International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:6, No:9, 2012

Analyzing the Relation of Community Group for Research Paper Bookmarking by Using Association Rule

P. Jomsri

***Abstract*—**Currently searching through internet is very popular especially in a field of academic. A huge of educational information such as research papers are overload for user. So community-base web sites have been developed to help user search information more easily from process of customizing a web site to need each specifies user or set of user. In this paper propose to use association rule analyze the community group on research paper bookmarking. A set of design goals for community group frameworks is developed and discussed. Additionally Researcher analyzes the initial relation by using association rule discovery between the antecedent and the consequent of a rule in the groups of user for generate the idea to improve ranking search result and development recommender system.

***Keywords*—**association rule, information retrieval, research paper bookmarking.

I. Introduction

THE recently, researching within the information retrieval has considered an alternative approach of retrieving the information based on community of users in the system. Many social bookmarking systems have been designed and implemented for improve systems. Especially, Social resource sharing systems are web-based systems that allow users to upload all kinds of resources.

Furthermore, Search engines are the important tools that people search document on internet. It can return search result by user query. Nowadays, social network has recently received a wide adoption by various web services such as social bookmarking systems. They provide functions that allow users to share content with one another. In a field of academic have a several work of research to regard one which use search engine for searching research paper and investigate the literature reviews such as *CiteULike*[1]. It 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.

As part of social research paper bookmarking system has community group which perhaps each community may concentrate on the same topic. In addition, user in social bookmarking system can join with another groups or communities that user interest. Those communities which users are members may related content or research topic.

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).

Therefore, in this paper proposed to analyze the relation of community group for research paper bookmarking by using association rule. The main point is study the relation of user group by using data mining techniques for optimize ranking.

The paper is organized as follows. Section II discusses related works. The framework of this paper is described in Section III. The association rule analysis 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

This section contain in to two parts: first is background of community based on social bookmarking and second is related research with CiteULike.

A. Community based on social bookmarking

In recent years, many studies of community-based on search engine have been carried out. The main techniques involved in community-based search engine include *recommendation*, *relevance feedback*, *personalization*, and their combinations. Many research try to measures the similarity or relation between groups for improve the performance of recommender system such as Senot and *et al.* build group profile of TV viewing data by combine with individual user for showing how group interest[4]. Therefore the group personnel relationship exists in social groups of all sorts, which can be researched using the knowledge of the complex social networks system. There are several specific research projects on community of social network such as Cohen, and Havlin studied the degree distribution of co-author research network in mathematics and the neuroscience domain. These distributions do not strictly follow the power-law distribution [15]. Zhang and Di described the clustering algorithms of co-author research network [16]. Chang, and Daren showed the results of proprietary Chinese medicine network in 2005 [23]. Hong, Wei-dong, and Wen analyze relation of group personnel relationship. By comparing the group personnel relationship models and the empirical models, the simulation results in according with the empirical findings quit well [17]. Some researchers applied association rule mining for improve the web performance such as Heymann, Ramage, and Garcia-Molina, [12] use association rule mining based in combinations with other measures for link prediction on social tags. Schmitz and *et al.* [13] describe the idea of using association rules to determine hyponymy and hyponymy relations between tags in social tagging data. They have a strong emphasis on formal concept analysis and its usage in context of social tagging data.

Although this paper are following a similar initial thought of utilizing classical data mining techniques for discovering

structures in social bookmarking. This paper focus is on structuring groups of user. Aim of this is suggest to the user group bundles for organizing information. The idea is that research paper was assigned by any given users which are a reflection of his interesting and share research paper with other user in the same group. In addition, relationships between group and their user perceived "paper" can be gained.

B. Citeulike

CiteUlike (www.citeulike.org) is a web-based social bookmarking services and traditional bibliographic management tools. It assists researchers and academics in storing, organizing, sharing and discovering links to academic research papers. Like many successful software tools, CiteUlike has a flexible filing system based on the tags. It has been available as a free web service since November 2004. As of September 2011, there are approximately 5,549,945 articles on CiteULike. Their metadata, abstracts, and links to the papers at the publishers' websites. Users can also add reading priorities, personal comments, and tags to their papers. CiteULike also offers the possibility of users setting up and joining groups that connect users sharing academic or topical interests. These group pages report on recent activity. The full text of articles is not accessible from CiteULike, although links to online articles can be added.

Toine Bogers [10] divide a type of metadata from the CiteULike website into five types. First is Topic-related metadata: including all metadata descriptive of the article's topic. Second is Person-related metadata: such as the authors of the article. Third is Temporal metadata: such as the year. Fourth is Miscellaneous metadata: such as the article type. Fifth is User-specific metadata: including the tags assigned by each user, comments by users on an article, and reading priorities. As CiteULike offers the possibility of users setting up groups that connect users that share similar academic and topical interests for each group we collected the group name, a short textual description, and a list of its members.

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: Jomsri, Sanguansintukul, and Choochaiwattana [6], [7] create three heuristic indexers: "tag"(T), "title, abstract"(TA) , "tag, title and abstract"(TTA) and compare with CiteULike. Experiment found that TTA is the best indexer. Furthermore they create a new algorithm for ranking method that is a combination of similarity ranking with paper posted time or *CSTRank* [5]. Capocci and Caldarelli [8] analyzed the small-world properties of the CiteULike folksonomy and the other researcher are [10], [11], [9], and [14].

This paper uses different views to re-ranking search results of research paper bookmarking with focus on the diversity and reliability.

This paper extends the method of association rule that is data mining technique to re-ranking search results.

III. Framework for Community Group of Research Paper Bookmarking

A framework for community group of research paper bookmarking is follows in Fig 1. General community of users who interesting in research paper bookmarking will post papers that they interest to server system of social bookmarking system such as CiteULike. This technique can provide paper with other users for search paper. The framework mechanism was designed in four steps:

- *Historical data of each user groups:* After process of user share all their public entries with user community and comment on other papers. Java programming is used to implement a crawler on the research documents. The crawler collects data from research paper bookmarking. The collected documents consist of research papers and each record in the paper corpus contains: article ID, article name, abstract, tag of each paper, link for viewing full text article, groups name, along with group are interest the same paper, book title that published paper, posted date, posted time, paper priority ,and etc.
- *Association rule:* This step is preparing and cleaning data for creating association rule model. The relation during users group that interested in the same paper was analyzed. This technique is recommending base on similarity and were describe in section IV.
- *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.
- *Re-ranking search result:* this step is effect after similarity measurement for improve search result. The ranking of search results are rearranged from the highest similarity score to the lowest similarity score.

IV. Association Rule Analysis

Association rule discovery is a popular data mining method and well researched method for discovering interesting relations between variables in large databases. Many the research lead data mining and association rule for analyzes and increase efficiency in searching result [18], [19], [20], [21], [22].

This paper analyzed basic data relation by using association rule discovery from personalized function for explore pattern to improve ranking. Researcher explored association of a set title name of paper and set groups of user. We expect that the article were posted more than one group will should significant for create ranking. The data set has over 64,320 rows.

Each row of the data set represents a user group that papers were appearing. There for, a single paper can have multiple rows in the data set.

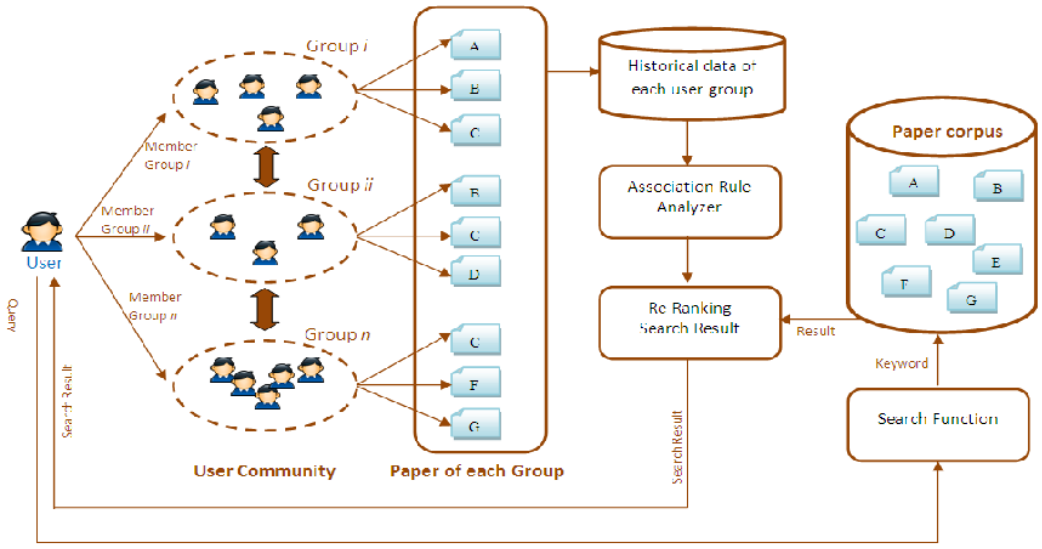

Fig. 1 A Framework for community group of research paper bookmarking

A. Association rule discovery

The rules tab in the form of $X \rightarrow Y$ is applied for extracting rules. Where X and Y are disjoint item sets of user group. For each rule of the form $X \rightarrow Y$, researcher defines the *supp* and *conf* as the *support* and *confidence* as follows.

$$conf(X,Y) = \frac{count(X,Y)}{count(X)} \quad (1)$$

such as $conf\,(group\ X, group\ Y)$

$$conf(group\ X, group\ Y) = \frac{count(group\ X\ , group\ Y)}{count(group\ X\ )} \quad (2)$$

$$sup(X,Y) = \frac{count(X,Y)}{count(All)} \quad (3)$$

such as $sup(group\ X, group\ Y)$

$$sup(group\ X, group\ Y) = \frac{count(group\ X, group\ Y)}{count(All)} \quad (4)$$

Table I shows examples of rules for predicting the group. Confidence and support value are used for rule selections. Because plenty of rules are generated, some simple concerns in rule selections include:

1) Select the rule with maximum confidence.
2) Select the rule with maximum support if confidence value is equal.
3) Select the rule that happens first when confidence and support values are equal.

TABLE I
EXAMPLES OF RELATION MODELS OF GROUP THAT USER POST WITH CONFIDENCE AND SUPPORT VALUES

Rule	*Conf (%)*	*Sup (%)*
microRNA→Bioinformatics	72.46	3.08

From table I, shows the rule explains:

- Support of $X \rightarrow Y$ is the probability that a paper has both *X* group and *Y* group

Confidence of $X \rightarrow Y$ is probability that a paper appear in *Y* group given that the paper appear in *X*

B. Result of Association rule discovery

Table II shows total association prediction model for group of users with confidence and support values.

TABLE II
RELATION MODELS OF GROUP THAT USER POST WITH CONFIDENCE AND SUPPORT VALUES

Rule	*Conf (%)*	*Sup (%)*
Genetics-of-Gambling→G4ID	53.95	8.39
G4ID→Genetics-of-Gambling	100	8.39
Philosophy_of_informatic→Blog_and_WikiResearch	82.46	5.16
Blog_and_WikiResearch→Philosophy_of_informatic	40.22	5.16
Statistics and Social Science→Biostatistics	86.09	3.40
Biostatistics→Statistics and Social Science	88.89	3.40
microRNA→Bioinformatics	72.46	3.08
Bioinformatics→microRNA	17.48	3.08
mgh_lcs→Blog_and_WikiResearch	90.30	2.13
Blog_and_WikiResearch→mgh_lcs	16.60	2.13
ReadingLab→Clinical_Psychology	45.63	2.12
Clinical_Psychology→ReadingLab	85.94	2.12
Social navigation→Adaptive-Web	69.83	1.63
Adaptive-Web→Social navigation	53.76	1.63
Automatic sumarization→ASR	86.64	1.31
ASR→Automatic sumarization	50.50	1.31
NLP→ASR	83.10	1.16
ASR→NLP	44.47	1.16

V. EXPERIMENTAL SETTING

The experimental setting is divided into two sections. Section *A)* describes the data set, section *B)* discusses 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 ,paper priority and the along with group.

B. Evaluation Matrix

The informal was conducted with twenty students that were recruited as experiment participants. In the step of measuring the system accuracy, we need to use information retrieval classification metrics, which evaluate the capability of the system to suggest a short list of interesting items to the user. The precision and recall are the standard measurement for the probability that the system makes a correct or incorrect decision about the user interest. With r_x being the research paper from randomly picked for user u and $D(u,r)$ is the set of recommended research papers, recall and precision are defined as Equation (5)and (6):

$$recall = (D(u,r)) = \frac{1}{|U|}\sum_{u\in U}\frac{|rp(u,r)\cap D(u,r)|}{|rp(u,r)|} \quad (5)$$

$$precision = (D(u,r)) = \frac{1}{|U|}\sum_{u\in U}\frac{|rp(u,r)\cap D(u,r)|}{|D(u,r)|} \quad (6)$$

Where

$|rp(u,r)|$ is the number of is relevant documents,

$|D(u,r)|$ is the number of retrieved documents.

$|rp(u,p)\cap D(u,p)|$ is the number of relevant documents from the number of retrieved documents.

Recall measures the percentage of interesting items suggested to the users, with respect to the total number of interesting items. Whereas, precision measures the percentage of interesting items suggested to the users, with respect to the total number of suggested items. The values precision and recall are shown in section VI. The twenty subjects were considered as experts in the field participated in the experiment. Therefore, their relevancy ratings are assumed to be perfect. In the study setting, each subject is assigned to investigate the research papers obtained from the r_x. The 10 documents for relevancy are displayed. Finally, the subjects were asked to rate the relevancy of the search results on a two-point scale: score 0 is not relevant at all and score 1 is relevant.

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 paper were described in two section first is result of association rule and second is result of evaluation by using precision and recall.

1) Result of association rule

Form table III, We choose the rule that have confidence value more than 60%. The strength rules were hold such as Social navigation with Adaptive-Web has Confidence 69.83%. Article which appears in Social navigation will appear in Adaptive-Web always. Therefore, the relationship of these rule may help to created ranking for optimize search results to user. However, Adaptive-Web with Social navigation has Confidence 53.76%. So article which appears in Adaptive-Web group will not appear in Social navigation always. Therefore, the relationship of these rule may not help to created ranking for optimize search results.

TABLE III
CONFIDENCE OF ASSOCIATION RULE WHERE $\alpha = 60\%$

Rule	*Conf (%)*	*Rule Hold*
Genetics-of-Gambling→G4ID	53.95	No
G4ID→Genetics-of-Gambling	100	Yes
Philosophy_of_informatic→Blog_and_WikiResearch	82.46	Yes
Blog_and_WikiResearch→Philosophy_of_informatic	40.22	No
Statistics and Social Science→Biostatistics	86.09	Yes
Biostatistics→Statistics and Social Science	88.89	Yes
microRNA→Bioinformatics	72.46	Yes
Bioinformatics→microRNA	17.48	No
mgh_lcs→Blog_and_WikiResearch	90.30	Yes
Blog_and_WikiResearch→mgh_lcs	16.60	No
ReadingLab→Clinical_Psychology	45.63	No
Clinical_Psychology→ReadingLab	85.94	Yes
Social navigation→Adaptive-Web	69.83	Yes
Adaptive-Web→Social navigation	53.76	No
Automatic sumarization→ASR	86.64	Yes
ASR→Automatic sumarization	50.50	No
NLP→ASR	83.10	Yes
ASR→NLP	44.47	No

In addition, we use link analysis to show the relation of group's users interested in the same paper.

Fig.2 shows example of similarity measurement from Adaptive-Web –Various kinds of user adaptive web system: hypermedia, IR, filtering — by using Link Analysis. We found that some articles were appear in Adaptive-Web will appear in ARTFL group, NET8-UAM group, Social Web group, Social Navigation group, and Philosophy of Information group. Form result of the similarity we can develop this model into paper recommendation mechanism.

2) Result of evaluation Matrix

Since the subject relevancy ratings, some users only rating one in a group and some users even did not rating any research paper in most groups. The experiment result is depicted in table IV. We use two different correct sets in the experiment. The first is correcting set of original method (not include association rule technique) and the second considers set by include association rule technique before re-ranking method (Our Method). The result is listed in the second and third column.

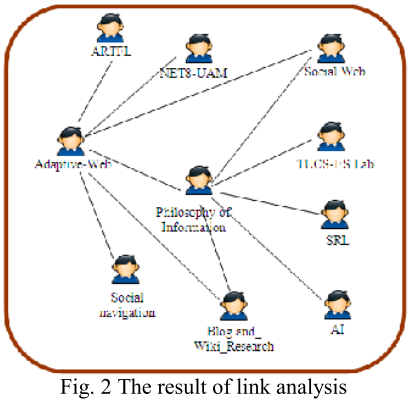

Fig. 2 The result of link analysis

TABLE IV
PERFORMANCE OF PROPOSED FRAMEWORK

Average Precision	*Original Method*	*Our Method*
P@1	55.0%	73.2%
P@2	43.1%	65.7%
P@3	40.2%	63.2%
P@4	38.7%	61.1%
P@5	35.2%	55.5%
P@10	27.0%	49.0%
P@15	25.5%	44.1%

B. Discussion

This paper presents techniques ranking search result for users based on the relation of user group. In the association rule step, the *support* and *confidence value* were used to determine the groups relation. The performance of system by include association rule technique tag based filtering recommendation has accuracy more than the original system**.** Therefore, the relation of user group has a potential and can use this technique for improve ranking search result.

VII. Conclusion and Future Work

In this study, a related of community group of research paper bookmarking framework is proposed.

This approach studies users' behavior from research paper bookmarking and then use association rule to analysis user's preference and can bring to improve ranking. The experiment has shown some interesting results and it is believed the research direction is promising.

In fact, during our study, it is becoming clear that only relying on one method to predict the preference. Furthermore, in this paper the ranking mechanism only considers one-to-one association rule like *group X*→*group Y*. This assumption is to simplify the problem.

In addition, This paper preliminary analysis a relation of the group uses that appear the same article by using association rule discovery .Result of preliminary analysis of there some rule is interesting and can bring to improve performance ranking.

In future, researcher plan to use this information to advance analysis for improve web searching.

Acknowledgment

The authors would like to thank Suan Sunadha Rajabhat University for scholarship support. The study is not possible without the data from CiteULike.

References

[1] CiteULike, http://www.CiteULike.org
[2] Connotea, http://www.connotea.org
[3] BibSonomy, http://www.bibsonomy.org
[4] C. Senot, D. Kostadinov, M. Bouzid, J. Picault, A. Aghasaryan, and C. Bernier ,"Analysis of Strategies for Building Group Profiles,"in *User Modeling, Adaptation, and Personalization 2010*, Lecture Notes in Computer Science, 2010, Volume 6075/2010, pp 40-51.
[5] P. jomsri, A Combination of Similarity Ranking and Time for Social Research Paper Searching, World Academy of Science, Engineering and Technology 78 2011,pp. 638-643
[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] P. Heymann, D. Ramage, and H. Garcia-Molina, "Social tag prediction," in SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 2008, pp. 531–538.

[13] C. Schmitz, A. Hotho, R. J'lschke, and G. S. and,"Mining association rules in folksonomies," in DataScience and Classification, ser. Studies in Classification,Data Analysis, and Knowledge Organization. SpringerBerlin Heidelberg, 2006, pp. 261–270. [Online].Available:http://www.springerlink.com/content/gmv832553g0x3673/

[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] R. Cohen, and S. Havlin, "Scale-free network are ultrasmall Physica", A311, p590

[16] Peng Zhang, Zengru Di, Complex System and Complexity Science, 2(3), pp.30-34

[17] W. Hong, W. Wei-dong , X. Na, H. Wen, "Group Personnel Relationship Analysis Based on Social Networks", in *IEEE International Symposium on IT in Medicine & Education, 2009*. (ITIME '09),pp.1003 – 1008

[18] X. Chen, and Y. Wu. Personalized Knowledge Discovery: Mining Novel Association Rules from Tex t. Available: http://www.siam.org/meetings/sdm06/proceedings/067chenx.pdf

[19] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. (2008, Oct). Mining Association rule in Folksonomies. *Journal of Information Science (JIS)* [Online].Available:*http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.9741&rep=rep1&type=pdf.*

[20] C. Haruechaiyasak, M. Shyu, and S. Chen, "A Data mining Framework for Building A Web-Page Recommender System", Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, IRI - 2004, November 8-10, 2004, Las Vegas Hilton, Las Vegas, NV, USA. pp. 357-362

[21] R. Forsati, M.R. Meybodi, A. Ghari Neiat, "Web Page Personalization based on Weighted Association Rules", *International Conference on Electronic Computer technology 2009*, pp. 130-135

[22] S. niwa , T. Doi, and S. Honiden, " Web Page Recommender System based on Folksonomy Mining for ITNG'06 Submissions", *Proceedings of the Third International Conference on Information Technology:New Generations (ITNG'06)*

[23] Hui Chang, Daren He, *Science and Technology Review* , 24(9),pp. 84-87