An Improved Framework for Tag-Based Academic Information Sharing and Recommendation System

Jyoti Gautam, Ela Kumar

Abstract— **The Internet and the World Wide Web provides methods of storing and sharing information, especially in academic fields. CiteULike is a community-based research paper sharing system, which is popular among researchers. This paper proposes a framework for a tag-based Academic Information Sharing and Recommender System which shares information such as question papers, assignments, tutorials and quizzes on a specific area. The approach is based on the set of tags for recommending academic information to each user on the area of his choice. User self-defined tags could be attached to academic information and it can be used further to calculate tag score. The improved index (based on the improved TFIDF algorithm) will be used further to calculate tag score to give the optimised results. This Sharing and Recommender System would be of great significance to the Academic Community.**

Index Terms— **academic information, improved TFIDF (term frequency inverse document frequency), recommender system, sharing system, tag-based**

I. Introduction

Search engine comes into being and continues to grow and develop in order to let people be able to get information from the web easily. People have generated various kinds of ranking algorithms and try to give user an optimised result list. However, because of the simpler expression format of the web information and user queries, there is less relevance between user queries and web information. The success and popularity of social network systems, such as del.icio.us, Face book, etc., have generated many interesting problems to the research community. This provides us with a new perspective on how to improve the quality of information retrieval.

Many popular Web services like Delicious and flickr.com rely on folksonomies (social-tagging) [4]. Research on folksonomies [13] is still at an early stage in spite of the rising popularity of these Web services. The focused areas have been on the study of the data properties, the analysis of usage patterns of tagging systems [15], the discovery of hidden semantics in tags, the using of annotations in enterprise search, and the user's interest in discovery for personalized search. So, one area that arises is to consider utilizing the semantic tag information with web page.

Recently, social-tagging has been widely adopted by various social bookmarking systems. Various systems have been evaluated for their tagging behaviour. Different metrics have been designed for the tagging behaviour and evaluated. These systems provide functions that allow users to share content with one another. CiteULike [28] is a search engine for academia that tries to give the best results for research papers and literature reviews. Scientists, researchers and academicians are able to store, organize, share and discover links to academic research papers.

Another area that arises is an improved method of term weighting for text classification. [8] Traditional term weighting approach is TFIDF (Term Frequency Inverse Document Frequency) that calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. The improved method is based on the fact that low frequency terms are important whereas high frequency terms are unimportant, so it designs higher weights to the rare terms frequently [22].

In the Information Retrieval community, we can think of forming community of users in the system. These communities can be of the users interested in different areas. In academics, these communities could be like computers, electronics, electrical, mechanical etc.

Many Recommender Systems have been designed and implemented for various items including newspaper, research papers and emails. The goal is to increase the accuracy of the recommendations. So, in order to improve the accuracy of recommendations, an improved hybridised approach has been suggested.

Presently, there are works concerning [4] collaborative filtering on tags to recommend personalized information to users. There are works on utilizing tagging information for assisting users to find interesting items easily, quickly and efficiently for academic papers. So, in order to continue with the previous work, an improved tag-based Academic Information Sharing and Recommender System has been proposed.

Manuscript received January 27, 2012; revised March 9, 2012.

Jyoti Gautam is with the Department of Computer Science and Engineering, JSS Academy of Technical Education, C-20/1, Sector- 62, NOIDA, Uttar Pradesh, India (phone: +919810416894; fax:01202400097; e-mail: jyotig@jssaten.ac.in).

Ela Kumar is with the School of Information and Communications Technology, Gautam Buddha University, Greater Noida, Uttar Pradesh, India (phone:+919873426162; e-mail: ela_kumar@gbu.ac.in).

ISBN: 978-988-19252-1-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

II. RELATED WORK

This section portrays the background of the framework. The section is divided into two parts: a community-based search engine and social bookmarking.

A. Community-Based Search Engine

In the year 2010[9], Pijitra Jomsri, Siripun Sanguansintukul, Worasit Choochaiwattana suggested a framework for tag-based research paper recommender system. It consisted of two modules. (a) Research paper sharing system – This system provides users with new ways to share their research interests. They can post and comment on papers. They can also discover interesting papers posted by other users who share the same interests. The system uses a tag which is a keyword to get attached to the paper. Tags are user-defined terms for a paper. The tags can be used to create a profile for each user. A recommender mechanism could take advantage of this created user profile in research paper recommendations. (b) Research paper recommender system – It analyses user's posted papers and post histories to extract user preference from posting patterns and recommend papers based on user preferences.

In the year 2006, Mislove [1] suggested a Web search framework enhanced by social networks, and studied the mechanisms for content publishing and location in social networks, a framework that led to considerable improvements in effectiveness.

In the year 2006, Xian Wu, Lei Zhang, Yong Yu [21] explored the technique of Social Annotations for the Semantic Web. These annotations are manually made by normal web users without a predefined formal ontology. Compared with the formal annotations, despite social annotations are coarse-grained, informal and vague, they are also more accessible to more people and better reflect the web resources' meaning from the user's point of views during their actual usage of the web resources. As an example of social bookmark service, it can be shown how emergent semantics can be statistically derived from the social annotations. Furthermore, the emergent semantics can be applied to discover and search shared web bookmarks. The evaluation of the approach shows that our method can effectively discover semantically related web bookmarks that current social bookmark service cannot discover easily. The method can be further improved by modeling to support incremental analysis of the social annotations data.

In the year 2007, Beydoun [7] presented a "semantic annotation approach" to support search in a social network.

Umer Farooq, Thomas G. Kannampallil, Yang Song [20] used six tag metrics – tag growth, tag reuse, tag non-obviousness, tag discrimination, tag frequency, and tag patterns in order to evaluate tagging behavior of a social bookmarking system in the year 2007. The tagging behaviour cannot be evaluated alone through these tag metrics. Some hybrid metrics can be developed for more exhaustive measurement schemes for tags and tagging behavior.

In the year 2008, Suchanek [6], who analyzed tag and found that tags are meaningful where the tagging process is influenced by tag suggestions.

Noel [14], in the year 2008, analyzed the tagging behavior of people who were describing four frequently entered references.

In the year 2008, S. Xu, Shenghua Bao, Ben Fei, Zhong Su [13] proposed a personalized search framework to utilize folksonomy for personalized search. Specifically, three properties of folksonomy, namely the categorization, keyword, and structure property are explored. The personalized search framework uses Weighted Borda-Fuse as the rank aggregation approach. More sophisticated rank aggregation methods can be developed to test the personalized search framework.

In 2010, Denis Parra-santander, Peter Brusilovsky[4] provided enhancements of user-based collaborative filtering algorithms to provide recommendations of articles on CiteULike, a social tagging service for scientific articles. Both the enhancements provided were beneficial. An improved precision was provided by incorporating the number of raters into the algorithms. BM25 similarity measure providing an alternative to Pearson correlation for calculating the similarity between users and their neighbours increases the coverage of the recommendation process. Scalability is a deficiency of the algorithm. The design of algorithms can be done to calculate local ranking of users and items around their network.

Toine Bogers [19], in the year 2008, 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.

B. Social Bookmarking

[26]Searches based on social-bookmarking have become increasingly popular, which lets users specify their keywords of interest, or tags on web resources. Social tagging, also known as social annotation or collaborative tagging is one of the major characteristics of Web 2.0. Social-tagging systems allow users to annotate resources with free-form tags. The resources can be of any type, such as Web pages (e.g., delicious), videos (e.g., YouTube), photographs (e.g., Flickr), academic papers (e.g., CiteULIke). A popular website in academia is CiteULike (www.CiteULike.org).[57] CiteUlike is a free service for managing and discovering scholarly references.

- Easily store references you find online
- Discover new articles and resources
- Automated article recommendations
- Share references with your peers
- Find out who's reading what you are reading
- Store and search your PDF's

CiteULike has a filing system based on tags. Tags provide an open, quick and user-defined classification model that can produce interesting new categorizations. Additionally, it is also capable to:

- 'tag' papers into categories
- Add your own comments on papers
- Allow others to see your library

ISBN: 978-988-19252-1-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

There are some websites such [27] as Delicious (online bookmarking), Flickr (online photo management and sharing application), Furl (File Uniform Resource Locators), Blinklist (links saver), Diigo (collect and organize anything e.g. bookmarks, highlights, notes, sceenshots etc.), Otavo (collaborative web search), Stumbleupon (discovery engine), Blummy (tool for quick access to favorite web services), and Folkd (saves bookmarks and links online) etc. which contain these tag information.

For example, Delicious is an online bookmark community site, all users can register free, establish their own accounts, set up and maintain bookmarks under their accounts. Of course, the users can share their bookmarks with each other on the network. The bookmarks refer to the words that users use to describe the page. For instance, for the page "http://www.google.com", some users describe it with the mark "search" whereas some describe it as "search engine". The bookmarks are all created by the users, which reflect the understanding and classification of web information from user's viewpoint. Therefore, these bookmarks can be considered as web tags.

Capocci [2], in the year 2007, analyzed the small-world properties of the CiteULike folksonomy.

In the year 2007, Santos-Neto[5] explored three main directions for presenting characterizations of CiteULike and Bibsonomy that target the management of scientific literature.

III. A FRAMEWORK FOR TAG-BASED ACADEMIC INFORMATION SHARING AND RECOMMENDER SYSTEM

The framework for tag-based academic information system is divided into two parts. The Framework is illustrated in Fig 1.

1. Academic Information Sharing System
2. Academic Information Recommender System

The system deals with academic information like question papers, assignments, tutorials and quizzes. These academic information will be grouped into communities of users depending on the subjects/area of interest .i.e. question papers, assignments, tutorials and quizzes of a particular subject fall into one community whereas of the other subject fall into the other one and so on.

1. Academic Information Sharing System

An academic information sharing system provides users with new methods to share their academic interests. They can post and comment on academic information. The users can also view information posted by other users of the same community. The users can create their own keywords for attaching to the academic information. These keywords are known as tags. Tags are defined as user-provided terms for the academic information. The tags can also be used to create a profile for each user. Academic Information Recommender System could take advantage of this created user profile in academic information recommendations.

2. Academic Information Recommender System

It consists of the following modules.

(a) Crawler: An academic information crawler is a computer program that browses directly to the academic information sharing systems of the WWW in a predetermined manner. The crawler is responsible for gathering academic information such as author, tags used, etc. related to the question papers, quizzes, assignments and tutorials in a specific community. This information helps the system to determine a user's interests and also helps the system to create an index for academic information. Java programming can be used to implement a crawler on this framework.

(b) Academic Information Corpus: In this corpus, we have a collection of information related to assignments, quizzes, tutorials and question papers on different areas. Different experts of the various communities can post and share their academic information.

(c) Indexer: Improved TFIDF (Term Frequency-Inverse Document Frequency) will be used for creating indices. TFIDF is a weight used in information retrieval and text mining. In the year 2009, Xin Hu, Hua Jiang, Ping Li and Shuyan Wang [46] proposed an improved method of term weighting for text classification. This approach simply thinks low frequency terms are important, high frequency terms are unimportant, so it designs higher weights to the rare terms frequently. The proposed method provides an effective term weighting approach to avoid the deficiency of the traditional approach, and makes use of kNN classifiers to classify over widely-used benchmark data set Reuters-21578.

The TFIDF approach gives equal weights to the different terms in the query text whereas this approach uses supervised term weighting approach. The proposed formula is defined as follows in "(1)".

$$W(t_k, d_j, c_i) = (1-\alpha).tfidf_{k,j} + \alpha.\ tfidf_{k,j}\ .A_i/C_i \qquad (1)$$

Where t_k refers to the term in the query, d_j refers to the document, c_i refers to the category, A indicates the number of documents belonging to category c_i where the term t_k occurs at least once and C denotes the number of documents belonging to category c_i where the term t_k does not occur at least once. Through the values of proportion of A and C, it can be shown that the three terms assign different discriminating power to text classification (TC).

α is called balance factor, usually,$0 \leq \alpha \leq 1$. There are two special cases.

When $\alpha = 0$, equation [1] becomes classic TFIDF approach, and when $\alpha = 1$, equation [1] becomes our newly improved approach. Using balance factor, we can get better classification results.

(d) Search function: Cosine similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them. The cosine of the angle between two vectors thus determines whether two vectors are pointing in roughly the same direction.

Given two vectors of attributes, A and B, the cosine similarity, Θ, is represented using a dot product and magnitude as "(2)".

$$\text{Similarity} = \cos(\Theta) = A.B/\|A\|\|B\| \qquad (2)$$

In the case of information retrieval, the cosine similarity of two documents will range from 0 to 1, since the term frequencies cannot be negative. The angle between two

ISBN: 978-988-19252-1-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

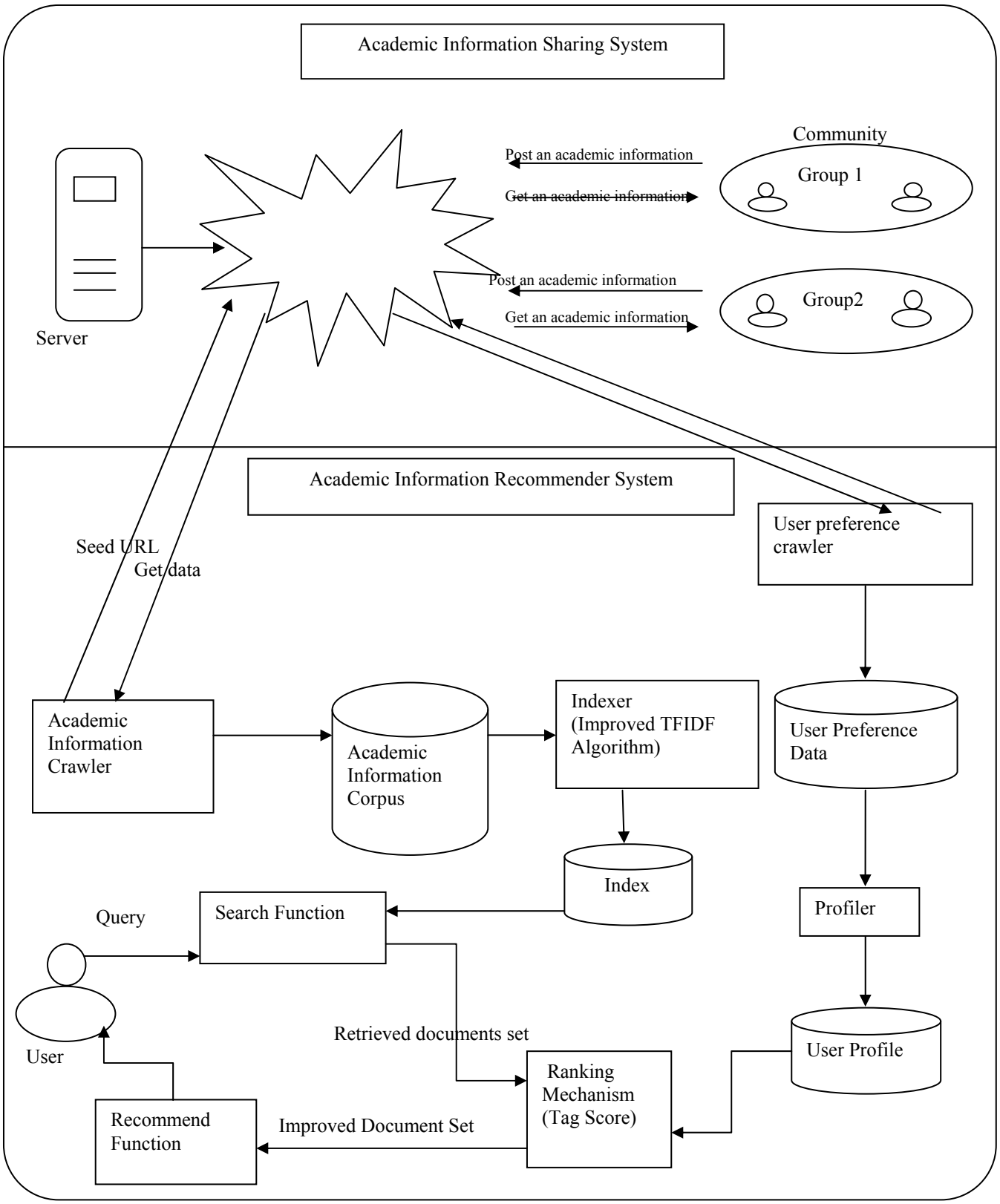

Fig. 1. A Framework for Tag-Based Academic Information Sharing and Recommender System

ISBN: 978-988-19252-1-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

frequency vectors cannot be greater than 90.

(e) Ranking: Once the search function has been used to retrieve the documents matching the query, social tagging has been used further to retrieve the set of documents which better match the query.

Let us assume that each user assigns a tag to each of the academic information (tutorials, quizzes, assignments and question papers). Now this tag makes a relationship between each user and the tag. This way, academic information gets a tag from all the users visiting it and a tag score can be calculated for academic information by calculating the scores from all the users visiting it.

For example, the tags are:

tag1= Outstanding (5),
tag2=Excellent (4),
tag3=Average (3),
tag4= Bad (2),
tag5= Very Bad (1).

So, the weighted score is like as "(3)".

$$\text{Tag Score: } \sum_{i=1}^{i=5} (\text{tag } i * (\text{weight})) / (\text{sum of weights of all the tags}) \quad (3)$$

Here, i can be from 1 to 5.

Where,tag1 is outstanding with the value 5
tag2 is excellent with the value 4
tag3 is average with the value 3 and like that.

Weight refers to the frequency of occurrence of a certain tag.

(f) User preference crawler and user preference data: This crawler is responsible for crawling user preference data, which detail the academic information posted by each individual user including a set of personally defined tags.

(g) Profiler: A profiler exploits the use of preference data in the recommender mechanism to suggest academic information that matches with user preferences.

(h) User profile: A collection of personal data related to a specific user. A profile refers to the precise digital representation of a person's identity. User profiles can be considered as the computer representation of a user model, delivering personalized information. The prototype of the system and preliminary results are presented.

(i) Recommender mechanism: The proposed method uses a set of tags and provides a ranking based on the search function and tag score to recommend academic information as shown in Fig 2.

Algorithm:

1. Analyse user's posted academic information along with the tags and calculate a tag score for academic information.
2. Using the seed URLs, create an academic information corpus.
3. Create an improved index by calculating the value of improved TFIDF algorithm for the academic information corpus.
4. Use the cosine function to calculate the similarity between the query vector and the document vector to retrieve the set of documents.
5. Using the ranking function, re-rank and obtain the set of retrieved documents.
6. Finally, recommend according to the ranking function.

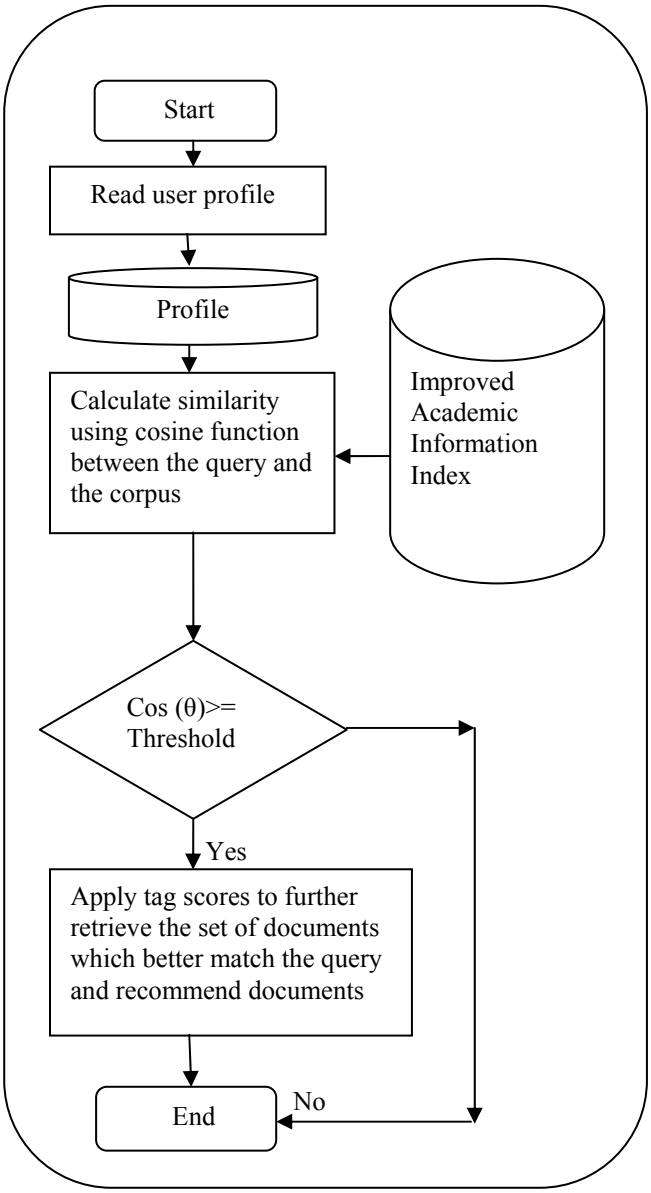

Fig. 2. Recommender Mechanism

IV. ILLUSTRATION

The sharing system shares academic information and the set of tags related to it. The academic information crawler obtains data from the World Wide Web. This is used to maintain academic information corpus. Improved TFIDF algorithm is used to create an index. User preference crawler maintains data specific to a user, i.e., his personal records, his name, author, title, tags used etc. A cosine function is used to compute similarity between the user's query vector and the document corpus. The retrieved URLs are given to the ranking module, which uses a ranking function based on the tag score. It ranks the documents and recommends the improved list of recommended documents to the user.

ISBN: 978-988-19252-1-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

V. CONCLUSION

This paper proposes a framework for Academic Information Sharing and Recommender System. Academic information includes paper, quizzes, assignments, tutorials on different subjects. The different subjects like computers, electronics, mechanical etc. make different communities in the system, in which different users can exchange the information and add tags. The tagging information is obtained from different resources which is the metadata created for every resource. Tag is used as a basis for calculating the tag scores. The index is created according to the improved TFIDF algorithm. The documents are retrieved using the cosine similarity. The resources retrieved are further re-ranked according to the tag scores. The papers are recommended to the user after application of the improved ranking algorithm. A similar kind of system has been developed for Research Paper Recommendations.

In the future, the proposed framework can be developed.

REFERENCES

[1] A. Mislove, K.P. Gummadi, and P. Druschel. "Exploiting social networks for internet search.", In *Proc. of the 5th workshop on Hot Topics in Networks* (Hot- nets-V), 2006.

[2] A. Capocci and G. Caldarelli. "Folksonomies and Clustering in the Collaborative System CiteULike. " ,eprint arXiv:0170.2835, 2007.

[3] Chongchong Zhao, Zhiqiang Zhang. "A New Keywords Method to Improve Web Search", *12th International Conference on High Performance Computing and Communications*, IEEE, (1-3rd Sept, 2010), pages 477-484.

[4] Denis Parra-Santander, Peter Brusilovsky. "Improving Collaborative Filtering in Social Tagging Systems for the Recommendation of Scientific Articles.", *in proc. International Conference on Web Intelligence and Intelligent Agent Technology*, vol.-01, IEEE /WIC/ACM, 2010, pages 136-142.

[5] E. Santos-Neto, M. Ripeanu, and A. Iamnitchi. "Tracking usage in collaborative tagging communities," In *'Workshop on Contextualized Attention Metadata.'*CAMA, 2007.

[6] F.M. Suchanek, M. Vojnovi'c, and D. Gunawardena, "Social Tags: Meaning and Suggestions," CIKM'08, ACM, Napa Valley, California, USA, 26-30 October 2008.

[7] G.Beydoun, R. Kultchitsky and G. Manasseh. "Evolving semantic web with social navigation." in Expert Syst. Appl., 2007, 32(2): 265-276.

[8] Oren, Nir. "Re-examining tfidf based information retrieval with Genetic Programming.", In *Proceedings of SAICSIT*, 2002, pp.1-10.

[9] Pijitra Jomsri, Siripun Sanguansintukul and W. Choochaiwattana "A Framework for Tag-Based Research Paper Recommender System: An IR Approach" in *proc. of the 24th International Conference on Advanced Networking and Applications Workshops*, IEEE, 2010, pages 103-108.

[10] P.Jomsri S. Sanguansintukul and W. Choochaiwattana. "A Comparision of Search Engine Using "Tag, Title and Abstract" with CiteULike – An Initial Evaluation" The *4th International Conference for Internet Technology and Secured Transactions* (ICITST-2009) London, UK, 9-12 November 2009.

[11] S. Xu, S.Bao, B. Fei, Z. Su and Y. Yu."Exploring Folksonomy for Personalized Search," In *Proc. Of SIGIR*, pp. 155-162, 2008.

[12] Sergey Brin and Larry Page. "The Anatomy of a Large–Scale Hypertextual Web Search Engine," in Proc. *of the Seventh International World Wide Web Conference (Brisbane, Australia)*, pp.107-117, Apr 14-18, 1998.

[13] S.Xu, S.Bao, Ben Fei, Zhong Su, "Exploring Folksonomy for Personalized Search.," in *proc. of the 32nd Annual ACM SIGIR Conference, ACM*, July 19-23, 2008, Singapore.

[14] S. Noel and R.Beale. "Sharing vocabularies: Tag Usage in CiteULike," *Proc. Of the 22nd Annual Conference of Interaction, a Specialist group of the British Computer Society (HCI)*, Liverpool, UK, 1-5 September (2008).

[15] S. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," Journal *of Information Science*, 2006, pages 198-208.

[16] Thorsten Joachims,"Optimizing Search engines using Click-through Data," KDD'02- *proc. of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2002.

[17] T. Joachims, L. Granka, B.Pang, "Accurately Interpreting Clickthrough Data as Implicit Feedback," In *Proc. of the ACM Conference on Research and Development on Information Retrieval(SIGIR)*, 2005.

[18] T Bogers, "Recommender systems for social bookmarking," PhD Thesis, Tilburg University, 2009.

[19] T. Toine Bogers, and A. van den Bosch, "Recommending Scientific Articles Using CiteULike", RecSys'08, Switzerland, 23-25 October 2008.

[20] Umer farooq, Thomas G. Kannampallil, Yang Song, "Evaluating Tagging Behaviour in Social Bookmarking Systems: Metrics and design heuristics," ACM, NOVEMBER 4-7, 2007.

[21] Xian Wu, Lei Zhang and Yong Yu., "Exploring Social Annotations for the Semantic Web." In proc. of the 15th International Conference on World Wide Web(WWW 06), ACM, 2006, pages 417-426.

[22] Xin Hu, Hua Jiang, Ping Li, Shuyan Wang., "An improved method of term weighting for text classification." ,in *proc. of International Conference on Intelligent Computing and Intelligent Systems(ICIS 2009)*, IEEE, 2009

[23] Yi Jin, Zhuying Lin, Hongwei Lin , "The Research of Search Engine Based on Semantic Web," in proc. of International Symposium on Intelligent Information Technology Application Workshops(IITAW), IEEE, 2008, pages 360-363.

[24] Yi-hong Lu, Yan Huang, "Document Categorization with Entropy based TF/IDF classifier", WRI Global Congress on Intelligent Systems (GCIS), IEEE, 2009, pages 269-273.

[25] Zhiqi Fang, Yue Ning, Tingshao Zhu, " Hot keyword Identification for extracting Web Public Opinion," in proc. of the 5th International Conference on Pervasive Computing and Applications(ICPCA), IEEE, 2010, pages 116-121.

[26] Caimei Lu, Xiaohua Hu and Jung-ran Park, "Exploiting the Social Tagging Network for Web Clustering," *in proc. IEEE Transactions on Systems, Man and Cybernetics* ,VOl. 41, NO. 5, 2011.

[27] http:// delicious.com/.

[28] "CiteULike: Everyone's library" Internet: http://www.CiteULike.org

ISBN: 978-988-19252-1-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)