International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:6, 2011

A Combination of Similarity Ranking and Time for Social Research Paper Searching

P. Jomsri

***Abstract*—**Nowadays social media are important tools for web resource discovery. The performance and capabilities of web searches are vital, especially search results from social research paper bookmarking. This paper proposes a new algorithm for ranking method that is a combination of similarity ranking with paper posted time or *CSTRank*. The paper posted time is static ranking for improving search results. For this particular study, the paper posted time is combined with similarity ranking to produce a better ranking than other methods such as similarity ranking or *SimRank*. The retrieval performance of combination rankings is evaluated using mean values of NDCG. The evaluation in the experiments implies that the chosen *CSTRank* ranking by using weight score at ratio 90:10 can improve the efficiency of research paper searching on social bookmarking websites.

***Keywords*—**combination ranking, information retrieval, time, similarity ranking, static ranking, weight score

I. Introduction

THE world wide web has become increasingly. Especially, Social resource sharing systems are web-based systems that allow users to upload all kinds of resources. The systems can be distinguished according to what kind of resources are supported such as Flickr [4] is social resource sharing system for photos while Delicious[5] is social resource sharing system for web page. One reason for their immediate success is the fact that no specific skills are needed for participating, and that these tools yield immediate benefit for the individual user without too much overhead.Furthermore, the social resource sharing systems for academic research paper are vital. Search engines on user social bookmarking systems are therefore increasingly popular alongside the functions that allow users to share content with one another, such as *CiteULike*[1]. This search engine helps scientists, researchers and academics store, organize, share and discover links to academic research papers. *Connotea*[2] is a free online reference management for all researchers, clinicians and scientists. *BibSonomy*[3] is a system for sharing bookmarks, lists of literature and BIBTEX based publication entries simultaneously. However, the best known in the academic and research paper arena is CiteULike.Within the information retrieval community, there has been considered an alternative approach for retrieving information based on the community of users in the system. Many social bookmarking systems have been designed to work using similarity ranking. Similarity ranking has focused on improving the order of search results returned to users by measuring the match between query terms and the content of the web resource. Various approaches for ranking the results have been studied such as the TTA engine resulted from our previous investigation using search engines with four different indexing methods, namely 1) tag, 2) title, abstract 3) tag, title and abstract and 4) CiteULike, were evaluated [6],[7]. The results suggested that the search engine using TTA performs the best. Furthermore, static ranking is important for a search engine in measuring the quality of the web documents for providing numerous benefits search results.In this paper, we propose to combine search results between similarity ranking from search engines which created indexes using “tag, title and abstract” (TTA) with static ranking of paper posted time because recent paper may be the subject interesting. This method is called *CSTRank*. The data sets were collected from the native academic social bookmarking of CiteULike. Discovering how to improve the capability of these search engines will help researchers to develop search engines that meet with users’ requirements the most. The paper is organized as follows. Section II discusses related works. The framework of this paper is described in Section III. The *CST algorithm* explained in Section IV, The experimental setting is shown in Section V. Results and discussions from the experiments are presented in Section VI. Finally, the conclusion and future work are given in Section VII.

II. RELATED WORK

Many previous works related to research paper searching focus on improving the efficiency of academic web resource searching. Researchers who studied in research paper searching such as CiteULike: Capocci and Caldarelli [8] analyzed the small-world properties of the CiteULike folksonomy. Toine Bogers and Van den Bosch [10] employed CiteULike to generate reading lists for scientific articles based on the user’s online reference library. They applied three different CF algorithms and found that user-based filtering performs the best. Santos-Neto, Ripeanu, and Iamnitchi [11] explored three main directions for presenting characterizations of CiteULike and Bibsonomy that target the management of scientific literature. The technique from CiteULike has been applied to other academic search such as Farooq et al. presented four novel implications for designing the CiteSeer [9],[14]. Jomsri, Sanguansintukul, and Choochaiwattana [6], [7] create three heuristic indexers: “tag”(T), “title, abstract”(TA) , “tag, title and abstract”(TTA) and compare

P. Jomsri is with the Faculty of science and technology, Suan Sunandha Rajabhat University, Dusit, Bangkok 10300 Thailand (phone: +6602-160-1111; e-mail: pijitra.jo@ssru.ac.th, pijitra_jom@hotmail.com).

with CiteULike. Experiment found that TTA is the best indexer.

There are currently two major categories of ranking algorithms based on similarity ranking (query-dependent ranking) and static ranking (query-independent ranking): In classical Information Retrieval [26], the system works to find documents corresponding to the user query. Information retrieval algorithms usually work based on matching words in documents. In other words, for each query the documents with the more similar content to the query will be selected as the more relevant ones. Examples of the content based ranking algorithms are TF-IDF [22]. Sun and Lee Giles [17] proposed popularity weighted ranking algorithm for academic digital libraries that uses the popularity factor of a publication venue to overcome the limitations of impact factors compare with PageRank. The algorithm is also evaluated by click through data from CiteSeer usage logs. Carmel et al. [21] propose a novel framework for social bookmark weighting which allows us to estimate the effectiveness of each of the bookmarks individually for several IR tasks. They show that by weighting bookmarks according to their estimated quality, they can significantly improve social search effectiveness.Several works has applied static ranking and combined both similarity and static ranking for improved search results. Heymann et al. [18] measured the document popularity according to the number of times it was bookmarked. A similar measurement was proposed by Yanbe et al. [19].Hotho et al. [20] proposed *FolkRank*, a PageRank-like measure that is devised for multi-entity graphs. *FolkRank* mutually reinforces entities that have many relationships with other (important) entities. A document that is bookmarked by "important" tags or by "important" users (judged by their *FolkRank* scores) is considered as more important. Bao et al. [15] proposed another version of a PageRank like measure,termed *SocialPageRank*. They showed that the combination of both textual similarity and SocialPageRank scores gains on search effectiveness.Craswell et.al. [13] presented a new query independent feature based on this applying sigmoid transformations to PageRank, indegree, URL Length and ClickDistance. Mohammad Zareh et.al [16] improved the A3CRank method based on the content, connectivity, and click-through. A3CRank outperforms other combinational ranking algorithms such as Ranking SVM in terms of P@n and NDCG metrics. However Dou et al. [23] worked on using click-through data directly for personalization. The utility of personalization is highly dependent on the ambiguity of the query. If the query is highly specific (unambiguous) then the personalization is likely to have a negative effect on the results. Some researchers applied the feature of time to improve ranking such as Berberichl, Vazirgiannis, and Weikum [12] who introduced T-Rank, a link analysis approach that takes into account the temporal aspects of freshness such as timestamps of most recent updates and activity such as update rates of pages and links. T-Rank results show that it can improve the quality of ranking web pages.This paper uses different views to rank search results of research paper searching with focus on the diversity and reliability. We extend the method of TTA indexing to create ranking based on paper post time.

III. Modification of A Framework for Social Research Paper

A modified framework for the research paper search engine is illustrated in this section. From our previous work [4] show a framework for social research paper bookmarking and in this work try to modify these by combine static ranking that is paper posted time to improve search result. The original framework is divided into a two part system: research paper bookmarking and a research paper searching. In this paper create the modified framework for develops into a part of a research paper searching.

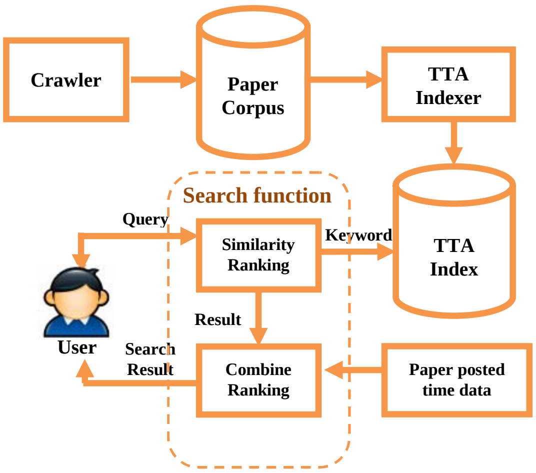

Fig. 1 A modified framework for research paper searching

A. Units Research paper bookmarking

A research paper bookmarking system provides users with new ways to share their research interests, such as with CiteULike. They can automatically share all their public entries with other users and comment on other papers. They can also discover interesting papers posed by other users who share the same interests. This kind of system allows users to create their own keywords for attaching to the posted papers. These keywords are known as tags. All public references can also be searched and filtered by tag. In addition, the site provides groups that users can join themselves or by invitation. Research paper bookmarking gives access to personal or shared bibliographies directly from the web. It allows seeing what other people posted. Also it is possible to browse the public libraries of people or group with similar interest to discover interesting papers.

B. Research paper searching

This paper concentrates on improving research paper searching. The modified framework for research paper searching is showed in Fig.1 and was described into four steps:

1) Crawler: A research paper crawler is a small computer program that browses directly to the paper sharing systems of the WWW in a predetermined manner. The research

paper crawler is responsible for gathering research paper information such as paper author, tags used, posted time, year, priority of paper, groups of similarity papers, etc. This useful information helps the system to determine the user's interests and also helps the system to create indexes for each paper. Java programming is used to implement a crawler on this framework.

2) Paper corpus: Paper corpus is a collection of research papers extracted from the research paper sharing system.

3) TTA Indexer and TTA index: TF-IDF (term frequency–inverse document frequency) will be used for creating indices. TF-IDF is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Jomsri, Sanguansintukul, and Choochaiwattana showed that research paper information like "tag, title and abstract" could be a useful source for creating indices for research papers [6],[7].

4) Search Function: Cosine similarity is a similarity measurement between two vectors of *n* dimensions. This involves finding the cosine of the angle between two vectors. This measurement is often used to compare documents in text mining. Two types ranking are similarity ranking and combine ranking were described as follows:

Similarity ranking: To compare a query with the research paper index, a cosine similarity measurement is used to retrieve and rank search results. The similarity score of query *q* for document *d* is defined as in equation (1).

$$score(q,d) = \sum_{t \in q} \left(tf(t \in d) \times idf(t)^2 \times B_q \times B_d \times L \right) \times C \quad (1)$$

Where $B_q = getBoost(t \;\; field\; in\; q)$

$B_d = getBoost(t \;\; field \;\; in \;\; d)$

$L = lenghtNorm(t \;\; field \;\; in \;\; d)$

$C = coord(q,d) \times queryNorm(s)$

Where B_q and B_d is the field boost and which is set during indexing. *L* is the normalization value of a field, given the number of terms with the field; *C* is a value from coordination factor, based on the number of query terms the document contains multiplied with the normalization value for a query, given the sum of the squared weights of each of query term. Note that getBoost is a function in Lucene [25], which is used to generate indexes for the experiments.

Combine ranking: is the combination of similarity ranking with the paper posted time factors. The detail will be discussed in Section IV.

IV. Ranking Method

This section describes a heuristic method for creating static ranking and combined similarity ranking with static ranking. The important static ranking factor is paper posted time. Similarity ranking filters on the high similarity score. Therefore, this paper proposes to combine the advantages of similarity ranking with static ranking from the paper posted time score. Also, the score value of the combined two methods is adjusted to be in the range of 0 to 1. This section is divided into two parts: 1) describe the detail of property factor, and 2) describe how to combine a similarity ranking with paper posted time.

A. Paper Posted Time

Posted time information for each paper is composed of the posted date and posted time, i.e. "2010-03-15 17:02:45". Firstly, the paper is sorted based on this information. Then, the posted time score (*T*) is calculated by the formula in Equation (2). T_r is the score of current rank.

$$T_r = T_{r-1} - 0.05 \quad (2)$$

Where, *r* =0, 1, 2, …, 19. The original value is T_0 =1.

B. Combining similarity ranking with paper posted time

CSTRank score use both similarity ranking (SimRank) and score of paper posted time (TimeRank). In addition, the weighting score is applied for each type of rank to find the optimal ranking. Equation (3) shows the combining similarity ranking with paper posted time. *Let* ω_c is combine weighting score.

$$CSTRank = (SimRank \times \omega_c) + (TimeRank \times (1 - \omega_c)) \quad (3)$$

V. Experimental Setting

The experimental setting is divided into three sections. Section *A)* describes the data set, section *B)* discusses research paper search engine settings and section *C)* describes evaluation metrics.

A. The data set

The crawler collected data from CiteULike during March to May 2010. The collected documents consist of 64,320 research papers. There are groups that are related to the computer science field. Each record in the paper corpus contains: title ID, title name, abstract, tag of each paper, and link for viewing full text article, book title within which the paper was published, posted date, posted time and paper priority.

B. Research paper searching setting

This section describes a methodology for heuristic indexer and ranking methods.

1) Indexer

In the experiments, an indexer was developed. Equation (4) shows a modified Term Frequency/Inverse Document Frequency (TF/IDF) formula for the indexer. Here, TTA corresponds to tag, title with abstract:

$$tfidf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|TTA|}{|\{d : t_i \in d\}|} \quad (4)$$

Where $n_{i,j}$ is the number of occurrences of the considered term in document d_j . | *TTA* | is the total number of documents in the corpus. $\left|\{d : t_i \in d\}\right|$ is the number of documents where the term t_i appears (that is $n_{i,j} \neq 0$). If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to use $1+\left|\{d : t_i \in d\}\right|$.

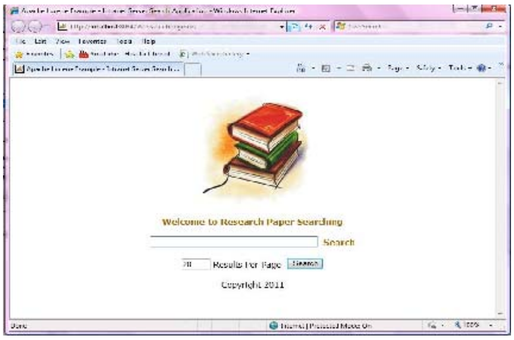
Welcome to Research Paper Searching
Search
20
Results Per Page
Search
Copyright 2011
Done
Internet | Protected Mode On
100%

Fig.2 Research Paper Searching web page.

We developed search engines based on the TTA indexer. Equation (4) is applied to the first search engine for creating the index. Fig. 2 shows an example of the interface web page developed in the experiment. Here, the subject can specify their search criteria and investigate the results from each search engine. The number of the results per page can also be defined. In addition, the subject can view the results by title, abstract and the full text. Seventy five queries were asked from fifteen subjects, who are considered as experts in the field, during their participation in the experiment. Therefore, their relevancy ratings are assumed to be perfect for each query. The top 15 search results for each search engine were displayed for relevancy judgment. Subjects can see the titleID of the document, title name for linking to the download of the full paper and the link to get information from CiteULike. However, the specific sources of results obtained from each search engine are hidden from the subjects.

2) *Ranking*

In this experiment, two type of ranking are developed: Similarity ranking (*SimRank*), and Combination Similarity and paper posted time (*CSTRank*). It is interesting to measure and compare the performance of these two rankings:

2.1) SimRank

This model applied similarity ranking based on the TTA indexing method. The formula appears in Equation (1).

2.2) CSTRank

Both *SimRank* and year of publication are applied in this step. Equation (5) shows the *CSTRank* score. In the experiment, five different weight values are chosen for the performance evaluation, where {ω_c = 0.5, 0.8, and 0.9}.

The value of 0.9 means that the combination of similarity and static rank in 90:10 ratio. The value of 0.80 denotes 80:20 ratios between similarity and static rank. Equation (5) shows examples of weight ratio 50:50 ratios (ω_c =0.5).

$$CSTRank = (SimRank \times 0.5)+(TimeRank \times (1\text{-}0.5)) \qquad (5)$$

3) *Relevancy setting*

In the study setting, each subject is assigned to investigate the research papers obtained from the search engines. Each subject specified three different queries. Each query is applied with all rankings. The first 20 documents for each search engine for the relevancy are displayed. Finally, the subjects were asked to rate the relevancy of the search results on a five-point scale: Score 0 is not relevant at all, Score 1 is probably not relevant, Score 2 is less relevant, Score 3 is probably relevant and Score 4 is extremely relevant.

C. *Evaluation Matrix*

NDCG (Normalized Discounted Cumulative Gain) as originally proposed by Jarvelin and Kekalainen [24], was used to evaluate the performance of each search engine. This metric is a retrieval measurement devised specifically for web search evaluation. The NDCG is computed as in the equation (6).

$$NDCG_q = M_q \sum_{j=1}^{k} \frac{\left(2^{r(j)}-1\right)}{\log(1+j)} \qquad (6)$$

Where *k* is a truncation or threshold level, *r(j)* is an integer representing the relevancy given by the subject, and M_q is a normalization constant calculated so that the perfect ordering would obtain a NDCG of 1. NDCG rewards relevant documents appearing in the top ranked search results and punishes irrelevant document by reducing their contributions to NDCG.

VI. Result and Discussion

This section separate in to two parts: first is results from the experiment and the second is the discussion.

A. *Results*

The results of the average NDCG score of the first 15 rank of *SimRank*, *CSTRank(50:50)*, *CSTRank(80:20)* and *CSTRank(90:10)* are shown in Table I.

Fig.3 shows the NDCG average score of four different rankings: *SimRank* and *CSTRank* with 4 different weights. The x-axis represents the first 15 documents of the search results, whereas the y-axis denotes the NDCG score.

The result from this figure suggests that *CSTRank(90:10)* seems to outperform other ranking methods.

TABLE I
AVERAGE OF THE NDCG SCORES FOR THE FIRST 15 RANKS OF FOUR DIFFERENT RANKINGS

K	Average of NDCG Score			
	SimRank	*CSTRank (50:50)*	*CSTRank (80:20)*	*CSTRank (90:10)*
1	0.72021	0.59746	0.70392	0.72233
2	0.66138	0.56049	0.65077	0.67647
3	0.64477	0.54234	0.63415	0.65470
4	0.63825	0.54540	0.63382	0.64224
5	0.63099	0.54482	0.63551	0.63489
6	0.63555	0.54995	0.62851	0.64220
7	0.63694	0.55682	0.62580	0.63881
8	0.63473	0.56350	0.62589	0.64031
9	0.64013	0.57243	0.63201	0.64290
10	0.64386	0.58090	0.63437	0.64568
11	0.64686	0.59329	0.64150	0.64826
12	0.65111	0.60283	0.64585	0.65708
13	0.65530	0.61645	0.65551	0.66437
14	0.66310	0.62560	0.66481	0.67295
15	0.67251	0.63828	0.67771	0.68472

K = ranking order, *SimRank* = similarity ranking, *CSTRank* = combination of similarity with paper posted time.

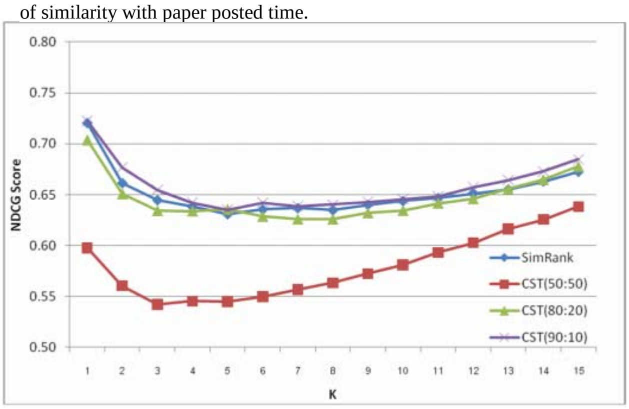

Fig. 3 Comparison of the NDCG average score of four ranking

Furthermore, we applied One Way ANOVA on NDCG for the top fifteen rank (K=1-15) to test whether there is a difference among the NDCG mean of the four different rankings. We found the evidence that not all of the means of NDCG of the four ranking are equal at α=0.05 levels of significance. In other words, the difference in the set of search results returned from four different ranking approaches were statistically significant. The multiple comparisons were used to find the differences among the four rankings method. Table II shows the result of multiple comparisons of four different rankings. The results from the multiple comparisons indicate that a set of mean difference search results provided by the *CSTRank (90:10)* combine weight ranking approach is statistically different from the set of search results provided by the *CSTRank (50:50)* at k=1-15. The *CSTRank (90:10)* is not statistically difference from the set of search results provided by the *SimRank* and *CSTRank (80:20)* approach. However, the mean difference value of *CSTRank (90:10)* is highest than the mean difference value of *SimRank* and *CSTRank (80:20).*

B. Discussion

There are some indications that results from the proposed heuristic ranking method, *CSTRank,* can improve research paper searching on social bookmarking. This might be because the method utilizes the information of user behavior. Especially, *CSTRank (90:10)* ,a combination of the similarity ranking 90 % and static ranking from paper posted time 10%, seems to outperform other weight ratios. In the study, a factor is considered: paper posted time. We observed that many researchers prefer to read more recent papers. However, the content of the paper, which is "tag, title and abstract" or TTA for this particular study is still important. Finally, the chosen experimental factor can help the system to adjust the ranking and improve search results of research paper searching.

TABLE II
RESULT OF MULTIPLE COMPARISONS OF THE TOP FIFTEEN RANKS.

Rank	*Indexing*		*Mean Difference*	*Std.*	*Sig.*
(K)	*(I)*	*(J)*	*(I-J)*	*Error*	*(2-tailed)*
1-15	*CST-Rank (90:10)*	*SimRank*	0.0062	0.00858	0.887
		CST-Rank (50:50)	0.0942	0.00858	0.000
		CST-Rank (80:20)	0.0142	0.00858	0.345

VII. CONCLUSION AND FUTURE WORK

This paper focuses on the combination ranking method. Here, the heuristic ranking implemented was *CSTRank.* fifteen subjects are assigned to investigate the research papers obtained from the search engines base on TTA indexer. Each subject specified three different queries. The first 15 documents for each search engine for relevancy are displayed. Finally, the subjects were asked to rate the relevancy of the search results on a five-point scale.The results show that *CSTRank(90:10)* returns a higher NDCG score than other rankings. This implies that *CSTRank (90:10)* has a better performance than other ranking. In order to confirm the result of the experiment, additional experiments should be conducted such as adding additional factors to improve search result. Future research in the area consists of extending the personalization; creating user profiles on research paper searching.

REFERENCES

[1] CiteULike, http://www.CiteULike.org

[2] Connotea, http://www.connotea.org

[3] BibSonomy, http://www.bibsonomy.org

[4] Flickr, http://www.flickr.com

[5] Delicious, http://www.delicious.com

[6] P.Jomsri, S. Sanguansintukul, W. Choochaiwattana, "Improve Research paper Searching with social tagging-A Preliminary Investigation," in *the Eight International Symposium on Natural Language Processing*, Thailand, 2009,pp.152-156.

[7] P.Jomsri, S. Sanguansintukul, W. Choochaiwattana, "A Comparison of Search Engine Using "Tag Title and Abstract" with CiteULike – An Initial Evaluation," *in the 4th IEEE Int. Conf. for Internet Technology and Secured Transactions (ICITST-2009),*United Kingdom,2009.

[8] A. Capocci, and G.Caldarelli, "Folksonomies and Clustering in the Collaborative System CiteULike," *arXiv Press*, *eprint No. 0710.2835*, 2007.

[9] U. Farooq, T.G. Kannampallil, Y. Song, C.H. Ganoe, M.C., John, L. Giles, "Evalating Tagging Behavior in Social Bookmarking Systems: Metrics and design heuristics," in *Proc. of the 2007 international ACM conference on Supporting group work (GROUP'07)*, Sanibel Island, Florida, USA, 2007,pp.351-360.

[10] T. Bogers, and A. van den Bosch, "Recommending Scientific Articles Using CiteULike," in *Proc. of the 2008 ACM conference on Recommender systems(RecSys'08),* Switzerland,2008 ,pp.287-290.

[11] E. Santos-Neto, M. Ripeanu, and A. Iamnitchi, " Tracking usage in collaborative tagging communities".

[12] K. Berberich, M. Vazirgiannis, , and G. Weikum, " T-Rank: Time-Aware Authority Ranking," in *WAW 2004*.

[13] N. Craswell ,S. Robertson, H. Zaragoza, ,and M. Taylor, "Relevance weighting for query independent evidence," in *Proc.of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, Salvador, Brazil,2005.

[14] U. Farooq, C.H. Ganoe, , J.M. Carroll, and C.L. Giles, "Supporting distributed scientific collaboration: Implications fordesigning the CiteSeer collaborator," in *IEEE Proc. of the Hawaii Int'l Conference on System Sciences*, Waikoloa, Hawaii,2007.

[15] S. Bao, X. Wu, B. Fei, G. Z. Xue, and Y. Yu, "Optimizing Web Search Using Social Annotations," in *Proc. of the 16th international conference on World Wide Web* (www2007), New York, USA,2007.

[16] A. Mohammad Zareh Bidoki , P. Ghodsnia, N. Yazdani, and F. Oroumchian, "A3CRank: An adaptive ranking method based on connectivity, content and click-through data", *J. Information Processing and Management*, Vol. 46, pp.159—169, 2010.

[17] Y. Sun, and C. Lee Giles, "Popularity Weighted Ranking for Academic Digital Libraries," in *ECIR 2007,* LNCS 4425, pp. 605-612, 2007.

[18] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Can social bookmarking improve web search?," in *WSDM '08: Proc. of the International Conference on Web Search and Web Data Mining*, New York, NY, USA. 2008,pp. 195–206.

[19] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, "Information retrieval in folksonomies: search and ranking," in *The Semantic Web: Research and Applications*, ,2006 ,pp. 411–426.

[20] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka, "Can social bookmarking enhance search in the web?," in *JCDL '07: Proc.* of *the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY, USA ,2007 ,pp. 107–116,.

[21] D. Carmel , H. Roitman, and E. Yom-Tov, "Social bookmark weighting for search and recommendation," in *The VLDB J.*, vol.19,pp 761–775, December 2010.

[22] M. Richardson, A. Prakash, and E. Brill, "Beyond PageRank: Machine Learning for Static Ranking," in *Proc. of the 15th international conference on World Wide Web***,** Edinburgh, Scotland (2006).

[23] Z. Dou, R. Song, and J.-R. Wen, "A large-scale evaluation and analysis of personalized search strategies," in *Proc. of the 16th international conference on World Wide Web*, 2007.

[24] K. Jarvelin, , and J. Kekalainen, "IR evaluation methods for retrieving highly relevant documents," in *Proc. of the International World Wide Web* ,2006.

[25] E. Hatcher, and O. Gospodnetic, "Lucene in Action," Manning Publications Co., United States of America 2006.

[26] R. Baeza-Yates, and B. Ribeiro-Neto, "Modern information retrieval," ACM Press/Addison-Wesley.