International Journal of Information Technology and Business Management
29th December 2012. Vol.8 No. 1

ISSN 2304-0777 www.jitbm.com

PROMOTING RESEARCH COLLABORATION BASED ON DATA MINING TECHNIQUES IN LIBRARY INFORMATION SYSTEMS

Elaheh Homayounvala[a], , Ammar Jalalimanesh[a]

[a] Information engineering department, Iranian Research Institute for Information Science and Technology, 1090, Enqelab St., Tehran, Iran, PO Box: 13185-137,1 Tehran, Iran,

Email: vala@irandoc.ac.ir

Abstract

Research collaboration connects distributed knowledge and competencies into new ideas and research institutes and has been the subject of many research projects. We argue that academic libraries, including libraries of universities and research institutes, hold a wealth of information regarding patrons' research interests hidden in their data bases. Mining these databases can provide better understanding of researchers' needs and interests. This paper has two main contributions. Firstly, it proposes a new methodology based on data mining techniques in library information systems to uncover patrons' research interests in order to facilitate research collaboration including interdisciplinary research. The proposed methodology, studies data mining techniques in a library information system as a case study and makes advantage of clustering algorithms to cluster researchers based on their library usage which is interpreted as their research interests. The second contribution of this paper is that, it presented a knowledge map as a visual representation of usage trends of an academic library to portray virtual interest groups based on item use information. The result of this study can support managers and decision makers for strategic decision making regarding future research directions and collaborations. The outcome of the case study confirms our hypotheses by revealing clusters of library users with similar research interests validated by their academic backgrounds.

Keywords: Research collaborations, data mining, Library Information Systems, academic library, patrons' research interests, researcher profiling, organizational knowledge map.

International Journal of Information Technology and Business Management
29th December 2012. Vol.8 No. 1

ISSN 2304-0777 www.jitbm.com

1. Introduction

Research collaboration has received a great deal of attention from governments and organizations recently and many research projects has been carried out in this subject (Bukvova, 2010). Research collaboration connects distributed knowledge and competencies into new ideas and research institutes (Heinze and Kuhlmann, 2008). Amabile et al. Refers to the definition of collaboration as "the coming together of diverse interests and people to achieve a common purpose via interactions, information sharing, and coordination of activities". They identify the core concept of this definition as "individuals who differ in notable ways sharing information and working toward a particular purpose"(Amabile et al., 2001). Research collaborations have been modelled in (Lee et al., 2011) by detecting dependency patterns in research collaboration environments based on co-authorship data at the organisation level. Dependency patterns are extracted in this investigation by a "cross-association clustering" technique. (Yang et al., 2010) models research collaborations between universities by exploring the link relation between their websites.(Cucchiarelli and D'Antonio, 2010) define a methodology for discovering potential collaboration through the investigation of the relation among actors in research networks. They used their methodology to analyse a research-oriented network that their users share potential research interests and paper co-authorship. Heinze and Kuhlmann analyze research collaboration in the growing domain of nanoscience within the German public research system. They based their research on multiple data sources, such as co-publications, macro research statistics, and in-depth interviews. They developed governance structures that support scientists' efforts to start teamwork across institutional boundaries (Heinze and Kuhlmann, 2008). Modelling research collaboration based on library usage however, seems not to be covered in the literature.

Data mining, on the other hand, has been applied in information science generally and in library information systems specifically, for many applications. The applications of data mining to the field of information science can be categorised in three main categories of personalised environments, electronic commerce and search engines (Chen and Liu, 2004). In library information systems, however, Papathedorou et. al summarise data mining applications in three main categories of service optimisation, decision support and personalization (Papatheodorou et al., 2003). Scott Nicholson has coined the term bibliomining for "use of data mining to examine library data records" (Nicholson and Stanton, 2003). He proposes utilizing data mining tools to datasets of libraries in order to aid organizational decision-making within the library or improving library services or external reporting and justification by tailoring services to meet the needs of user groups (Nicholson, 2003, Nicholson and Stanton, 2003).
He also suggests researchers to use bibliomining to generalise their findings about one library to other libraries by creating data warehouses for multiple libraries in order to help obtaining deeper understanding of institutions with the same tools previously used with only one library (Nicholson, 2006). The bibliominig process, he describes, consists of the following steps: "determining areas of focus", "identifying internal and external data sources", "collecting, cleaning, and anonymizing the data into a data warehouse", "selecting appropriate analysis tools", "discovery of patterns through data mining and creation of reports with traditional analytical tools" and finally "analyzing and implementing the results" (Nicholson, 2003).

Data mining can reveal patterns of behaviour among library users and staff and patterns of information resource use throughout the organization. Sources of data for data mining process in libraries includes, but are not limited to, user information, circulation information and searching and navigation information in case of digital libraries (Nicholson and Stanton, 2003). Data mining has been also applied in other literatures for clustering scientific journals in Digital Library (Lee et al., 2010) and for reader classification or constructing user communities with common interests by analyzing queries posed to a digital library (Chang and Chen, 2006, Papatheodorou et al., 2003). User communities are defined as "groups of users who exhibit common behaviour in their interaction with an information system" (Orwant, 1994) in (Papatheodorou et al.,

International Journal of Information Technology and Business Management
29th December 2012. Vol.8 No. 1

ISSN 2304-0777 www.jitbm.com

2003). (Kim et al., 2006) present a visual user-model data mining tool based on user tracking information such as queries and browsing result sets.

The application of data mining techniques in library information systems for research collaboration seems to be a new and novel application. We argue that academic libraries including libraries of universities and research institutes hold a wealth of information regarding patrons' research interests hidden in libraries data bases. Mining these databases can provide better understanding of researchers' needs and interests. This paper specifically studies applying data mining techniques in Library Information Systems of a research institute. It makes advantage of an academic library usage to uncover patrons' research interests in order to facilitate research collaboration including interdisciplinary research. The result of this study can help managers for strategic decision making regarding future research directions and also initiation of interdisciplinary research projects in order to promote research collaborations.

The organisation of this paper is as follows. First, the proposed methodology in this paper is explained in section 2. Then, a case study, namely IRANDOC case study, is described in section 3 and finally section 4 concludes the paper and makes suggestions for future work.

2. Proposed Methodology

In order to uncover library users' research interests, especially to promote interdisciplinary research, we made the assumption that researchers' borrowing history from the research institute or university library is an important source of information. Based on this assumption, our proposed methodology consists of four main steps as explained bellow and depicted in figure 1:

1. Create a data warehouse based on library operational data.
This step includes data collection, refinement and pre-processing of raw data and creation of a data warehouse.

2. Create knowledge map for library subjects and their usage.
This step is designed in order to gain better understanding of most studied subjects and their relationships to other subjects by creating a visual representation of library usage. Creation of this organisational knowledge map is the result of our former study reported in detail in (Jalalimanesh and Homayoun vala, 2011). The created knowledge map is a network in the form of an undirected graph. Each node represents a subject based on library of congress categorisation and the weight of each node shows the total number of books borrowed in that subject. Each edge between two nodes of *S1* and *S2* represents total number of books borrowed by all patrons for both subjects *S1* and *S2*. Since one person may have borrowed different numbers of books for subject *S1* and subject *S2*, the minimum of these two numbers is used for each person in total calculation. Some parts of the knowledge map for IRANDOC library is represented in figure 2.

3. Analyse knowledge map to choose specific subjects for further analysis.
Analysis of the knowledge map created in step two will help decision makers to choose some of the subjects for creation of research teams or initiation of interdisciplinary research projects. Decision makers can choose number of subjects for further analysis, considering their organizations policies in addition to the created knowledge map. Analysis of selected subjects will result in better understanding of researchers' needs and interests who are active in those subjects.

4. Apply data mining techniques to find active researchers in selected subjects from knowledge map analysis.

Data mining can be applied to cluster researchers based on their similarities in borrowing books in selected subjects. Cluster analysis will consequently create better understanding of patrons' research

International Journal of Information Technology and Business Management
29th December 2012. Vol.8 No. 1

ISSN 2304-0777 www.jitbm.com

interests which assists decision makers regarding research collaborations.

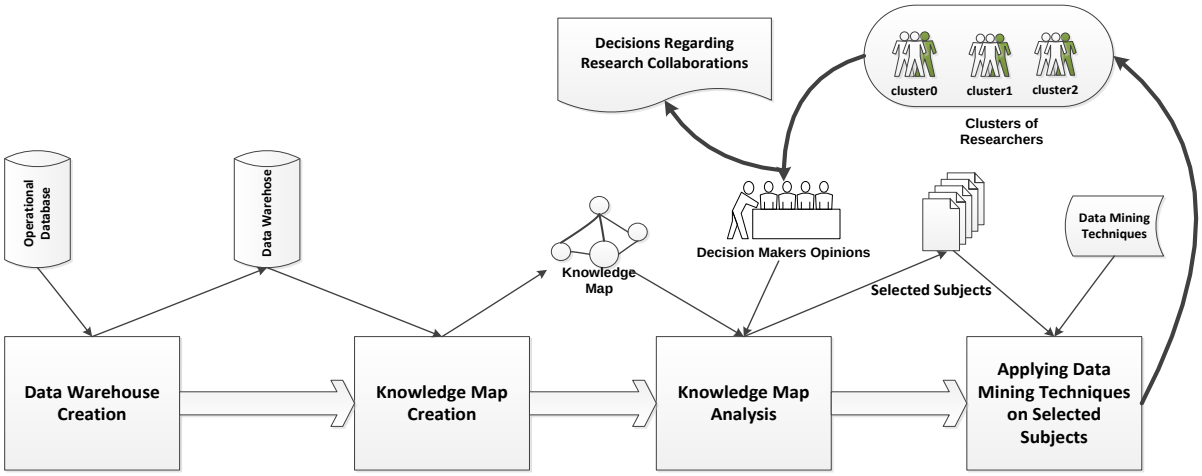

Figure 1 – Visual representation of the proposed methodology.

Organisational knowledge map, as it can be seen in figure 2, can offer a visual representation of most studied subjects as well as inter-connections of subjects to identify interdisciplinary research fields in the research institute under study. Then clustering techniques such as *k*-means or *k*-medoids can be applied in order to find researchers with similar research interests.

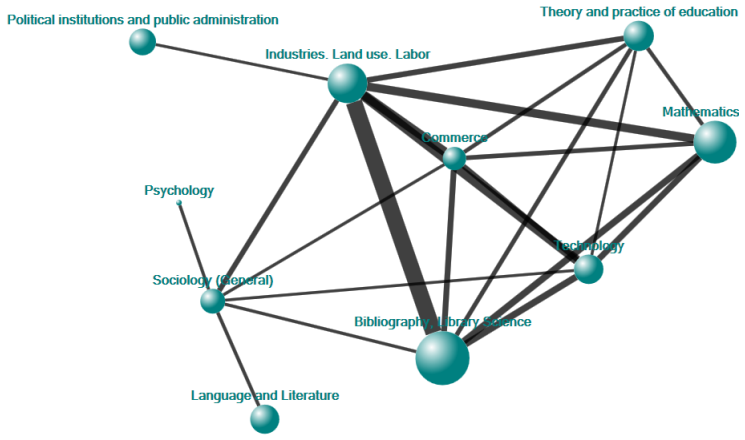

Figure 2 – Organisational knowledge map based on library information system.

k-means (MacQueen, 1967) is one of the simplest and most popular unsupervised clustering algorithms and among top ten data mining algorithms(Wu et al., 2008). Each point, in this algorithm, is assigned to the cluster whose centre is the nearest. k in *k*-means algorithm is fixed a priori and k centroids (centres)

International Journal of Information Technology and Business Management
29th December 2012. Vol.8 No. 1

ISSN 2304-0777 www.jitbm.com

are placed randomly at the beginning of the algorithm. *k*-means algorithm then recalculates the centroid or mean of each cluster and repeats assigning points to new centroids until centroids do not change anymore(Witten et al., 2011). *k*-means algorithm measure distance between two points based on squared Euclidean distances, but another unsupervised clustering algorithm which is called *k*-medoids minimises a sum of pair wise dissimilarities. Both *k*-means and *k*-medoids use expectation-maximisation strategy to converge to a minimum error condition. Centres in *k*-medoids algorithm must be one of the points to be clustered in contrary to k-means that centres could be any point in the space. "*k*-medoids is more robust to noise and outliers as compared to *k*-means" (Han and Kamber, 2006) and works well for small data sets. In our proposed methodology, a comparison of *k*-mean and *k*-medoids is presented for clustering.

In order to evaluate clustering algorithm, Davis-Bouldin index (Davies and Bouldin, 1979) is applied. The smaller the index the better the clustering algorithm, because Davis-Bouldin index calculates the ratio of the sum of within-cluster scatter to between-cluster separation as follows:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max\{\frac{S_n(Q_i) + S_n(Q_j)}{S_n(Q_i, Q_j)}\}$$

Here n is the number of clusters equal to k in *k*-means and *k*-medoids algorithms. $S_n(Q_i)$ is the ratio of the average distance of all objects in cluster Q_i to their cluster centre and $S(Q_i, Q_j)$ is the distance between cluster Q_i and cluster Q_j centres. Therefore Davis-Bouldin is a small number if the clusters are dense and also distant from each other.

3. IRANDOC Case Study

Iranian Research Institute for Information Science and Technology (IRANDOC) is an institute affiliated with the Ministry of Science, Research, and Technology (MSRT) which was established to work in the field of science and technology of Information and Librarianship. IRANDOC library has about 14000 Latin books. IRANDOC library books are organized based on the Congress Classification System and is run on the basis of the Open-shelf System. The users of the library are comprised of university professors, students, researchers, and the IRANDOC staff. Based on the Library Collection Policy, the IRANDOC Library, at present, provides the following subjects: Information Science and subjects related to the Library Science, Information Systems Management, Information Technology, Information Analysis, Knowledge & Information Management, Linguistics, Computerized Terminology and Technology.

We used the methodology, we have proposed in our earlier study, for drawing knowledge map based on library information system user logs (Jalalimanesh and Homayoun vala, 2011). IRANDOC knowledge map drawn based on this methodology is partly illustrated in figure 2. The size of sphere in each node shows the amount of books that studied by IRANDOC researchers totally. The width of bars that connect subjects together shows the amount of studies. IRANDOC knowledge map shows most studied subjects and also interrelation between them which are invaluable source of knowledge for IRANDOC decision makers in order to initiate research projects.

This figure shows the high interests of library users in bibliography, library science and general information resources. Assuming that decision makers in IRANDOC decides to initiate a research project in interdisciplinary subjects of Z (writing, book industries and trade, libraries, bibliography), HD (industries and land use) and QA (Mathematics), then by applying data mining techniques for these subject Z, HD and QA, we can cluster library users according to their interests in mentioned subjects. Figure 3 demonstrates a three dimensional scatter plot of library users who are interested in these three subjects based on the number of books studied in each subjects.

International Journal of Information Technology and Business Management
29th December 2012. Vol.8 No. 1

ISSN 2304-0777 www.jitbm.com

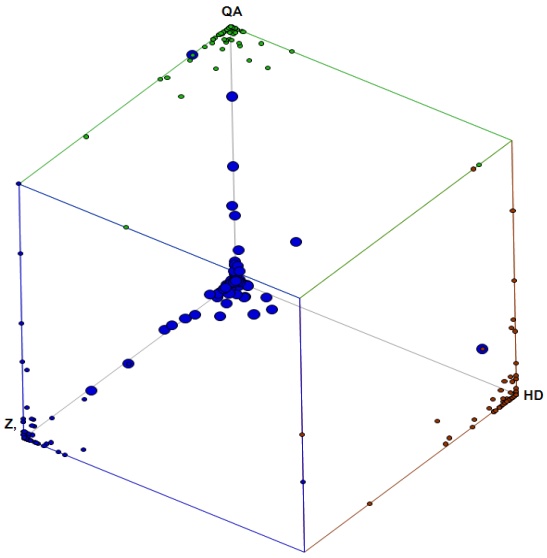

Figure 3 – 3D Scatter plot of subject Z, QA and HD.

Figure 4 demonstrates clustering of library users based on their similarities in books they have studied in mentioned subjects.

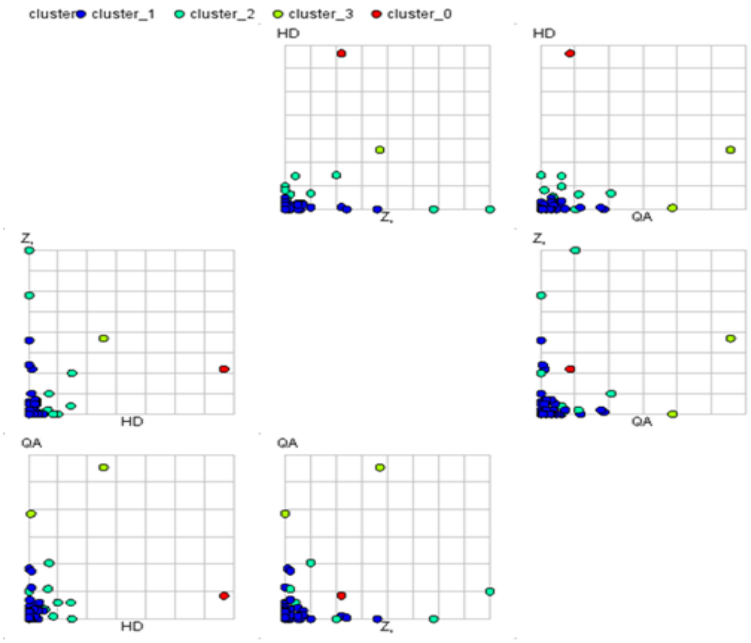

Figure 4 – Two dimensional Scatter plot for subjects Z, QA and HD clustered by *k*-means algorithm.

k-means clustering algorithm in Rapidminer software is used for clustering process. Figure 5 represents same users, clustered by *k*-medoids algorithm. Number of clusters in both clustering is four.

International Journal of Information Technology and Business Management
29th December 2012. Vol.8 No. 1

ISSN 2304-0777 www.jitbm.com

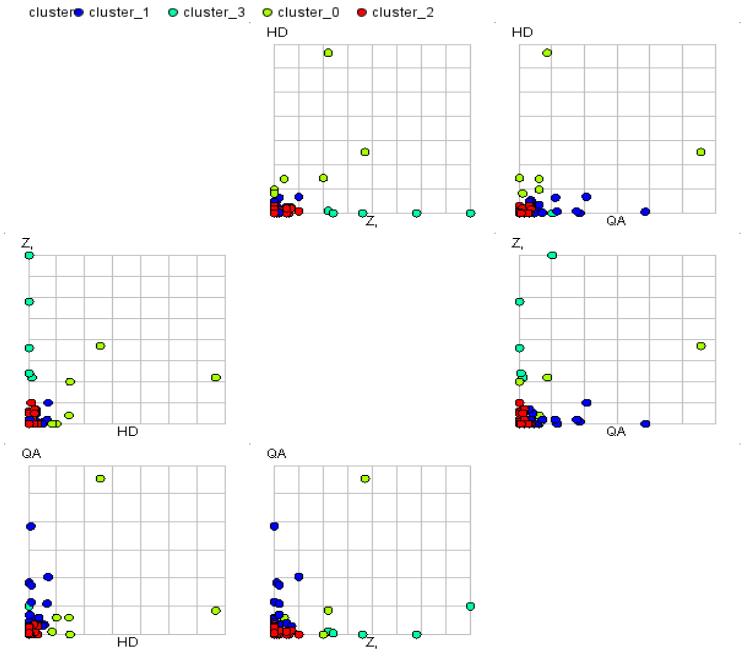

Figure 5 – Two dimensional Scatter plot for subjects Z, QA and HD clustered by K-medoids algorithm.

Table 1: Summarises performance of two algorithms.

Performance Table	K-means Algorithm	K-Medoids
Average within centroid distance	400.869	1354.053
Average within centroid distance cluster_0	0.000	14395.167
Average within centroid distance cluster_1	116.533	567.944
Average within centroid distance cluster_2	1530.346	38.478
Average within centroid distance cluster_3	4475.250	638.000
Davies Bouldin index	0.946	1.527

Table 1 – Performance tables for two algorithms

International Journal of Information Technology and Business Management
29th December 2012. Vol.8 No. 1

ISSN 2304-0777 www.jitbm.com

Although Davies-Bouldin index for *k*-means algorithm is better than k-medoids algorithm, analysis of created clusters in two algorithms reveals that *k*-medoids clusters are more meaningful than clusters in *k*-means algorithm. This supports the fact that *k*-medoids is more suitable for small data sets and is more robust to noise and outliers (Han and Kamber, 2006). Clustering with *k*-means algorithm resulted in clusters with just one and two members whose average within centroid distance are zero or very small and hence have smaller Davies-Bouldin index.

Analysis of clusters created by *k*-medoids algorithm, on the other hand, proves to be more meaningful in our special case study as discussed here. Cluster_2 members (red points) consist of library users who studied very limited number of books in all three subjects under investigation. This presumably indicates that they are not interested in mentioned subjects and we can call them "uninterested cluster". Cluster_0 members (light green points) are far from all three x, y and z axes. These members borrowed substantial number of books in all three subjects, although they might be much more interested in one or two subjects than others, but they all are good candidates for creating research teams on interdisciplinary subjects of Z, QA and HD. We can call this cluster "cross disciplinary active researchers". The average within centroid distance of this cluster is the highest in k-medoids algorithm which confirms the fact that members of this cluster are far from each other in the space, but for our special case study they form a meaningful cluster. Cluster_3 members (green dots) are mostly close to z-axis. They have borrowed many books from library science (Z) subject and few books from HD and QA class. This cluster is named "single disciplinary active researchers". Cluster_1 (blue dots) members borrowed books from all subjects but not as much as those borrowed by cluster_0.

In order to evaluate the quality of the clustering results, samples of cluster members were further investigated based on their research backgrounds reflected in their academic resumes and their research projects. *k*-medoids clusters are proved to be more meaningful clusters. As an example, by choosing three library users from cluster_1 in *k*-medoids algorithm randomly, it was figured out that all of them have been researching in library science and information resource, but they used system engineering techniques to handle their research projects. This confirms our findings regarding their research interests. Investigating number of researchers in cluster_0 against their research backgrounds also validate the fact that they are very active researchers who do not restrict their studies in one subject and are interested to bring other subjects to their major field of study.

4. Conclusions and Future Work

This paper had two contributions. Firstly, it proposed a new methodology for promoting research collaboration based on data mining techniques. The proposed methodology is applicable in the case of academic institutes such as universities and research institutes with active library usage. We argue that information hidden in databases of academic libraries is a key resource for eliciting patrons' research interests. The case study, in which this hypothesis is studied, is a library information system in a research institute, namely IRANDOC research institute. Library databases were mined to gain better understanding of patrons' research interests based on their book borrowing history in order to assists managers in making decisions regarding promoting collaborations including adding new features to library information system. These features may include creating reports for managers on potential research teams, based on library usage. Such reports assists decision makers in research team formation, especially in interdisciplinary research subjects. It can also support sending messages to patrons introducing researchers with similar research interests in order to promote collaborative research. Another important feature can be consulting active researchers for ordering new books.

The second contribution of this paper is that, it presented a visual representation of usage trends of IRANDOC library to portray virtual interest groups based on item use information. The organizational knowledge map presented here is a valuable visual tool for supporting decision makers and offer

ISSN 2304-0777 www.jitbm.com

substantial contribution facilitating analysis process in our proposed methodology.

In summary, this study demonstrated the application of data mining for research collaboration and built a case study to extract research collaboration strategies based on library usage. Two types of strategies can be extracted by this methodology: library users' research collaboration strategies and also managers or decision makers' strategies for research collaborations. Therefore, the result of this study has two folds; it can provide better service provision for individuals and also provides important information for library managers as well as research institute managers and decision makers. The case study approves validity of discovered patrons' research interests compared to their research background.

The research reported in this paper may be further continued by working on sub-categories of each library subject which gives a more detailed analysis of patrons' research interests.

Acknowledgment

The authors gratefully acknowledge the support of the Iranian Research Institute for Information Science and technology, Specially Dr Babak Seyfe, Dr. Gholam Ali Montazer, Dr. Hamid Reza Jamali and Dr. Omid Fatemi.

5. References

1. AMABILE, T. M., PATTERSON, C., MUELLER, J., WOJCIK, T., ODOMIROK, P. W., MARSH, M. & KRAMER, S. J. 2001. Academic-practitioner collaboration in management research: A case of cross-profession collaboration. *The Academy of Management Journal*, 44**,** 418-431.
2. BUKVOVA, H. 2010. Studying Research Collaboration: A Literature Review. *Sprouts: Working Papers on Information Systems*, 10(3).
3. CHANG, C. C. & CHEN, R. S. 2006. Using data mining technology to solve classification problems: A case study of campus digital library. *The Electronic Library*, 24**,** 307-321.
4. CHEN, S. Y. & LIU, X. 2004. The contribution of data mining to information science. *Journal of Information Science*, 30**,** 550.
5. CUCCHIARELLI, A. & D'ANTONIO, F. Mining Potential Partnership through Opportunity Discovery in Research Networks. Advances in Social Networks Analysis and Mining conference, 2010. IEEE, 404-406.
6. DAVIES, D. L. & BOULDIN, D. W. 1979. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on***,** 224-227.
7. HAN, J. & KAMBER, M. 2006. *Data mining: concepts and techniques*, Morgan Kaufmann.
8. HEINZE, T. & KUHLMANN, S. 2008. Across institutional boundaries?:: Research collaboration in German public sector nanoscience. *Research Policy*, 37**,** 888-899.
9. JALALIMANESH, A. & HOMAYOUN VALA, E. Organizational knowledge mapping based on library information system (IRANDOC case study) IADIS International Conference, Collaborative Technologies, 2011 Rome, Italy.
10. KIM, S., LELE, S., RAMALINGAM, S. & FOX, E. 2006. Visualizing user communities and usage trends of digital libraries based on user tracking information. *Digital Libraries: Achievements, Challenges and Opportunities***,** 111-120.
11. LEE, B., KWON, O. & KIM, H. 2011. Identification of dependency patterns in research collaboration environments through cluster analysis. *Journal of Information Science*, 37**,** 67.
12. LEE, J. Y., KIM, H. & KIM, P. J. 2010. Domain analysis with text mining: Analysis of digital library research trends using

ISSN 2304-0777 www.jitbm.com

profiling methods. *Journal of Information Science,* 36**,** 144.

13. MACQUEEN, J. Some methods for classification and analysis of multivariate observations. 5-th Berkeley Symposium on Mathematical Statistics and Probability, 1967 Berkeley. California, USA.
14. NICHOLSON, S. 2003. The bibliomining process: Data warehousing and data mining for library decision making. *Information technology and libraries,* 22**,** 146-151.
15. NICHOLSON, S. 2006. The basis for bibliomining: frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. *Information processing & management,* 42**,** 785-804.
16. NICHOLSON, S. & STANTON, J. M. 2003. Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries. *In:* NEMATI, S. & BARKO, C. D. (eds.) *Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance.* Idea Group Publishing.
17. ORWANT, J. 1994. Heterogeneous learning in the Doppelgänger user modeling system. *User Modeling and User-Adapted Interaction,* 4**,** 107-130.
18. PAPATHEODOROU, C., KAPIDAKIS, S., SFAKAKIS, M. & VASSILIOU, A. 2003. Mining user communities in digital libraries. *Information technology and libraries,* 22**,** 152-157.
19. WITTEN, I. H., FRANK, E. & HALL, M. A. 2011. *Data mining: practical machine learning tools and techniques,* Morgan Kaufmann San Francisco.
20. WU, X., KUMAR, V., ROSS QUINLAN, J., GHOSH, J., YANG, Q., MOTODA, H., MCLACHLAN, G. J., NG, A., LIU, B. & YU, P. S. 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems,* 14**,** 1-37.
21. YANG, B., LIU, Z. & MELOCHE, J. A. 2010. Visualization of the Chinese academic web based on social network analysis. *Journal of Information Science,* 36**,** 131.