th Conference on Hypertext and Hypermedia. Odense, Denmark, pp: 111-114.

Yanbe, Y., A. Jatowt, S. Nakamura and K. Tanaka, 2007. Can social bookmarking enhance search in the web? Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, pp: 107-116.