DESIDOC Journal of Library & Information Technology, Vol. 30, No. 2, March 2010, pp. 26-39

Visualisation of Relationships Among Library Users Based on Library Circulation Data

Sumit Goswami, Anandarup Mukherjee, Mansi Kharbanda, Abhinav Gupta* and Pulkit Soni*

DESIDOC, Defence R&D Organisation, Metcalfe House, Delhi-110 054
E-mail: sumit_goswami@desidoc.drdo.in

IMPS *College of Engineering & Technology, Malda, West Bengal*
E-mail: soab87@yahoo.com

Indira Gandhi Institute of Technology,Kashmere Gate, Delhi-110 006
E-mail: mansi126750@gmail.com

**Northern India Engineering College, Shastri Park, New Delhi-110 053*
E-mail: abhinavniec@gmail.com; pulkit_indian@yahoo.co.in

ABSTRACT

A library has a large number of databases from which information can be extracted. But data retrieved is superficial and indicates only about the transactions of library. To predict and anticipate the future issue patterns such as what kind of books might be in demand more or what is the taste of majority of users, can only be found by mining the information from the data of library. The library circulation data does not indicate much about the relationship between the users based on the books got issued by them. The aim of this paper is to develop communities based on the issue patterns and find how many similar users are there and which type of books a particular group of users prefer to read. So, the idea is to form a community-based on relationships among the library users and books and visualise the generated relationship through a graph to present the extracted information from the database more effectively. This is possible by finding the hidden relationships in data through community mining process, which has been done by identifying the relationship of users from library circulation data.

Keywords: Data visualisation, circulation data, community identification, data mining

1. INTRODUCTION

In today's world, with the availability of large volumes of data, it is impossible for human beings to interpret it and find useful information. Data mining enables to find answers and discover patterns in data. Data mining relates to the discovery of unknown facts. It is the knowledge discovery process that relates to the finding of hidden information, patterns and relationships from large set of databases. Data mining is, "the efficient discovery of valuable, non-obvious information from a large collection of data"[1]. It aims towards knowledge and pattern discovery. It also tries to discover that side of the knowledge which was never visualised to exist[2]. Another important utilisation of data mining is prediction of forthcoming data in an event where one tries to look for an association between the event and its operating condition. Data present at any academic or research library is vast and versatile. To find the knowledge lying behind the data which is not implicit, the concepts of data mining are used[3]. For example, in a library user database, to predict which of the members would move up in their issue patterns, the data is mined and then based on archived data and knowledge present the future course of events is predicted. This helps in building strategies and anticipation of adversities as well as necessities, which an organisation may face in future[4,5]. Thus data mining helps fulfill two different needs, which are knowledge discovery and prediction[6,7].

1.1 Motivation

Many things on this earth can be given a name according to the community it belongs to. Therefore,

Received on 09 October 2009; Revised on 27 November 2009

researchers started to take these to the complex network, to visualise these according to their needs[8,9]. The need of making communities was to understand the complexity behind the concept, as earlier it was not coming out in any of the picture form but forming communities made it easier. The following systems can be expressed by a network in which nodes represent the objects and arcs signify the relationships between them.

- Social networks such as scientific collaboration network, food network, transport network, etc.
- Technological network such as web network, software dependency network, IP address networks, etc.
- Biological network such as protein interaction networks, metabolic networks, gene regulatory networks, etc.

In this study, authors propose a method to discover library user communities. The basic concept of the study is to provide an interactive visualisation of communities of users and their books[10]. While data mining, the authors ought to find the latent part of information, which cannot be resolved using SQL/ACCESS queries. Such kind of information may include identifying the communities of books read by the users or communities of users with respect to the specific books.

The way helped in finding: which books are more interrelated or which of the readers are more related to each other; how the books and readers are interrelated; if a book is read by the user then how large is the probability that the next book in the community would be read by him/her on studying the communities formed. Several possible ways are present to form communities. In the present study, authors used two factors on which the community will be formed. These are the users, and the books issued to each user. The first factor is a measure of total number of users, which will give the total number of cliques. The second factor gives the number of books with each user that tells us the size of the clique of each user. Both these factors decide the final view of the communities. The communities are identified based on the edges between the common books with each user. These edges give the relationship between various users and it forms a community, which can be easily identified.[11]

2. MINING OF CIRCULATION DATA

The database of books, users, and their associated transactions is important for managing a library. A library database has all the information about its users, the books issued to them, all kind of transactions between the library officials and the publishers, and the different subjects of books in the library. It still does not help in establishing any useful relationship in the data which is not possible through SQL queries. For analytical purposes, one has to find hidden relationships in the data, which can be done using various mining techniques to uncover the hidden patterns (Fig.1).

2.1 Structure of the Data

The study reported in this paper was conducted on a standard library database. “ISSUES” table in the database contains the record of the issue and return of a book by a library user. This table is the sole source of circulation data on which authors accomplished their study for finding communities. The issue table contains an attribute “CODE” which has the unique library membership code of a library user and the “TITLE” which has the title of the issued book. Authors used only two attributes for members and books, respectively for the analysis. The data has to be put in a proper format before it becomes suitable for visual analysis. The errors should be removed from it to the maximum possible extent. Pre-processing the data before visualising includes scanning the entire data and bringing it into a form which can be enabled to be visualised through a visualisation tool. It has heuristics for dividing member records because the member records in ISSUES table sometimes contain redundant member names, or redundant entries. Besides, these may have empty abstract entries, which might not be of any use. Visualisation process creates a graphical

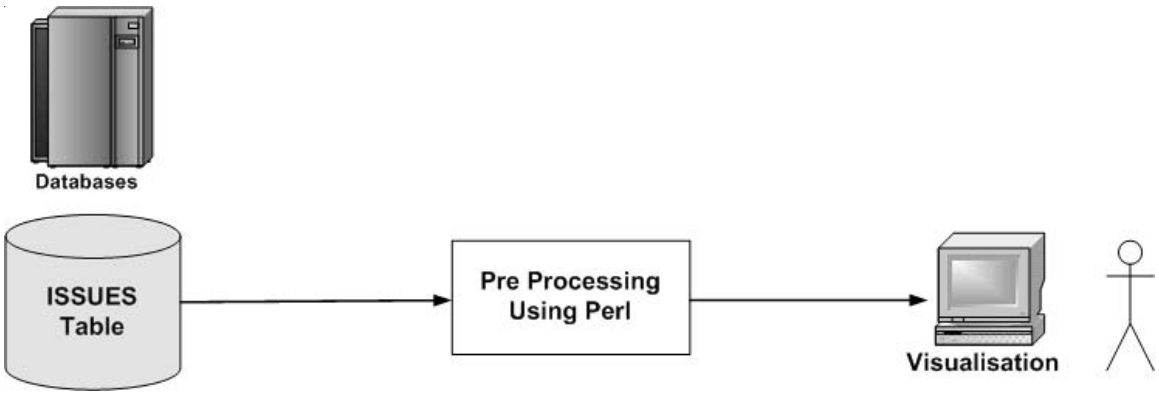

Figure 1. Schematic representation of mining of circulation data.

screen for visualising communities easily[12]. To find a community, the user must be able to browse the data interactively. For this authors used a suitable graph tool that helps in visualising communities through diagrammatic representations via graphs. After establishing a particular relationship amongst the users, the nature can be easily predicted the behavior of the users as well as their for the books. However, mining of the circulation data is an ongoing process, as new patterns are discovered new questions can be raised, which may lead to further mining.

2.2 Data Staging

After extracting the suitable data from the entire database, the next step is to prepare data storage for mining. The extracted data has to be changed, converted, and made ready in a format that is suitable for query and analysis. Three major functions performed during the data staging part are: data extraction, data transformation and data loading (ETL). Data staging provides an area and a place where data can be changed, cleaned, combined, and prepared for actually storing in warehouse for analysis[2].

2.2.1 Data Extraction

Extraction forms the primary stage of data staging component of data mining. Data staging helps in extracting the raw data from the source which is modified to act as input for transformation stage. In this study, the data from 'Circul.mdb', an MS Access file, of case study library database for finding hidden relations as well as existing patterns amongst users of the library (schema shown in Fig. 2). The database contained the following fields:

(i) Type

(ii) Data source

(iii) Code

(iv) Title

(v) Author

(vi) Remarks

(vii) A900

(viii) Isuue date

(ix) Return status

(x) Return date

(xi) Issue officer

(xii) Return officer

This final table was created by making an SQL query for the distinct code and the title from the database. This

	Field Name	Data Type
	Type	Text
	Data Source	Number
	Code	Number
	Title	Text
	Author	Text
	Remarks	Text
	A900	Text
	Issue Date	Date/Time
	Return Status	Text
	Return Date	Date/Time
	Issue Officer	Text
	Return officer	Text

Figure 2. Schema of document issue table in Circul.mdb.

query was made to give the title a unique serial number to make the implementation part easier to visualise. Next step in data extraction is to identify the attributes that can be used for data mining. CODE and TITLE were choosen to make communities of users based on similarity of books issued to them.

2.2.2 Data Transformation

Transformation is an intermediate process in data staging component of data mining. This is the refining stage in which the data is filtered according to the requirements of the study and then used for knowledge discovery. Each book is represented by a vertex and the books which are issued to a common user are connected by the arcs forming a mesh. So the entire data received after extraction is converted into arcs and vertices. Vertices are represented by the instances of TITLE. Arcs represent the relation between them. For example, if CEE0001 user is issued books or TITLE, "Let us C", "Computer Networks", "Organisational Behaviour", then the arcs and vertices would be as shown in Fig. 3. The different books are numbered and a relationship is formed between them by their pair-wise representation under the Edges head.

2.2.3 Data Loading

This is the final stage of the ETL process of data staging component. In this stage transformed data is input into the graph visualisation tool, which processes it

```
CEE0001
*Vertices 3
1 Let us c
2 Computer networks
3 Organisational behavior
*Edges
1 1
1 2
1 3
2 3
```

Figure 3. Representation of books issued by a user in terms of vertex and arcs.

to draw graph for the given data. Entire data is represented into a form shown in Fig. 3 and saved in a suitable format which can be read by the graph visualisation tool. For example, the graph visualisation tool used in this study is 'Pajek' which takes the input file in .net format. The graph visualisation tool thereafter represents the data in the form of a suitable graph which can be used for finding the unobvious piece of information.

3. COMMUNITY IDENTIFICATION

Community is the group of people that share some common features. The word usually refers to a segment of people identified around common features. Community makes us to identify similar things easily. In community identification using library database, communities of people reading similar books and how these are related with respect to similar books read by them are identified. This concept was implemented using circulation data, which envisaged using a graph (Fig. 4). Different colors represent the distinct books read by each user. These books are related to each other and are shown in form of mesh. The common book between the different users is connected to form a community of users.

3.1 Graph Representation

Graph is always an easy mode to understand a complex network, problem etc. It is extensively used in almost every field of computer science. It elaborates a complex system in an easy way, and to visualise it better. Graphs are not only used in computer science but also in other fields to deal with the complex problems. Graphs represent a network in terms of vertices, edges, etc. Vertices are drawn as node, and edges are as a line connecting vertices. The purpose of making graph was to make communities of books borrowed by users. This graph helps in understanding the relation between different communities of users who borrow similar books. Communities can be made between any fields of the database to see the graphical relationship between the respective fields. Earlier it was difficult to make any relationship between the fields, but graphs have sorted this out and the latent concept is understood. The advantages of defining any system in terms of graph are:

- The system comes to the level that is easy to get through the algorithm used if any in the representation.
- The representation can also be changed if required.

3.2 Graph Visualisation

Graph visualisation tool, Pajek[13] was used for the study. Pajek is a software for visualisation of large networks. This software is free open source and can be downloaded from Internet. It gives the graphical representation of the data. This software helps in understanding that how individual community can be identified. The in-built algorithms used represents the graph in a different way. It accepts the file in .net, .mat and few other similar formats. Some other formats like .txt are also acceptable but they all had been generated through Perl language coding which has been used in the study. The basic reason of using the Perl language is that there was a requirement to display the edges in the form so that community can be easily identified. The .net file is obtained form a converter available on the Pajek website. This converter converts the excel file into the .net file.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The pre-processing steps involved in the study have been explained with the help of a flowchart represented in Fig. 5. Perl was used during pre-processing to prepare the

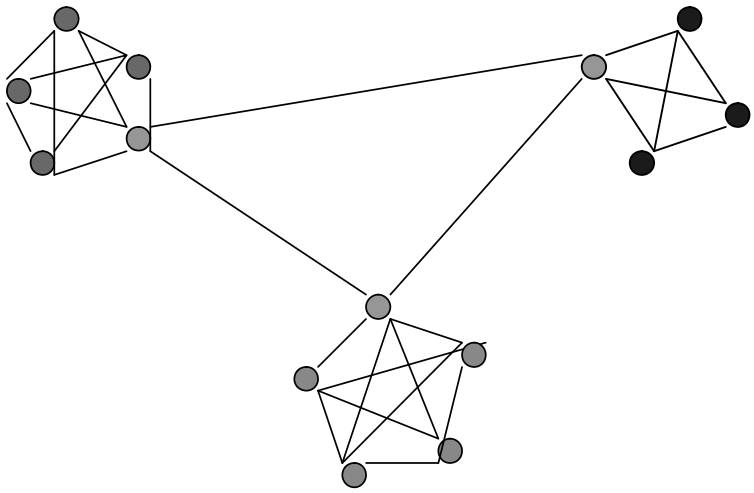

Figure 4. A graph representing the books issued by a user and the common books.

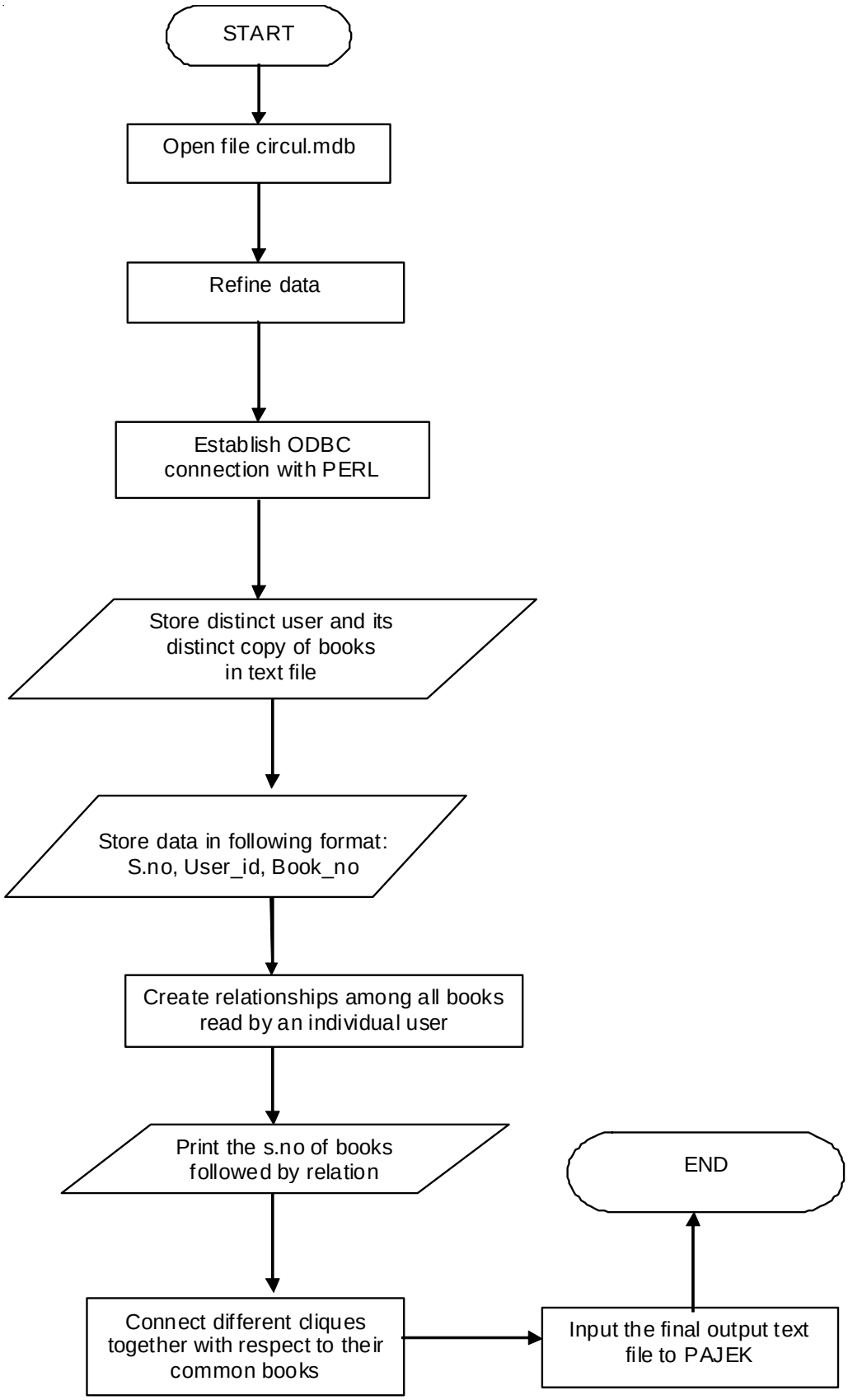

Figure 5. Flow chart representing pre-processing of data before visualisation.

data to make it acceptable in .net format, which can be used as an input to Pajek. A sample Perl code used in the study is represented in Fig. 6. This portion of the code written in Perl demonstrates the use of ODBC connection, named 'test1', created between the database provided and the program, to fetch data from the database according to the users' IDs one-by-one and also fetching their respective books and storing them in a file. Figure 7 gives an overview of the graph of cliques obtained from different users. Red color vertices represent distinct books of each user. Users with different books represent the individual clique in the graph. A clique in a graph is a set of pair-wise adjacent vertices. In the final graph cliques were obtained in a similar way. The common books between the different users display a community of users. Each issued book is represented by an edge, the data for which is an input to the Pajek from a .net file (Fig. 8). All the books issued to a user are connected to

```
use strict;
use DBI;

my $dbh = DBI-> connect('dbi:ODBC:test1');

my @data;
my @data1;
my $i=0;
my $a=1;

open(MYFILE,">compound.txt");
open(TEST,">codes.txt");
my $sth = $dbh->prepare('SELECT DISTINCT code FROM  final ') ;
$sth->execute();

while((@data=$sth->fetchrow_array())&& ($i<=20) )
 {
my $firstname=$data[0];
   print"\n\n\n$firstname\n";
   print TEST "\n$firstname\n";
   my $sql="SELECT DISTINCT sno FROM final WHERE code='$firstname'";

my $sth1 = $dbh->prepare($sql);
$sth1->execute() ;
```

Figure 6. A sample perl code used in the study to extract data from the database.

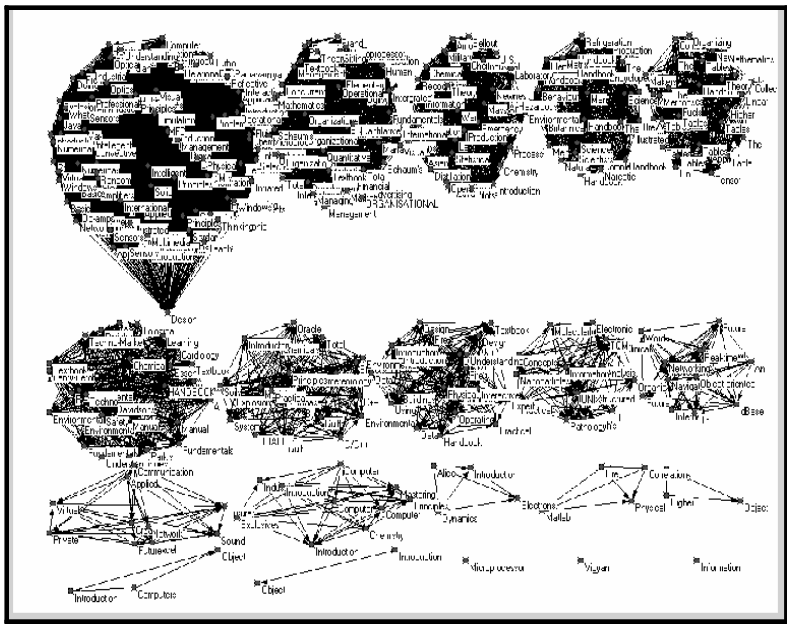

Figure 7. Cliques obtained for different users.

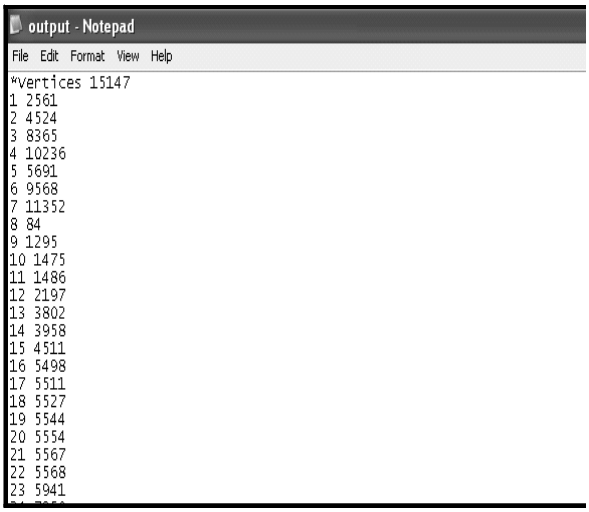

Figure 8. Representation of books identified by acession No. under 'Vertices' as input to Pajek.

form a mesh. The information of arcs for the mesh is available in the input .net file (Fig. 9). Figures 10-14 represent the graph based on the issue pattern of the books of 5, 10, 20, 30, and 50 library users. Looking at these graphs, the books that had been read by more than one user and the number of books issued to a particular user can easily be visualised. Graph also help us to correlate the taste of each user as closeness of their links represent closeness in their subject choice. This clusters the users with a similar choice nearer to each other.

Figure 15 depicts the connectivity among 100 library users who were connected based on the usage of the books on similar subjects. The arc formed on the top of Fig. 15 depicts the users who were not connected to the selected 100 users. Figures 16 and 17 depict the relationship among the library users from a particular lab only, i.e. ISSA and DESIDOC, respectively. The relationship is based on the connectivity of the users who had books from a common subject area. Accession

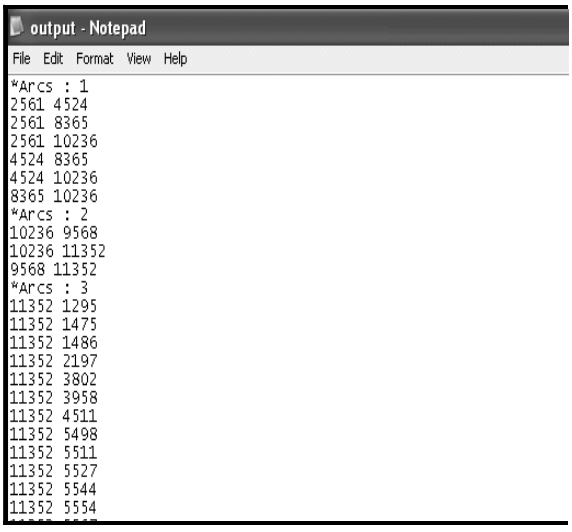

Figure 9. Representation of arcs between books issued by each individual user.

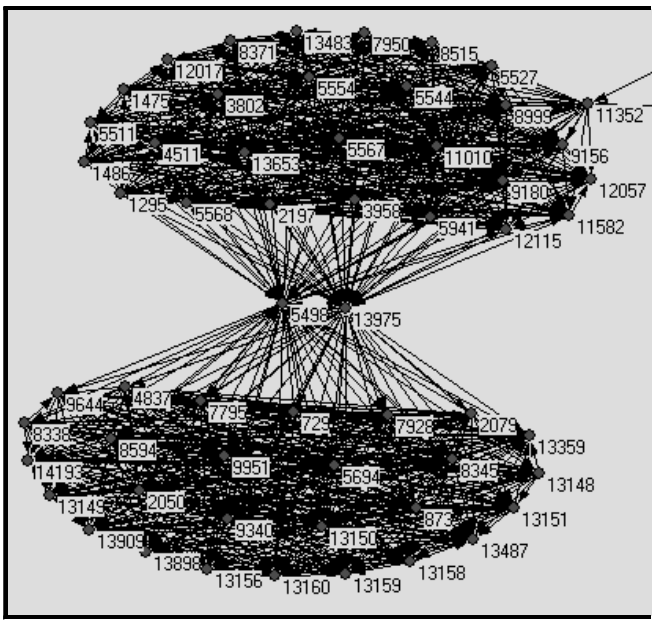

Figure 10. Graph for five users' books.

number does not fulfill the purpose for community identification in this case as there may be multiple copies of the same book, which may hamper clustering all the users of a particular book. Furthermore, there were various books on the same micro subject published by different authors. So clustering based on Class number was more effective. Figures 18 and 19 show the graph of all the users with books on Philosophy and Psychology, and Religion and Theology respectively based on these two macro subject usage pattern. Figure 20 is an alternate separated out view of Fig. 19 to give a better visualisation.

5. CONCLUSION AND FUTURE WORK

After making community of users through books as the node, it can be said that the members who use books form a mesh are related to mesh of another user if and only if they have any books in common with them. Also, if a particular group of users have some common books among them then these members may have some generalised subject of books. Uncommon books of each member can be suggested for other member of the group. For example, if A and B are the books read by a member and A and C are read by another member, then these two members are related to each other through a common node A. If these two can be generalised based on the subject area then book B can be suggested to the member who has not read that book. Other member can be suggested book C. Thus, in spite of such a large database of library user can be suggested his choice of books based on his needs. This can be also be used to trace all the researchers working in a common field of research subject as they would have interlinked mesh. Thus, library circulation data can be of strategic importance in huge research organisation's or defence

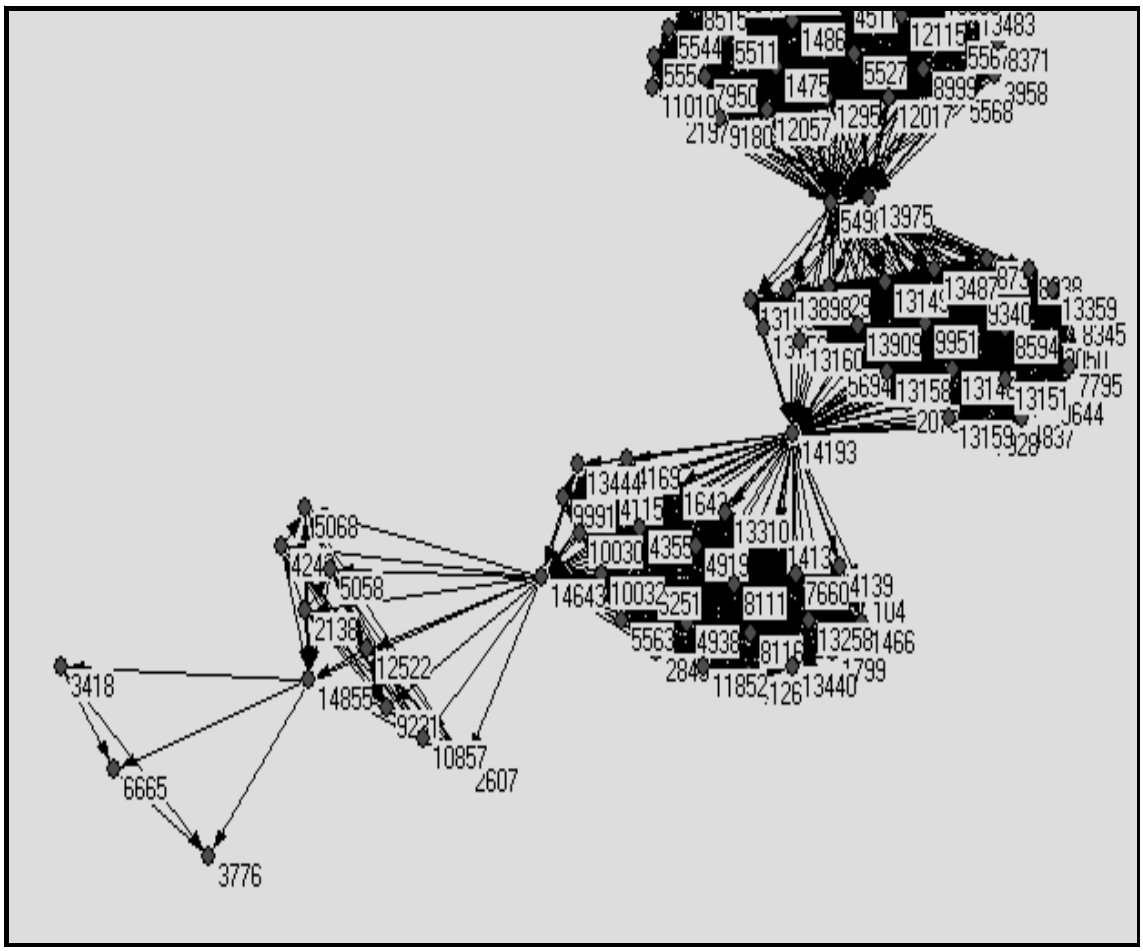

Figure 11. Graph for 10 users' books with labelled nodes.

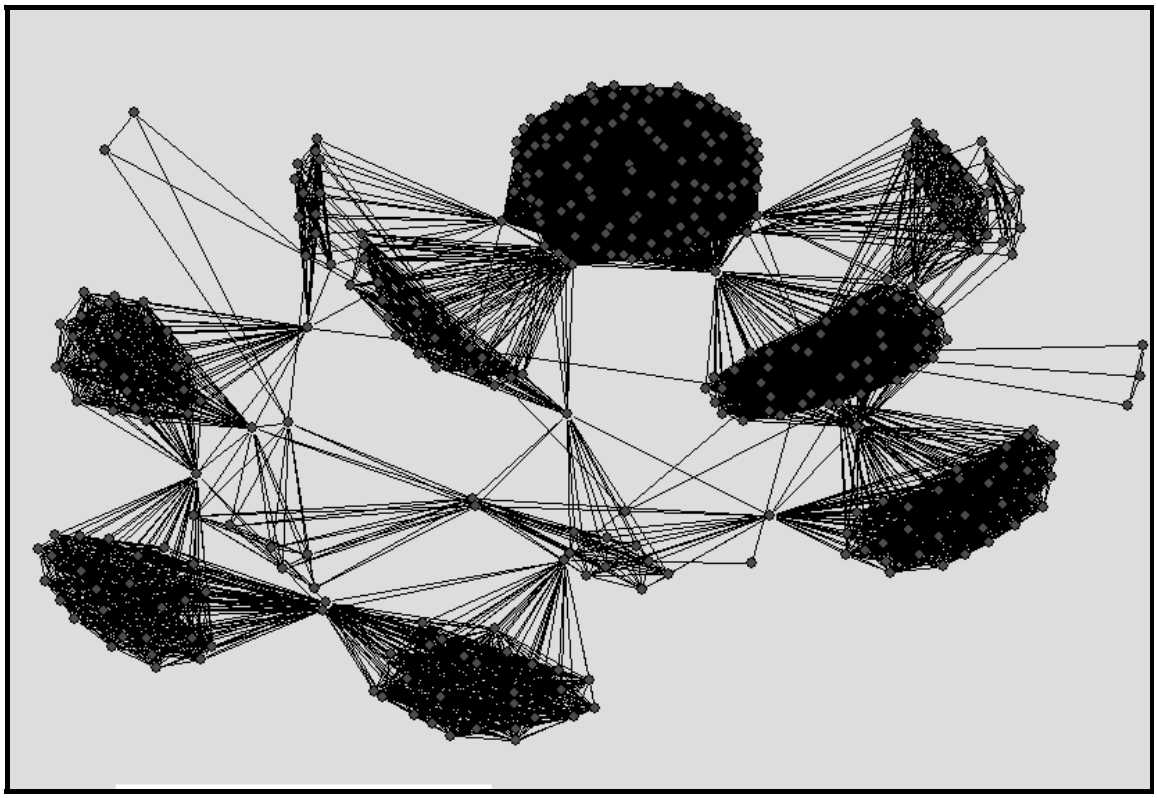

Figure 12. Graph for 20 users' books.

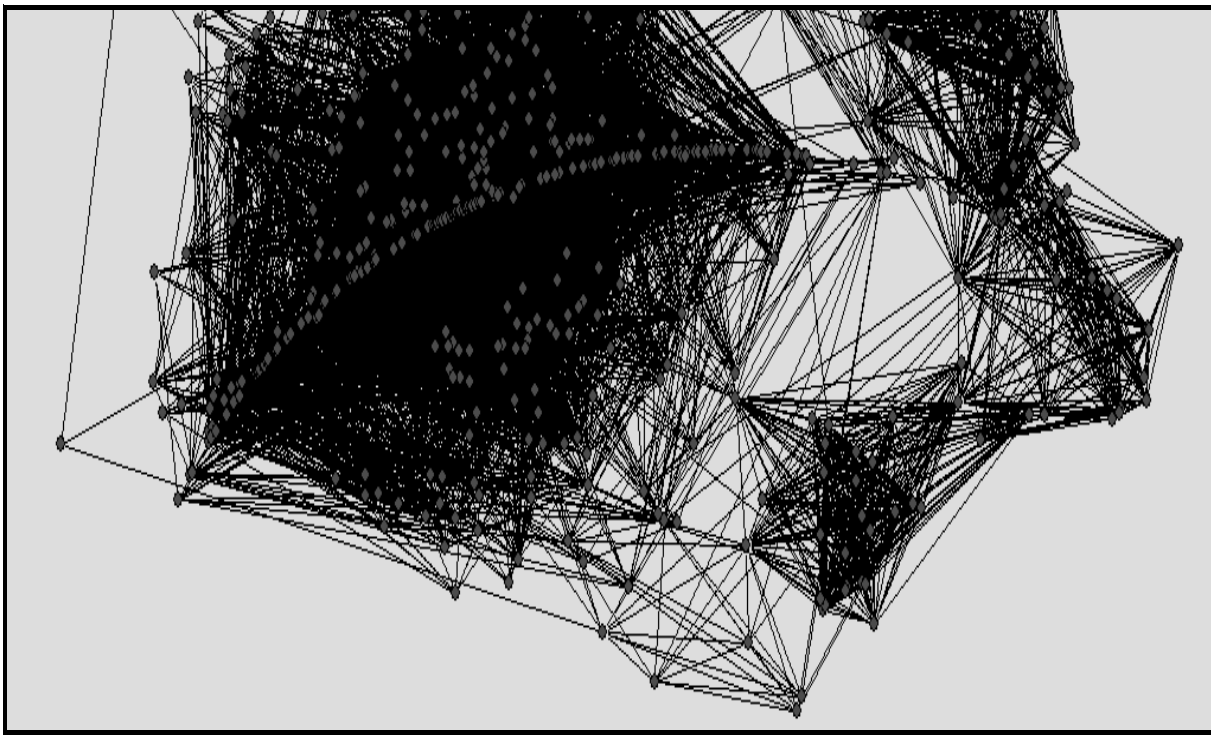

Figure 13. Graph for 30 users' books.

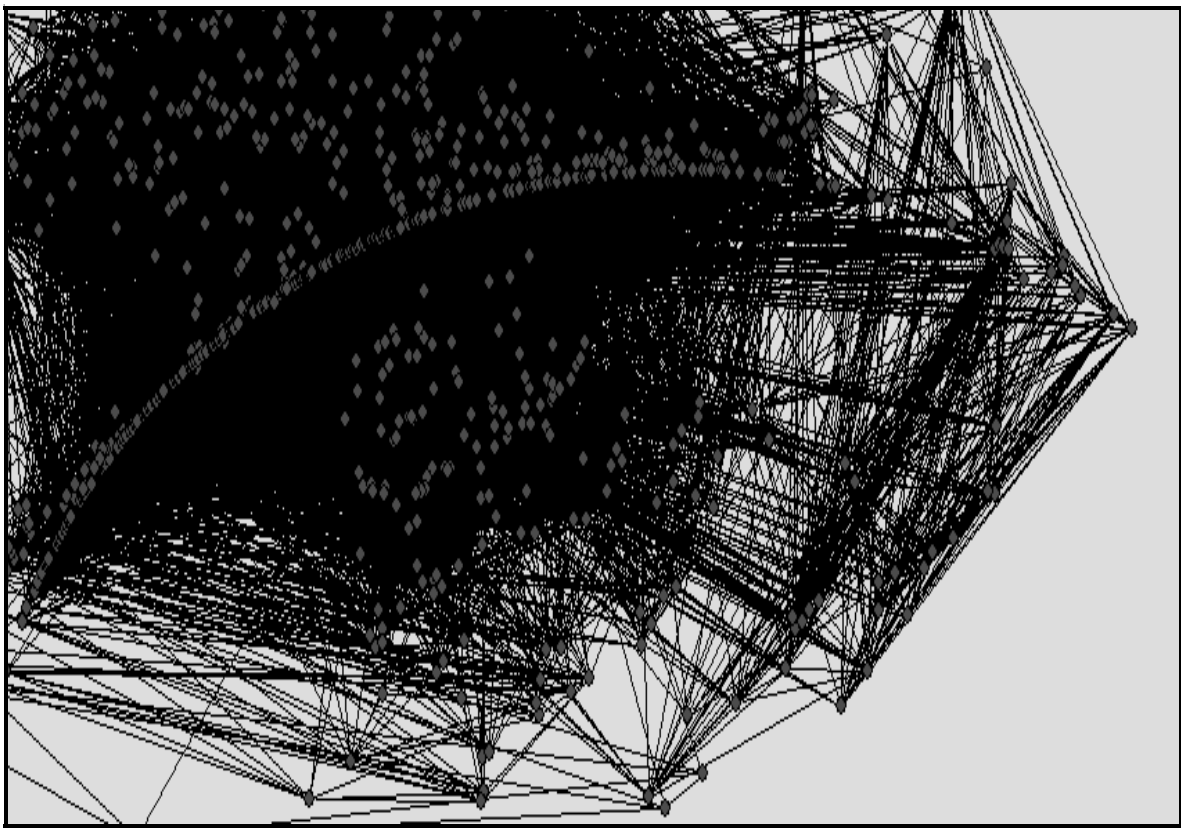

Figure 14. Graph for 50 users' books.

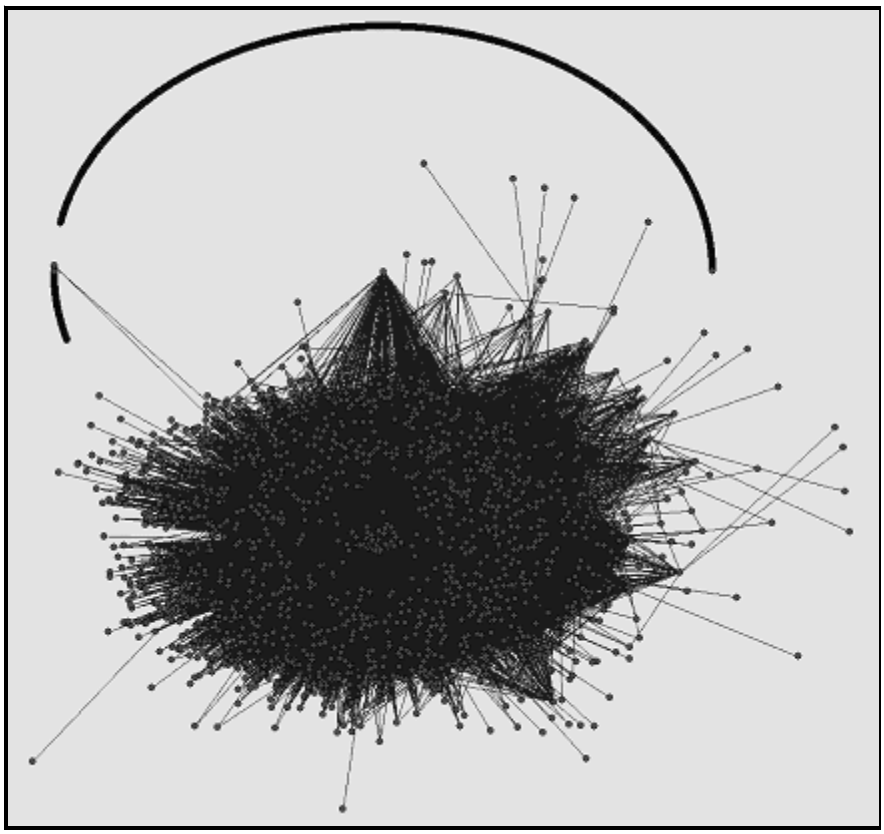

Figure 15. Graph depicting relationship between 100 users connected based on class number.

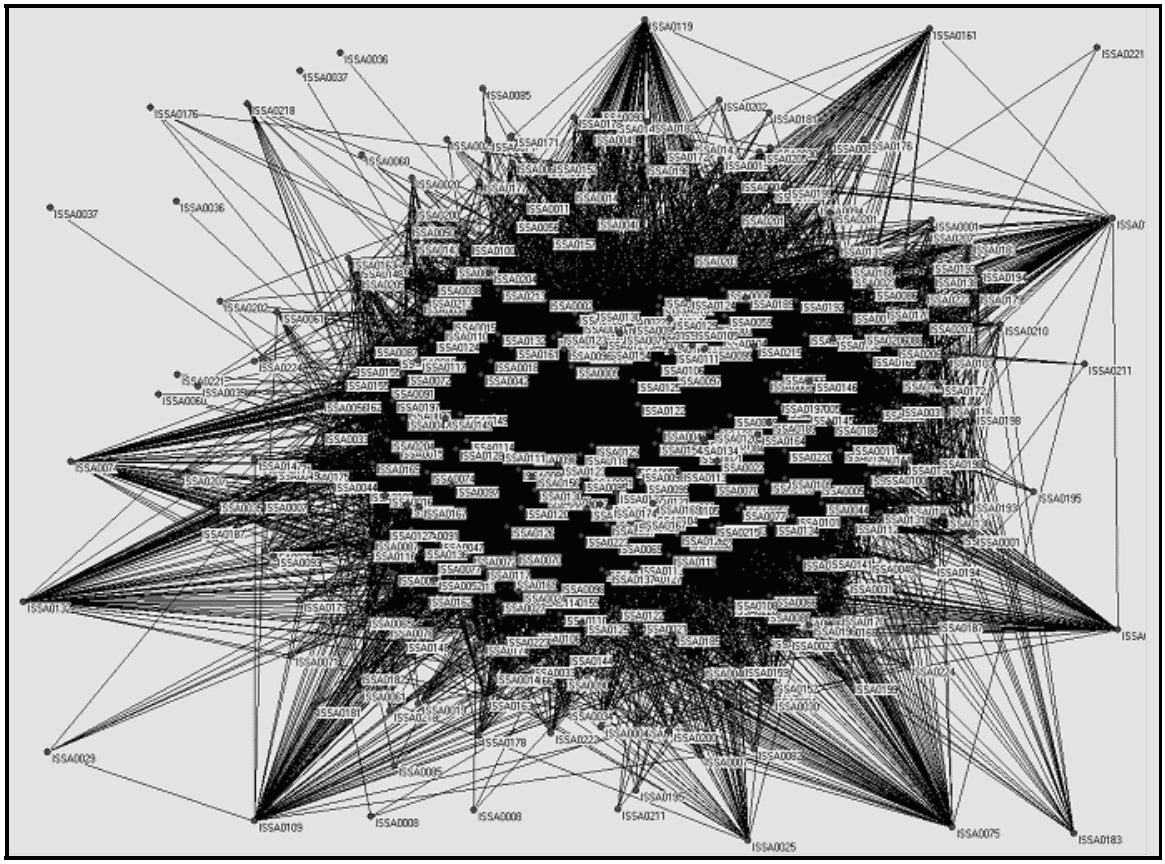

Figure 16. Users from one particular lab connected on circulation pattern.

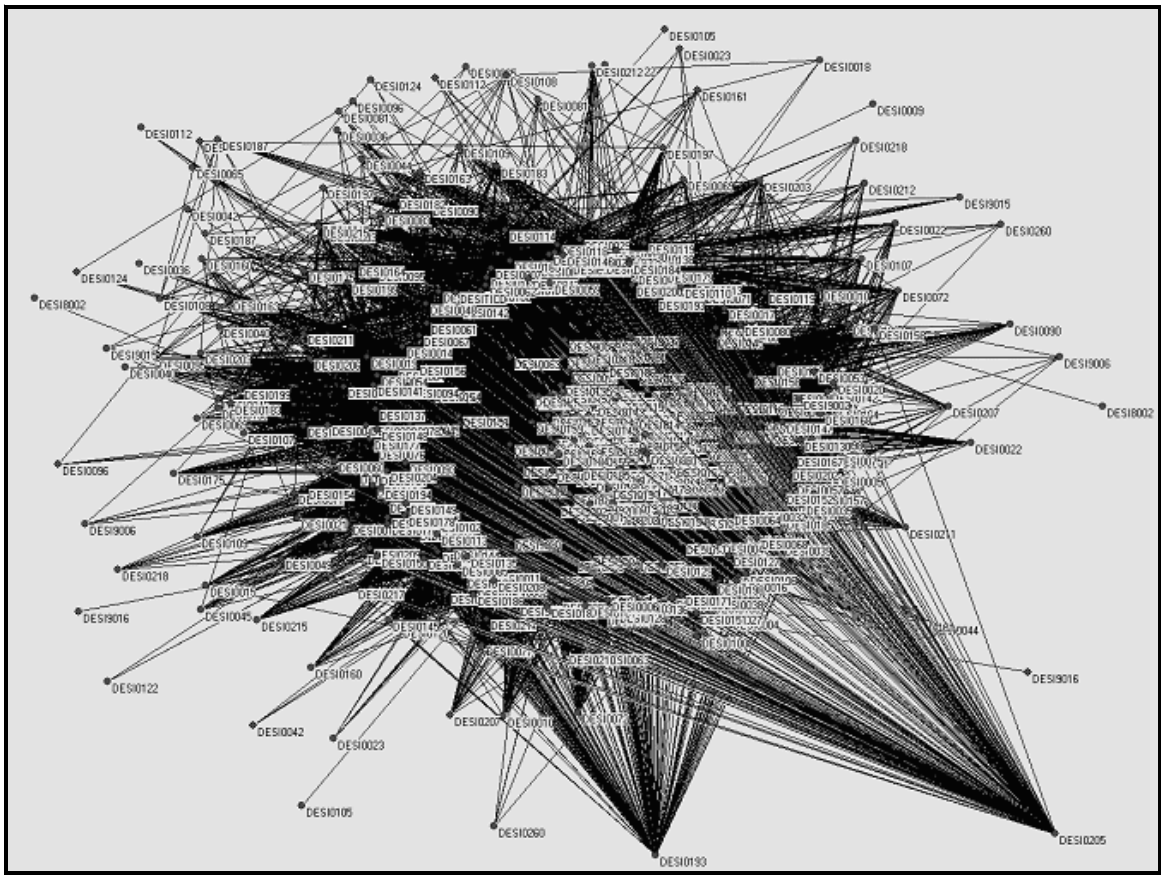

Figure 17. Subject-based community formed among all DESIDOC library users.

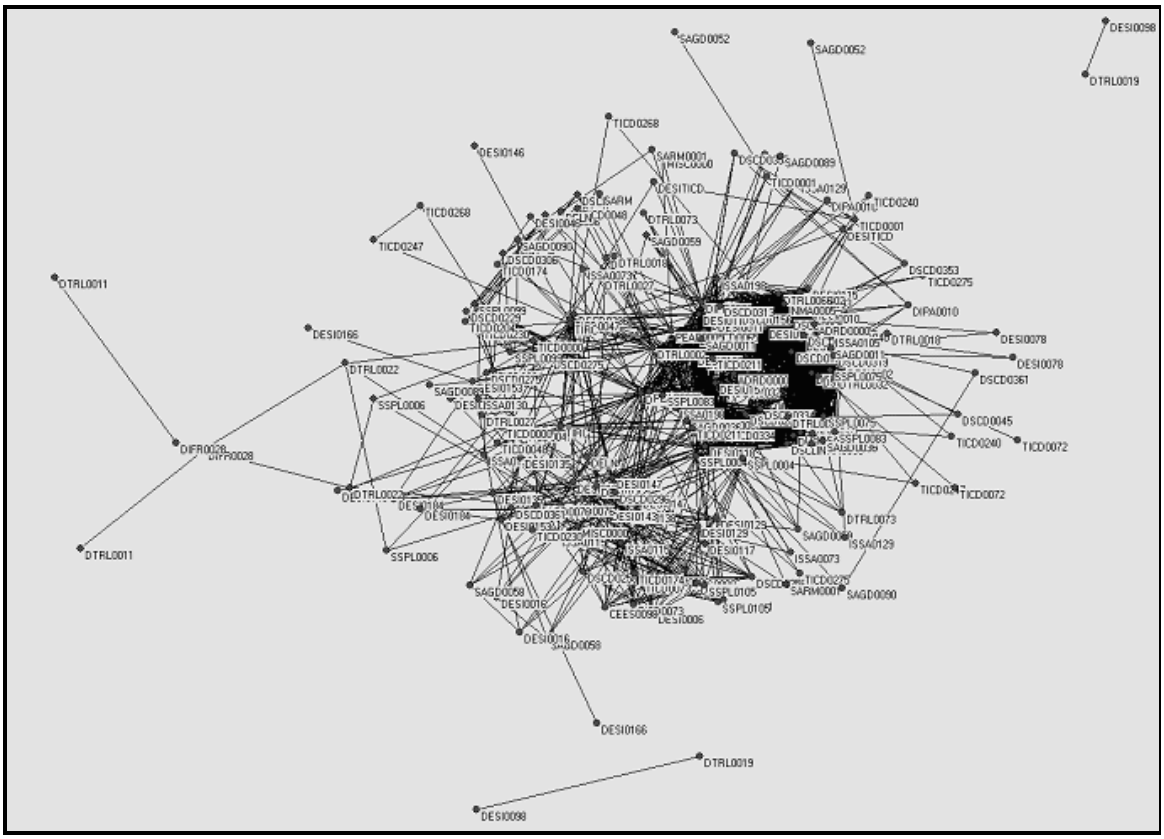

Figure 18. Relationship of users based on the usage of books on Philosophy and Psychology.

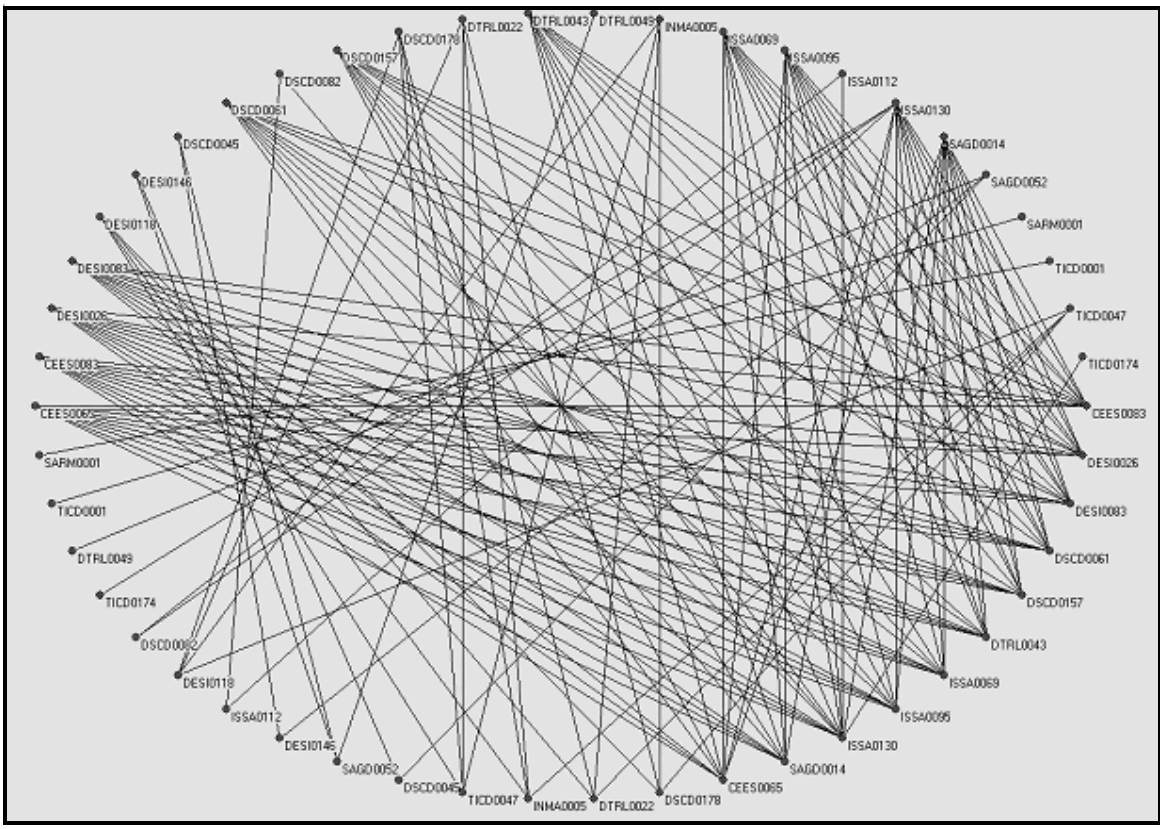

Figure 19. Relationship among users based on issue of books on Religion and Theology.

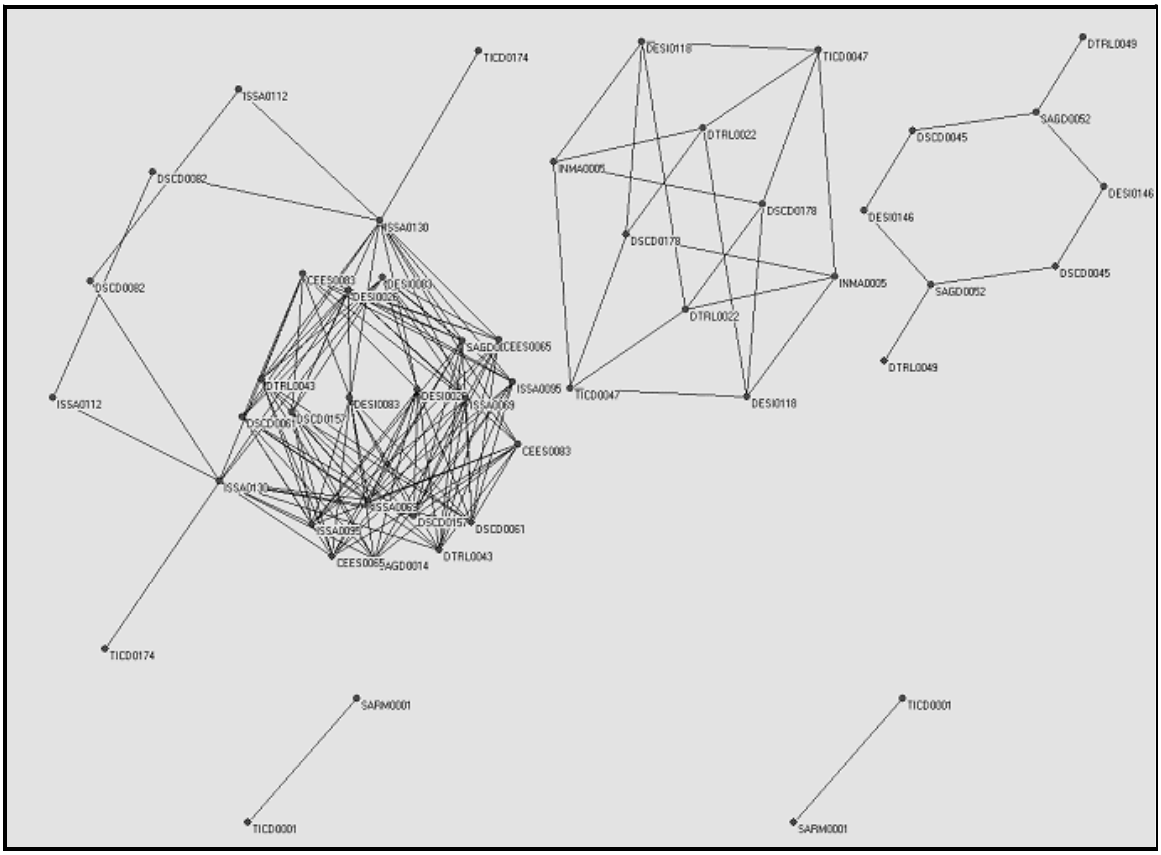

Figure 20. A separated out view of Figure 19.

establishments as it can be used to trace the researchers in a common field.

Authors intend to further visualise the entire users data and extract the patterns from the issue list. A similar study is also being done by forming a graph representing each library user by a node and connecting all those users which have a common book issued among them. This may also lead to a similar clustering of library users on the choice of their subject area or research imitativeness.

REFERENCES

1. Bigus, J.P. Data mining with neural networks, McGraw-Hill Inc, 1996.
2. Ponniah, P. Data warehousing fundamentals: A comprehensive guide for IT professionals. John Wiley and Sons, 2001. pp. 399-426.
3. Papamichail, G.P. & Papamichail, D.P. Real-time concept creation in distributed e-learning systems. *In* IASTED Int. Conf. On Web Education, Innsbruck, Austria, 2004. pp. 631-36.
4. Levenson, A.R. Leveraging adversity for strategic advantage. *Organisational Dynamics,* 2002, **31**(2), 165-76.
5. Genestre, A.; Herbig, P. & Shao, A.T. What does marketing really mean to the Japanese? *Market. Intelli. & Plann.* 1995, **13**(9), 16-27.
6. Ellatif, A. & Mohamed, M. Association rules technique to evaluate software users satisfaction. http://ssrn.com/abstract=1078506, (accessed on 07 September 2009).
7. Singh, S.; Vajirkar, P. & Lee, Y. Context-based data mining using ontologies. *In* International Conference on Conceptual Modeling, Chicago, 2003, **2813**. pp. 405-18.
8. Magnusson, C. & Vanharanta, H. Visualising sequences of texts using collocational networks. *In* Proceedings of Third International Conference of Machine Learning and Data Mining in Pattern Recognition, MLDM 2003, 5-7 July 2003, Leipsig, Germany. pp. 276-83.
9. Shibata, N.; Kajikawa, Y.; Takeda, Y. & Matsushima, K. Detecting emerging research fronts based on topological measures in citation networks of scientific publications, *Technovation*, 2008, **28**(11), 758-75.
10. Ichise, R.; Takeda, H. & Ueyama, K. Community mining tool using bibliography data. *In* Proceedings of Ninth International Conference on Information Visualisation, IV'05, London, 2005. pp. 953-58.
11. Khambatti, M.; Ryu, K. & Dasgupta, P. Structuring peer-to-peer networks using interest-based communities. *In* Proceedings of the International Workshop On Databases, Information Systems and Peer-to-Peer Computing, P2PIR'06, Berlin, Germany, 2003. pp. 1-16.
12. Ichise, R.; Takeda H. & Ueyama, K. Exploration of researchers' social network for discovering communities. *In* Proceedings of JSAI workshops, Kitakyushu city, Japan, 2005. pp. 458-69.
13. Pajek—Program for large network analysis. http://vlado.fmf.uni-lj.si/pub/networks/pajek/ (accessed on 07 September 2009).

About the Authors

Mr Sumit Goswami holds an MTech degree in Computer Science and Engineering from IIT Kharagpur, a Post Graduate Diploma in Journalism and Mass Communication, a Bachelors degree in Library and Information Science, and BTech in Computer Science and Engineering. He has been working as a Scientist with Defence Research and Development Organisation (DRDO), Delhi, India since 2000. His areas of interest include network centric operations, mobile ad hoc and sensor networks, web-hosting security, text mining and machine learning. He has published 41 papers/chapters in various journals, books, conferences and seminars. Sumit also chaired a session on the theme Computer Architecture in National Conference on Emerging Principle and Practice of Computer Science held at GND Engineering College, Ludhiana, India in August 2006.

Mr Anandarup Mukherjee is a final year student of BTech in Electronics and Communications Engineering from IMPS College of Engineering and Technology (under West Bengal University of Technology), Malda, West-Bengal. He has also done diploma courses in C, C++, Java, Visual Basic from Centre for Development of Advanced Computing(C-DAC), a certificate course in microcontrollers from Eastern Regional Testing Labs (East), Kolkata, and Digital Image Processing from IIT Kharagpur. His areas of interest include microcontrollers and microprocessors. He also takes keen interest in programming with knowledge of multiple languages. He is also constantly creating microcontroller-based projects and developing various projects in MATLAB and Visual Basic.

Ms Mansi Kharbanda is a final year student of BTech in Computer Science and Engineering from Indraprastha University, Delhi. Her areas of interest include software engineering, mobile communication and computer networking. She has attended workshops on Open Solaris Introduction and Java Beans organised by SUN CLUB, IGIT. Manshi is an active participant in IEEE and CSI society as a student member. She takes keen interest in programming and algorithm implementation.

Mr Abhinav Gupta is a final year student of BTech in Computer Science from NIEC, GGSIPU, Delhi. He has a good hand in handling programs related to C, C++, and Perl programming languages. Abhinav was ranked among top 0.1 per cent students who participated in class Xth CBSE Board Examination (2004) securing 100 per cent in Mathematics. He was also awarded a Merit-cum-Scholarship by AISSE under National Scholarship Scheme. Recently, his works related to Weka-Data Mining Tool got space as an article in leading IT magazine PC-Quest. He is keen on learning new things related to the field of data mining and knowledge extraction.

Mr Pulkit Soni is pursuing BTech in Information Technology from Northern India Engineering College, (affiliated to Guru Gobind Singh Indraprastha University), Delhi. His areas of interest include Data Mining, Computer Networking, Mobile Communication, and ERP. Pulkit has coordinated Technical Events (FNORD'09) in his college and has also attended seminars like Interactions'08 and Technophilia organised by Northern India Engineering College.