

Visualizing Users, User Communities, and Usage Trends in Complex Information Systems Using Implicit Rating Data

Seonho Kim

Dissertation submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Advisory Committee:

Edward A. Fox, Chair
Weiguo Fan
Christopher North
Deborah Tatar
Ricardo da Silva Torres

April 14, 2008
Blacksburg, Virginia

Keywords: digital library, information system, visualization, 5S, social network, user modeling, personalization, implicit data, quantity-based evaluation

Copyright © 2008 Seonho Kim
All Rights Reserved

Visualizing Users, User Communities, and Usage Trends in Complex Information Systems Using Implicit Rating Data

Seonho Kim

ABSTRACT

Research on personalization, including recommender systems, focuses on applications such as in online shopping malls and simple information systems. These systems consider user profile and item information obtained from data explicitly entered by users. There it is possible to classify items involved and to personalize based on a direct mapping from user or user group to item or item group. However, in complex, dynamic, and professional information systems, such as digital libraries, additional capabilities are needed to achieve personalization to support their distinctive features: large numbers of digital objects, dynamic updates, sparse rating data, biased rating data on specific items, and challenges in getting explicit rating data from users. For this reason, more research on implicit rating data is recommended, because it is easy to obtain, suffers less from terminology issues, is more informative, and contains more user-centered information.

In previous reports on my doctoral work, I discussed collecting, storing, processing, and utilizing implicit rating data of digital libraries for analysis and decision support. This dissertation presents a visualization tool, VUDM (Visual User-model Data Mining tool), utilizing implicit rating data, to demonstrate the effectiveness of implicit rating data in characterizing users, user communities, and usage trends of digital libraries. The results of user studies, performed both with typical end-users and with library experts, to test the usefulness of VUDM, support that implicit rating data is useful and can be utilized for digital library analysis software, so that both end users and experts can benefit.

To my father *Kiro Kim* and mother *Heeja Lee*
To my brothers

Acknowledgements

First and foremost, I thank my advisor Dr. Edward A. Fox, for his warm guidance and advice not just on academic knowledge but also about the right way of thinking and studying. I admire his passion for new ideas and enthusiasm for work. I admire his being gentle in appearance, but sturdy in spirit. I would have never survived my graduate program without his endless support. I also give heartfelt thanks to my other committee members, Dr. Fan, Dr. North, Dr. Tatar, and Dr. Torres for their excellent suggestions regarding research directions. They offered many helpful critiques and thoughtful comments on my work that lead to improving the quality of this dissertation.

Also, I thank members of the Digital Library Research Laboratory (DLRL) for their words of encouragement, and helpful suggestions. These people from the DLRL include, but are not limited to, Venkataraghavan, Yi, Yifei, Xiaomo, Xiaoyu, Pengbo, Ming, Bing, Dr. Weon, Dr. Chakraborty, Srinivas, Ryan, Yuxin, Wensi, Baoping, Lokeya, Hussein, Douglas, Johnny, Vikram, Rohit, and Unni. I specially thank Uma, Seungwon, Xiaoyan, and Rao for their valuable discussion, comments and being domain experts to verify my experiments. I don't want to forget those who have worked together with me. These people include, Nithiwat, Nicholas, Kapil, Subodh, Sandi, and Sheng. Their co-work formed the basis of my doctoral work.

In addition, I thank my good Korean friends in Torgersen Hall and the CS department, Jaehoon, Ji-Sun, Jason, Yongju, Yoon-Soo, Hannah, Si-Jung, Taehyuk, Joon Seok, Sunghee, Seung In, Myoungkyu, Jae-seung, and Haeyong, for helping my experiment. Special thanks go to Kibum, Pilsung and Dong Kwan. They took care of me brotherly and encouraged me to overcome many difficulties.

Also, I must thank all my advisors that helped me until now, Dr. Jong-Hyeok Lee, and Dr. Geunbae Lee, at POSTECH, Dr. Dong-Yul Ra, Dr. Jong Hyun Kim, and Dr. Jung-Bong Suk, at Yonsei University, Dr. Sun-Hwa Hahn, and Dr. Taehee Kim (now moved to Yongsan Univ.) at KISTI. I also thank Dr. Kandogan, at IBM Almaden Research Center, for providing me a great research opportunity and for his effort toward patent work.

Finally, I thank my parents and my brothers for their love and patience. Their continuous support and trust were chief factors in my success.

Table of Contents

ABSTRACT	ii
Acknowledgements	iii
Table of Contents	v
List of Figures.....	viii
List of Tables	x
Chapter 1. Introduction.....	1
1.1 Overview and Motivation	1
1.2 Problem Statement.....	3
1.2.1 Needs for Personalized User Interfaces	3
1.2.2 Difficulties in Classifying Digital Library Materials.....	3
1.2.3 Limitations of Available Data.....	4
1.2.4 Lack of Analysis Tools for Digital Libraries	5
1.2.5 Needs for User-Centered Evaluation	6
1.3 Research Objectives.....	6
1.4 Research Question and Hypotheses	7
1.5 Contributions.....	8
1.6 Organization of this Dissertation	8
Chapter 2. Review of the Literature.....	11
2.1 Data Mining Applied to the Web.....	11
2.2 Digital Library Behavior and Usage Analysis	12
2.3 Recommender System for Digital Library.....	13
2.4 Visualizing Social Networks, Documents, and Topics.....	13
2.5 User Modeling	16
2.6 Implicit Rating Data.....	16
2.7 Evaluation of Information Visualization	17
Chapter 3. 5S Perspective of Digital Library Personalization	18
3.1 Society.....	19
3.2 Scenario.....	19

3.3	Structure	20
3.4	Space	20
3.5	Stream	21
Chapter 4.	Collecting Data and Preprocessing.....	22
4.1	User Tracking System: Collecting Implicit Rating Data	22
4.1.1	About NDLTD	23
4.1.2	User Tracking System.....	23
4.2	User Survey: Collecting Explicit Rating Data	24
4.3	User Models: Storing Raw Data and Statistics	24
4.4	Using Legacy Data: VTLS NDLTD log data	26
Chapter 5.	Visualization and Knowledge Finding	28
5.1	VUDM: A Visualization Tool	28
5.1.1	Need for a Visualization Tool for Digital Library Users	28
5.1.2	Data Description and Preprocessing	29
5.2	Visualization Strategies	31
5.3	Loose Grouping Algorithm.....	36
5.4	Knowledge Finding.....	38
5.4.1	User Characteristics and Relations	38
5.4.2	Virtual Interest Group and Relations	40
5.4.3	Usage Trend	41
5.4.4	Concept Drift	42
5.5	User Evaluation.....	43
Chapter 6.	Formative Evaluation	44
6.1	Evaluation Design.....	44
6.2	Results and Conclusions	45
Chapter 7.	Algorithm Evaluation and Enhancing Scalability	46
7.1	Community Finding Performance.....	46
7.2	Building a Test Collection	47
7.3	Document Classification and Noun Phrasing	49
7.4	Application: Measuring Accuracy of Community Finding Algorithms	52
7.5	Enhancing Scalability of FSWMC.....	57

7.6	Results and Conclusions	63
Chapter 8.	Summative Evaluation.....	67
8.1	Experiment Design.....	67
8.2	General Data Obtained.....	69
8.3	Rating Data Obtained.....	71
8.3.1	Accuracy	71
8.3.2	Efficiency	76
8.4	Conclusions.....	82
Chapter 9.	Experts Review	84
9.1	Study Design	84
9.2	General Information Obtained	85
9.3	Ratings and Comments for Enhancement.....	86
9.4	Conclusions.....	88
Chapter 10.	Conclusions and Future Work.....	89
10.1	Deliverables	89
10.2	Conclusions.....	89
10.3	Future Work	91
References		92
Appendix A-1: Call for Participation Advertisement.....		99
Appendix A-2: Questionnaire		100
Appendix A-3: Institutional Review Board Approval Letter		106
Appendix B-1: Call for Participation Advertisement.....		107
Appendix B-2: Questionnaire		108
Appendix B-3: Online Tutorial of VUDM for Library Experts Survey		113
Appendix B-4: Institutional Review Board Approval Letter		129

List of Figures

Figure 1. A tag cloud with terms related to “5S Model”	14
Figure 2. A Themescape with documents related to the Virginia Tech April 16 incident	15
Figure 3. A GalaxyView with documents related to the Virginia Tech April 16 incident.....	16
Figure 4. 5S perspective of digital library personalization	18
Figure 5. Structure of user model data	20
Figure 6. A JavaScript based user interface.....	22
Figure 7. System diagram of user tracking system.....	24
Figure 8. XML schema of user model data	26
Figure 9. Iteration of data mining process using visualization tools	29
Figure 10. The main window of VUDM shows an overview of virtual user communities	32
Figure 11. Windows for detailed information about users and user communities	33
Figure 12. Zoomed user space	34
Figure 13. Filtering user space	35
Figure 14. Detailed information on demand	36
Figure 15. Illustration of the kNN (top) and FSWMC (bottom) algorithms	37
Figure 16. Fixed-size window multi-classification algorithm	38
Figure 17. User and user community characteristics and relations	39
Figure 18. Visualizing usage trends of digital library	41
Figure 19. Detecting drift of concepts makes it possible to provide timely recommendation.	42
Figure 20. Workflow to build a collection of research interests and learning topics.....	48
Figure 21. Hierarchical structure of the collection	49
Figure 22. An example of ETD metadata	50
Figure 23. Visualization of user communities in the NDLTD collection as of June 2006. User communities were found by kNN (top), FSWMC (middle), and Tanimoto (bottom) algorithm.	53
Figure 24. EstimateCategories Function	55
Figure 25. Counting shared interests using the output of EstimateCategories Function	56
Figure 26. Change of the number of total users and active users.....	58
Figure 27. Change of the number of active users in terms of DOA, zoomed-in from Figure 24.	58
Figure 28. Completion time of FSWMC runs in $O(n^2)$	59
Figure 29. Completion time of FSWMC runs in $O(n\sqrt{n})$	60
Figure 30. Completion time of FSWMC runs in $O(n)$	61
Figure 31. Comparison of completion times of the FSWMC and its two variations	62
Figure 32. Completion times of the kNN , Tanimoto, FSWMC and its two variations	62
Figure 33. Accuracies of five community finding algorithms	63

Figure 34. Accuracy and size of community.....	65
Figure 35. Academic level of participants	70
Figure 36. Age distribution of participants	70
Figure 37. Preferred statistical analysis software package	71
Figure 38. Average ratio of correct, given-up, and wrong answers of VUDM (left bar), and other software packages (right bar)	72
Figure 39. Average ratio of correct, given-up, and wrong answers of VUDM (left bar), and Excel (right bar)	72
Figure 40. Average accuracy and confidence intervals of overall user tasks	73
Figure 41. Averages accuracy and confidence interval of each user task.....	75
Figure 42. Averages of overall completion time and confidence intervals.....	76
Figure 43. Average completion time and confidence interval of each user task	78
Figure 44. Average scores and confidence intervals of two different academic levels in terms of software used (left), and in terms of academic levels of participants (right)	81
Figure 45. Average completion times and confidence intervals of two different software within academic levels (left), and academic levels within two different software (right)	82
Figure 46. IRB approval letter for the summative evaluation	106
Figure 47. Java version checking	115
Figure 48. VUDM demo is installed under C:\VUDM	116
Figure 49. Invoke VUDM.....	116
Figure 50. Open new data in current window	117
Figure 51. Open new data in a new window.....	117
Figure 52. Select a data folder	118
Figure 53. Data loading	118
Figure 54. User data of February 2006 is successfully loaded and visualized	119
Figure 55. VUDM overview	119
Figure 56. VUDM visualization strategy	121
Figure 57. Zoomed image of a community	122
Figure 58. Control the strictness of community finding by adjusting the θ value	123
Figure 59. User community of January 2006 ($\theta = 0.2$)	124
Figure 60. User community of January 2006 ($\theta = 0.18$)	125
Figure 61. User community of January 2006 ($\theta = 0.15$)	126
Figure 62. User community of February 2006 ($\theta = 0.15$)	127
Figure 63. Details on demand	127
Figure 64. IRB approval letter for the experts review	129

List of Tables

Table 1. Data set: User model data used for VUDM formative evaluation, from 1,200 users	30
Table 2. 102 categories of research interests and learning topics.....	50
Table 3. Raw data for results of analysis.....	51
Table 4. Raw data for evaluation	56
Table 5. Two task sets containing five user tasks.....	68
Table 6. t-test result on accuracy of overall user tasks between VUDM and other software packages (top), and between VUDM and Excel (bottom)	73
Table 7. Average accuracy and confidence interval of each user tasks	75
Table 8. t-test result on the accuracy of each user task between VUDM and other software packages	76
Table 9. t-test result on overall completion time between VUDM and other software packages (top), and between VUDM and Excel (bottom)	77
Table 10. Average completion time and confidence interval of each user task.....	79
Table 11. t-test result on the completion time of each user task between VUDM and other software packages.....	79
Table 12. Results Interpretation.....	80
Table 13. Pros and Cons of VUDM.....	80
Table 14. The effect of academic level	82
Table 15. Importance and uses of information about users, user communities, and usage trends	85
Table 16. Average rating of VUDM in providing three types of information	87
Table 17. Average rating of VUDM in providing information for five user task types.....	87
Table 18. Comments for enhancement of VUDM.....	87

Chapter 1. Introduction

In this chapter, we review traditional methods for analyzing users, communities, and usage of online shopping malls and small information systems. We describe the problems in applying these methods to complex information systems, which motivated the need for this research. We then present the research objectives, hypotheses, and expected contributions of this research.

1.1 Overview and Motivation

The World Wide Web (WWW) was invented for easy publishing and easy access to content, from anywhere, through the Internet. WWW technologies had been developed to support two distinct user groups, content providers and content consumers. However, with the huge success and popularity of the WWW, the boundary between content providers and consumers is shrinking, leading to dynamic two-way WWW services – such as blogs, wikis, online journals, online forums, etc. – where people can publish and access content easily. This trend changed WWW users from passive anonymous observers to active visible individuals with personalities [51, 79].

Studying users of such dynamic services is providing opportunities for research on collaborative filtering, personalization, user modeling, and recommender systems, which will eventually enhance the quality of service of the systems. The basic ideas of these technologies are to analyze users' properties and logs, to classify the users, to predict their goals for using the systems, and to help them to accomplish their goals. Many information systems, such as travel guide systems, DVD-based information systems, and online shopping malls, are already conducting such user studies based on log analysis and online or offline user surveys. However, studying users in complex information systems like digital libraries is difficult; this cannot be solved by the techniques used for simple information systems and online shopping malls.

Many patrons of digital libraries are researchers and distance learners, and their information needs correspond to very narrow topics, which are hard to classify and which change dynamically as time passes. For example, in a small information system, e.g., a

DVD search system, user's preferences for movies can be analyzed easily from available data – such as rental history, purchase history, search logs, and user supplied information (e.g., from registering) – because movies are well classified by their genre, director, language, actors, year of production, Motion Picture Association of America (MPAA) rating [56], producer, etc. Thus, features of users can be estimated from the features of movies they accessed. However, the contents of digital libraries, such as dissertations, theses, conference papers, journal papers, project reports, and technical books, are not classified precisely enough to help describe users. This is not only because there is no proper classification system for such detailed levels of all scholarly areas, but also because the classes in the detailed levels change dynamically and overlap with other classes.

Characterizing users of digital libraries with the features extracted from roughly classified contents is not useful for personalized digital library services. For example, two users, who are classified under “Computer Science > Data & Information > Digital Library”, may have different research interests because there are so many detailed topics under this class. Similarly, a paper useful to one user may not be interesting to another user. Also, the user preference information explicitly provided by users may not be informative because of terminology differences and/or the breadth of topical coverage of digital libraries. Even people with the same research interests express their interests with different terms, while the same terms sometimes represent different research fields. Also, as Jones [37] observed in his digital library analysis research, explicitly entered information may contain spelling errors or unusual abbreviations that make it difficult to analyze. Further, collecting users' ratings of each item they access in digital libraries is expensive or infeasible.

Much research on personalized services in digital libraries, however, follows the methods used for small information systems and online shopping malls, ignoring the distinctive characteristics of complex information systems explained above. Therefore, we have an urgent need to develop techniques to analyze users, user communities, and usage of digital libraries, which will allow us to detect dynamic changes of users' characteristics and features, without depending on the users' explicit ratings and traditional static classification systems.

1.2 Problem Statement

Analyzing users, communities, and usage for personalization is more difficult in digital libraries than in online shopping malls and small information systems, as mentioned in the previous section. Below we introduce further the features and problems that motivate this dissertation work.

1.2.1 Needs for Personalized User Interfaces

Digital libraries provide large amounts of information. One of the goals of user interface design for digital libraries is to reduce users' efforts in finding useful items that match their purpose. There have been two especially popular directions of research in user interface design for information systems. One aims to increase the degree of precision and reduce the effort in searching, by using: intricate searching algorithms, ranking methods, classification and clustering techniques, visualization techniques, etc. The other direction involves supporting various types of content and media. Examples include querying interfaces for searching images, music, speech, chemical formulas, or DNA [93].

As information systems become bigger and more dynamic and complex, like digital libraries, personalization becomes ever more important. Personalized user interfaces, often integrated with recommender systems, are needed, because users in complex information systems want to obtain up-to-date, newly arrived, hot documents as well as ordinary documents that are relevant to their queries. Such interfaces are especially helpful for researchers and distance learners, because they often need information outside their sphere of knowledge. This dissertation work should lead to important improvements in personalized recommender-capable user interfaces for digital libraries.

1.2.2 Difficulties in Classifying Digital Library Materials

There are many standard library classification systems aimed to help organize library materials. The Library of Congress Classification [21], Dewey Decimal Classification [26], and Dickinson Classification [18] are among the most common classification systems in the English speaking world. More classification systems are used elsewhere.

Examples include the Korean Decimal Classification (KDCP) in Korea [63], Nippon Decimal Classification (NDC) in Japan [11], *Principles de Classement des Documents Musicaux* in France, Chinese Library Classification (CLC) in China [98], etc. However, these library classifications are not suitable for our purposes. To accomplish our goal of implementing collaborative filtering and recommendation in digital libraries, where the preferences of other persons in the same group are referred to for personalized service, we need to learn users' preferences for information from the properties of library materials they access. User preferences and interests in research and learning often are very narrow, complex, comprehensively overlapping over multiple disciplines, and changing dynamically moment by moment. More detailed classifications will provide more information about the materials. However, because the library classifications (such as those mentioned above) are not designed for this purpose, they do not classify library materials down to the very detailed level. Also, other user requirements we must accommodate, such as the tendency for dynamic change and the overlapping of classes, make it more difficult to use the library classifications for our research.

1.2.3 Limitations of Available Data

A number of different types of data are required to analyze users, user communities, and usage trends in information systems and online shopping malls. Web log data [72], which is generated by HTTP daemons, server side, contains useful system-oriented information, such as size, name and type of transferred object, access time, error code, client address, object location in the file system, etc. However, web log data shows only system information and doesn't show anything on the user side, such as research interests, learning topics, community involvement, preferences, etc.

Studying user centered digital libraries can support research on collaborative filtering [64, 67], personalization [28, 67], user modeling, and recommender systems. Most such studies consider users' ratings of the information they select, as well as users' preferences – e.g., on research areas, majors, learning topics, or publications – which are entered explicitly [53, 59, 96]. This explicit rating data is collected through online or offline user surveys or by asking users to rate materials whenever they finish using them. However,

obtaining explicit rating data is expensive and burdensome to users. Further, explicit rating data has a shortcoming caused by terminological confusion [68, 70] and digital libraries' broad topical coverage. This problem poses serious challenges regarding the identification of users' research and learning interests. Even people with the same research interests express those interests with different terms, while the same terms sometimes represent different research fields. For these reasons, we need other evidence to help distinguish users' research interests, without depending on their written comments. Another approach is using implicit rating data, which avoids the problems of explicit rating data described above. However, implicit rating data has been regarded meaningful only when used together with explicit rating data as an auxiliary source of information [50, 58], and its effectiveness has not been proven yet.

1.2.4 Lack of Analysis Tools for Digital Libraries

There are many challenging research questions of interest to digital library administrators and users, such as:

- What are the current trends of information seeking for this digital library?
- What kinds of people are using this digital library?
- Who is a potential mentor for whom?
- How has the focus of retrieval changed for a particular user?
- What academic areas are emerging as popular attractions?
- How many people are interested in which topics? How many are experts?
- How many virtual groups of users, who share interests, exist in the digital library?
- Which topics are related to which other topics?

In the Business Intelligence field, On Line Analytical Processing (OLAP) tools, which are usually configured with databases, are commonly used in reporting about sales, marketing, management reporting, business performance management, budgeting, forecasting, financial reporting, etc. [9]. OLAP tools can provide specific quantity-based calculation and data analysis to support decision making. However, normal OLAP tools are not able to answer the comprehensive questions on complex information systems

listed above. To answer these questions, analysis tools for digital libraries should be able to support knowledge finding tasks related to the needs for personalization.

We may benefit from data mining and unsupervised learning techniques applied to the large volume of usage data from community-driven information systems like blogs [51], wikis [92], and other types of online journals. Visualization techniques also support direct involvement of users in exploration and data mining, so they can utilize their creativity, flexibility, and general knowledge. With these techniques, we can go beyond what is possible if only examining data from OLAP systems [40, 74].

1.2.5 Needs for User-Centered Evaluation

Evaluating digital libraries is a multi-faceted problem, which involves measuring the quality and quantity of collections, user interfaces, services, network connections, storage systems, staff actions, protocols, users and community support mechanisms, management support schemes, etc. Evaluating each factor of these requires different methods. Most of them currently are handled through quantitative approaches. However, some factors, like those involving services, staff, protocols, users, user communities, etc., are not easily evaluated by quantitative methods. Further, quantitative methods tend to downplay some of the most important evaluation factors, such as user satisfaction about services, efficiency and effectiveness in addressing users' needs, and support for users, communities, and managers.

Bollen suggested that a user centered approach to digital library evaluation, using a document relation network obtained from the retrieval patterns of a user and his communities, would solve many parts of the problems [14]. However, a user centered approach to digital library evaluation may lead to other intricate problems regarding designing, collecting, and processing of the data. There are few studies of these problems.

1.3 Research Objectives

Key goals of this dissertation work are to propose and evaluate techniques which utilize implicit rating data in complex information systems, so as to obtain useful knowledge for personalization and analysis of digital library usage, such as the relations

between users and user communities, and overview knowledge of digital library usage trends.

In particular, the objectives are to:

1. Propose techniques:
 - A. For utilizing implicit rating data for digital library analysis
 - B. For visualizing digital library information about users, user communities, and usage trends
 - C. For systematic evaluation of the accuracy of user communities, based on the similarity of their interests, found by various clustering algorithms
2. Evaluate:
 - A. The usefulness of the knowledge obtained from implicit rating data and related visualization techniques
 - B. The scalability of implicit rating data processing techniques
3. Develop:
 - A. A tool that helps with personalization of digital libraries, supporting advanced users as well as digital library engineers and administrators, that demonstrates the utility of using implicit information, and that is shown through evaluation to be useful and usable.

1.4 Research Question and Hypotheses

The research objectives lead to the following question:

Are the effects of implicit rating data great enough to help personalization problems, such as characterizing users and user communities not only in online shopping malls, but also in complex information systems, such as digital libraries and search engines, where individual interests are very narrow and specific, and items in the collection are hard to classify?

We hypothesize that:

1. Implicit rating data is effective in characterizing users, user communities, and revealing usage trends in digital library.
2. We can make use of implicit rating data to design and implement useful and usable personalization-support applications.
3. Digital library engineers and administrators can benefit from the techniques and tools we propose.

1.5 Contributions

The outcomes of this research will have a powerful impact on the fields of personalization and usage analysis of information systems, such as digital libraries, blogs, wikis, search engines, etc. The following items will be contributions of this work:

1. We will provide empirical evidence that shows the effectiveness of implicit rating data. Researchers in personalization and information system usage analysis will be able to extend their available data from explicitly obtained data, such as user survey and interactive questioning, to include implicit rating data, which is collectable without interfering with users' system use.
2. We will identify the types of implicit rating data feasible to collect in complex information systems and the methods to collect and store them.
3. We will implement a visualization tool that helps digital library administrators and service designers to find useful knowledge for personalized services from the implicit rating data.
4. We will evaluate the usefulness of the visualization tool by measuring the accuracy and efficiency of the tool. Also, we will show how to enhance the scalability, applicability, and privacy protection of the visualization tool.

1.6 Organization of this Dissertation

This dissertation consists of nine chapters and two appendixes. In this chapter, Introduction, we have discussed some current trends of web technologies, have listed problems that motivated this research, and have described our research objectives,

questions, and hypotheses. We also have presented the planned contributions of this research.

Chapter 2, *Review of Literature*, presents a detailed look at previous work that has influenced this research. Areas of coverage include: 5S digital library framework, digital library evaluation and usage analysis, user modeling techniques, personalization and recommending issues, difficulties in clustering and classification, social networks and document visualization, data mining, user tracking, and implicit rating data.

Chapter 3, *5S perspective of Digital Library Personalization*, explains digital library personalization from the 5S model perspective. Especially, personalization, usually including recommender systems, is explained through the Society and Scenario models.

Chapter 4, *Collecting Data and Preprocessing*, describes how we defined our two types of data, implicit and explicit rating data, and how we collect the data. Our approaches to user tracking and user surveying are presented. We also explain how we stored and processed the data involved.

Chapter 5, *Visualization and Knowledge Finding*, introduces our Visual User model Data Mining tool (VUDM), with which implicit rating data is used to visualize user behavior, user relations, user communities, and usage trends for digital libraries. This chapter also explains how useful knowledge can be extracted from the visualization.

Chapter 6, *Formative Evaluation of VUDM*, describes a preliminary experiment and user evaluation based on which the VUDM was modified. The evaluation was performed with graduate students majoring in computer science, most of whom have interests in the digital library research field.

Chapter 7, *Algorithm Evaluation and Enhancing Scalability*, suggests a methodology to evaluate the accuracy of user communities. The results of an evaluation of community finding algorithms, including *k*NN, Tanimoto, and FSWMC, which are explained in

detail in Chapter 5, are presented. Also, the scalability of FSWMC is tested and scalable variations are proposed.

Chapter 8, *Summative Evaluation of VUDM*, presents an experiment and its results, constituting a summative evaluation of the usefulness of VUDM. The usefulness of VUDM is assessed by measuring the accuracy and efficiency for five types of user tasks.

Chapter 9, *Experts Review* presents the procedure and results of VUDM reviews performed by library experts. Subjects are those among the faculty and staff of university libraries. Participants were provided an executable VUDM, an online tutorial, and a questionnaire to complete. They were asked to evaluate the VUDM and to compare it with their preferred library analysis tools.

Chapter 10, *Conclusions* summarizes the achievements of this doctoral work and suggests possible research directions for future work.

Finally, the *Appendixes* list documents used for experiments performed for this dissertation, such as questionnaires, call for participation advertisements, and IRB approval letters.

Chapter 2. Review of the Literature

This chapter presents a detailed look at previous work that has influenced this research. Our study ranges over various research areas, from digital library design to visual data mining.

The 5S model [30], developed for digital libraries, also can be considered as a framework for the design and deployment of a recommender system and its components. Efforts for web log standardization are reviewed and related to designing with user model data and integration with user tracking systems. Techniques for development of user interfaces, user modeling, personalization, and recommendation, suitable for online shopping malls and information systems, are surveyed with regard to their application to complex information systems like digital libraries. Document clustering and classification techniques are studied for clustering and classifying the users of digital libraries, and virtual user communities, represented by their interests in research topics, to implement collaborative filtering. Research is reviewed about social networks and information visualization techniques for complex information systems. Data mining and pattern recognition techniques are looked into, to extract important hidden knowledge from log, cache, and user model data in digital libraries. User tracking techniques, a.k.a., motion tracking, usually adapted for studies on interface design in virtual reality (VR), environments [15, 16], are used to trace users' behavior through web user interfaces. In our research, a user tracking system is considered a good source of implicit rating data. Studies on implicit rating data are reviewed to help in the design and analysis of raw data collected by the user tracking system and stored in the user model.

2.1 Data Mining Applied to the Web

Data mining is a technique for knowledge discovery, finding patterns and models in data in trade, financial, communication, and marketing organizations, which is normally used in knowledge management systems [24]. Data need to be cleaned for data mining to eliminate error and for consistency, but mining from web data, which is semi-structured, or unstructured, information, poses significant challenges because that is more

sophisticated and dynamic than what is contained in structured commercial databases [8] [90] [22]. Also, unlike normal data mining, web mining usually involves natural language processing [39].

Srivastava [76] classified 4 aspects of data for web mining.

1. Content: text, image, audio, video, metadata, and link
2. Structure: HTML, XML, link structure in the Web
3. Usage: secondary data from web log, proxy log, cookies, and users' feedback
4. User profile: demographic information of users, which is normally obtained explicitly

2.2 Digital Library Behavior and Usage Analysis

Log analysis is a key source of information about activities of digital libraries and patrons' information seeking behavior. Analyzing behavior of the digital library and the usage patterns of patrons facilitates evaluation, and opens opportunities to improvement and enhanced services of digital libraries. Transaction logs are used to analyze patterns related to information seeking, such as the type of Boolean operators, ranking and sorting mechanisms, type of queries, and selected collections for search. The log also is useful for evaluating the information system. However, despite the fact that most digital libraries have logging systems, the format and information in them vary. This causes problems in interoperability, reusability of logging systems, and comparability of the results of log analysis. Gonçalves et al. defined an XML based log standard for digital libraries [32] to capture a rich, detailed set of system and user behaviors supported by current digital library services. Klas et al. tried to collect logs from three different tiers, system, user behavior, and user-system interaction [49]. The log from the user-system interaction tier can be detailed into four levels, i.e., UI, interface, service, and conceptual. The Klas et al. logging scheme facilitates the comparison of different evaluations and supports the re-use of evaluation data. However, without extension, these log suggestions are not suitable for personalization and recommender systems, because they are focused on capturing events and behaviors of digital libraries and users rather than users' research interests and searching concepts.

2.3 Recommender System for Digital Library

Common expected roles for digital libraries are the active support of researchers, distance learners, and social communities. Many studies of collaborative filtering and adaptive techniques for recommender systems have been conducted for online shopping malls and personalized websites [62, 85]. However, classifying items and users is easy in those domains, so techniques have emphasized the support of exploring learning methods to improve accuracy instead of focusing on 1) scalability, 2) accommodating new data, and 3) comprehensibility – three important aspects of recommender systems in digital libraries [19].

Recommender systems have been studied with regard to Ungar's system for online media shopping malls for Usenet netnews systems (e.g., GroupLens [50]), and for digital libraries: PASS [97], Renda's system [67], Fan [28], and DEBORA [59]. The main idea of these studies is the sharing of user profile and preference information, which was entered explicitly by users, with other users within the system, to provide them with more productive services. PASS provided personalized service by computing the similarity between user profiles and documents in pre-classified research domains. Renda's system emphasized collaborative functions of digital libraries by grouping users based on their profiles. Fan proposed a genetic programming based two stage system for an information push service. The DEBORA project developed a user interface for collaboration in digital libraries.

2.4 Visualizing Social Networks, Documents, and Topics

Visualization techniques aid end-users of social networking services, to facilitate discovery and increased awareness of online communities. For example, visualization of networks of criminals and criminal events can help unearth hidden patterns in crime data as well as detect terrorist threats [95]. Especially, a network visualization technique, using node-link depictions, is used frequently. Heer, in Vizster [33], and Friendster [5], visualized relationships between members in an online data site by using network visualization technique. Connections among members are decided by explicitly entered information about "friends". Distance between members is given by a spring-embedding

(force-directed) layout algorithm. The main idea of this algorithm is to connect members with lower connectivity with short links and members with higher connectivity with long links at the center of the community. Boyd, working with Social Network Fragments [17], visualized clusters of contacts derived from the *to* and *cc* lists in email archives.

Visualization techniques also are used to characterize documents and topics. Most document visualizations consist of three states, analysis, algorithm operation, and visualization [75]. Generally, in the *analysis* stage, first-order statistics such as frequency-based measurement, and higher-order statistics such as user interests expressed as a set of words, as well as semantic data, are extracted. *Algorithms* generate efficient and flexible representations of documents, such as by clustering and projecting to a lower dimensional space. In the *visualization* stage, the data produced by the algorithms are encoded and presented and made sensitive to interaction. The tag cloud in Figure 1 is an example of a visualization technique to depict tagged web documents. The importance of a tag, such as the number of documents or number of hits, is visualized with font size and color. A tag cloud allows users to jump to document sets covering a desired topic or for a particular popularity.

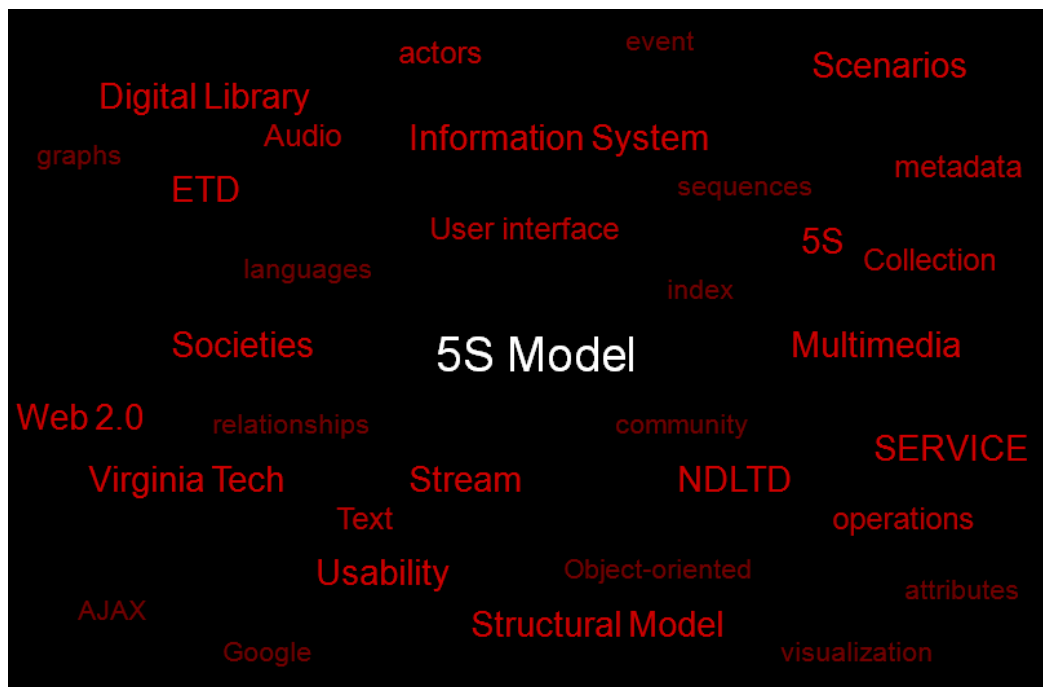


Figure 1. A tag cloud with terms related to “5S Model”

Wise's IN-SPIRE Themescape [94] uses a three dimensional landscape metaphor to visualize documents. Documents are represented by small points and those with similar content are placed close together. A peak in the Themescape represents the documents around it which are concentrating on a closely related topic. Labels on a peak are two or three topics shared by documents in that area.

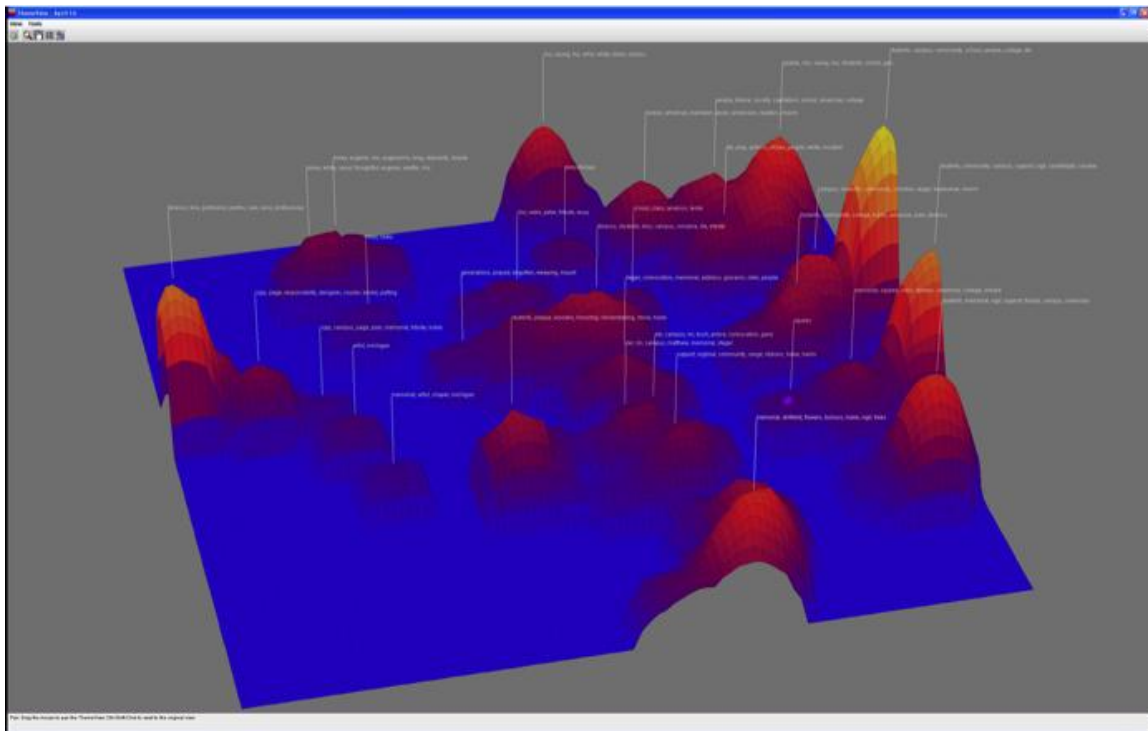


Figure 2. A Themescape with documents related to the Virginia Tech April 16 incident

Figure 2 and Figure 3 show the result of analysis (using the IN-SPIRE software) of nouns, and noun phrases, in the Virginia Tech April 16 Archive, with full size images available in the DL-VT416 Digital Library [82]. Figure 2, the Themescape view, shows clusters of topics of documents in the form of a 3D landscape, while Figure 3 shows a GalaxyView that presents the same data as Figure 2, where coloring is used instead of 3D effect. Clusters covering similar topics are located close to each other, which illustrates the popularity of topics.

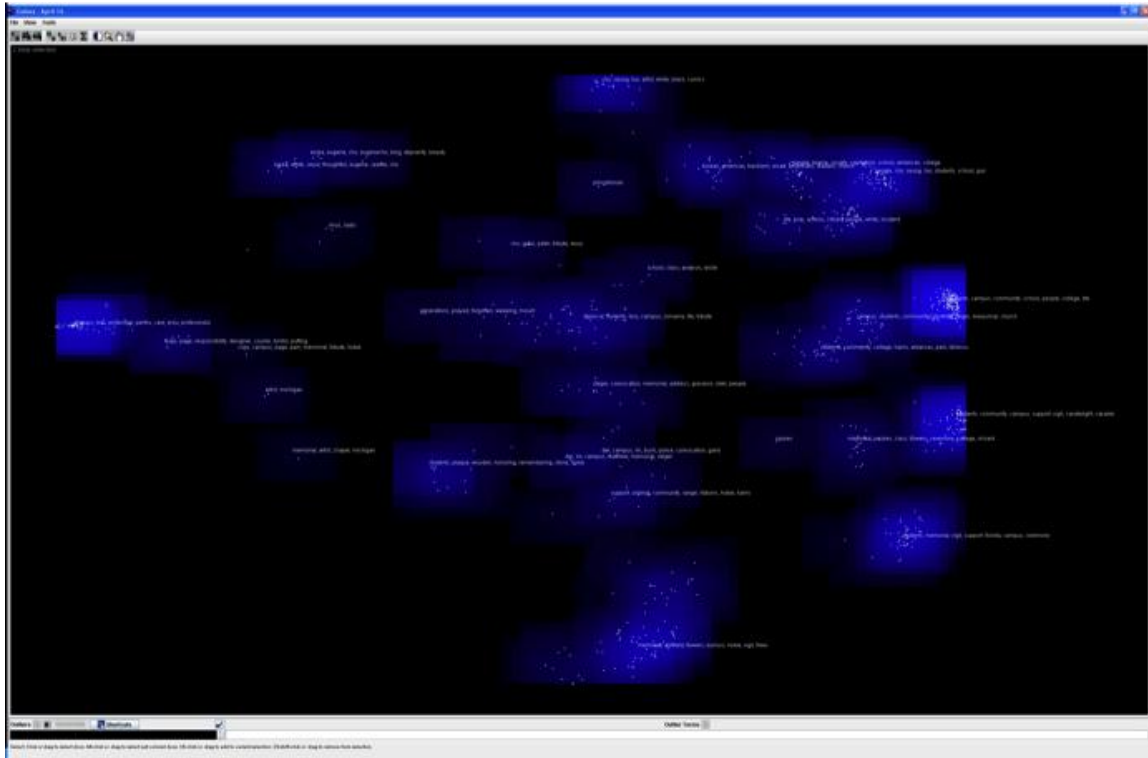


Figure 3. A GalaxyView with documents related to the Virginia Tech April 16 incident

2.5 User Modeling

Probabilistic approaches to user modeling have made it possible to learn about user profiles, as well as to revise them based on additional data [54, 65]. Tang utilized users' browsing patterns for collaborative filtering [80]. Webb examined challenging user modeling approaches like data rating, concept drift, data sparseness, and computational complexity [89].

2.6 Implicit Rating Data

GroupLens [50] employs a time used factor, which is a type of implicit rating data, together with explicit rating data. Nichols's research [58] emphasized the potential of the implicit rating technique and suggested the use of implicit data as a check on explicit ratings.

2.7 Evaluation of Information Visualization

The main purpose of information visualization is to generate *insight* regarding the data. Insight often leads to new discovery of knowledge, which is the main consideration for any science. The primary function of information visualization tools is to make it easy for the user to gain insight regarding the data. Therefore, the evaluating effect of visualization is measuring how well the visualization tool generates unpredicted insights, beyond predefined data analysis tasks. Most evaluation of visualization tools is performed by controlled measurements of accuracy and efficiency on predetermined questions and answers. Unlike typical evaluations of visualization, focused on controlled measurements, Saraiya et al. [69] proposed an insight-based methodology for evaluating bioinformatics visualizations by defining several characteristics of insight that make it possible to recognize and quantify degrees of insight in real-world, uncontrolled, scenarios in a real time manner. Chen provided a result of a meta-analysis for the accuracy and efficiency of visualization in information retrieval [23]. Plaisant [66] summarized four trends and several challenges of current evaluation studies in information visualization, and suggested next directions of research for this area.

Chapter 3. 5S Perspective of Digital Library Personalization

5S is a formal framework supporting the modeling of digital library applications [31]. Because this doctoral work includes developing a personalization-support application for digital libraries, it will be meaningful to include in this dissertation an explanation of personalization in digital libraries from a 5S perspective, which is covered in this chapter. Especially, personalization, including recommender systems, is explained through the Society and Scenarios models, which are emphasized in our personalized-support system. Figure 4 illustrates how digital library personalization will be associated with the 5 parts of the 5S framework. The following sub-sections explain Figure 4.

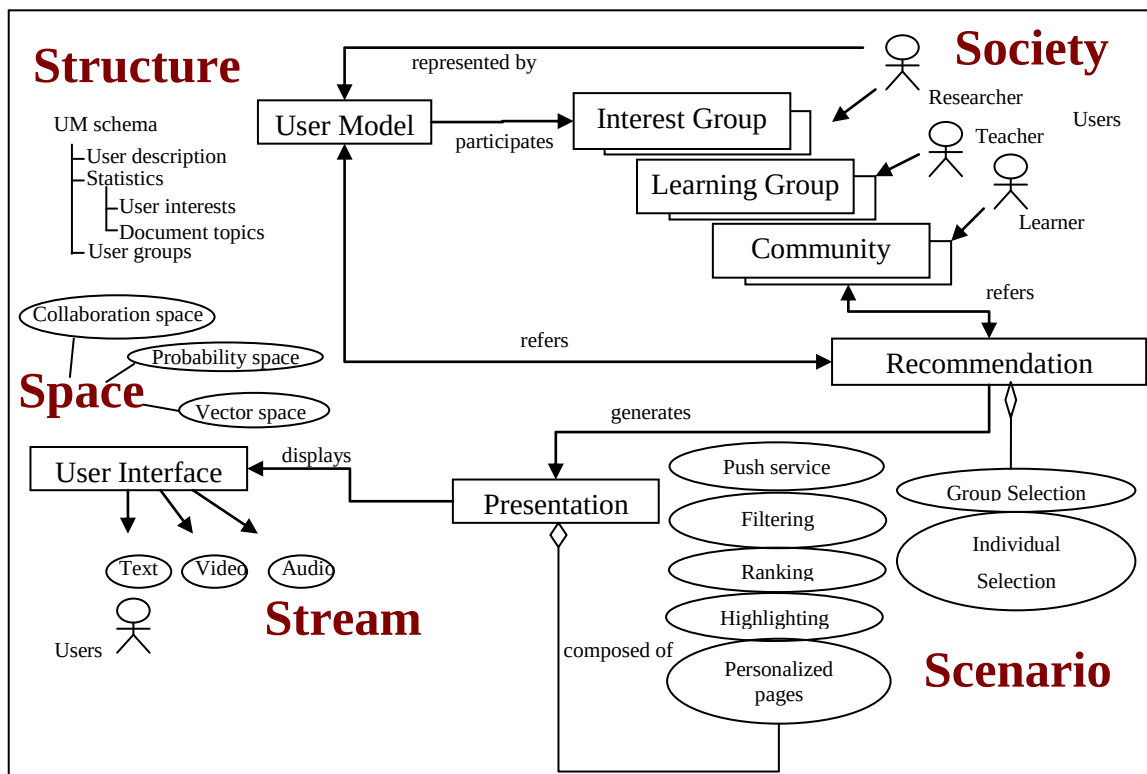


Figure 4. 5S perspective of digital library personalization

3.1 Society

The society information is the most important cue for a recommender system and for personalization. A society model in digital library personalization includes the learning communities, virtual research interest groups, and local communities. In this doctoral work, we focus on finding the hidden societies in digital libraries based on tracking patrons' research interests and learning topics which are obtained from implicit rating data. Our approach is different from other previous similar research which depends on explicit rating data. Because we use an unsupervised grouping algorithm, instead of assigning a user to pre-defined classified communities, we can find more specific and unpredicted user communities. Reverse interpretation of virtual interest group information can provide good clues to trace the drift of concepts for each member of the group, because belonging to a group means the user has interest in the topics shared by the members of the group. That is, if a user is assigned to a new interest group, it is understandably said she has a new research interest or learning topic represented by the interest group.

3.2 Scenario

The scenario model defines how personalization is actually realized in the user interface. A 'push service' is an aggressive personalization service that delivers newly arrived and hot information, in which the user may be interested, even while the user is in an offline state. 'Filtering' prevents both less interesting and irrelevant information from being delivered to the user. 'Ranking' determines the order of documents based on the estimated degree of relevance to the user's research interests and learning topics. 'Highlighting' is to identify outstanding relevant items among non-relevant items, to increase the chance of being selected, without modifying the original order and layout. 'Personalized pages' provide a customizable web layout to let the user decide which services will be provided.

'Group selection' is a process wherein the recommender decides which user groups should benefit from the current recommendation, and 'Individual selection' involves deciding which individuals will benefit.

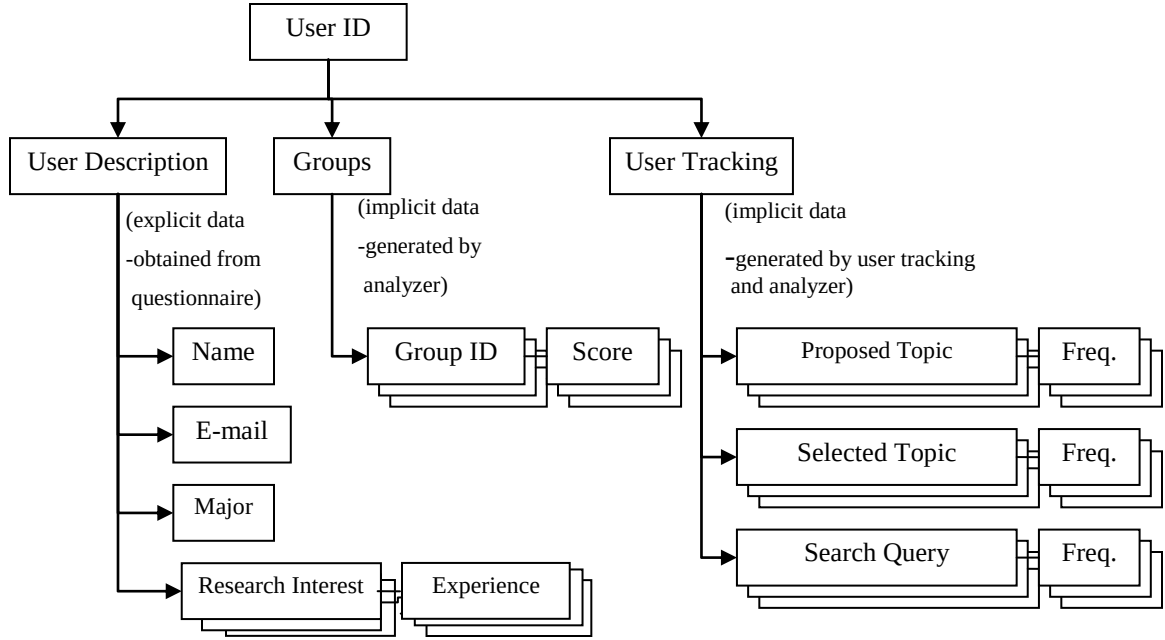


Figure 5. Structure of user model data

3.3 Structure

The structure model of the 5S perspective of digital library personalization defines the data structure of the user model. User modeling techniques help represent users' personalities in using digital libraries. Each user's model is realized by data which stores the demographic, communities, and preferential information regarding digital libraries. User model data is recorded in XML format. Figure 5 shows the structure of user model data. Three kinds of data, which are user description, group information, and user tracking information, are included. These parts are obtained or generated by a user questionnaire, the user model analyzer explained in Section 2.4 of our technical report [45], and the tracking system, respectively. Section 2.3 of that report provides a detailed description, with an example.

3.4 Space

We represent a user's research interests and learning topics with high dimensional user tracking data, i.e., implicit rating data, which is stored in the "User Tracking" part of the user model structure illustrated in Figure 5. In order to calculate similarities, which

provide information for figuring out possible collaboration and recommendation, between users and between interest groups, we employ the vector space model [12] which is popular in information retrieval. Also, in visualizing users and user communities, we apply spatial distance between users and interest groups. Section 5.3 explains our community finding algorithm. Our recommendation algorithm [44] is performed based on calculating the probability that the user or user group, of the potential recipient, will give a positive rating for the current recommendation.

3.5 Stream

The stream model defines data streams, such as text, video, music, image, etc., that will be provided to users in personalized digital libraries.

Chapter 4. Collecting Data and Preprocessing

In this chapter, we describe how we have defined two types of data we used, implicit rating data and explicit rating data, and the method we used to collect the data. A user tracking interface and a user survey are presented. We also explain how we stored and processed the data when it was needed.

4.1 User Tracking System: Collecting Implicit Rating Data

A user tracking system, embedded in the web user interface [42], see Figure 6, is the key source of gathering implicit rating data. It is implemented with JavaScript and collects information on a user's activities while she is logged in. Cookies are used for storing temporary data and for communication with the processing module in the server.

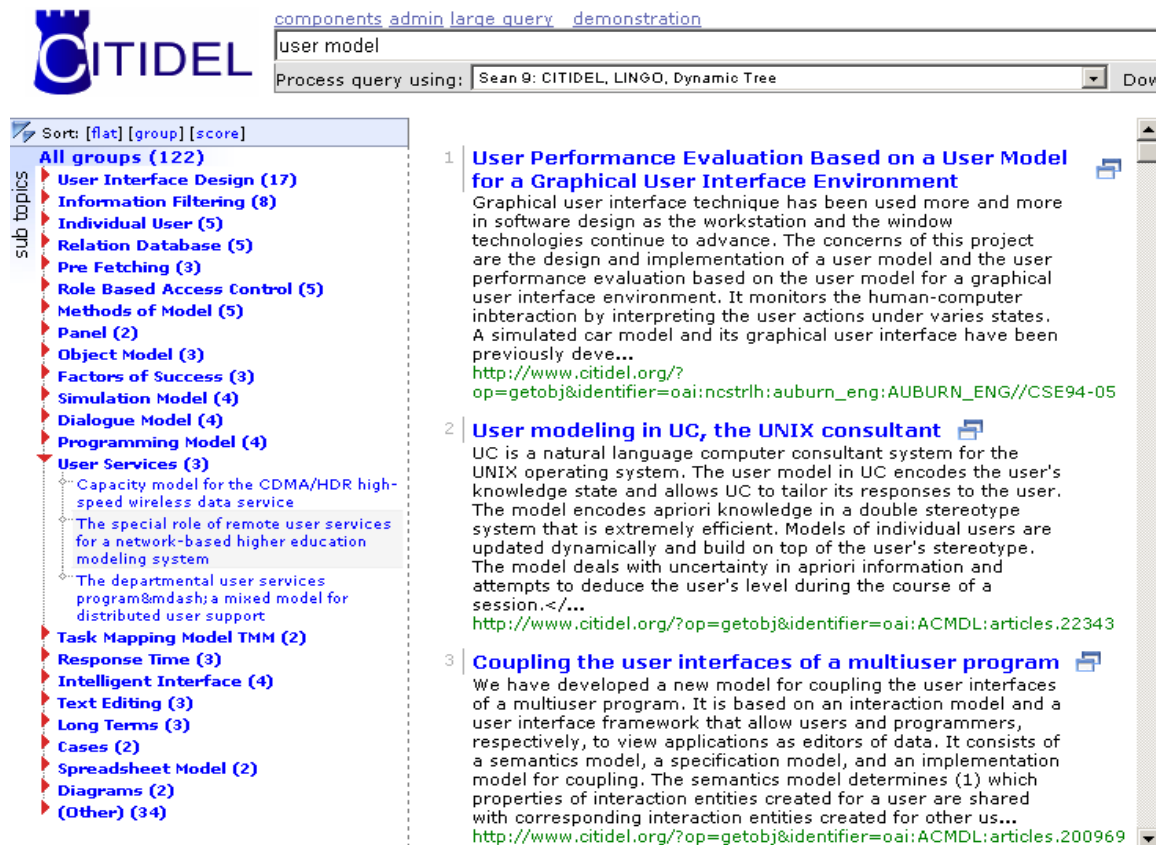


Figure 6. A JavaScript based user interface

4.1.1 About NDLTD

The Networked Digital Library of Theses and Dissertations (NDLTD) union catalog [6] contains entries for over 400,000 electronic theses, dissertations (ETDs), and important documents provided by more than 325 member institutions, such as libraries and universities.

4.1.2 User Tracking System

The standard HTTP log, unfortunately, does not include the title of an anchor, which is treated as a document topic in our system. For instance, when a user selects the anchor “Statistics for Librarians” on a web page, we would like the title “Statistics for Librarians” to be stored in the log file along with the data gathered by the standard HTTP protocol such as URL, current time, error codes, IP addresses, etc. Therefore, we had to develop a special interface that has an embedded user tracking system.

Our system is implemented with JavaScript and embedded in the web search interface of NDLTD, see Figure 6 and Figure 7, which is based on Carrot² [2]. When a user logs into NDLTD, her profile information is loaded, and the user tracking system starts to collect the user’s activities, such as sending queries, browsing the dynamic tree, opening document clusters, etc. Also, it collects information from the result documents shown on the screen to the user. User data is collected during one login session, which will be closed by logging out or terminating the Internet Explorer, and stored in local disk temporary space as a cookie file. The cookie is transferred to the server-side process module, for updating the user data when the session is closed.

Once the temporary rating data is transferred to the digital library server, it is merged into the user model, which is already stored in the user model database. We use XML formatting for all messages exchanged and stored. The cookie size limit of 4000 bytes is large enough to store the user’s behavior for a single login session.

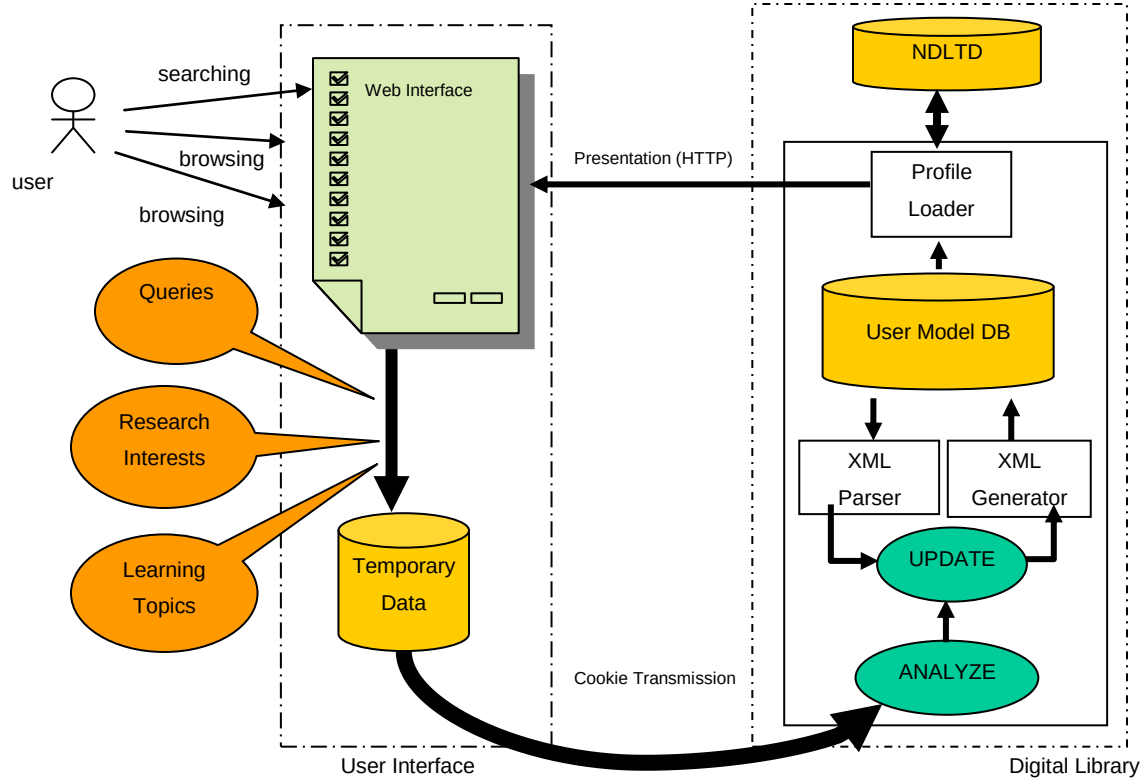


Figure 7. System diagram of user tracking system

4.2 User Survey: Collecting Explicit Rating Data

Even though our focus in this doctoral work is proving the effectiveness of implicit rating data and utilizing it, we also collect explicit data about user's research interests, learning topics, and basic demographic information – for evaluation purposes. Our visual data mining tool (VUDM), see Chapter 5, uses explicit rating data solely, except user IDs for labeling user icons. [45] describes details of our online survey system and processes. Using this survey, we collect explicit data such as user's name, email, major, broad research interest, detailed research interest, and years of experience for each research topic.

4.3 User Models: Storing Raw Data and Statistics

Generally, user models are implemented with shared feature equations and parameters for the equations that represent the users' characteristics [38, 86]. However, our user

models consist of common feature equations, raw data items, and statistics collected by the user tracking system. Parameters for feature equations can be calculated from the raw data and statistics when they are needed. Therefore, our user models are more interoperable and transferable, and contain more potential knowledge.

The *analyzer* module, server-side, see at the bottom of the right rectangle of the system diagram in Figure 7, is in charge of three functions: XML handling, updating, and analyzing. The analyzer consists of six Java classes for implementing these functions and representing data objects. Two classes, TestSAX and WriteXML, are for XML handling and updating, and four classes, Item, GroupItem, Research, and UserModel, are for analyzing and representing the User Model.

Figure 5 in Chapter 3 shows the logical structure, while Figure 8 provides the XML schema of the user model. The information in the user model can be broken into two categories: personal information and search history information. The *proposed* element contains a set of terms and their frequencies which are identified by the search engine and document clustering system of the digital library, and “proposed” to the user through the user interface to select if she is interested in the terms. Also, the *selected* element contains a set of terms and their frequencies, which are actually “selected” by the user to obtain more information by updating the current page or moving to a linked page.

```
<?xml version="1.0"?>
<schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">

  <element name="user">
    <complexType>
      <sequence>
        <element name="UserID" type="string"/>
        <element name="email" type="string"/>
      </sequence>
    </complexType>
  </element>

  <element name="major" type="string"/>
  <element name="research" minOccurs="0" maxOccurs="unbounded" nillable="true">
    <complexType mixed="true">
      <sequence>
        <element name="interest" type="string" minOccurs="0" maxOccurs="unbounded"/>
        <complexType>
          <simpleContent>
            <extension base="double">
              <attribute name="experience" type="decimal" minInclusive="0" maxInclusive="100"/>
            </extension>
          </simpleContent>
        </complexType>
      </sequence>
    </complexType>
  </element>
</schema>
```

```

        </complexType>
    </element>
    <element name="group" nillable="true">
        <complexType>
            <sequence>
                <element name="GroupID" minOccurs="0" maxOccurs="unbounded">
                    <complexType>
                        <simpleContent>
                            <extension base="double">
                                <attribute name="prob" type="decimal" minInclusive="0" maxInclusive="1"/>
                            </extension>
                        </simpleContent>
                    </complexType>
                </element>
            </sequence>
        </complexType>
    </element>
    <element name="query" nillable="true">
        <complexType>
            <sequence>
                <element name="item" type="ItemType" minOccurs="0" maxOccurs="unbounded"/>
            </sequence>
        </complexType>
    </element>
    <element name="proposed" nillable="true">
        <complexType>
            <sequence>
                <element name="item" type="ItemType" minOccurs="0" maxOccurs="unbounded"/>
            </sequence>
        </complexType>
    </element>
    <element name="selected" nillable="true">
        <complexType>
            <sequence>
                <element name="item" type="ItemType" minOccurs="0" maxOccurs="unbounded"/>
            </sequence>
        </complexType>
    </element>
</sequence>
</complexType>
</element>
<complexType name="ItemType">
    <simpleContent>
        <extension base="string">
            <attribute name="freq" type="integer"/>
        </extension>
    </simpleContent>
</complexType>
</schema>

```

Figure 8. XML schema of user model data

4.4 Using Legacy Data: VTLS NDLTD log data

Besides the user tracking system, to collect implicit rating data, and the user survey to collect explicit rating data, a legacy web log data convertor is used in this doctoral work. The purpose of the data convertor is to transform the legacy web log data into data in user model format. Using legacy web log data is important, because most currently available data is already collected in legacy web log form, and the data of the user model is just starting to be collected and thus is insufficient.

In the converted user model data, only the “query” field may contain the list of items referred to by the user, while the originally proposed user model format contains it in the “query”, “selected items”, and “proposed items” fields. Even though the user model data, converted from legacy web log data, doesn’t have all of the features of the proposed user model format, it is still usable for this doctoral work.

The most disputable issue in converting legacy web log data into user model data is how to decide the user identity for each record in the legacy web log data. Unlike the user model data, web log data usually doesn’t contain information about the subject of the data. In our converting process, we used the IP-address as an identifier of the data, assuming that all users have unique IP-addresses and don’t share their addresses with others. In most cases, this assumption is true. However, there are many exceptions; a computer in a public place has an IP-address shared among multiple users. In spite of the problems mentioned above, it is still worthwhile to convert the legacy web log data because of the large amount of available data.

Chapter 5. Visualization and Knowledge Finding

This chapter introduces our Visual User model Data Mining tool (VUDM) [43, 46]. We explain how we present information about user behavior, user communities, user relations, and usage trends of a digital library using implicit rating data, and how useful knowledge could be extracted from our visualization techniques. Results of formative evaluation for VUDM are described in Chapter 6.

5.1 VUDM: A Visualization Tool

5.1.1 Need for a Visualization Tool for Digital Library Users

According to the newest trends regarding web technology, in keeping with so-called Web 2.0 [61], web applications should become more flexible and evolutionary, leveraging the collaboration of users. In this situation, analyzing a large volume of user data and usage data in community driven information systems, such as blogs, wikis, online journals, etc., can help with research challenges in the social computing field as well as for information systems and data mining [51]. Also, in digital libraries, which are complex information systems, research investigations regarding community services and intelligent personalized user services require those kinds of data analysis. Thus, a goal of developing VUDM is to visualize user communities and digital library usage trends to support decision makers and domain experts in designing and building personalized digital library services. Visualization helps us answer complex and comprehensive questions, which cannot be answered using normal OLAP systems [74]. Several top questions on digital libraries and its patrons in those categories are listed below (reproduced here for convenience from Section 1.2.4):

- What are the current trends of information seeking in this digital library?
- What kinds of people are using this digital library?
- How have the points of focus for retrieval changed, for a certain user?
- What academic areas are emerging attractions?
- How many people are interested in what topics?

- How many virtual groups, who share particular interests, exist in the digital library?
- Who is a potential mentor for whom?
- How many experts are in which areas?
- Which topics are related to which other topics?

Visualization also provides visual exploration support to its users so that they feel involved directly in the data mining process, and can utilize their own creativity, flexibility, and general knowledge [40]. Figure 9 depicts how users explore knowledge through information visualization. Knowledge finding with a visualization tool is achieved by repetitive hypothesis generation, data exploration, and hypothesis verification. That is, through exploring the data, we can verify an hypothesis, and at the same time it arouses another hypothesis. Statistics and machine learning techniques, embedded in the visualization tool, support the user to verify her hypothesis.

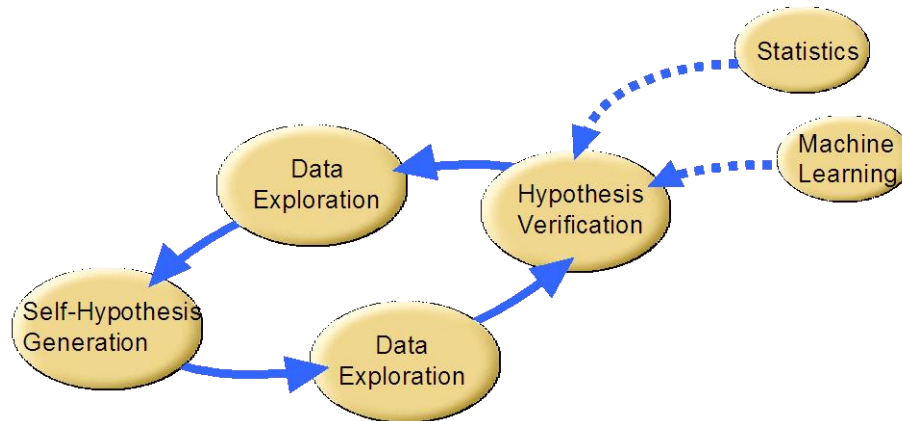


Figure 9. Iteration of data mining process using visualization tools

5.1.2 Data Description and Preprocessing

Our initial data set consists of 1,200 user models, describing those who registered to use our search result clustering service in NDLTD, which was introduced in Section 4.1.2, between August 2005 and May 2006. During the registration process, new users explicitly provide data, called “explicit data”, such as their specialty, major (area of

interest), and number of years worked in each such area. Table 1 describes the initial data set used for VUDM.

Table 1. Data set: User model data used for VUDM formative evaluation, from 1,200 users

	Data Type	Description
Implicit Data	Query Log	Avg. 4.59 queries per user. (Estimated from the Query Log generated by the Web Service demon) Avg. 4.52 distinct queries per user. (Obtained from the Query Log stored in User Data)
	Topics (Browsing Activities)	Avg. 19.3 topics per user (User's browsing activities were tracked by recording the Noun Phrases traversed.)
Explicit Data	Number of Years of Experience	N/A
	Demographic Information	N/A

Explicit data is easy to analyze with normal analysis tools. However, such data is insufficient when addressing the comprehensive questions listed in Section 5.1.1. Further, user interests and behavior change over time, so it is important to enhance user models with data from a user tracking system, i.e., “implicit (rating) data” (so-called because the data was not entered explicitly in answer to questions). Our implicit data consists of a “query log” and two types of interest “topics” which have the form of noun phrases. The user tracking system runs on an NDLTD service that provides document clustering, and collects the cluster names that users traverse. It records positively rated, as well as ignored, hence negatively rated, “topics” [47]. Our 1,200 user models contain both explicit data and implicit rating data that grow with the use of NDLTD, but our focus is on visualizing such user models mainly using implicit rating data. The data allows us to characterize users, user groups, and broader user communities. At the same time, we can characterize topics and (scholarly) areas of interest. Combining the two types of information allows identification of areas of user expertise, mentoring relationships among users, and changes/trends related to the data and information considered.

5.2 Visualization Strategies

With our Visual User model Data Mining (VUDM) tool, we aimed to visualize the following information.

- Virtual Interest Groups
- Individual Interests
- Relations between Groups
- Relations between Individuals
- Interest Density
- Search Topic Trends

The visualization strategies of our Visual User model Data Mining (VUDM) tool fully follow Ben Shneiderman’s information visualization mantra [20, 71], “Overview first, zoom and filter, then details on demand”. The main window presents an overview of all users (shown as icons) and communities (i.e., groups, shown as spirals) as illustrated in Figure 10. In this figure, 1 displays an overview of users, virtual interest groups, and their relationships. The multiple windows, 2, present different data from different months at the same time to make it easy to find any changes made between months. The slide bar, 3, controls the correlation threshold (θ). The small tables at the bottom, 4, 5, and 6, show detailed information about groups, topics, and highlighted users, respectively. When using the right mouse button, dragging up and down, 7 and 8, and free dragging, 9, cause: zoom, un-zoom, and panning.

The visualization of users and topic-based groups aims to summarize high dimensionality data, in order to support key tasks (see Section 5.4). Three degrees of freedom (three dimensions) are shown, since one can vary the position (x , y coordinates) of a spiral center, as well as the distance (of a user icon) from the center.

The positions of spirals (groups) are not controlled absolutely because the dimensionality of data is too high. It is only important to maintain relative distances among spirals (interest groups). For laying out the spirals, a “grid layout” method [34] is used. That is, the whole space is divided into equal-sized rectangles and the “groups of

similar groups” are centered in each rectangle. Each “group of similar groups” consists of a representative (largest) group at the center and satellite similar groups around it at a distance based on the group similarity with the representative group.

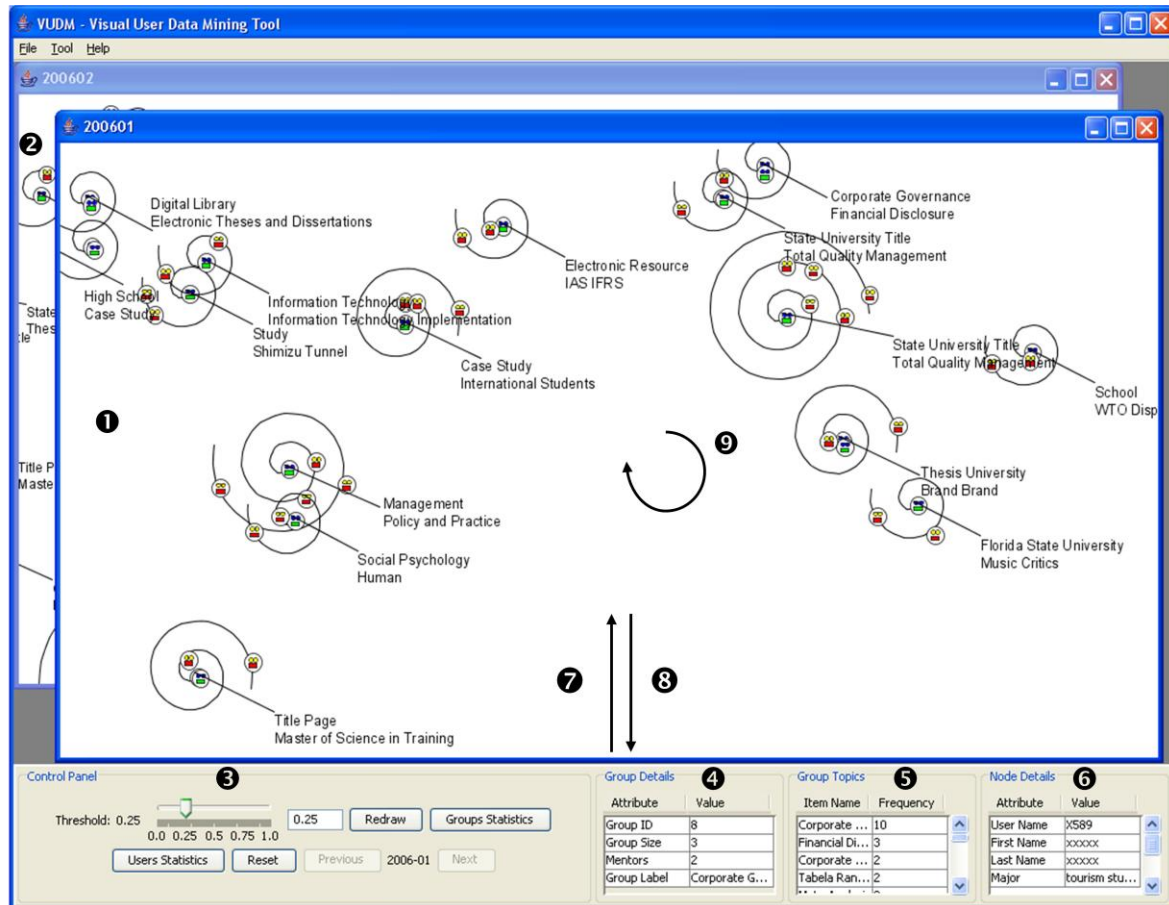


Figure 10. The main window of VUDM shows an overview of virtual user communities

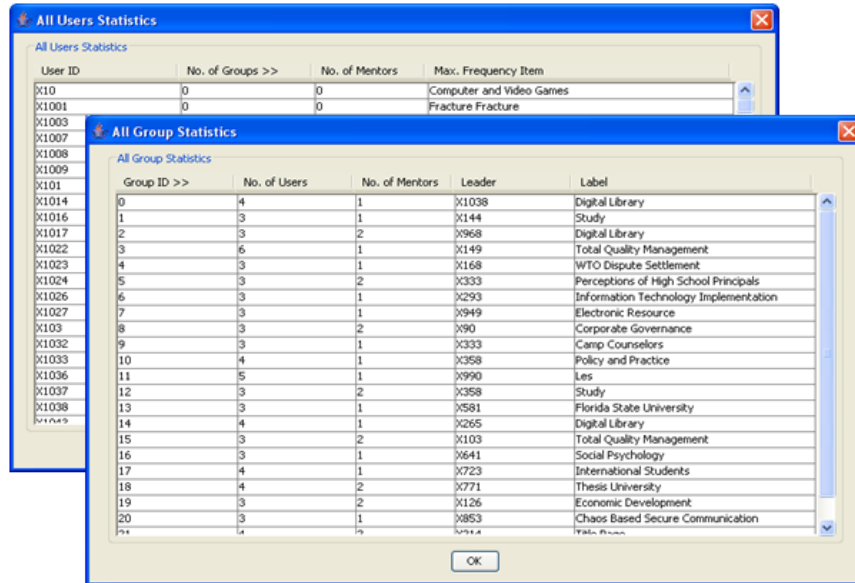


Figure 11. Windows for detailed information about users and user communities

Figure 11 shows windows for detailed information about users and user communities. These windows provide basic analysis functions such as sorting by data field. Figure 12 is a snapshot of VUDM when it operates in “zooming” mode. Because VUDM should visualize thousands of users in overview mode, it is necessary to zoom into the desired area to make it easy to distinguish, locate, and select users and user communities. Some user icons in different communities are connected with lines to show they are identical. Zooming is achieved by dragging the mouse upward while pressing the right button on the desired area.

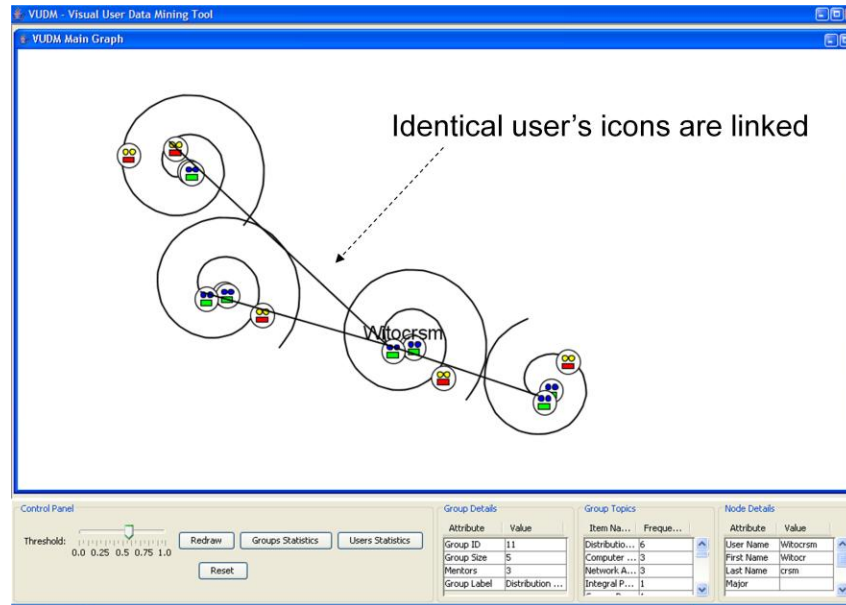


Figure 12. Zoomed user space

VUDM also provides a filtering feature. Filtering is achieved by moving a slider which is associated with the user correlation threshold θ , which will be explained later in this chapter, on the user interface. With higher θ , a stricter community finding is performed, so that less probable user communities are filtered out; therefore, finding user communities becomes strict too. Figure 13 shows how the correlation threshold θ influences finding user communities. The left window shows that more users and user communities were found when a low correlation threshold was used, and the right window shows that fewer users and user communities were found because of a high correlation threshold. In addition, all user icons and group spirals can be dragged with the mouse, e.g., to examine a congested area.

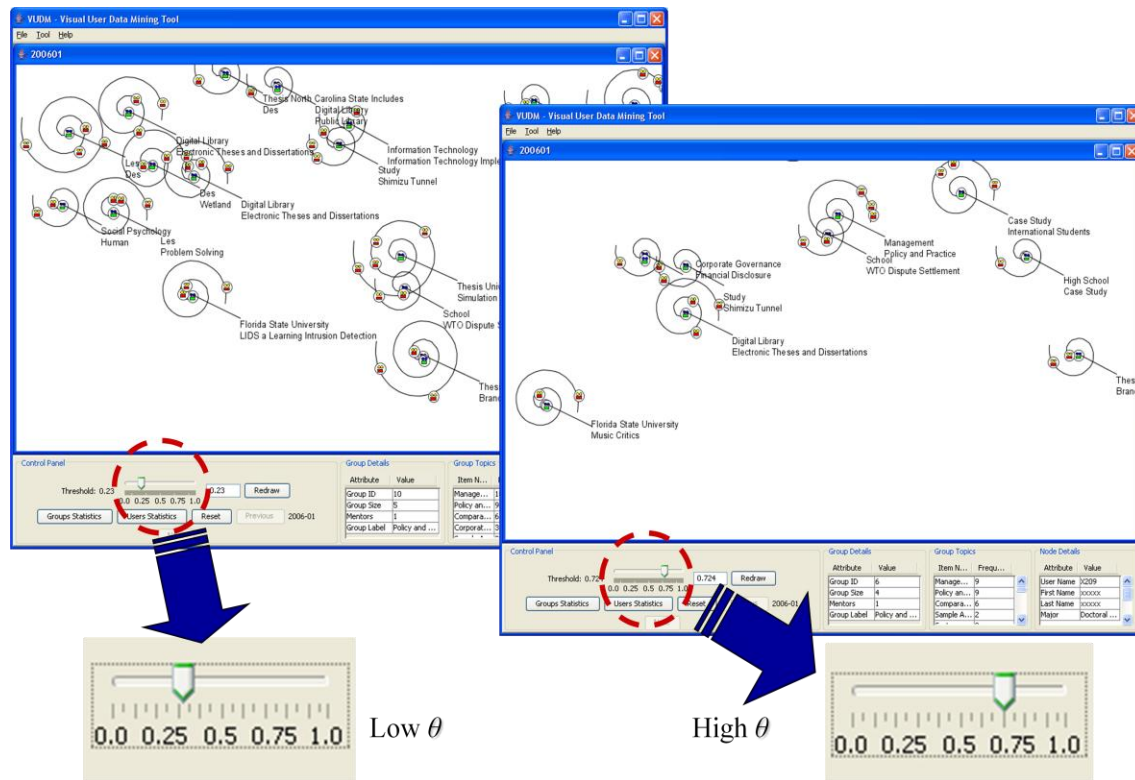


Figure 13. Filtering user space

VUDM provides detailed information about users and user communities on demand – see the two windows at the bottom of Figure 14. The table panel, right top, contains three information tables about the selected user or group. Each table shows group details, group topics, and user details, respectively. Two sub-windows of the bottom row present detailed information about all groups and all users in the system. In the window for user details, the user’s ID, number of groups to which she belongs, and research interests are listed. In the window for group details, the group’s ID, number of users in the group, and a list of topics for the group are shown. These detail information tables support basic OLAP functions, such as sorting and counting. Thus, VUDM services combine the strengths of graphical and text-oriented presentations.

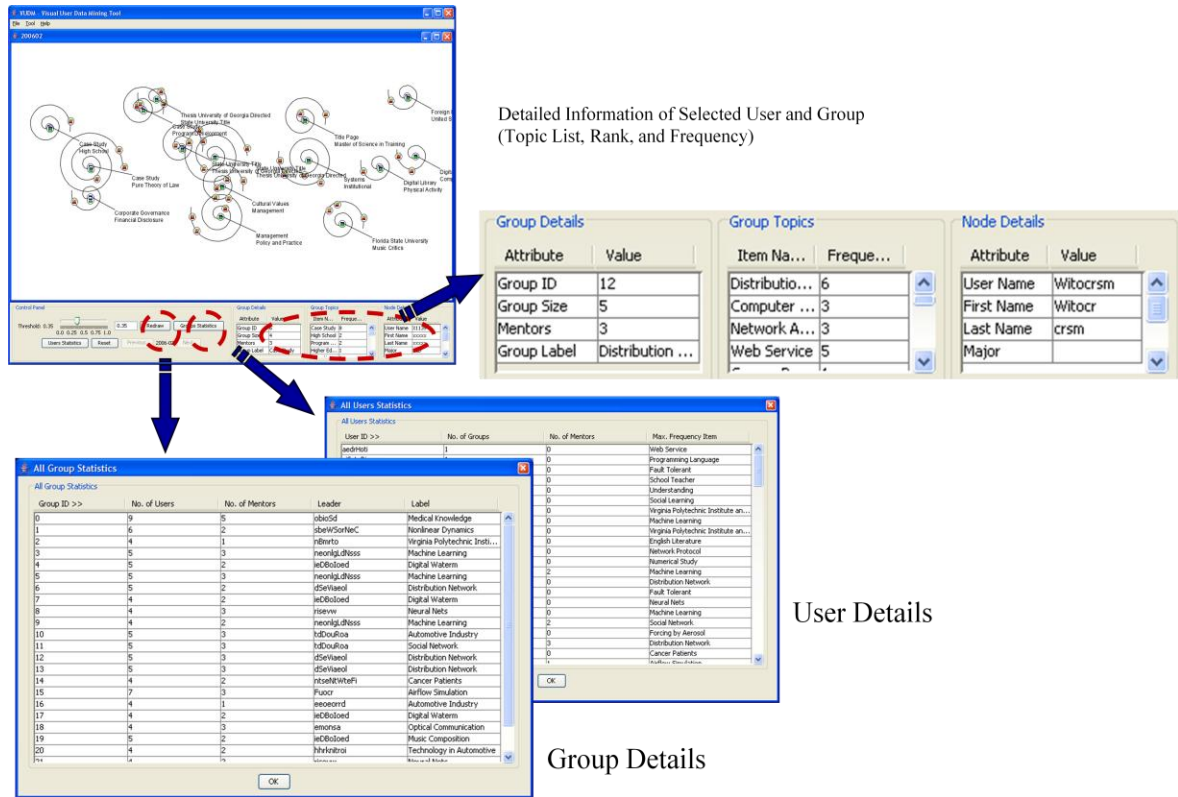


Figure 14. Detailed information on demand

5.3 Loose Grouping Algorithm

For classifying users into virtual interest groups and finding “groups of similar groups”, we use the same algorithm. Because any statistical information about distribution and underlying densities of patrons, such as sample mean and standard deviation, are not known, nonparametric classification techniques, such as Parzen Windows and k -Nearest-Neighbor (k NN) [25, 27], should be used. But k NN is inappropriate since it assigns the test item into only one class, it needs well-classified training samples, and its function depends on the size of the sample. For these reasons we devised a modified k NN algorithm: “fixed-size window multi-classification” (FSWMC) algorithm. Figure 15 illustrates the difference between k NN and FSWMC. In this figure, Top Row: The k NN rule starts at the test point (red spot) among classified samples, and grows the surrounding circle until it contains 'k' samples. The size of the circle depends on the number of samples, Then, it classifies the test point into the most dominant class in the circle. Bottom Row: The fixed-window multi-classification rule classifies all samples

enclosed by the fixed sized, $r=\theta$, circle, surrounding the test point (red spot), into a new class. Once a point is classified, its color is changed to grey. Then, if the new generated class is a sub- or super-class of an already found class, we remove the redundant sub-class. In this figure, two classes, $\{c\}$ and $\{a,b,d,e\}$, are found up to stage $n=16$.

Distances between samples (the spots in the hyperspace) are calculated using Equation (1) in Section 5.4.1. While the window size, r , of the k NN is dependent on ‘ n ’ (the total number of samples), the window size of FSWMC is fixed to the correlation threshold θ . The θ value is entered from the user interface. In this algorithm, a test sample will be assigned to 0 or more classes, depending on the number of neighbors within the distance θ . Theoretically, a maximum of ‘ n ’ classes, one class for each sample, can be found. However, we reduce the number by the “removing subclass rule”: a class whose elements are all elements of another class can be removed to ensure there are no hierarchical relationships among classes. Also, we remove trivial classes, where the number of elements is smaller than a specified value. Even though Parzen Windows also uses a fixed-size window, our algorithm is more similar to k NN because k NN and FSWMC estimate directly the “a posteriori” probabilities, $P(\text{class}|\text{feature})$, while the Parzen Windows estimates the density function $p(\text{feature}|\text{class})$. We also use our algorithm to find “groups of similar groups”. However, in that case we assign the testing sample to the most dominant class among samples within the surrounding region, because a group should be assigned to only one “group of similar groups”.

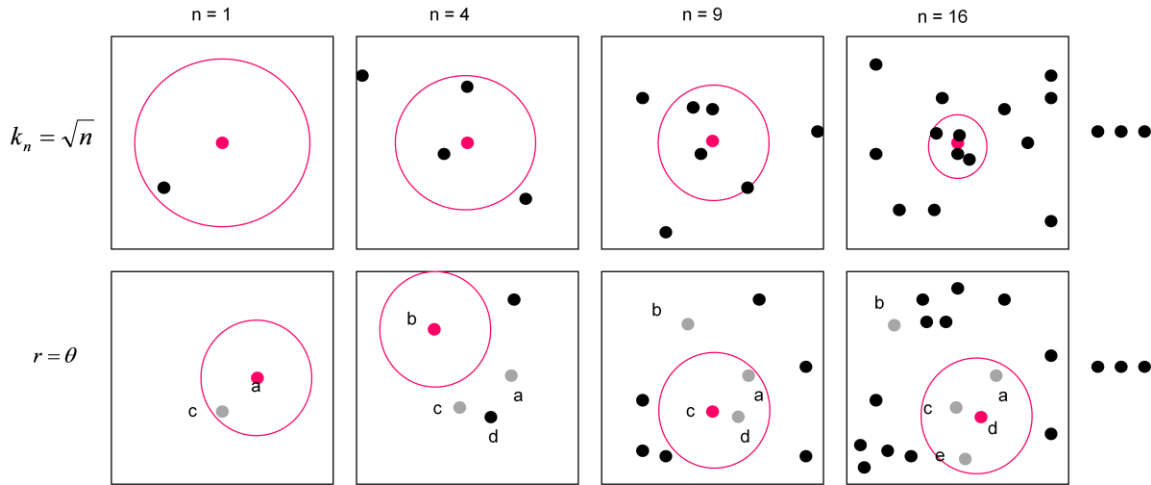


Figure 15. Illustration of the k NN (top) and FSWMC (bottom) algorithms

Figure 16 describes pseudo code for the fixed-size window multi-classification algorithm. The value ' θ ' is the correlation threshold, which is given by the VUDM user interface shown in Figure 13.

```

for each item  $i$  in the system {
    generate a new class  $c$ ;

    for each item  $j$  in the system
        if distance  $(i, j) \leq \theta$     assign item  $j$  into  $c$ ;

        for each class  $t$  in the system {
            if  $c \supseteq t$  discard  $t$ ;
            else if  $c \subset t$  discard  $c$ ;
        }
}

```

Figure 16. Fixed-size window multi-classification algorithm

5.4 Knowledge Finding

The goal of our visualization is to support understanding about users, user groups, and topics – and their interrelationships. We consider three categories of knowledge: user characteristics and relationships, virtual interest groups and relationships, and usage trends. These are discussed in detail in the following three subsections.

5.4.1 User Characteristics and Relations

User characteristics are the most important information for personalization. Many commercial online shopping malls, such as amazon.com and ebay.com, are already utilizing user characteristics for personalized services. VUDM visualizes each user's interest topics and expertise level by putting his icon on spirals in a 2D user space (see Figure 17 left). Each spiral represents a set of closely related topics and, thus, forms a virtual interest group with the users on the spiral who share the topics. Small face icons on the spirals are users. The size of a spiral is proportional to the size of the group. Distance between user icons within a group reflects their similarity with regard to topics. Because a user may be interested in multiple topics / scholarly areas, VUDM puts copies

of his icon on all spirals that match his interests, linking copies together with connection lines when the user is highlighted (see Figure 17 right). The distance between two spirals reflects the similarity between the two groups.

The amount of expertise on a topic for a user is used to determine the distance from the center of the spiral to that user's icon. The closer to the center of the spiral, the more expertise the person has about the topic. Expertise is computed as a function of the number of years the user has worked in the area, and of the length of usage history. High-ranked persons in a group are colored differently, and are classified as mentors; novice users may be encouraged to collaborate with them.

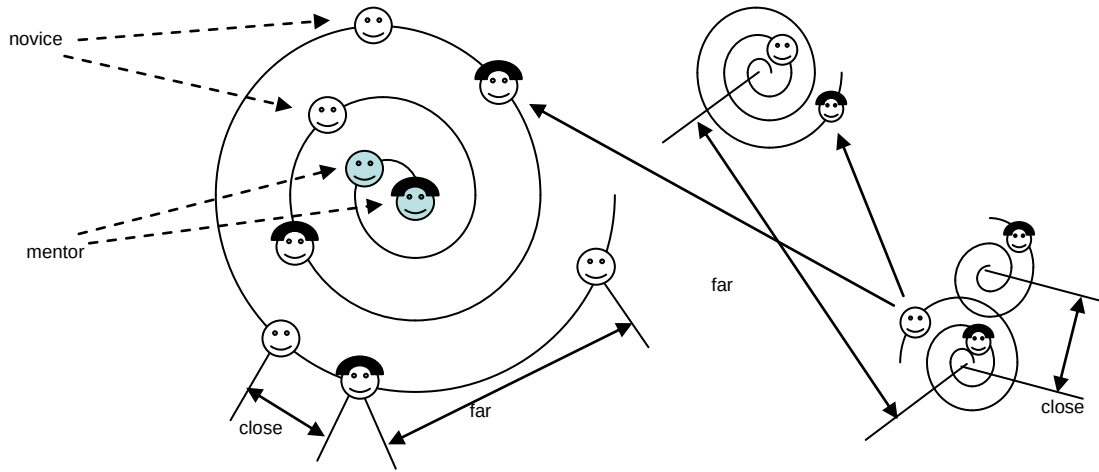


Figure 17. User and user community characteristics and relations

Decisions about the: formation of a virtual interest group, selection of users who make up a group, and location of each member icon's distance from the center of a spiral, are made by calculating correlations between users according to Equations (1) and (2). We used mainly implicit data rather than explicit data, because collecting implicit data is more practical than collecting explicit data, and it helps us avoid terminology issues (e.g., ambiguity) which are common in information systems [47].

$$\text{correlation}(a,b) = \frac{\sum_j (v_{a,j} - \bar{v}_a)(v_{b,j} - \bar{v}_b)}{\sqrt{\sum_j (v_{a,j} - \bar{v}_a)^2 \sum_j (v_{b,j} - \bar{v}_b)^2}} \quad \text{Equation (1)}$$

$$\bar{v}_a = \frac{\text{number of topics positively rated by 'a' + number of used queries by 'a'}}{\text{number of topics proposed to 'a' + number of used queries by 'a'}} \quad \text{Equation (2)}$$

(1) represents the correlation of users ‘a’ and ‘b’. ‘ v_{aj} ’ is the rating value of item ‘j’ of user ‘a’ which means the number of positive ratings on ‘j’ made by ‘a’. ‘j’ represents common topics or research interests which are rated by users ‘a’ and ‘b’. ‘ $\bar{v}_a$ ’ is the average probability of positive rating of the user, as obtained by (2) [44].

5.4.2 Virtual Interest Group and Relations

Virtual Interest Groups are virtual clusters of digital library users who share specific research interests and topics. Visualizing virtual interest groups helps us understand the characteristics of digital library patrons, may help patrons identify potential collaborators, and may aid recommendation. From this visualization, it is possible to figure out distributions of users, preferences regarding research interests / topics, and potential interdisciplinary areas. The VUDM finds virtual interest groups by connecting user pairs with high correlation values (above a threshold). The higher the threshold, the more precise will be the virtual interest group.

VUDM arranges virtual interest groups in two dimensional user space according to their degree of relationship (similarity) with other groups. Relative distance between groups reflects the degree of relationship; more highly related groups are closer. We assume that in two highly related groups, users in one group will share interests with users in the other. We measure the degree of relationship between two groups either by calculating the vector similarity between the two group representatives (a union of the model data for all members), using Equation (3), or by calculating the Tanimoto Metric (4), a.k.a. Jaccard similarity coefficient [27, 81], which uses the number of members in common [27]. Compared to vector similarity, the Tanimoto Metric has lower computational cost but still is effective.

$$\text{groupsim}(A, B) = \frac{\sum_{i \in T} v_{A,i}}{\sqrt{\sum_{i \in T} v_{A,i}^2}} \frac{\sum_{i \in T} v_{B,i}}{\sqrt{\sum_{i \in T} v_{B,i}^2}} \quad \text{Equation (3)}$$

$$D_{Tanimoto}(A, B) = \frac{n_A + n_B - 2n_{AB}}{n_A + n_B - n_{AB}} \quad \text{Equation (4)}$$

Equation (3) represents the group similarity between two virtual interest groups ‘A’ and ‘B’. ‘ $v_{A,i}$ ’ is the sum of the frequencies of positive rating on topic ‘i’ made by all users in group ‘A’. ‘T’ is the set of all topics in the system which are rated positively at least once. Equation (4) represents the similarity distance between two groups ‘A’ and ‘B’. ‘ n_A ’ and ‘ n_B ’ are the number of users in A and B, respectively. ‘ n_{AB} ’ is the number of users in both groups A and B.

5.4.3 Usage Trend

In addition to characteristics and relationships among individual users and virtual interest groups, general usage trends also are of interest. Visualizing usage trends in VUDM is accomplished by providing overviews over time. Thus, Figure 18 shows VUDM results for three months.

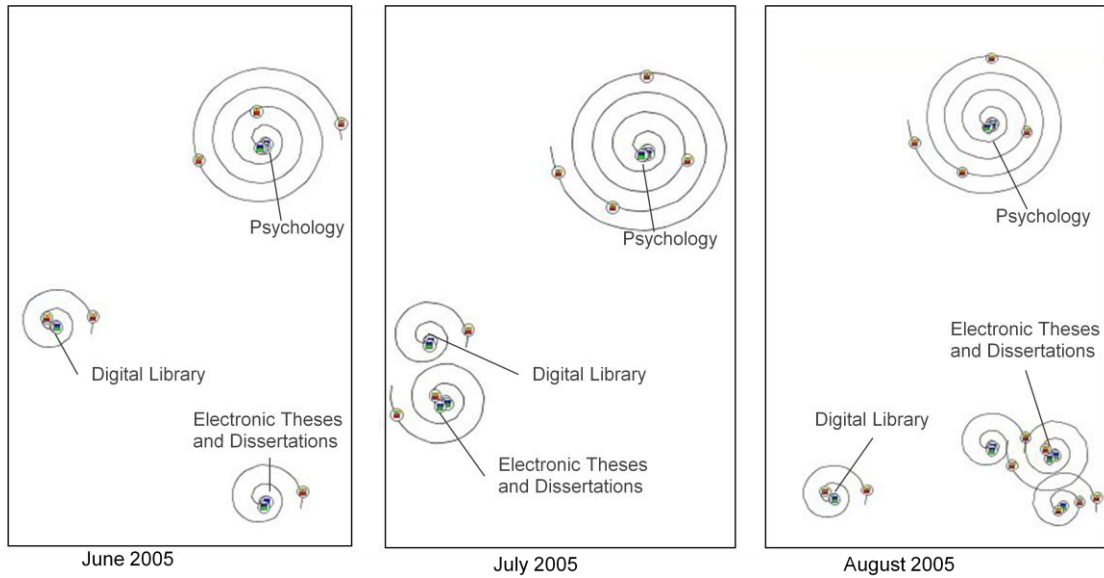


Figure 18. Visualizing usage trends of digital library

In June we see a cluster of small groups at the bottom. In July we see those are attracting more users and groups, and seem to be merging, while an old topic, the large spiral at the top, adds one more user. That large group shrinks in August, at the same time

as there are further shifts among the small groups (now three) at the bottom. Thus, we see which areas emerge, are sustained, shrink, or grow. Further, we may extrapolate from series of changes to make predictions.

5.4.4 Concept Drift

Detecting concept drift is a well known problem in the machine learning area, that involves user models dynamically adjusting to user changes quickly, as the real attributes of a user are likely to change over time [89, 91].

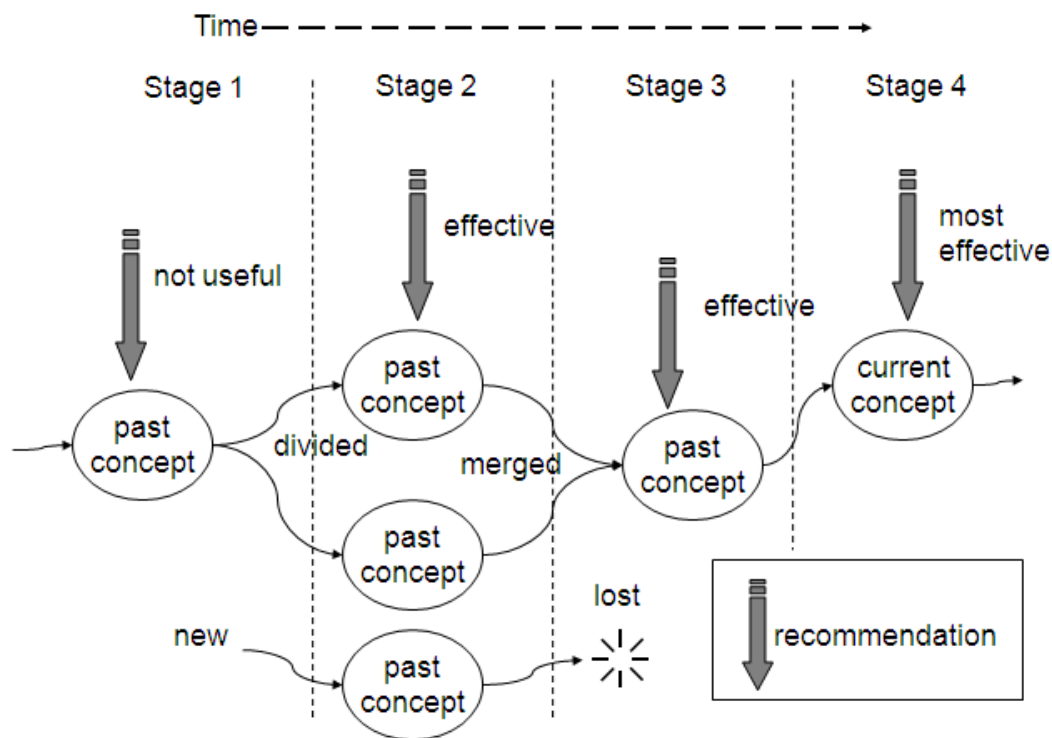


Figure 19. Detecting drift of concepts makes it possible to provide timely recommendation.

In recommender systems, detecting the concept drift of a user allows making recommendations at the proper times. As a user community, portrayed by a spiral, represents a set of closely related topics and interests, it also can be regarded as a concept, which is usually a research interest or a learning topic, describing the people on the spiral. If a concept of a user drifts to another new concept, a clone of her icon appears on the new spiral and a connection line links the new icon together with other previous instances

of her icon, to represent that they are one person. Therefore, by tracing connection lines over time, it is possible to detect new drifts of concepts.

Figure 19 shows how an effective recommendation can be made by tracking the drift of concepts. The transition from stage 1 to stage 2 shows the number of concepts of a user is increased. This is the case when she develops her previous concept into more specific multiple concepts and when she gets a totally new concept. Also, the number of concepts can be decreased as the transition from stage 2 to stage 3 shows. This may happen when the user puts her multiple concepts into a shape, or when she forgets or loses interest in her previous concept. In stage 4, it is possible to tell which concept is her latest concept, and recommending regarding to the concept will be most effective.

5.5 User Evaluation

Evaluating a visualization tool objectively is difficult, and VUDM is a data mining tool to support digital library administrators or decision makers, not to support normal users. Therefore, we conducted an analytic formative evaluation based on user interviews, instead of a summative evaluation. Most participants agreed about its usefulness in finding important knowledge about users and user groups, while a small portion of the participants pointed out problems. Section 6.1 provides a detailed description about participants of the formative evaluations, and the results of the survey, as well as several critiques and concerns raised about VUDM.

Chapter 6. Formative Evaluation

This chapter describes a formative evaluation based on which the VUDM was modified. The evaluation is performed with graduate students majoring in computer science, most of whom have interests in or are working in the digital library field.

6.1 Evaluation Design

It is difficult to evaluate a visualization tool objectively. Therefore, we conducted an analytic formative evaluation including user interviews. Our evaluation consists of two classes of sessions, answering sessions and interview sessions. Unlike a summative evaluation, where the goal is to prove the effectiveness of software statistically with many random participants, formative evaluation often aims to collect professional suggestions from several domain knowledgeable participants, as the product is being developed [35]. Eight Ph.D. students majoring in computer science were recruited, making sure they have basic knowledge on the topics of digital library, data mining, and information visualization. Participants were given enough time to become familiar with the VUDM and then were allowed to ask any questions that came to mind. After this process, they were asked to evaluate the effectiveness of VUDM with regard to providing each of five types of knowledge that might be sought by digital librarians:

- a. Information seeking trends
- b. Virtual interest group distributions
- c. User characteristics
- d. Trends in the near future
- e. Drift of concepts

In the answering session, participants could answer either ‘negative’ or ‘positive’ for each question. If they selected ‘positive’, they were asked to select the degree of agreement from 1 to 10. During the interview session, participants were asked to comment on VUDM’s problems and to make any suggestions that came to mind.

6.2 Results and Conclusions

All participants answered positively for all questions, except two questions were answered negatively by one participant (see below). The average (non-negative) scores for each question were 89%, 85.5%, 86.2%, 75.8%, and 69%, respectively. During the interview session, most participants indicated difficulties with understanding some of the features of the visualization. For example, some were confused about groups and their locations. Some didn't understand the reason that there are no labels for groups and users. The fact is that VUDM characterizes user and group based on set of topics (the user and group involved with), and provides topic tables which consisted of hundreds of topics ordered by frequencies, instead of labels. One negative answer was about the question 'c', using the topic tables. The participant commented that the topic tables don't work with visualization because they contain too much detail information. The other negative answer was about question 'd'. It is difficult for VUDM users to spot changes in usage trends since they must see multiple pictures about usage trends for the past several months to predict the next month. The participant commented that VUDM should provide better visualization for this task, such as animation or colored traces showing changes. Since our approach is new, it is not surprising that some users were confused about the novel features of VUDM. Further testing, with more time allowed for users to become familiar with our approach, is needed. Another problem we identified is that our user model data is just cumulative. It is not easy to determine if and when a topic goes out of favor. If we worked with sliding windows covering different time periods, we might solve such problems.

Also, because the NDLTD union catalog covers all scholarly fields, and we only had 1,200 registered users, finding virtual interest groups was hard. Adding more user data or applying VUDM to subject-specific digital libraries, like CITIDEL [3] or ETANA-DL [4], should solve this problem.

Finally, privacy issues were identified. Devices and modifications were requested to secure personal sensitive information, such as user IDs.

Chapter 7. Algorithm Evaluation and Enhancing Scalability

This chapter presents the results of an evaluation which was performed to compare the precision of FSWMC, and its two variations, with two of the most related popular clustering algorithms, *k*NN and Tanimoto, in finding virtual user communities, and their properties. Also, a further analysis of the FSWMC algorithm regarding scalability, and its modifications to solve the problem, will be discussed.

7.1 Community Finding Performance

The challenging fact of comparing the performance of community finding is to devise an objective method to measure the correctness of virtual user communities. Asking all users, whether their virtual user community assignments are correct or not, is very expensive and subjective. Also this way is likely to intrude on their privacy [88].

Our approach is to build a test collection which is a collection of “research interests” and “learning topics” from well-classified Electronic Theses and Dissertations (ETDs) [6, 48] and ACM Transactions [1], so that the correctness of assigning users in a group into a certain user community can be calculated by matching users’ profiles and this collection using set theory.

Research interests and learning topics of patrons in complex information systems constitute basic information to help provide them with advanced personalized services. Although this information usually is provided by the users explicitly, e.g., through filling out registration forms and answering online surveys and questionnaires, some researchers have emphasized the importance of implicit ways to collect the information. Kelly, in her dissertation [41], tried to figure out users’ document preferences based on their implicit feedback tendencies. Nichols [58] and GroupLens [50] showed the great potential of implicit data from users. Human behavior analysis and modeling in information systems, such as by Fidel [29], Belkin [13], and Lin [52], also are techniques to characterize users and communities for personalized service. Bollen suggested user centered approaches to digital library evaluation, using retrieval patterns of a user and her communities, to solve many parts of problems [14]. However, there are few methods and tools developed to

evaluate these efforts. Therefore, the goal of this study is to develop such an evaluation tool in support of personalization.

Research interests and learning topics of a person are normally represented by a set of noun phrases, for example, "digital library", "complexity analysis", and "web log standardization". Anick [10] also emphasized the usefulness of noun phrases in information retrieval, such as succinctness of describing "concepts", ease of detection and extraction, tightness of expression, appearance in actual queries, etc.

Therefore, we will identify a digital library user's research interests and learning topics by examining the noun phrases extracted from documents she referenced in the digital libraries, because the title, abstract, and body consist of noun phrases of the documents that provide reasonable hints about her research interests and topics.

7.2 Building a Test Collection

The overall goal of this study is to build a collection of research interests and learning topics for various scholarly areas through analyzing well-classified documents, and to make it useful by storing it in a database and providing tools, such as Java Application Programming Interfaces (APIs) and web-based user interfaces. Figure 20 depicts the overview of all stages to achieve the goal. In stage (1), a large set of documents is collected. This set includes the Networked Digital Library of Theses and Dissertations (NDLTD) ETD Union Catalog [6], which is a collection of ETD metadata provided by 325 member institutions [57], such as universities and libraries. Although this collection covers all scholarly areas, we added the ACM DL (Digital Library) Transactions [1] to our data analysis to help guide our focus on computing, because according to our preliminary research [48], many users and information demands in our target digital library, NDLTD, are in the computing field. The NDLTD ETD Union Catalog is collected by using the OAI/ODL Harvester 2.0 [77, 78], and the ACM DL Transactions are crawled with a web crawler, SPHINX [7, 55, 60]. NDLTD ETDs are classified into 76 categories of research interests and learning topics, henceforth referred to as "categories". ACM Transactions are selected because they are already well-classified into 26 specific topics of computer science and engineering by experts; see the row labeled 8

in Table 2. In stage (2), 102 categories are predefined for classification as shown in Table 2. The method we used to classify the documents is specified in the next section.

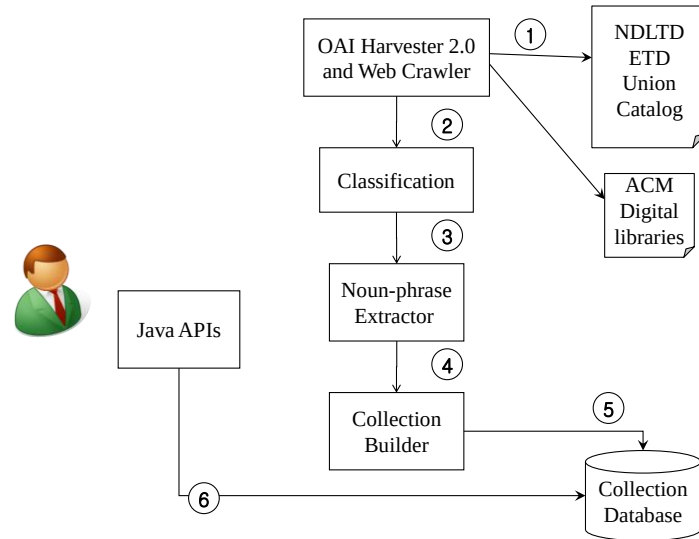


Figure 20. Workflow to build a collection of research interests and learning topics

In stage (3), a noun phrase extractor, developed at the University of Arizona [84], is employed to extract meaningful noun phrases that represent research interests and learning topics. The noun phrase extractor is tuned to focus on title and subject elements of each ETD's metadata, see Figure 22, because these parts were entered by its author directly and contain carefully selected phrases that are expected to represent the ETD. The description element, which contains the abstract of the paper, was excluded because it turned out that including the description just adds noise, reducing the accuracy of the whole result. For documents of ACM Transactions, title and keywords fields were analyzed.

In stages 4 and 5, the results of the document classifier and noun phrase extractor were classified into research interests and learning topics. They were merged and converted to SQL queries, and entered into the database. Figure 21 is a logical view of the collection database after stage 5. Research interests and learning topics are stored with their category names along with their frequency in the source documents.

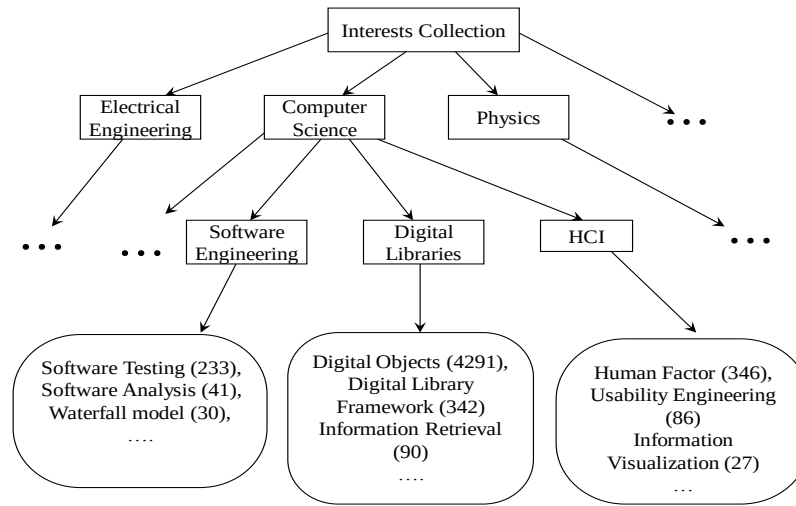


Figure 21. Hierarchical structure of the collection

To facilitate the use of this collection, some Java APIs are implemented for developers and web applications, at stage 6. These APIs include several presentation functions, such as an extraction function dealing with statistical information about the collection, interactive functions implemented with SQL to manipulate the collection, and calculation functions to estimate categories for given user data.

7.3 Document Classification and Noun Phrasing

ACM Transaction documents are already well-classified according to 26 specific topics in the computing fields, but, classifying the NDLTD ETD union catalog is a challenge. Each ETD record is in XML format and contains meta-information describing the document, such as Dublin Core information, title, subjects, description, publisher, date, etc., like the sample given in Figure 22. Classification of an ETD was achieved by taking advantage of classification information entered by its author, subject elements, and using a keyword-mapping-based ad-hoc algorithm for the title and subject elements.

```

<dc oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" dc=http://purl.org/dc/elements/1.1/
xsi="http://www.w3.org/2001/XMLSchema-instance" schemaLocation="http://www.openarchives.org/
OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd"><title>Composer-Centered Computer-Aided
Soundtrack Composition</title><creator>Vane, Roland Edwin</creator><subject>Computer Science</subject>

```

```

<subject>human computer interaction</subject><subject> music composition</subject><subject>soundtracks
</subject><subject>creativity </subject><description>For as long as computers have been around, people have
looked for ways to involve them in music.... </description><publisher>University of Waterloo</publisher>
<date>2006</date><type>Electronic Thesis or Dissertation </type><format>application/pdf</format><identifier>
http://etd.uwaterloo.ca/etd/revane2006.pdf</identifier><language>en</language><rights>Copyright: 2006

```

Figure 22. An example of ETD metadata

Classification is targeted to classify the source documents into a predefined set of 102 categories, including categories from the ACM DL, as listed in Table 2. The category names and keyword sets for each category are identified by surveying faculty/college systems of universities. We referred to the websites of five universities in Virginia: Virginia Tech, University of Virginia, George Mason University, Virginia Commonwealth University, and Virginia State University. This table has an extended version of categories which we used for our preliminary work [48], which aimed to reveal the amount of information supply and demand in each scholarly field in ND LTD, by adding the ACM DL collection to specify the computing field. The goal of classification in this study is finding representative documents for each category, rather than classifying all documents precisely. Therefore, ambiguous or inter-category documents were excluded.

Table 2. 102 categories of research interests and learning topics

	8 high levels categories	102 categories
1	Architecture and Design	Architecture and Construction, Landscape and Architecture
2	Law	Law
3	Medicine, Nursing and Veterinary Medicine	Dentistry, Medicine, Nursing, Pharmacy, Veterinary
4	Arts and Science	Agriculture, Animal and Poultry, Anthropology, Apparel and Housing, Archaeology, Art, Astronomy, Biochemistry, Biology, Botany, Chemistry, Communication, CropSoil and Environment Sciences, Dairy Science, Ecology, Engineering Science, English, Entomology, Family, Food, Foreign Language Literature, Forestry, Geography, Geology, Government International Affair, History, Horticulture, Hospitality Tourism, Human Development, Human Nutrition and Exercise, Informatics, Interdisciplinary,

		Library Science, Linguistics, Literature, Meteorology, Mathematics, Music, Naval, Philosophy, Physics, Plant, Politics, Psychology, Public Administration Policy, Public Affair, Sociology, Statistics, Urban Planning, Wildlife, Wood, Zoology
5	Engineering and Applied Science	Aerospace, Biological Engineering, Chemical, (<i>Computer Science and Engineering</i>)*, Electronics, Environment, Industrial, Materials, Mechanics, Mining and Mineral, Nuclear, Ocean Engineering
6	Business and Commerce	Accounting and Finance, Business, Economics, Management
7	Education	Education
8	Computer Science and Engineering	Algorithms, Applied Perception, Architecture and Code Optimization, Asian Language Information Processing, Autonomous and Adaptive Systems, Computational Logic, Computer Systems, Computer Human Interaction, Database, Automation of Electronic Systems, Embedded Computing, Graphics, Information Systems, Information and System Security, Internet Technology, Knowledge Discovery, Mathematical Software, Modeling and Simulation, Multimedia, Programming Language, Sensor Networks, Software Engineering, Speech and Language Processing, Storage, Bioinformatics, Networking

* The Computer Science and Engineering category is expanded in row 8.

The next step is extracting noun phrases from the classified documents. University of Arizona's AZ Noun Phraser [84] is used and tuned to exclude out-of-interest nouns. Along with noun phrases, term frequency, tf , is extracted to obtain the frequency in the whole document set, $\sum tf$. Brief statistics of source data, as well as results of classification and noun phrasing, are given in Table 3.

Table 3. Raw data for results of analysis

Source	Number of Records	Number of Categories	Number of Distinct Noun Phrases Extracted
NDLTD ETD Union Catalog	242,688	76	77,831
ACM DL	3395	26	1,645

7.4 Application: Measuring Accuracy of Community Finding Algorithms

A major application of our research is an evaluation tool for user clustering, or community finding, which is an essential procedure for almost all personalization techniques and recommender systems. The VUDM [43, 46], Visual User model Data Mining tool, is a visualization tool to provide both an overview and details of users, user communities, and usage trends of digital libraries. Community finding also plays an important role in VUDM.

Figure 23 shows how much the result of VUDM depends on various community finding algorithms. Three figures present user model data from 21,900 users of June 2006, collected by VTLS.com [87]. Users and user communities are represented with icons and spirals, respectively. The distances between user icons, or spirals, represent how similar they are.

There are many user clustering methods studied, but, there was no known method to evaluate the performance, and accuracy, of community finding algorithms, except asking the users directly, or conducting a user survey, whether it is online or offline. However, requiring users to enter information explicitly is expensive, hard to achieve when the communities are large, and can lead to violation of privacy [88]. In this section, we describe an experiment we conducted to see how the collection of research interests and learning topics could be used as an evaluation tool for measuring the accuracy of community finding algorithms: k NN, Fixed Window Size Multi-Classification (FSWMC), and Tanimoto.

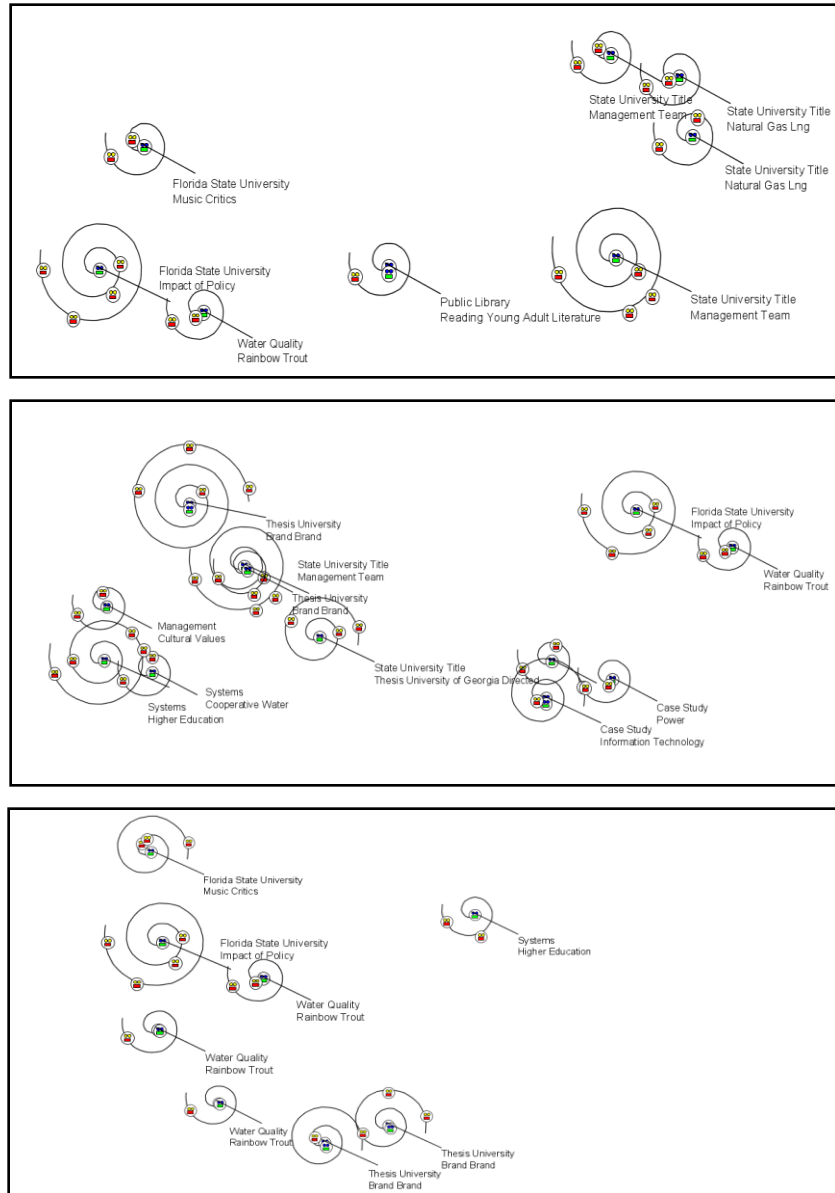


Figure 23. Visualization of user communities in the NDLTD collection as of June 2006. User communities were found by *kNN* (top), *FSWMC* (middle), and *Tanimoto* (bottom) algorithm.

There are many user clustering methods studied, but, there was no known method to evaluate the performance, and accuracy, of community finding algorithms, except asking the users directly, or conducting a user survey, whether it is online or offline. Requiring users to enter information explicitly is expensive, and hard to achieve when the communities are large. In this section, we describe an experiment we conducted to see

how the collection of research interests and learning topics could be used as an evaluation tool for measuring the accuracy of community finding algorithms: k NN [25, 27], Fixed-Size Window Multi-Classification (FSWMC) [45], and Tanimoto, a.k.a. Jaccard similarity coefficient, [27, 81].

- k NN: Similarity is based on the cosine correlation. We find the user community in the nearest k neighbors of a given user. Performance depends on the number of users in the whole system [27].
- FSWMC: Similarity is based on cosine correlation. In contrast to k NN, we find the user community within a given threshold of grouping strictness. Performance depends on the given threshold [45].
- Tanimoto: Similarity is based on simple set comparison. This has a low computational cost [27].

In order to measure the accuracy of community finding algorithms, we implemented a function, which is based on the cosine coefficient, that estimates categories to which a given user belongs, and their probabilities. Using this function, the accuracy of a community finding algorithm is measured by calculating the average of the overlap ratio of matched categories among users within a group, see Equation (5). If all members in each group are estimated to belong in the same categories, this value will be 1, otherwise it will approach 0 as the number of matched categories is reduced.

$C_n = \langle \text{Vector} \rangle \text{EstimateCategories}(\text{UserModel } n);$

$$\text{Accuracy} = \frac{|G|}{\sum_{g \in G} \frac{|\bigcup_{n \in g} C_{n,w}|}{w}} \quad \text{Equation (5)}$$

In this equation, G is the set of all interest sharing groups among users, and $|G|$ is the cardinality of set G , which is the total number of elements of set G . g represents an interest sharing group in G , and n represents a user in group g . $C_{n,w}$ is the first w most highly rated research interests and learning topics categories, i.e., “categories”, of user n ,

where w is the size of the window. The size, or tolerance, of the window is the range of categories considered meaningful, from the estimated list of belonged to categories. Two users are considered to be grouped correctly if both share a category within the tolerant window of size w within their estimated categories. For example, let user a 's categories and user b 's categories be estimated as $\{c_1, c_2, c_3, c_4\}$, and $\{c_3, c_1, c_2, c_4, c_{99}\}$ in descending order of probability, respectively. Then, if w is 1, user a and b will not be considered to be sharing any research interests and learning topics because their first estimated categories c_1 and c_3 are not the same, and we assume only that the first ranked categories are meaningful. However, if w is 2, they will be considered as sharing some research interests and learning topics because c_1 exists in both users' first two estimated categories. In this paper, we set w to the value 3. Because most users in the raw data we used belonged to more than 35 categories, selecting 3 for w is strict enough. One limitation of this method is that it calculates only the correctness of user communities which are already found but doesn't consider the user communities that should be found.

Figure 24. EstimateCategories Function

Figure 24 and Figure 25 depict how the accuracy of measuring a user community works using the `EstimateCategories()` function. `EstimateCategories()` considers the collection of research interests and learning topics which is described in Section 7.2 to identify a set of categories that the user belongs to. The input of this function is a user model, which is described in Section 4.3, and the output is a list of categories and their probabilities of matching in descending order.

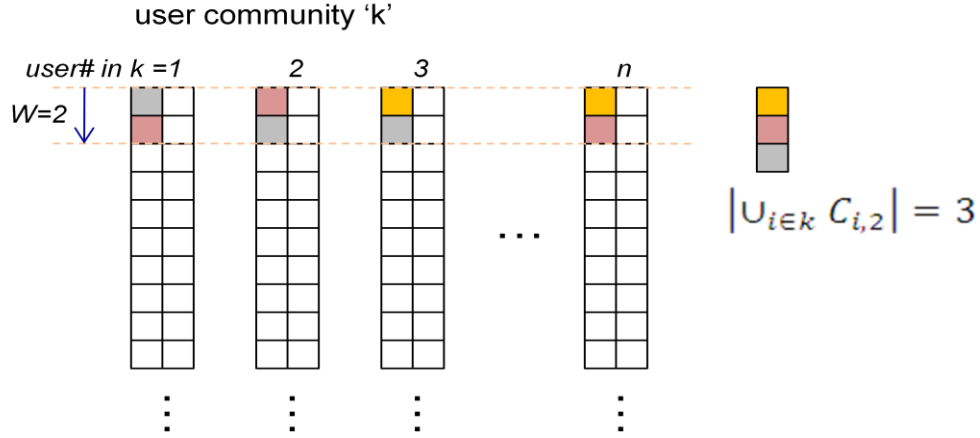


Figure 25. Counting shared interests using the output of EstimateCategories Function

Figure 25 shows how the output of the EstimateCategories() function is used to calculate the number of shared research interests and learning topics among users in a community. Only the first w highly ranked categories for each user are used for calculation. If all users in a community share the same w categories, the cardinal number of the union of w categories of all users in the community will be w , which means the accuracy value of the community is 1. If all users in a community have different sets of w categories, the cardinal number of the union of w categories of all users in the community will be 'n' times w , where the 'n' is the number of users in the community, and the accuracy value of the community will approach 0.0 as 'n' increases.

The raw data used for this evaluation is described in Table 4.

Table 4. Raw data for evaluation

Source	Number of Records	Data Size	Collection Period
NDLTD [6] log data collected by VTLS [87]	267,790 records (30,600 users)	29,250 KB	June 2005 to December 2006

The raw data consists of user models of 30,600 users in NDLTD, collected by VTLS, which were converted from standard web log into user model data format as described in Section 4.3. The data were anonymized by removing user names, emails, IP addresses, etc., to protect the privacy of NDLTD users.

7.5 Enhancing Scalability of FSWMC

Besides the three algorithms described in the previous section, two more FSWMC variations, which are modified to run in $O(n \sqrt{n})$ and $O(n)$, will be included in this evaluation. The original algorithm for FSWMC consists of two loops, one of which is located in the other. Each runs in $O(n)$, as listed below, and thus its asymptotic complexity is $O(n^2)$. Below is the rough algorithm structure of FSWMC.

```
1:      for each user {                                ← outer loop
2:          generate a new empty class;
3:          for each user {                            ← inner loop
4:              assign all users enclosed by radius =  $\theta$  circle surrounding this item
5:              into the new class;
6:              if this class is a sub-class or super-class of an already found class
7:              then discard the sub-class;
8:          }
9:      }
```

This way of community finding needs enhancement because most users have used the digital library just once or twice, and including all the inactive users for community finding is computationally complex and noisy. In order to support this, we analyzed the distribution of users and their activities as follows.

Figure 26 and Figure 27, below, show the distribution of users in terms of the “Degree of Activity”, henceforth DOA, which is defined as the sum of the number of queries and the number of items in the library used and referred to by the user, as represented by Equation (6).

$$\text{Degree of Activity (DOA) of user 'a' = number queries used by 'a' +} \\ \text{number of items referred to by 'a'}$$

Equation (6)

Even though there are many exceptions, it is understandable to assume that regular and active customers will have higher DOA. A curve labeled as “over N ” shows the increasing number of users whose DOA is greater than N at the time shown on the horizontal axis of the point. The figure shows the total number of users increased rapidly

from 3,500 to 30,600 during the period, while the number of active users increased slowly.

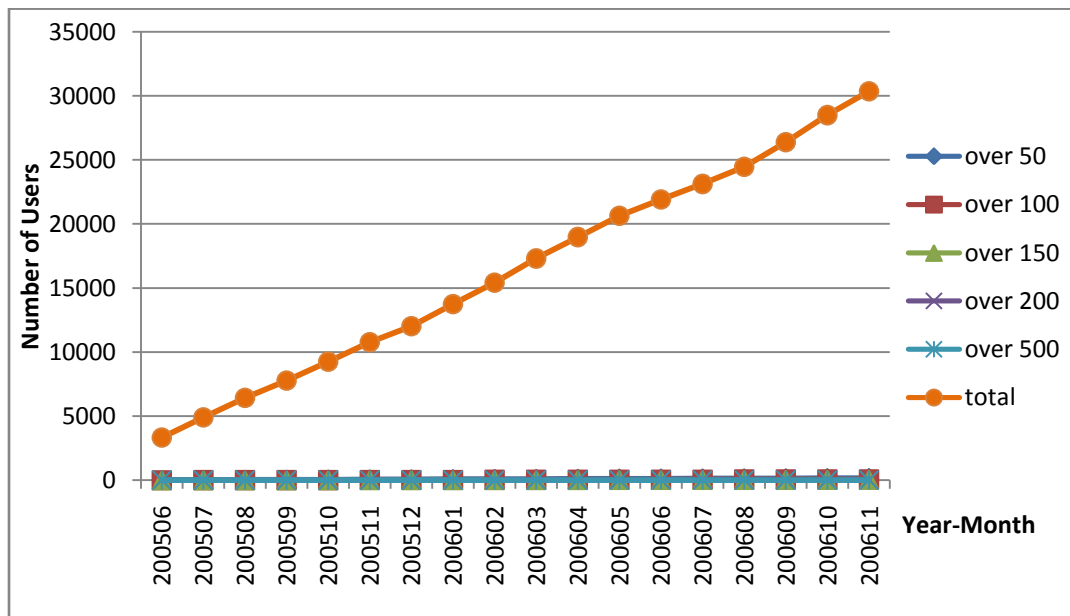


Figure 26. Change of the number of total users and active users

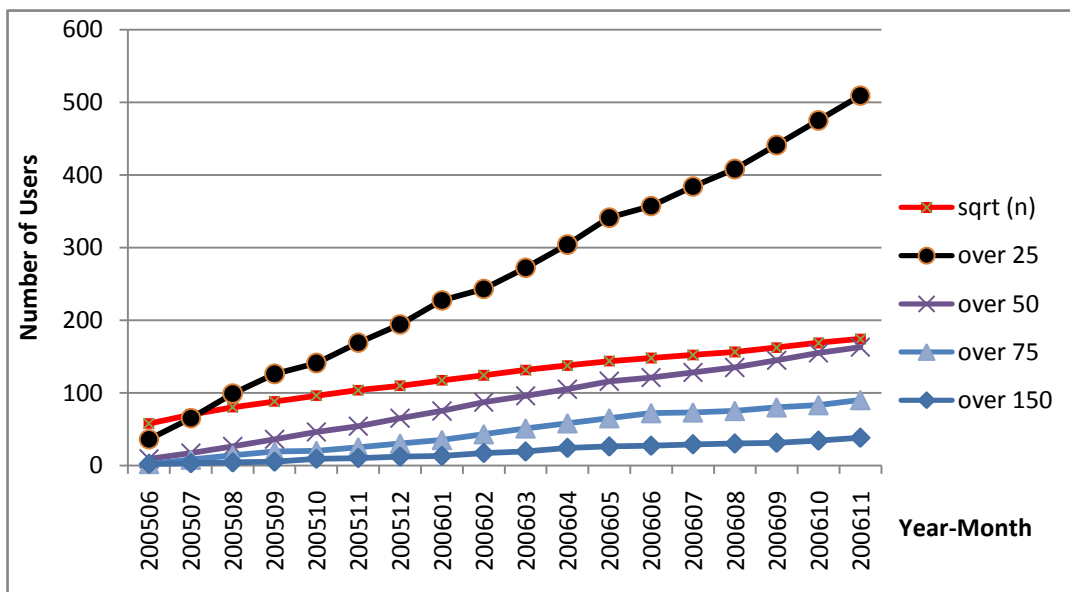


Figure 27. Change of the number of active users in terms of DOA, zoomed-in from Figure 24.

In Figure 26, the number of active users is too small to be shown in the same scale with the number of total users. Figure 27 provides a zoomed-in figure of the curves of active users which were hard to distinguish in Figure 26. The curve labeled as “sqrt (n)” shows the change of sqrt of the number of whole users, n. This curve is added to find out a threshold, γ , which makes the complexity of each loop lower than $\sqrt{n}$. The way to find γ using this curve will be described later when we talk about the first variation of FSWMC.

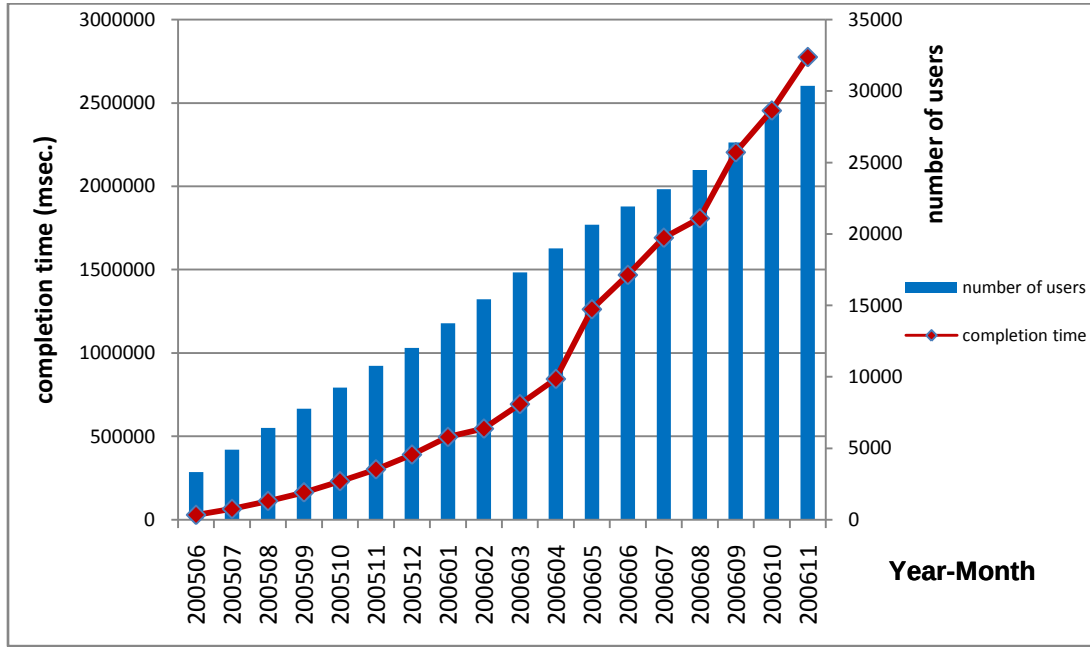


Figure 28. Completion time of FSWMC runs in $O(n^2)$

Figure 28 shows how the completion time of community finding, the polygonal line, increases as the number of user data, the bars, increase linearly as time passes. The completion time increases in speed with complexity $O(n^2)$ and it takes almost 50 minutes to complete the community finding at the last month of the data, December 2006, when the number of users was about 30,600. Throughout this chapter, the completion time is measured on a Pentium-4, 2.0 GHz Windows system.

The first variation of FSWMC, which runs in $O(n \sqrt{n})$ time, is made by adding a condition to line 1, such as “for each user, whose ‘degree of activity (DOA)’ is greater than γ ”, so that the iteration of the loop is performed only for active users (whose number is always less than $\sqrt{n}$); which is the necessary condition of γ to

guarantee that the number of iterations of the loop is less than $\sqrt{n}$. In the increasing rates of active users and $\sqrt{n}$, as seen in Figure 27, the number of users, whose DOA is greater than 75, will never exceed $\sqrt{n}$ with an assumption that the population of whole users will increase, in such a way that maintains the current, or observed, increased rates of users in each DOA. Therefore, selecting 75 or a greater value for γ is safe, so we can say “the number of active users, whose DOA is greater than 75, is always less than $\sqrt{n}$ ”. In this paper, we select 75 for γ .

This modification is reasonable because it prevents community finding among non-active users and puts active users at the center of communities. Moreover, this variation doesn't exclude any users from being considered in community finding. Figure 29 shows the completion time of the first variation of FSWMC, which runs in $O(n \sqrt{n})$. The completion time is reduced to about 10% of the original FSWMC, which runs in $O(n^2)$.

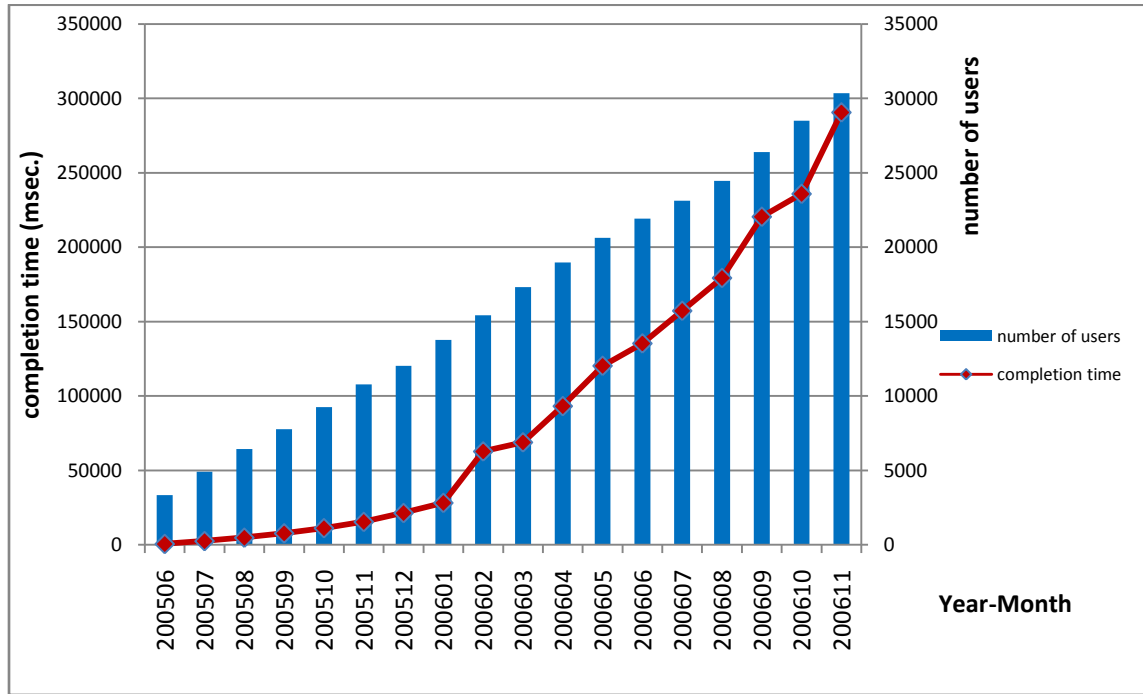


Figure 29. Completion time of FSWMC runs in $O(n \sqrt{n})$

The second variation of FSWMC is achieved by applying the same condition to the inner loop, which starts from line 3 of the algorithm structure. Because both the inner and the outer loop run for $\sqrt{n}$ active users, the total asymptotic complexity is $O(n)$. This

variation will make community finding very fast because it concentrates only on highly active patrons of the digital library and their number is much smaller than the number of total users. Figure 30 shows that the completion time of community finding of this variation is just about 4% of that of the $O(n \sqrt{n})$ variation.

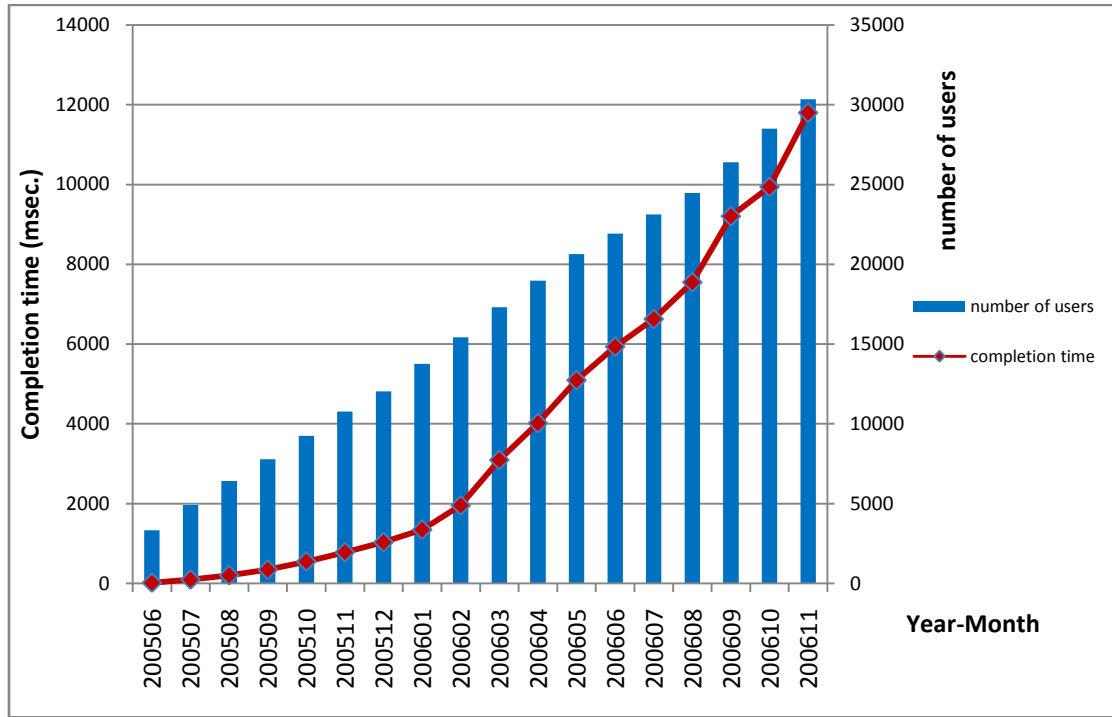


Figure 30. Completion time of FSWMC runs in $O(n)$

However, since now most users are excluded from being considered for community finding, and the effects of active users' behaviors are not known, the accuracy of this variation needs to be evaluated.

The difference of the completion time of the three FSWMC algorithms described above is clear when the three completion time curves are shown in a single chart (see Figure 31).

For a comparison of efficiency, the completion time of k NN and Tanimoto algorithms were overlapped with that of FSWMC and its two variations as shown in Figure 32. This chart shows that the efficiency of original FSWMC, which runs in $O(N^2)$, is similar to that of k NN and Tanimoto. This result is explained by the fact that both the k NN and Tanimoto algorithms run in $O(N^2)$, like FSWMC. 10 is user for the 'k' value of k NN throughout the evaluations of this chapter.

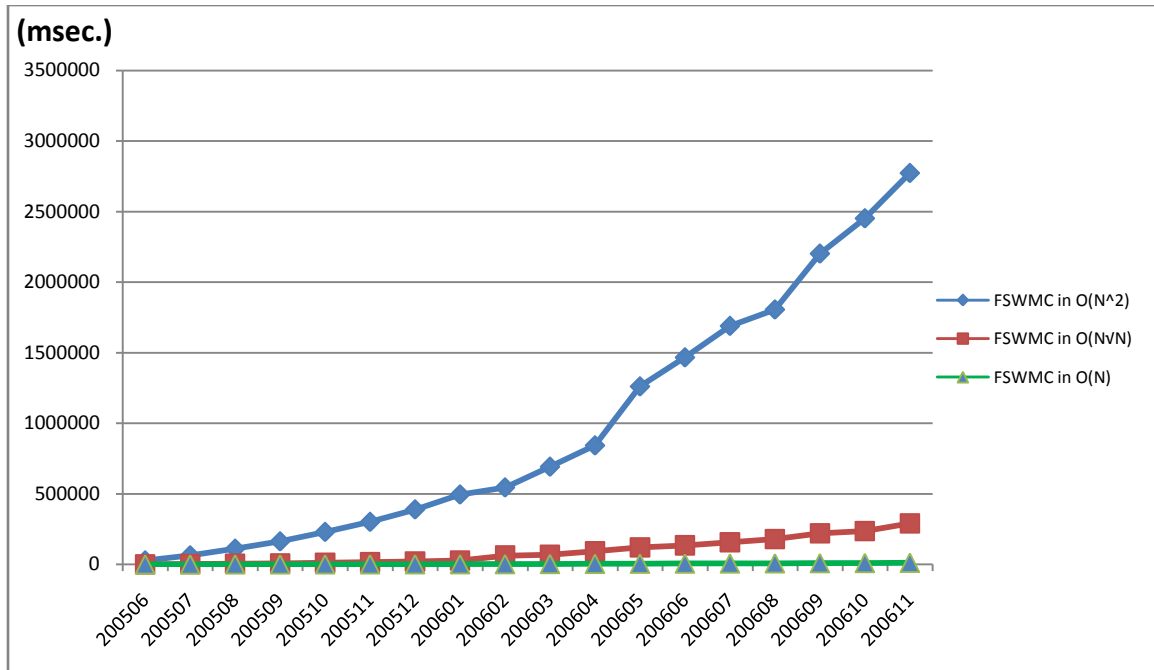


Figure 31. Comparison of completion times of the FSWMC and its two variations

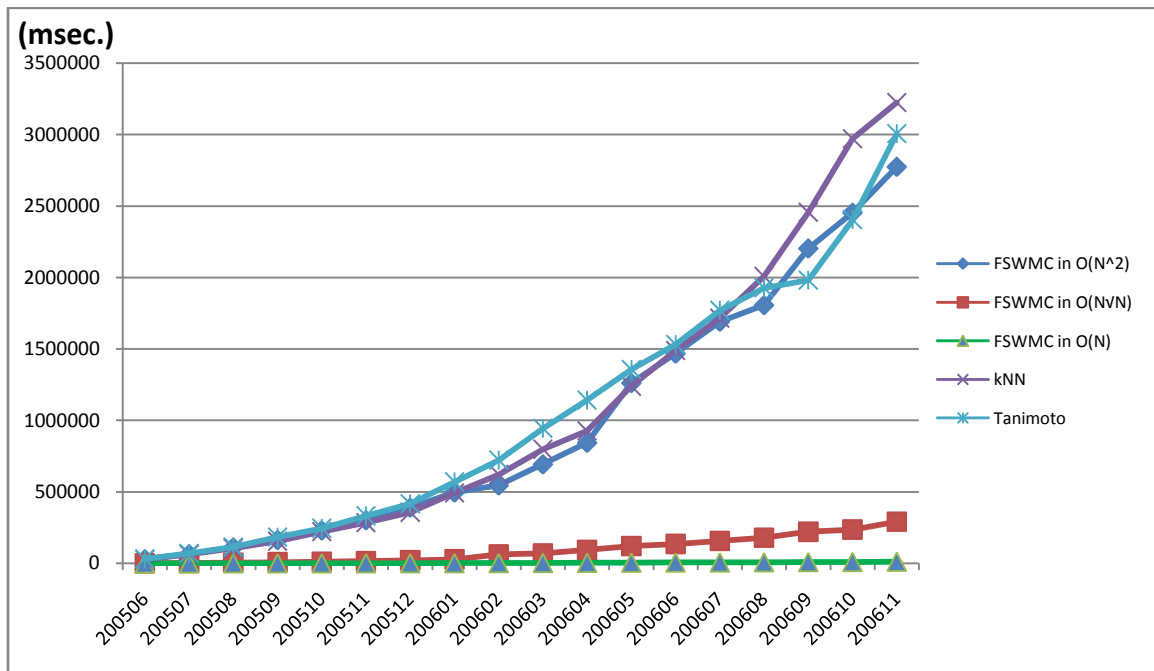


Figure 32. Completion times of the kNN, Tanimoto, FSWMC and its two variations

7.6 Results and Conclusions

Figure 33 shows the accuracy, calculated using Equation (5), of user communities found by two popular clustering algorithms, kNN and Tanimoto, and three FSWMC variations, as the number of discovered user communities changes. kNN showed good accuracy when it found more than about 100 user communities, however, in that case, most groups consist of just one or two users, see Figure 34, which has little practical value.

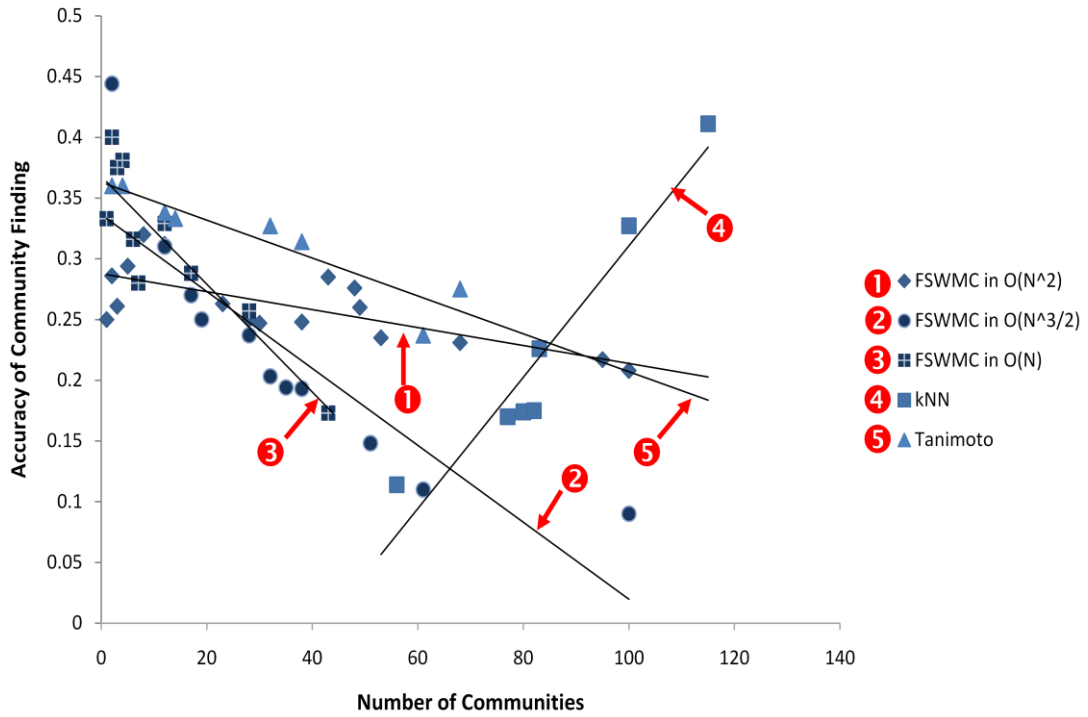


Figure 33. Accuracies of five community finding algorithms

Tanimoto performed well when the number of communities is small, but the accuracy deteriorates rapidly as the number of groups increases. FSWMC in $O(n^2)$, the original, shows relatively good and stable performance, even though the accuracy was not the best among the three algorithms, regardless of the change in number of communities. FSWMC's $O(n \sqrt{n})$ variation, which avoids clustering among non-active users, was more accurate than its original when the number of communities is less than about 20 but it deteriorates more rapidly than its original. The accuracy of FSWMC in $O(n)$ showed the

most rapid dropping of performance as the number of groups increases. The results of two FSWMC variations are as predicted because 1) these two consider only, or are focused on, highly ranked active users for community finding, and 2) a normal user's research interests and learning topics will vary as she become a more active, and regular patron, while the decision window size w in Equation (5) is fixed.

Based on this evaluation, we can use different algorithms according to the properties of the raw data and user communities we desire to find. That is, 1) the Tanimoto, or FSWMC in $O(n)$ or $O(n\sqrt{n})$ is good to find a small number of accurate user communities with high θ option, the strictness of community finding, and 2) the FSWMC in $O(n^2)$ is proper when we try to find large number of user communities with a generous θ option. The boundary number of communities that makes the decision for the above cases 1) or 2) will increase as the amount of raw data increases.

Figure 34 shows the accuracy of communities regarding to their sizes, which are represented by the number of users in a community. This figure explains why the kNN is not very practical in community finding even though it shows good performance when the number of communities was over 100 as Figure 33 shows. All tested community finding algorithms present similar performance except the kNN , which shows good accuracy when the sizes of communities are two or less. The sizes of communities were two or less when the number of communities was over 100 in Figure 33; we typically don't consider such a small user group, which consists of one or two users, as a user community.

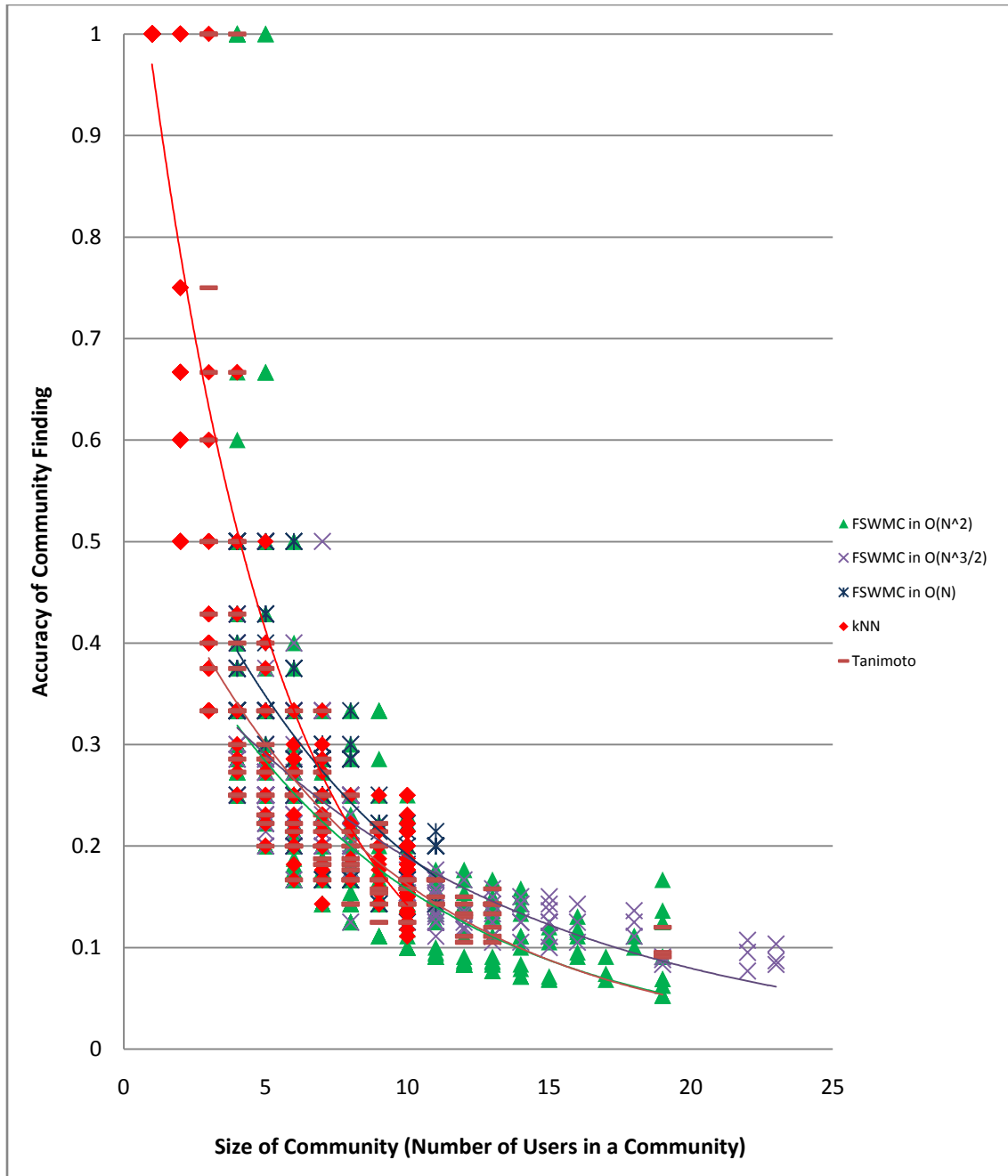


Figure 34. Accuracy and size of community

In this chapter, we proposed a methodology of measuring the accuracy of user communities, which is very expensive, time consuming, likely to lead to violation of privacy, and difficult in other methods. Using the proposed method, it is possible to test any clustering algorithms applied to user community finding, as well as to test their many parameters to increase the precision of user communities – quickly, efficiently, and

without violating privacy. The proposed methodology consists of a collection of research and learning topics extracted from a well classified ETD collection, to cover all scholarly areas, and the ACM Transactions, to cover the computer science field specifically. It is implemented with a database manager and Java API, to make it applicable to other social network related applications. Also, it uses an equation to utilize the collection to evaluate a given set of user communities.

As an example of the application of the methodology, we performed an experiment to compare the accuracies and characteristics of five community finding algorithms, k NN, Tanimoto, FSWMC, and its two variables, using the proposed methodology. For further study, other possible applications of this collection and methodology, such as detecting concept drifts of certain users, visualizing document topics, or visualizing user communities based on interests and topics, also should be studied.

Chapter 8. Summative Evaluation

This chapter presents the procedure and results of evaluations we conducted to evaluate the usefulness of VUDM. The subject set of the evaluation is normal users, who are not library experts, rather graduate or undergraduate students. The evaluation with normal users is performed by measuring accuracy and efficiency of several user tasks using VUDM. The accuracy is measured by checking the answers for user tasks, while the efficiency is measured using the time to complete the tasks. In Sections 8.1, the experimental design is described. In Section 8.2, general data is analyzed while in Section 8.3, the rating data is analyzed. Section 8.4 concludes the discussion of the results of evaluation.

8.1 Experiment Design

As Chen [23] observed from his meta-analysis research, evaluating the effect of visualization techniques normally is achieved by measuring *accuracy* and *efficiency* of user tasks, realistically associated with use of the visualization tool.

Accuracy includes precision, error rate, the average number of incorrect answers and the average number of correct answers.

Efficiency includes average time to complete, and performance time [35].

The purpose of this experiment is to assess the usefulness of VUDM by measuring accuracy and efficiency of VUDM and comparing it with participants' preferred data analysis tools. The IRB approval letter and questionnaires of this experiment are attached in Appendix A. The accuracy of VUDM is measured by testing the accuracy of user tasks, and the efficiency is measured by testing how it helps users to finish user tasks, that is, measuring the time to complete user tasks.

Although some research, such as [69], proposed techniques to measure these variables in the real environment, with open-ended user tests, it is typical to evaluate visualization techniques in controlled studies that measure user performance with pre-prepared tasks and answers. Therefore, we will conduct our evaluation in a controlled environment.

Participants were recruited through the Virginia Tech experiment management system, SONA [73], in a first come first reservation manner until the desired number of participants, 60, had been recruited. Besides that a participant must be at least 18 years old, it is necessary that a participant must be familiar with at least one data analysis software package, for basic analysis functions such as sorting, adding, and drawing bar charts, etc.

In the experiment, participants performed five user tasks using the two types of software, the VUDM and their preferred data analysis software package, allowing us to measure and compare the correctness and completion time of user tasks. To avoid memory effects, two sets of user tasks were prepared, so for half the cases one was done first. One set was for VUDM and the other for the participant's preferred software package. The two task sets are mostly similar in type but the required information to solve the task and the answers were different. Each task set consists of five user tasks, each of which represents one of the five compound knowledge types about the whole population, as mentioned in Section 6.1.

Table 5. Two task sets containing five user tasks

	Task	Type	Question
Set 1	Task 1	Information seeking trends	"In June 2006, more users are interested in Music than Knowledge Management" (True/False)
	Task 2	Interest group distribution	"In June 2006, the 'Power System' related area has more interest-groups than 'Library' related area" (True/False)
	Task 3	User characteristics	"In June 2006, users who are interested in Knowledge Management are also interested in Library Science" (True/False)
	Task 4	Trends in near future	"From the trends from June 2006 to August 2006, there will be more information needs for 'Power System Control' in the next month, September 2006" (True/False)
	Task 5	Drift of concepts	"Someone, who has interest in 'Management' in January 2006, developed a new interest in 'Culture Value' in February 2006" (True/False)
Set 2	Task 1	Information seeking trends	"In June 2006, more users are interested in Economics than Psychology" (True/False)

	Task 2	Interest group distribution	“In June 2006, the ‘Power System’ related area has more interest user groups than ‘Library’ related area” (True/False)
	Task 3	User characteristics	“In June 2006, users who are interested in Economics are also interested in Database” (True/False)
	Task 4	Trends in near future	“From the trends from June 2006 to August 2006, there will be more information needs for Psychology in the next month, September 2006” (True/False)
	Task 5	Drift of concepts	“Someone, who has interest in ‘Case Study’ in January 2006, developed a new interest in Management in February 2006” (True/False)

Table 5 lists the two sets of five user tasks and actual task questions for each task type. VTLS user-model data, described in Table 4 in Section 7.4, is provided for user tasks using VUDM, and the same data in CSV (Comma Separated Value) format is provided for user tasks using other software packages. Data in CSV format is usable by almost all statistical analysis software packages, including Excel, MiniTab, SPSS, Matlab, and SAS. Completion time is measured for each separate task.

The experiment includes answering a pre-test questionnaire, learning VUDM, answering task set 1, and answering task set 2. The orders of task set to answer, and association between task set and software, are deliberately shuffled before the experiments were scheduled, to prevent any possible bias. The learning session includes an oral explanation (by the experimenter) of VUDM and a demo, and the participant’s task activity followed by question answering. Completion time is measured by the participants themselves because pilot subjects argued that being watched during an experiment is uncomfortable.

8.2 General Data Obtained

Using a call-for-participation email and an advertisement posting, attached in Appendix A-1, in the SONA system [73], a total of 60 volunteers were recruited and scheduled for the experiment in a first come first reservation manner. 43 of the participants were undergraduate students and 17 of the participants were graduate students. All of the participants were students of Virginia Tech. Figure 35 shows the

distribution of academic level of participants, and Figure 36 shows the distribution of ages of participants.

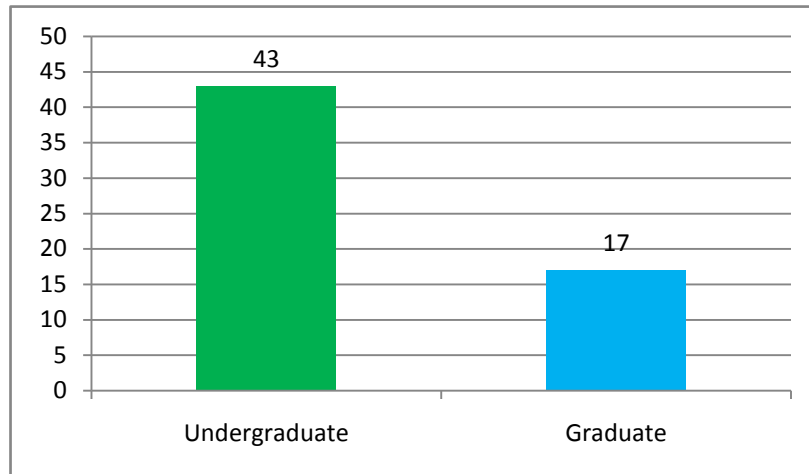


Figure 35. Academic level of participants

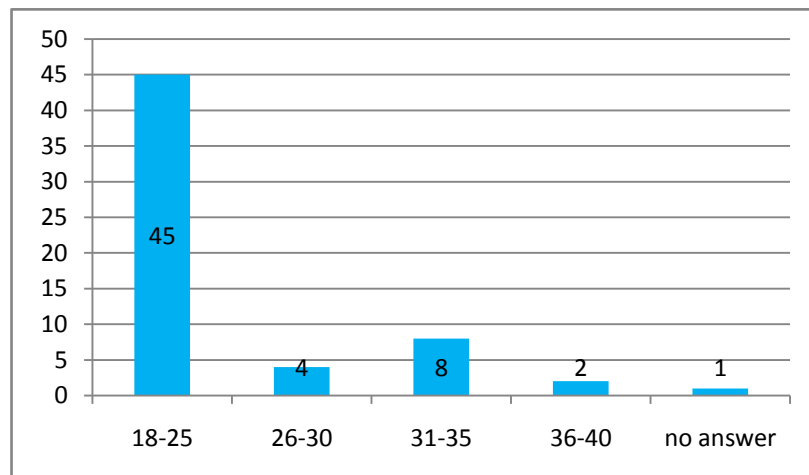


Figure 36. Age distribution of participants

As the recruiting advertisement email clearly requested, participants must be familiar with at least one statistical analysis software package. At the pre-test questionnaire session, participants were asked to select their preferred software. They used both their preferred software and VUDM during their experiments to compare the two. Figure 37 shows the distribution of statistical software packages participants selected. Most participants selected Excel. Two chose MiniTab and two SPSS. Several participants

wanted to change their software in the middle of experiment, from SPSS to Excel and from MiniTab to Excel. In that case, their last selected software was counted.

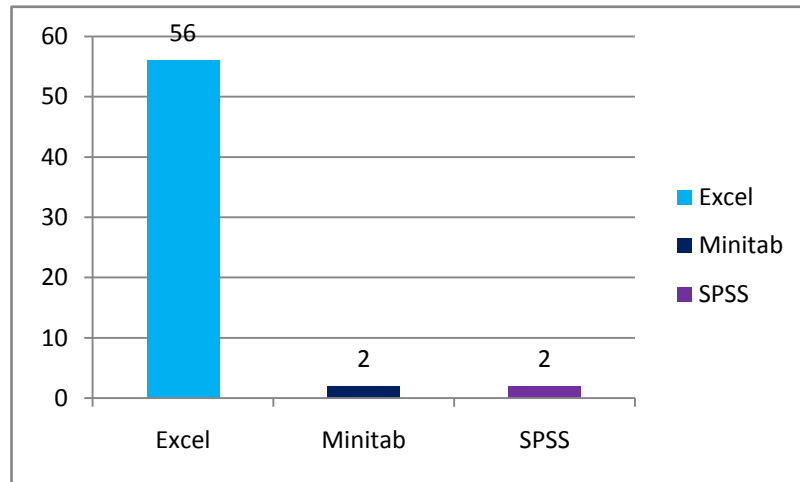


Figure 37. Preferred statistical analysis software package

8.3 Rating Data Obtained

8.3.1 Accuracy

Accuracy is measured by checking the correctness of user tasks. Each task set consists of five user tasks, including one dummy task at the beginning to allow the participant to get familiar with the data and software. 20 points is assigned for each right answer and no points for wrong answers. The user tasks are questions that request the participant to explore a certain part of the data and to answer whether a given statement is true or false. Participants were discouraged from guessing the answer to a task when it is too hard to find an answer with the software they are using for the task. Instead the participant can select “giving up” as an answer. Only correct answers earn 20 points. Giving up answers were considered wrong; wrong answers earned 0 points. Figure 38 shows the average ratio of correct, given-up, and wrong answers for VUDM and other software packages, and Figure 39 shows the average ratio of correct, given-up, and wrong answers for VUDM and Excel. The higher give-up rate of other software packages in this chart can be explained if other software packages are less suitable to solve complex and comprehensive tasks like the given user tasks.

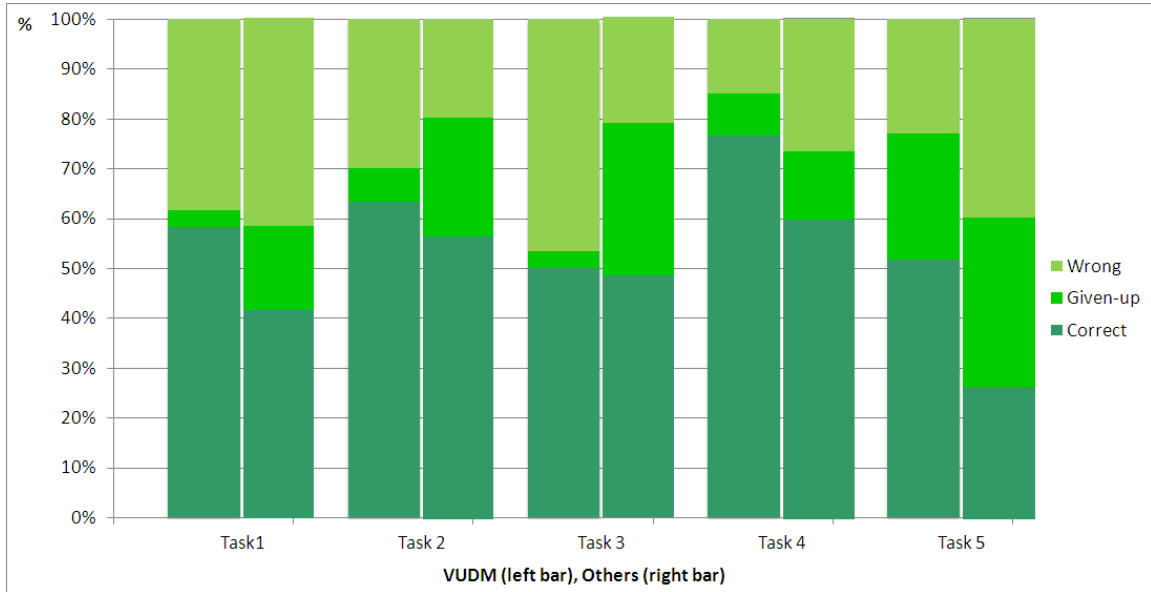


Figure 38. Average ratio of correct, given-up, and wrong answers of VUDM (left bar), and other software packages (right bar)

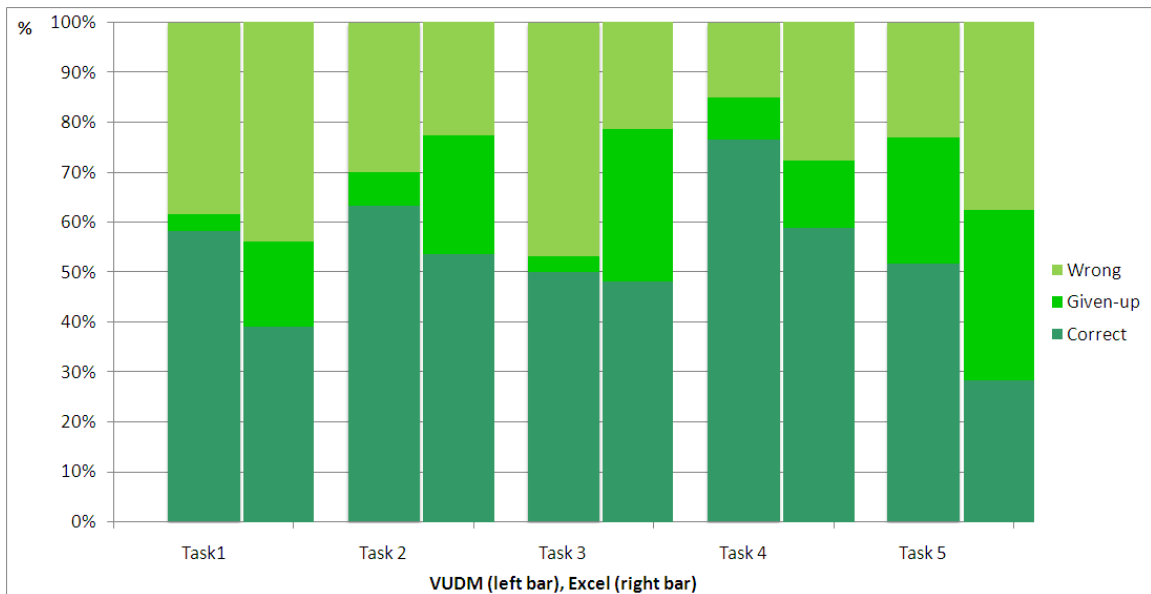


Figure 39. Average ratio of correct, given-up, and wrong answers of VUDM (left bar), and Excel (right bar)

Figure 40 shows average accuracies and confidence intervals of five user tasks of VUDM and other software packages (left), VUDM and Excel (right), and Table 6 presents the result of the t-test. 0.05 is used as the significance level of the t-test

throughout all the analyses in this chapter. The average accuracies of VUDM and other software are 60.00 and 46.67, respectively. The confidence intervals of VUDM and other software are ± 3.46 and ± 3.28 , respectively. The result of the one-tail t-test, $p = 0.001272$, means the difference between two average means, one from VUDM and the other from the other software, is statistically significant. Therefore, it is safe to say that VUDM users are more accurate in performing the given user tasks over other software package users.

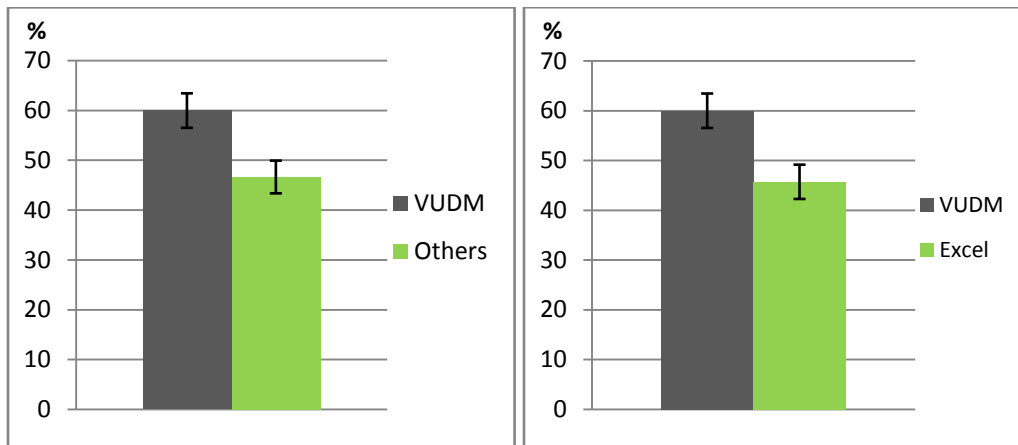


Figure 40. Average accuracy and confidence intervals of overall user tasks

Table 6. t-test result on accuracy of overall user tasks between VUDM and other software packages (top), and between VUDM and Excel (bottom)

t-Test: Paired Two Sample for Means

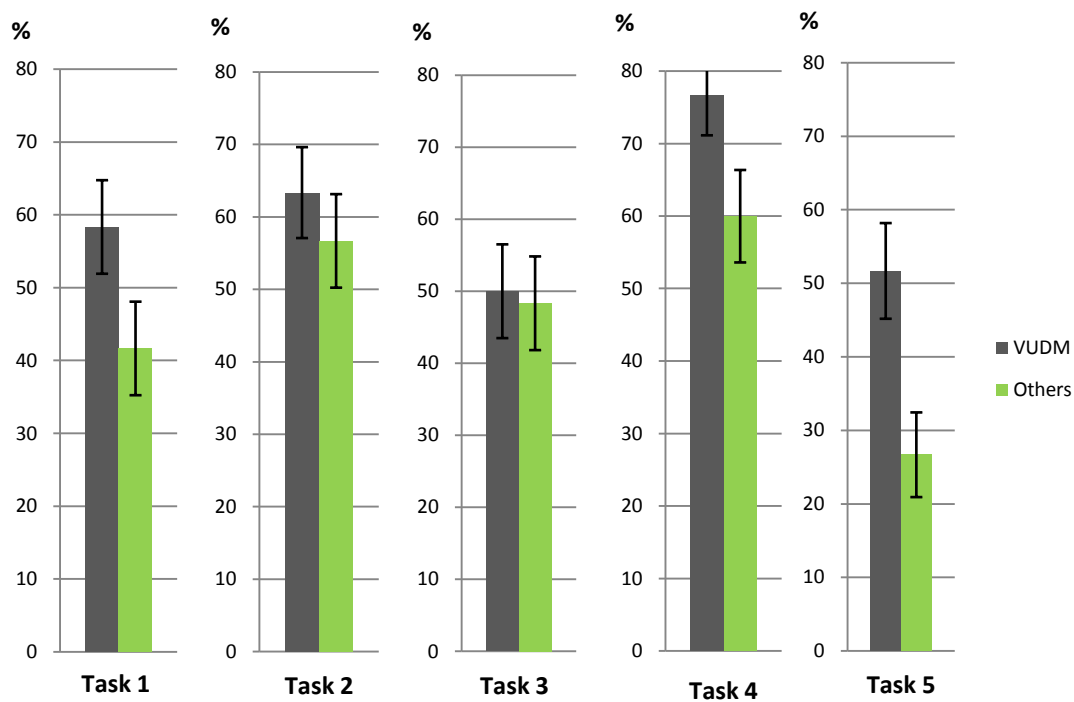
	Variable 1	Variable 2
Mean	60	46.66667
Variance	718.6441	646.3277
Observations	60	60
Pearson Correlation	0.318328	
Hypothesized Mean Difference	0	
df	59	
t Stat	3.384717	
P(T<=t) one-tail	0.000636	
t Critical one-tail	1.671093	
P(T<=t) two-tail	0.001272	
t Critical two-tail	2.000995	

t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	Variable 2
Mean	60	45.71429
Variance	718.6441	664.9351
Observations	60	56

Hypothesized Mean Difference	0
df	114
t Stat	2.925139
P(T<=t) one-tail	0.002078
t Critical one-tail	1.65833
P(T<=t) two-tail	0.004155
t Critical two-tail	1.980992

Figure 41 and Table 7 show average accuracies of VUDM and other software package (top), VUDM and Excel (bottom) for each user task.



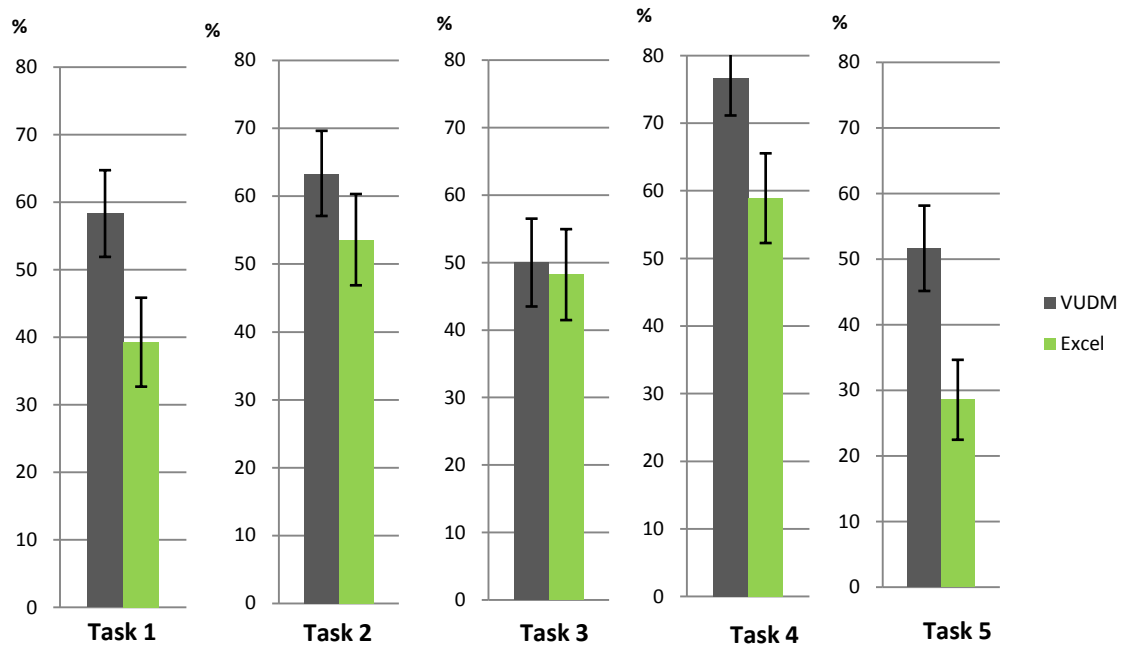


Figure 41. Averages accuracy and confidence interval of each user task

Table 7. Average accuracy and confidence interval of each user tasks

		Task 1	Task 2	Task 3	Task 4	Task 5
VUDM	Average Accuracy (%)	58.33	63.33	50.00	76.67	51.67
	Confidence Interval (%)	6.42	6.27	6.51	5.51	6.51
Others (Excel, SPSS, and MiniTab)	Average Accuracy (%)	41.67	56.67	48.33	60.00	26.67
	Confidence Interval (%)	6.42	6.45	6.51	6.38	5.76
Excel	Average Accuracy (%)	39.29	53.57	48.21	58.93	28.57
	Confidence Interval (%)	6.59	6.72	6.74	6.63	6.09

Table 8 presents the results of one-tail t-test p values, comparing the accuracy of VUDM and other software packages (top row), and VUDM and Excel (bottom row), for each user task. From the result of the t-test, it is observed that VUDM performed better

than other software packages in accuracy for task 4, task 5 and for the overall average of five tasks. Shaded numbers indicate that the difference is statistically significant.

Table 8. t-test result on the accuracy of each user task between VUDM and other software packages

Accuracy (t-test p-value)	Task 1	Task 2	Task 3	Task 4	Task 5	Overall
VUDM and Others	0.033725	0.234921	0.418422	0.015901	0.003011	0.000636
VUDM and Excel	0.020292	0.145377	0.424588	0.021013	0.005405	0.002078

8.3.2 Efficiency

Efficiency is measured by timing user tasks. Completion time is measured in seconds by the participants themselves. The completion times of given-up user tasks are not included in the analysis. Seven users missed recording the completion times for some tasks, and successfully recorded them for some tasks. In that case, only successfully recorded times are included in the analysis. Figure 42 shows the average completion time and confidence interval of five user tasks of VUDM and other software packages, and Table 9 presents the result of the t-test.

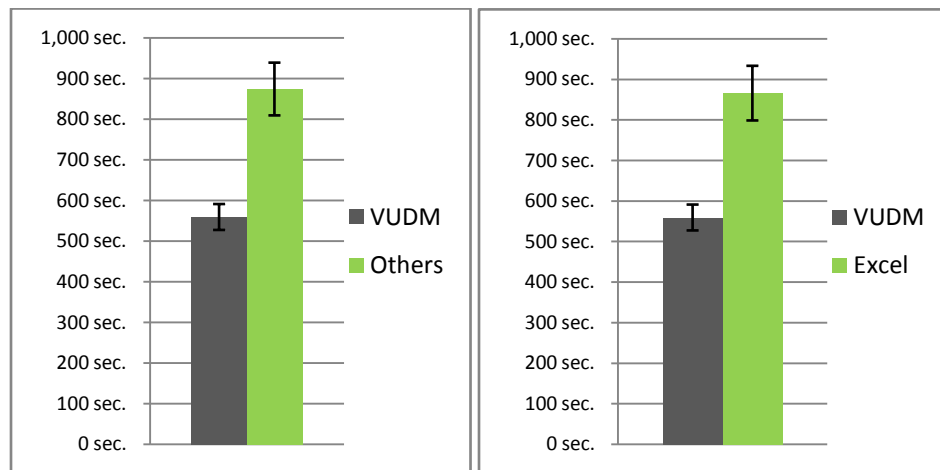


Figure 42. Averages of overall completion time and confidence intervals

The average completion times of other software packages are greater than that of VUDM. The t-test tells that the difference between the two averages is statistically significant, thus users of VUDM take less time to complete the given five user tasks.

Table 9. t-test result on overall completion time between VUDM and other software packages (top), and between VUDM and Excel (bottom)

t-Test: Two-Sample Assuming Unequal Variances

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	559.1731	874.0755
Variance	52834.58	223189.3
Observations	52	53
Hypothesized Mean Difference	0	
df	76	
t Stat	-4.35555	
P(T<=t) one-tail	2.05E-05	
t Critical one-tail	1.665151	
P(T<=t) two-tail	<u>4.1E-05</u>	
t Critical two-tail	1.991673	

t-Test: Two-Sample Assuming Unequal Variances

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	559.1731	865.8776
Variance	52834.58	222481.9
Observations	52	49
Hypothesized Mean Difference	0	
df	69	
t Stat	-4.11452	
P(T<=t) one-tail	5.28E-05	
t Critical one-tail	1.667239	
P(T<=t) two-tail	0.000106	
t Critical two-tail	1.994945	

Figure 43 and Table 10 show average completion time and confidence interval of each user task. For all of the user tasks, using VUDM takes less time.

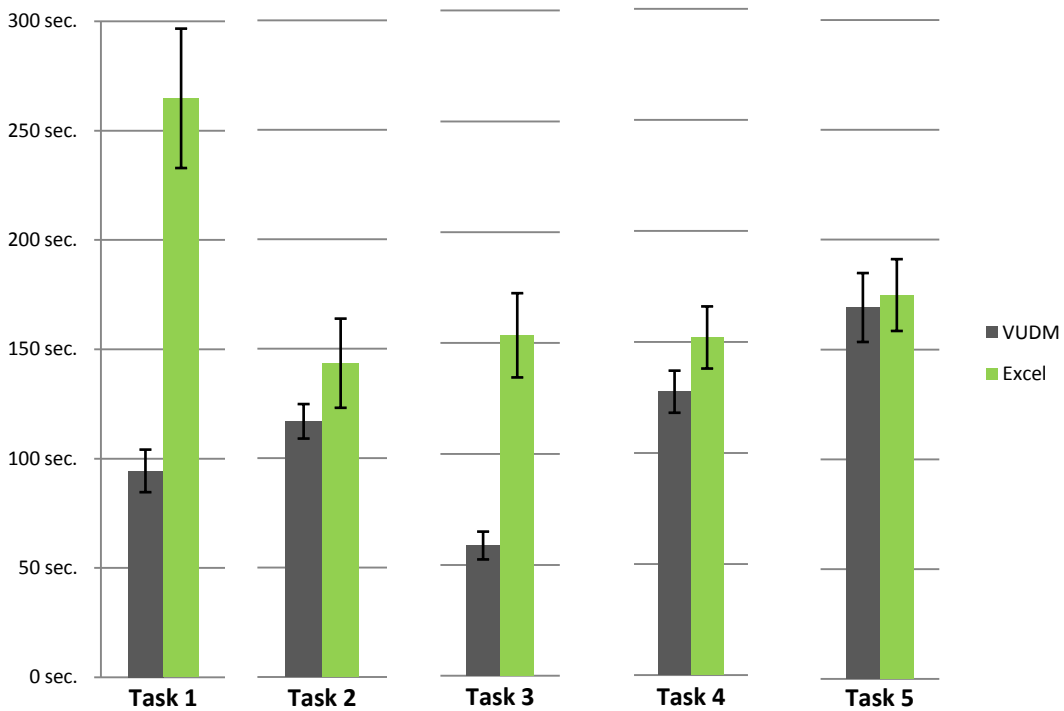
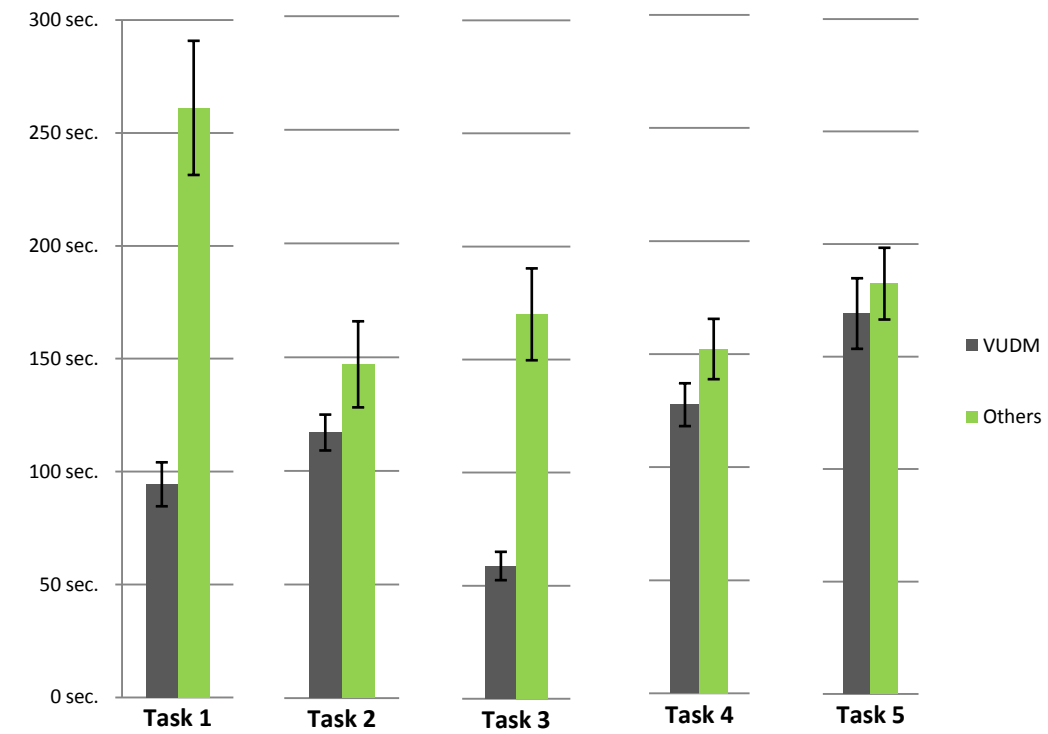


Figure 43. Average completion time and confidence interval of each user task

Table 10. Average completion time and confidence interval of each user task

		Task 1	Task 2	Task 3	Task 4	Task 5
VUDM	Average Completion Time (second)	94.36	116.81	58.72	127.64	169.13
	Confidence Interval (second)	9.74	7.89	6.27	9.47	15.69
Others (Excel, SPSS, and MiniTab)	Average Completion Time (second)	261.09	146.82	170.02	152.26	182.43
	Confidence Interval (second)	29.72	18.94	20.32	13.34	15.95
Excel	Average Completion Time (second)	264.76	143.36	153.48	152.06	174.8
	Confidence Interval (second)	31.88	20.38	18.99	16.35	69.78

The results of one-tail t-test p-values, between the average completion time of VUDM and that of other software packages, are presented in Table 11. Shaded numbers in the table indicate that the difference is statistically significant. Therefore, the result says that VUDM is statistically significantly more efficient for task 1, task 3, and for the overall set of five tasks.

Table 11. t-test result on the completion time of each user task between VUDM and other software packages

Efficiency (t-test p-value)	Task 1	Task 2	Task 3	Task 4	Task 5	Overall
VUDM and Others	6.77E-07	0.074004	1.01E-06	0.067896	0.276823	1.63E-05
VUDM and Excel	1.87E-06	0.114442	6.83E-06	0.075957	0.401492	4.08E-05

From Table 8 and Table 11, we can conclude that the VUDM is either more accurate or more efficient in performing the user tasks except for task 2, in which VUDM doesn't show any significant difference over other software packages. These results were

interpreted, and the positive effects and negative effects of VUDM that led to these results were inferred, as shown in Table 12 and Table 13.

Table 12. Results Interpretation

	Task 1	Task 2	Task 3	Task 4	Task 5
Accuracy	-depend on cognition power -contingency	+visualized group relations -depend on cognition power -some communities were not shown	-no specific statistical info. provided	+visualized the amount of information needs on certain topics +multiple windows display time series of trends +visualized community clusters	+visualized multiple interest of a user +user detailed info. table +time series of individual interests
Efficiency	+community size visualized +community labels +frequency of accessed items +sorted communities	+group labels +visualized community clusters -lack of searching	+visualized multiple interests of a user	-lack of searching -depend on cognition power -contingency	-lack of searching -incomplete interaction

Table 12 shows the positive and negative effects of VUDM on each of five types of user tasks we tested. The items with a ‘+’ bullet in this table are positive effects of VUDM, while the items with a ‘-’ bullet are negative effects of VUDM. Table 13 aggregates the positive features and negative features in Table 12 into a pros and cons table of VUDM.

Table 13. Pros and Cons of VUDM

	Accuracy	Efficiency
Pros of VUDM	- Visualized community relations - Visualized the amount of information needs in certain topic - Multiple windows display time	- Communities are labeled based on the topics shared among users in them - Visualized community size

	series of trends - Visualized community clusters - Visualized multiple interests of a user - Detailed user information tables - Visualized time series of individual interests	- Detailed community information table -Sort function on communities - Visualized community clusters - Visualized multiple interests of a user
Cons of VUDM	- Depend on cognition power - contingency - Some communities were not shown because of filtering by both the user and visualization strategy of VUDM - No specific statistical information provided	- Lack of searching - depend on cognition power - contingency - incomplete interaction

Because some level of skill in using any statistical analysis software packages is needed to participate in this experiment, the effect of academic level needs to be examined; it is possible that the participants with higher academic level may be more skilled with the statistical software as they had more experience with them. Therefore, one more test is conducted to check if the academic levels of participants have any effect on the accuracy or efficiency in performing the five user tasks. Unfortunately, there were few graduate students, so this test may not show much.

Two charts in Figure 44 show the comparison of average accuracies in terms of software used, and academic levels, respectively.

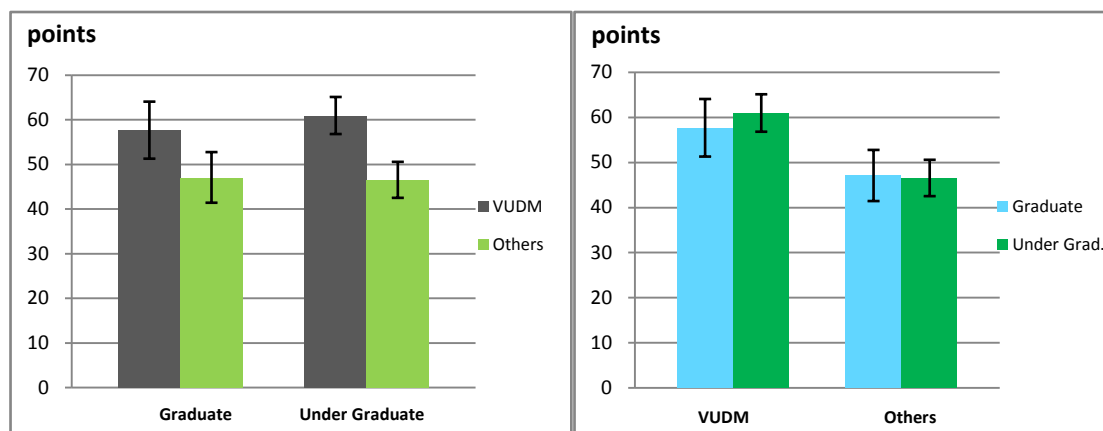


Figure 44. Average scores and confidence intervals of two different academic levels in terms of software used (left), and in terms of academic levels of participants (right)

Two charts in Figure 45 show the comparison of completion times in terms of software used, and academic levels, respectively.

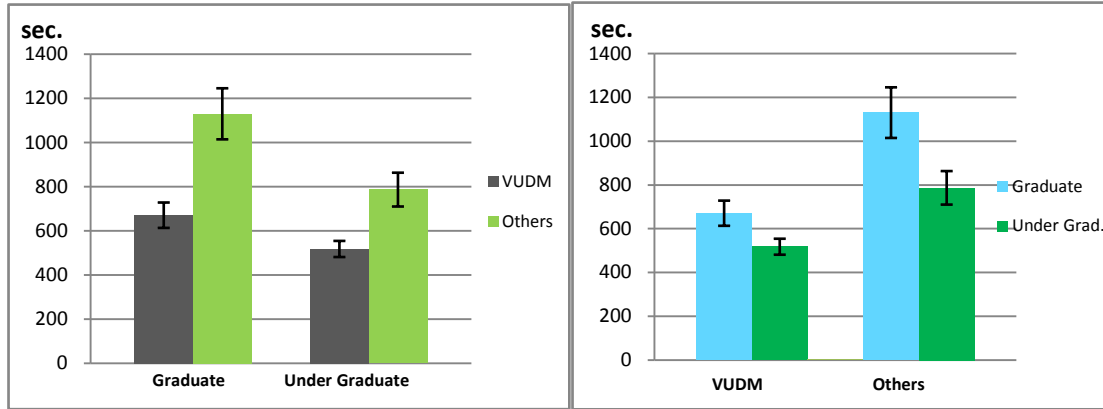


Figure 45. Average completion times and confidence intervals of two different software within academic levels (left), and academic levels within two different software (right)

Table 14 shows the result of a one-tail t-test about the effect of academic level on accuracy and efficiency. The result says the efficiencies in using both VUDM and other software packages are affected by the academic level of the participant, but accuracies are not affected. For more reliability regarding the answer, more experimentation is recommended.

Table 14. The effect of academic level

(t-test p-value)	Accuracy	Efficiency
VUDM	0.3349	0.0162
Others	0.4689	0.0095

8.4 Conclusions

In this experiment, the usefulness of VUDM is evaluated by normal, non-expert, library end users. Usefulness is evaluated by measuring accuracy and efficiency in conducting 5 types of user tasks, in which VUDM performs better than other statistical analysis software packages. Accuracy is measured by checking the percentage of right and wrong answers for the tasks, and efficiency is measured by time to complete each user task. We found that users perform more accurately with VUDM than with other

software packages for some types of user tasks, such as figuring out the trends for the near future and tracking the drift of concept of certain users. Also, VUDM shows better efficiency for some types of user tasks, such as discovering information seeking trends and detecting characteristics of certain users. That is, VUDM is either accurate or efficient for user tasks we are testing, except that users with VUDM don't show statistically significant better performance in perceiving the distribution of interest sharing communities. For more confidence regarding the generalizability of this evaluation, further experiments with more participants and various other user tasks are needed.

Chapter 9. Experts Review

This chapter presents the procedure and results of VUDM reviews performed by library experts. The subject set is selected faculty and staff of libraries. Participants were provided an executable VUDM, an online tutorial and questionnaire to complete. They were asked to evaluate the VUDM and to compare it with their preferred library analysis tools. In Sections 9.1, the detailed study design is described. In Section 9.2 general data is analyzed. In Section 9.3, the summary of rating data and comments for enhancement are presented. Section 9.4 concludes the discussion of the results of experts review.

9.1 Study Design

In this study, library experts, drawn from the faculty and staff of libraries, were targeted. The purpose of this study is 1) to learn the current status of software being used to analyze library data, 2) to figure out what features they are looking for in their analysis software, 3) to collect what features of VUDM satisfy them and how, and 4) to discover how VUDM should be enhanced to be practically deployed. This study is performed in the form of an online survey using a web-based survey system [83] at Virginia Tech. The IRB approval letter and questionnaires of this experiment are attached in Appendix B. The call-for-participation emails were sent to selected library experts to recruit participants.

The questionnaire consists of four sections. The first section of the questionnaire contains general questions asking about the participant's library. The second section of the questionnaire is for those who are not using library analysis tools, and contains questions about their plan about using library analysis tools and important features they are considering for their choice. The third section of the questionnaire is for those who are using library analysis tools. This section asks how they obtain the information about users, user communities, and usage trends of their library from their library analysis tools. Also, it asks how and why they are satisfied with their library analysis software regarding getting the information. The fourth section of the questionnaire is to ask the library experts about VUDM. In this section, an online tutorial of VUDM is provided. Participants can download and execute the VUDM and experiment with it for a while

until they feel used to it. The VTLS user model data, which is described in Table 4 in Section 7.4, is provided for their experiment. Then, participants are asked to rate and describe how and why they like or don't like VUDM, and what they want from VUDM so they can use it to analyze their own libraries.

9.2 General Information Obtained

The call-for-participation emails for this user study were sent to the targeted audience. Thus, the number of participants is very small. A total of 13 library experts completed the questionnaire. Average length of their careers in library related work, as a faculty or a library staff, is 9 years and 8 months.

All of the participants, except one who didn't answer, said their libraries are computerized and provide web-based search service for their content.

To the question that asks about the analysis tools or software being used to analyze their library data, four participants answered that they use web-based tools, such as Apache web-server and emails, one participants named the "Millenium", which is a Web-Java-based automated library system developed by Innovative Interfaces [36]. Also, one participant answered that she uses various home-grown tools. The remaining participants answered that they don't use any specific software.

All participants, except one, agreed that the information about users, user communities, and usage trends is important for their libraries. Table 15 summarizes some possible uses of the information, provided by participants. One participant, who disagreed, commented that the information is not useful for her/his library because it is not practical and difficult to analyze.

Table 15. Importance and uses of information about users, user communities, and usage trends

	information about users	information about user communities	information about usage trends
will be useful for	- Improving services and resources - Justifying cost (including staff time)	- Improving services and resources - Justifying cost (including staff time)	- Improving services and resources - Justifying cost (including staff time)

- planning services - determining where to spend time to build collection - determining what needs to be digitized and offered - selecting content - personalizing the service - planning future services - understanding users - determining best use of resources - finding problems	- identifying short-term trends or usage spikes - planning hours, services, collections - understanding the connection between various topics and groups - providing community services - understanding patrons - customizing community services - customizing web content to specific patron groups - targeting acquisitions	- pleading our case for more funds - planning staffing, facilities, and collections - projecting what is of growing and declining importance - purchasing media and books - planning future - finding news - analyzing success - future projects, maintenance, and purchases - financial planning, reporting, and equipment planning - personalization
--	--	---

The most important features that the participants expect for their next library analysis software, is web-based operability. Other features include time savings and a constant, up-to-date report ability to see short-term trends in their library data.

For the questions asking the current source of information about users, user communities and usage trends, most participants answered they are not getting the information. A few participants answered that they are getting the information from a web based survey system, a web reporting system, and from email feedbacks. These few also said they are basically satisfied with the quality of the information from the sources.

9.3 Ratings and Comments for Enhancement

Besides providing general information about their libraries and tools for library data analysis, participants rated the VUDM in terms of providing the information of users, user communities, and usage trends. The result is presented in Table 16 on a scale from 0 to 100. The value 0 represents worst while 100 means best. This result shows that participants mostly rated highly the VUDM for the features of providing information

about user communities, but gave low ratings for the features of providing information about usage trends.

Table 16. Average rating of VUDM in providing three types of information

	information about users	information about user communities	information about usage trends
average rating (from 0 to 100)	70.9	87.3	65.5

For more details, they also rated VUDM regarding providing information to perform the five comprehensive user tasks, listed in Section 6.1, which are used for the normal user summative evaluation in the previous chapter, presented in Table 17 on the same scale used in Table 16.

Table 17. Average rating of VUDM in providing information for five user task types

	information seeking trends	distribution of interest groups	user individual characteristic	trend in near future	drift of concepts
average rating (from 0 to 100)	69.1	76.4	54.5	58.2	61.8

This table shows that experts gave low ratings for VUDM's ability to provide information about user individual characteristics. From their comments, presented later in this section, they expect the VUDM provides more specific information, not just insight through visualization, for each user individual as compared to normal statistical software packages. Also, this table shows VUDM needs enhancement in presenting trends for the near future and drift of concepts of individuals.

Table 18 summarizes experts' comments for enhancing VUDM in terms of providing information about users, user communities, and usage trends.

All participants, except one, answered that they would be willing to use VUDM if it is enhanced as they commented. The participant who didn't want to use VUDM, even after it is modified, commented that visualizations make data more confusing.

Table 18. Comments for enhancement of VUDM

	information about users	information about user communities	information about usage trends	overall
will be improved by	- providing more specific information - providing better visualization - aggregating statistics of individuals	- adding a searching function	- pulling out each community and plot their population over time - adding single page time line graph	- converting into web-based software - adding more analysis features - making it tweakable for each institution using it - adding more features - providing figure-oriented visualization, instead of interactivity

9.4 Conclusions

A study of library experts, such as faculty and staff, was performed to collect ratings and opinions for enhancement of VUDM. To answer the questionnaire, participants downloaded and executed the VUDM and compared it with their preferred analysis tools.

The current VUDM is a prototype visualization tool for library information. It lacks various analysis features that statistical software packages normally provide. Thus, many participants asked us to add more analysis features to VUDM.

It appears from their comments that VUDM still has room to be enhanced with regard to its visualization capabilities. For example, the way that VUDM provides the trends of library usage depends too much on users' cognitive power. Participants asked that VUDM should find and visualize important knowledge automatically. Another visualization problem is with regard to presenting specific information. Most respondents preferred traditional ways, such as using numbers, text and tables, to present specific information, over visualizations used in VUDM.

Further, VUDM needs to be converted into web-based software to be more useful and popular. Finally, it is necessary to make VUDM customizable according to each institutes' needs.

Chapter 10. Conclusions and Future Work

This chapter summarizes the achievement of this Ph.D. research and suggests possible research directions for further study. It begins with a discussion of deliverables.

10.1 Deliverables

Below are the deliverables that I have produced for my doctoral work and intend to make available to the Digital Library Research Laboratory and research community.

1. Software to track users' behavior and collect implicit rating data at the NDLT web interface. It is available at http://rocky.dlib.vt.edu/~shkim/VUDM_tutorial/tools/tracking
2. Anonymized VTLS log data and NDLT user tracking data in both raw format and user model data format
3. Software to visualize the users, user communities, and usage trends of digital libraries. The code can be found at http://rocky.dlib.vt.edu/~shkim/VUDM_tutorial/src/
4. A collection of research interests and learning topics in XML format, which is used as a test basis to evaluate the accuracy of user communities. It is available at http://rocky.dlib.vt.edu/~shkim/VUDM_tutorial/tools/evaluation/collection/
5. Software to evaluate the accuracy of user communities. It is available at http://rocky.dlib.vt.edu/~shkim/VUDM_tutorial/tools/evaluation/src/

10.2 Conclusions

The focus of this research is on implicit rating data and its utilization in a digital library system by using interaction through an information visualization system. Implicit rating data is easier to collect than explicit rating data. Accordingly, it would be helpful to know if it could be useful.

This work shows the utility of implicit rating data through a type of existence proof, i.e., demonstrating that a system can be developed that uses that data to good effect. The demonstration system developed is VUDM: Visual User-model Data Mining tool.

One part of the demonstration was to develop that system. Chapter 5 describes the development of the system and it is available as one of the deliverables of this work.

Another part of the demonstration was to show that the system was scalable. Algorithms were developed and runs with varying amounts of data were made to show that not only can effective results be achieved with a computationally expensive algorithm, but also variants of that algorithm were developed that had much lower computational complexity and yielded satisfactory levels of effectiveness.

To show that implicit rating data could be helpful, the demonstration system was tested in a variety of studies. First, it was iteratively refined through a formative evaluation process. Second, it was tested in a controlled study with a group of computer scientists familiar with digital libraries. Third, it was tested in a controlled study with a small group of digital librarians.

In the first study, the system was repeatedly improved. Some of the enhancements include: Enhancing visualization method for displaying representative topics of a user community, for displaying time series of user data, and animated visualization for locating a user community and a user.

In the second study, there were measurements of accuracy and time, and qualitative assessment questions were answered. A set of 5 tasks that seemed to be representative were assigned to the 60 users. Users worked with both VUDM and their preferred statistical software system, typically Excel. On average, VUDM performance was better than the statistical system, with regard to both efficiency and accuracy. However, in some tasks VUDM did less well than the statistical system with regard to accuracy; similarly, it did less well with regard to efficiency for a set of the tasks. Qualitative assessments showed that VUDM had a number of positive features, but some negative features also were uncovered.

In the third study, there was a survey, targeting library experts, which subjects are selected from faculty and staff of libraries. The main purposes of the survey are to ask if they think VUDM could be useful for their library, and to collect enhancement ideas to make VUDM useful in a real digital library. Most librarians answered VUDM would be useful, or could be useful after some modifications, and commented that they are willing to try to use it once enhanced.

VUDM is a demonstration or prototype system, that clearly can be improved. Nevertheless, even in its current form it appears to be useful enough that it might be employed to help in some situations by those working with digital libraries.

Since VUDM uses implicit rating data solely, we can conclude that there are situations in which the use of implicit rating data can be beneficial to some digital library users.

10.3 Future Work

There is a great deal of future work possible to extend this study, such as: improving VUDM based on comments received, testing the improved VUDM in further experiments with the content collections previously studied, building upon new data collection regarding use, and testing the improved VUDM with other content collections and other user communities.

Besides improving VUDM, a lot of research directions are possible regarding developing the applications of VUDM and its basis techniques. This may include applications such as: personalization, decision making support, personal library visualization, and social network support.

Research for personalization of digital libraries, including recommender systems, shall receive benefits from this dissertation. Methods for more specific and timely recommendation utilizing the personal research interests and learning topics and drifts of concepts stored in the individual user model data could be studied. Collaborative filtering will be possible from the information on user communities found by VUDM.

Regarding supporting decision making for digital library operations, more research is needed. The knowledge finding in VUDM depends too much on the personal opinion of the VUDM user, and on contingency. To obtain consistent and reproducible knowledge from VUDM and user model data, methods for objective and descriptive knowledge finding need to be studied.

Also, the visualization techniques used in VUDM would be applicable for a personal digital library or personal information manager. For these purposes, the techniques to visualize the time series of usage trends and drift of concepts need to be refined. Also, it would be meaningful to integrate the features of social network support of VUDM, with other online-offline social network support systems.

References

1. ACM Transactions, ACM DL: Transactions, available at http://portal.acm.org/browse_dl.cfm?linked=1&part=transaction&coll=ACM&dl=ACM&CFID=19463219&CFTOKEN=17951970, 2007.
2. Carrot2 Project, A research framework for experimenting with automated querying of various data sources, processing search results and visualization, available at <http://project.carrot2.org/>, 2006.
3. CITIDEL, Computing and Information Technology Interactive Digital Education Library, available at <http://www.citidel.org>, 2006.
4. ETANA-DL, Managing Complex Information Application: An Archaeology Digital Library, available at <http://digbase.etana.org:8080/etana/servlet/Start>, 2006.
5. Friendster, Online Social Network Portal, available at <http://www.friendster.com/>, 2007.
6. ND LTD, Networked Digital Library of Theses and Dissertations, available at <http://www.ndltd.org>, 2008.
7. Charu C. Aggarwal, Fatima Al-Garawi and Philip S. Yu, On the design of a learning crawler for topical resource discovery. In *ACM Transactions on Information Systems (TOIS)*, 19 (3), <http://portal.acm.org/citation.cfm?id=502115.502119&coll=ACM&dl=ACM&CFID=57273276&CFTOKEN=28115067>, 2001, 286-309.
8. Rajendra Akerkar and Pawan Lingras, *Building an Intelligent Web: Theory and Practice*. Jones and Bartlett Publishers, Sudbury, 2008.
9. Steven Alter, *Information Systems: Foundation of E-Business* Prentice Hall, Upper Saddle River, NJ, 2002.
10. Peter G. Anick and Shivakumar Vaithyanathan, Exploiting Clustering and Phrases for Context-based Information Retrieval. In *Proceedings of the 20th annual international ACM SIGIR conference on research and development*, Philadelphia, PA, 1997, 314-323.
11. Japan Library Association, Nippon Decimal Classification (NDC), available at <http://www.jla.or.jp/index-e.html>, 1956.
12. Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley, Essex, England, 1999.
13. Nicholas J. Belkin, Interaction with Texts: Information Retrieval as Information-Seeking Behavior. In *Information Retrieval*, 1993, 55-66.
14. Johan Bollen and Rick Luce, Evaluation of Digital Library Impact and User Communities by Analysis of Usage Patterns. In *D-Lib Magazine*, 8 (6), <http://www.dlib.org/dlib/june02/bollen/06bollen.html>, 2002
15. Doug A. Bowman, Jian Chen, Chadwick A. Wingrave, John F. Lucas, Andrew Ray, Nicholas F. Polys, Qing Li, Yonca Haciahmetoglu, Ji-Sun Kim, Seonho Kim, Robert Boehringer and Tao Ni, New Directions in 3D User Interfaces. In *International Journal of Virtual Reality (IJVR)*, 5 (2), 2006, 3-14.

16. Doug A. Bowman, Ernst Kruijff, Jr. Joseph J. Laviola and Ivan Poupyrev, *3D User Interfaces: Theory and Practice*. Addison-Wesley Professional, Boston, MA, 2004.
17. Danah Boyd and Jeffrey Potter, Social Network Fragments: An Interactive Tool for Exploring Digital Social Connections. In *Proceedings of the International Conference of Computer Graphics and Interactive Techniques (SIGGRAPH 2003)*, San Diego, 2003, 1.
18. Carol June Bradley and George Sherman Dickinson, *The Dickinson classification: A cataloguing & classification manual for music*. Carlisle Books, Carlisle, PA, 1968.
19. John S. Breese, David Heckerman and Carl Kadie, Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, 1997, 43-52.
20. Stuart K. Card, Jock D. Mackinlay and Ben Shneiderman, *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers, San Francisco, 1999.
21. Lois Mai Chan, *A Guide to the Library of Congress Classification*. Library Unlimited, 1999.
22. George Chang, Marcus J. Healey, James A. M. McHugh and Jason T. L. Wang, *Mining the World Wide Web: An Information Search Approach*. Kluwer Academic Publishers, Boston / Dordrecht / London, 2001.
23. Chaomei Chen and Yue Yu, Empirical Studies of Information Visualization: A Meta-Analysis. In *Int. J. Human-Computer Studies*, 53, 2000, 851-866.
24. Hsinchun Chen, *Knowledge Management Systems: A Text Mining Perspective*. The University of Arizona, Tucson, 2001.
25. Belur V. Dasarathy, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, CA, 1991.
26. Melvil Dewey, *Dewey Decimal Classification and Relative Index*. Forest Press, Albany, New York, 1996.
27. Richard O. Duda, Peter E. Hart and David G. Stork, *Pattern Classification*. A Wiley-Interscience Publication, New York, 2000.
28. Weiguo Fan, *Effective Personalized Delivery of Information: A Two-Stage Model and Empirical Analysis*. University of Michigan, Ann Arbor, MI, 2002.
29. Raya Fidel and Annelise Mark Pejtersen, From Information Behaviour Research to the Design of Information System: the Cognitive Work Analysis Framework. In *Information Research*, 10 (1), 2004, paper 210.
30. Edward Fox, The 5S Framework for Digital Libraries and Two Case Studies: NDLTD and CSTC. In *Proceedings of the the 11th International Conference on New Information Technology (NIT'99)*, Taipei, Taiwan, 1999.
31. Marcos André Gonçalves, Edward A. Fox, Layne T. Watson and Neill A. Kipp, Streams, structures, spaces, scenarios, societies (5S): A formal model for digital libraries. In *ACM Transactions on Information Systems (TOIS)*, 22 (2), 2004, 270 - 312.
32. Marcos André Gonçalves, Ming Luo, Rao Shen, Mir Farooq and Edward A. Fox, An XML Log Standard and Tools for Digital Library Logging Analysis. In

- Proceedings of the Sixth European Conference on Research and Advanced Technology for Digital Libraries*, Rome, Italy, 2002, 16-18.
33. Jeffrey Heer and Danah Boyd, Vizster: Visualizing Online Social Networks. In *Proceedings of the 2005 IEEE Symposium on Information Visualization (INFOVIS '05)*, Washington, DC, 2005, 5.
 34. Ivan Herman, Guy Melançon and M. Scott Marshall, Graph Visualization and Navigation in Information Visualization: A Survey. In *IEEE Transactions on Visualization and Computer Graphics*, 6 (1), 2000, 24-43.
 35. Deborah Hix and H. Rex Hartson, *Developing User Interfaces: Ensuring Usability Through Product & Process*. Wiley Professional Computing, New York, NY, 1993.
 36. Innovative Interfaces, Trusted Library Technology and Service, available at <http://www.iii.com>, 2008.
 37. Steve Jones, Sally Jo Cunningham and Rodger McNab, An Analysis of Usage of a Digital Library. In *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries (ECDL '98)*, Lecture Notes in Computer Science 1513, Springer-Verlag, London UK, , 1998, 261-277.
 38. Fink Josef and Kobsa Alfred, User Modeling for Personalized City Tours. In *Artif. Intell. Rev.*, 18 (1), 2002, 33-74.
 39. Anne Kao and Stephen R. Poteet, *Natural Language Processing and Text Mining*. Springer, London, 2007.
 40. Daniel A. Keim, Information Visualization and Visual Data Mining. In *IEEE Transactions on Visualization and Computer Graphics*, 7 (1), 2002, 100-107.
 41. Diane Kelly, Understanding Implicit Feedback and Document Preference: A Naturalistic User Study, Doctoral dissertation, Rutgers University, 2004.
 42. Seonho Kim, NDLTD, Search Interface Embedded User Tracking System, available at <http://www.ndltd.org/>, 2006.
 43. Seonho Kim, VUDM, Visual User-model Data Mining tool, Online Tutorial, available at http://rocky.dlib.vt.edu/~shkim/VUDM_tutorial/, 2008.
 44. Seonho Kim and Edward A. Fox, Interest-based User Grouping Model for Collaborative Filtering in Digital Libraries. In *Proceedings of the 7th International Conference of Asian Digital Libraries (ICADL' 04)*, Shanghai, China, Lecture Notes in Computer Science 3334, Springer-Verlag, Berlin Heidelberg New York, 2004, 533-542.
 45. Seonho Kim, Edward A. Fox, Weiguo Fan, Chris North, Deborah Tatar and Ricardo da Silva Torres, Design and Evaluation of Techniques to Utilize Implicit Rating Data in Complex Information Systems, *Computer Science Technical Report*, Blacksburg, VA, Virginia Tech, <http://eprints.cs.vt.edu/archive/00000980/>, TR-07-20, 2007.
 46. Seonho Kim, Subodh Lele, Sreeram Ramalingam and Edward A. Fox, Visualizing User Communities and Usage Trends of Digital Libraries based on User Tracking Information. In *Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL '06)*, Kyoto, Japan, Lecture Notes in Computer Science 4312, Springer-Verlag, Berlin Heidelberg New York, 2006, 111-120.
 47. Seonho Kim, Uma Murthy, Kapil Ahuja, Sandi Vasile and Edward A. Fox, Effectiveness of Implicit Rating Data on Characterizing Users in Complex

- Information Systems. In *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL '05)*, Vienna, Austria, Lecture Notes in Computer Science 3652, Springer-Verlag, Berlin Heidelberg New York, 2005, 186-194.
48. Seonho Kim, Seungwon Yang and Edward A. Fox, Supply and Demand Analysis in NDLTD Based on Patron Specialty Survey and Contents Statistics. In *Proceedings of the 9th International Symposium on Electronic Theses and Dissertations (ETD '06)*, 2006
 49. Claus-Peter Klas, Hanne Albrechtsen, Norbert Fuhr, Preben Hansen, Sarantos Kapidakis, Laszlo Kovacs, Sascha Kriewel, Andras Micsik, Christos Papatheodorou, Giannis Tsakonas and Elin Jacob, A Logging Scheme for Comparative Digital Library Evaluation. In *Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL '06)*, Lecture Notes in Computer Science 4172, Springer-Verlag, Berlin Heidelberg New York, 2006, 267-278.
 50. Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon and John Riedl, GroupLens: Applying Collaborative Filtering to Usenet News. In *Communications of the ACM*, 40 (3), 1997 77-87.
 51. Ravi Kumar, Jasmine Novak, Prabhakar Raghavan and Andrew Tomkins, Structure and Evolution of Blogspace. In *Communications of the ACM*, 47 (12), 2004, 35-39.
 52. Shin-jeng Lin and Nicholas J. Belkin, Validation of a model of information seeking over multiple search sessions. In *Journal of American Society for Information Science & Tech.(JASIST)*, 56 (4), 2005, 393-415.
 53. Thomas W. Malone, Kenneth R. Grant, Franklyn A. Turbak, Stephen A. Brobst and Michael D. Cohen, Intelligent information sharing systems. In *Communications of the ACM*, 30 (5), 1987, 390-402.
 54. Eren Manavoglu, Dmitry Pavlov and C. Lee Giles, Probabilistic User Behavior Models. In *Proceedings of the the Third IEEE International Conference on Data Mining (ICDM'03)*, Melbourne, FL, 2003, 203-210.
 55. Robert C. Miller and Krishna Bharat, SPHINX: A Framework for Creating Personal, Site-Specific Web Crawlers. In *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, Software available at <http://www.cs.cmu.edu/~rcm/websphinx/>, 1998.
 56. MPAA, Motion Picture Association of America, Information on rating system for both movies and television, press releases, and legislation, available at <http://www.mpaa.org/>, 2008.
 57. NDLTD, NDLTD Membership, available at <http://www.ndltd.org/members/>, 2008.
 58. David M. Nichols, Implicit Rating and Filtering. In *Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering*, Budapest, Hungary, 1997, 31-36.
 59. David M. Nichols, Duncan Pemberton, Salah Dalhoumi, Omar Larouk, Clair Belisle and Michael B. Twidale, DEBORA: Developing an Interface to Support Collaboration in a Digital Library. In *Proceedings of the the Fourth European*

- Conference on Research and Advanced Technology for Digital Libraries (ECDL '00)*, Lisbon, Portugal, 2000, 239-248.
60. Alexandros Ntoulas, Junghoo Cho and Christopher Olston, What's new on the web?: the evolution of the web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web*, New York, USA, ACM, 2004, 1-12.
 61. Tim O'Reilly, What Is Web 2.0, available at <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>, 2006.
 62. Georgios Paliouras, Vangelis Karkaletsis, Christos Papatheodorou and Constantine D. Spyropoulos, Exploiting Learning Techniques for the Acquisition of User Stereotypes and Communities. In *Proceedings of the International Conference on User Modelling (UM '99)*, 1999, 169-178.
 63. Bong-seok Park, *Korean Decimal Classification (KDCP)*, Seoul: National Library of Korea, 1976.
 64. Dmitry Pavlov, Eren Manavoglu, David M. Pennock and C. Lee Giles, Collaborative Filtering with Maximum Entropy. In *IEEE Intelligent Systems*, 19 (6), 2004, 40-48.
 65. Michael Pazzani and Daniel Billsus, Learning and Revising User Profiles: The Identification of Interesting Web Sites. *Machine Learning*, Kluwer Academic Publishers, 1997, 313-331.
 66. Catherine Plaisant, The Challenge of Information Visualization Evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, 2004 109-116.
 67. M. Elena Renda and Umberto Straccia, A Personalized Collaborative Digital Library Environment. In *Proceedings of the 5th International Conference on Asian Digital Libraries (ICADL'02)*, 2002, 262-274.
 68. Fred W. Riggs, Concepts, Words and Terminology. In *Proceedings of the Committee on Conceptual and Terminological Analysis, Working Paper No. 13*, Honolulu, University of Hawaii, Social Sciences Research Institute, 1971, 66.
 69. Purvi Saraiya, Chris North and Karen Duca, An Insight-Based Methodology for Evaluating Bioinformatics Visualizations. In *IEEE Transactions on Visualization and Computer Graphics*, 11 (4), 2005.
 70. Giovanni Sartori, Concept misinformation in comparative politics. In *American Political Science Review*, 64 (4), 1970, 1033-1053.
 71. Ben Shneiderman, The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations In *Proceedings of the IEEE Symposium on Visual Languages*, IEEE Computer Society Press, 1996, 336-343.
 72. The Internet Society, Hypertext Transfer Protocol - HTTP/1.1, available at <http://www.w3.org/Protocols/rfc2616/rfc2616.html>, 1999.
 73. SONA, Virginia Tech, SONA Experiment Management System, available at <http://www.psyc.vt.edu/research/sona.html>, 2008.
 74. Tom Soukup and Ian Davidson, *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*. Wiley Computer Publishing, New York, NY, 2002.
 75. Robert Spence, *Information Visualization*. Addison-Wesley, New York, 2001.

76. Jaideep Srivastava, Robert Cooley, Mukund Deshpande and Pang-Ning Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. In *SIGKDD Explorations*, 1 (2), <http://citeseer.ist.psu.edu/srivastava00web.html>, 2000, 12-23.
77. Hussein Suleman, Introduction to the Open Archives Initiative protocol for metadata harvesting. In *Proceedings of the ACM/IEEE 2nd Joint Conference on Digital Libraries (JCDL 2002)*, Portland, OR, Software available at <http://oai.dlib.vt.edu/odl/software/harvest/>, 2002, 414.
78. Hussein Suleman, Using the Repository Explorer to Achieve OAI Protocol Compliance. In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital Libraries*, Roanoke, VA, ACM, 2001, 459.
79. Cass R. Sunstein, Democracy and Filtering. In *Communications of the ACM*, 47 (12), 2004, 57-59.
80. Tiffany Ya Tang and Gordon McCalla, Mining Implicit Ratings for Focused Collaborative Filtering for Paper Recommendations. In *Proceedings of the Workshop on User and Group models for Web-based Adaptive Collaborative Environments (UM'03)*, Online proceeding Available at <http://www.ia.uned.es/~elena/um03-ws/>, 2006.
81. T.T. Tanimoto, IBM Internal Report Series, *IBM Technical Report Series*, November, 1957.
82. Virginia Tech, The 4/16 Research Library, A digital library for research, resources, analysis and community relating to 4/16/2007, and after, at Virginia Tech, available at <http://www.dl-vt-416.org/>, 2008.
83. Virginia Tech, A Web-based Survey Tool at Virginia Tech, available at <http://survey.vt.edu/>, 2008.
84. Kristin M. Tolle and Hsinchun Chen, Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools. In *Journal of American Society for Information Science*, 51 (4), 2000, 352-370.
85. Lyle H. Ungar and Dean P. Foster, A Formal Statistical Approach to Collaborative Filtering. In *Proceedings of the the Conference on Automated Learning and Discovery (CONALD '98)*, Carnegie Mellon U., Pittsburgh, PA, 1998.
86. Zarikas Vasilios, Modeling decisions under uncertainty in adaptive user interfaces. In *Univers. Access Inf. Soc.*, 6 (1), 2007, 87-101.
87. VTLS, Visionary Technology in Library Solutions, available at <http://www.vtls.com>, 2008.
88. Lingyu Wang, Sushil Jajodia and Duminda Wijesekera, *Preserving Privacy in On-Line Analytical Processing (OLAP)*. Springer, Berlin, 2007.
89. Geoffrey I. Webb, Michael J. Pazzani and Daniel Billsus, Machine Learning for User Modeling. *User Modeling and User-Adapted Interaction*, Kluwer Academic Publisher, 2001, 19-29.
90. Sholom M. Weiss, Nitin Indurkha, Tong Zhang and Fred J. Damerau, *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer, New York, 2005.

91. Gerhard Widmer and Miroslav Kubat, Learning in the Presence of Concept Drift and Hidden Contexts. *Machine Learning*, Kluwer Academic Publishers, 1996, 69-101.
92. Wikipedia, the free encyclopedia, available at <http://www.wikipedia.org/>,
93. Ken Winter, 'MyLibrary' Can Help Your Library (Personalize Web Pages). In *American Libraries*, 30 (7), 1999, 65-67.
94. James A. Wise, James J. Thomas, Kelly Pen-nock, David Lantrip, Marc Pottier and Anne Schur, Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of the First Information Visualization Symposium (InfoVis '95)*, Atlanta, GA, IEEE Computer Society Press, 1995, 51-58.
95. Jennifer Xu and Hsinchun Chen, Criminal Network Analysis and Visualization. In *Communications of the ACM*, 48 (6), 2005, 101-107.
96. Kai Yu, Anton Schwaighofer, Volker Tresp, Xiaowei Xu and Hans-Peter Kriegel, Probabilistic Memory-based Collaborative Filtering. In *IEEE Transactions on Knowledge and Data Engineering*, 16 (1), 2004, 56-69.
97. Chun Zeng, Xiaohui Zheng, Chunxiao Xing and Lizhu Zhou, Personalized Services for Digital Library. In *Proceedings of the 5th International Conference on Asian Digital Libraries (ICADL'02)*, 2002, 252-253.
98. Wenxian Zhang, Classification for Chinese Libraries (CCL): Histories, accomplishments, problems and its comparisons. In *Journal of Educational Media & Library Sciences*, 41 (1), 2003, 1-22.

Appendix A-1: Call for Participation Advertisement

Subject: Participants Needed for Digital Library Research Lab.

Experiment

Hello,

We are seeking participants for a usability test of our new digital library analysis tool. We will compare it with other popular statistical software packages such as Microsoft Excel, Minitab, SPSS and MATLAB.

To participate in the test, some prior experience of using any “one” of these software packages and familiarity with its basic functions are needed. For example, loading a data file, calculating sums and drawing a bar chart. This experiment will take about 80 minutes including 15 minutes during the learning session.

Location: Digital Library Research Lab, 2030 Torgersen Hall, Blacksburg, VA

Date: From October 15th to October 31st

Time: See the schedule table at <http://people.cs.vt.edu/~haebang/schedule.html>

Please select your convenient date and time and contact Seonho Kim (email: shk@vt.edu, phone: 540-921-7530).

If you have any questions regarding my research or need further information, please do not hesitate to contact me.

Thank you very much in advance for the attention paid to my request. I am looking forward to hearing from you.

Kind Regards,

Seonho Kim

shk@vt.edu

540-921-7530

Appendix A-2: Questionnaire

Pre-test General Questions

Participant #: ()

Please help us to categorize participants by completing the following items.

1. Age: Select one

- ☐ 18-25 ☐ 26-30 ☐ 31-35 ☐ 36-40
☐ 41-45 ☐ 46-50 ☐ 51-55 ☐ 56-60
☐ 61-65 ☐ 66-70 ☐ 71-

2. Occupation (if you are a student, indicate “graduate” or “undergraduate”):

3. Major / Area of specialization (only for students):

4. Rate your familiarity with computers on a scale from 0 to 5 (0: not familiar to 5: very familiar)

0 1 2 3 4 5

5. Select all the statistical software packages that you have used and rate your familiarity with them in 0-5 scale. (0 is lowest and 5 is highest familiarity)

- | | | | | | | |
|----------------------------------|---|---|---|---|---|---|
| ☐ Excel | 0 | 1 | 2 | 3 | 4 | 5 |
| ☐ Minitab | 0 | 1 | 2 | 3 | 4 | 5 |

- | | | | | | | |
|--|---|---|---|---|---|---|
| ☐ SPSS | 0 | 1 | 2 | 3 | 4 | 5 |
| ☐ MATLAB | 0 | 1 | 2 | 3 | 4 | 5 |
| ☐ Other (name:) | 0 | 1 | 2 | 3 | 4 | 5 |

6. From your answer of question 5, which software package(s) do you use most frequently?

Software Name:

Frequency of use: (e.g. once a week)

User Task Set 1

Participant #: ()

Software: ()

Please read the following statements and infer whether they are true or false, using the software.

Task 1: See the window 200606.

“In June 2006, more users are interested in Music than Knowledge Management.”

- 1) True
- 2) False
- 3) I give up.

Task 2: See the window 200606.

“In June 2006, the “Power System” related area has more interest user groups than “Library” related area.

- 1) True
- 2) False
- 3) I give up.

Task 3: See the window 200606.

“In June 2006, users who are interested in “Knowledge Management” are also interested in Library Science”.

- 1) True
- 2) False
- 3) I give up.

Task 4: See the windows 200606, 200607, and 200608.

“From the trends from June 2006 to August 2006, there will be more information needs for “Power System Control” in the next month, September 2006”.

- 1) True
- 2) False
- 3) I give up.

Task 5: See the window 200601 and 200602.

“Someone, who has interest in “Management” in January 2006, developed a new interest in “Culture Value” in February 2006”

- 1) True
- 2) False
- 3) I give up.

Any comment and suggestion:

User Task Set 2

Participant #: ()

Software: ()

Please read the following statements and infer whether they are true or false, using the software.

Task 1: See the window 200606.

“In June 2006, more users are interested in Economics than Psychology.”

- 1) True
- 2) False
- 3) I give up.

Task 2: See the window 200606.

“In June 2006, the “Power System” related area has more interest user groups than “Library” related area.

- 1) True
- 2) False
- 3) I give up.

Task 3: See the window 200606.

“In June 2006, users who are interested in “Economics” are also interested in “Database”.

- 1) True
- 2) False
- 3) I give up.

Task 4: See the windows 200606, 200607, and 200608.

“From the trends from June 2006 to August 2006, there will be more information needs for “Psychology” in the next month, September 2006”.

- 1) True
- 2) False
- 3) I give up.

Task 5: See the window 200601 and 200602.

“Someone, who has interest in “Case Study” in January 2006, developed a new interest in “Management” in February 2006”

- 1) True
- 2) False
- 3) I give up.

Any comment and suggestion:

Appendix A-3: Institutional Review Board Approval Letter

 VirginiaTech		Office of Research Compliance Institutional Review Board 2000 Kraft Drive, Suite 2000 (0497) Blacksburg, Virginia 24061 540/231-4991 Fax 540/231-0959 e-mail moored@vt.edu www.irb.vt.edu FWA000005721 expires 1/20/2010 IRB # is IRB000000667
DATE: October 9, 2007		
MEMORANDUM		
TO:	Edward A. Fox Seonho Kim	Approval date: 10/8/2007 Continuing Review Due Date: 9/23/2008 Expiration Date: 10/7/2008
FROM:	David M. Moore 	
SUBJECT: IRB Expedited Approval: "User Study for Measuring Accuracy and Efficiency of Visual Use Model Mining Tool (VUDM)", IRB # 07-490		
This memo is regarding the above-mentioned protocol. The proposed research is eligible for expedited review according to the specifications authorized by 45 CFR 46.110 and 21 CFR 56.110. As Chair of the Virginia Tech Institutional Review Board, I have granted approval to the study for a period of 12 months, effective October 8, 2007. As an investigator of human subjects, your responsibilities include the following: 1. Report promptly proposed changes in previously approved human subject research activities to the IRB, including changes to your study forms, procedures and investigators, regardless of how minor. The proposed changes must not be initiated without IRB review and approval, except where necessary to eliminate apparent immediate hazards to the subjects. 2. Report promptly to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others. 3. Report promptly to the IRB of the study's closing (i.e., data collecting and data analysis complete at Virginia Tech). If the study is to continue past the expiration date (listed above), investigators must submit a request for continuing review prior to the continuing review due date (listed above). It is the researcher's responsibility to obtain re-approval from the IRB before the study's expiration date. 4. If re-approval is not obtained (unless the study has been reported to the IRB as closed) prior to the expiration date, all activities involving human subjects and data analysis must cease immediately, except where necessary to eliminate apparent immediate hazards to the subjects. Important: If you are conducting federally funded non-exempt research, this approval letter must state that the IRB has compared the OSP grant application and IRB application and found the documents to be consistent. Otherwise, this approval letter is invalid for OSP to release funds. Visit our website at http://www.irb.vt.edu/pages/newstudy.htm#OSP for further information. cc: File		
<i>Invent the Future</i> VIRGINIA POLYTECHNIC INSTITUTE UNIVERSITY AND STATE UNIVERSITY <i>An equal opportunity, affirmative action institution</i>		

Figure 46. IRB approval letter for the summative evaluation

Appendix B-1: Call for Participation Advertisement

Subject: Participants Needed for Digital Library Research

Laboratory User Study

Hello,

We are seeking library experts for a usability test of our new digital library analysis tool, VUDM (Visual User-model Data Mining tool), which provides visualization aids about users, user communities and usage trends of digital library. You will be provided VUDM and an online tutorial to test it by yourself. This study is performed in online-survey manner, so you can participate any time and at any place according to your convenience.

To participate in this study, experience with working in a traditional or digital library is needed. This experiment will take about 60 minutes including 15 minutes during the learning session.

To participate in this study, visit the online survey page at

<https://survey.vt.edu/survey/entry.jsp?id=1202402156982>

and the online tutorial of VUDM at

http://rocky.dlib.vt.edu/~shkim/VUDM_tutorial/

Date: From February 20th to February 26st

If you have any questions regarding my research or need further information, please do not hesitate to contact me.

You can reach me at shk@vt.edu or 540-921-7530. Thank you very much in advance for the attention paid to my request.

Regards,

Seonho Kim

Digital Library Research Laboratory

Virginia Tech

Appendix B-2: Questionnaire

Thank you for assisting with this survey. The goal of this survey is to collect your opinion about your (preferred) library analysis tool(s) and our digital library analysis tool prototype, VUDM (Visual User-model Data Mining) tool.

Section 1 (General Questions)

Note: The term “user” represents “patron” of libraries throughout this survey.

1. How long have you been engaged in library related job?

2. Is your library computerized for patrons, books, and media?

3. Does your library provide online content through the internet?

4. What kind of analysis tools or software, are you or your team, using for your library analysis?

5. Do you think that information about users, user communities, and usage trends of your library will be useful in operating your library? (Yes/No)
 - a. If you answered “Yes”, please answer the following questions.
 - i. Information about users will be useful for

 - ii. Information about user community will be useful for

iii. Information about usage trends of your library will be useful for

b. If you answered “No”, please explain why you think this information will not be useful for your library.

6. Are you, or your team, using any software or tools to analyze users, user communities, or usage trends of your library? (Yes/No)

If you answered “Yes”, go to Section 3. Otherwise, go on Section 2.

Section 2 (for those not using library analysis tools)

1. Do you have any plans to use, or purchase, software for library analysis in the near future? (Yes/No)

If you answered “Yes”, please list them and explain.

2. What features do you consider most important (i.e., expect to be of greatest value) for the analysis software for your library?

Please skip Section 3 and jump to Section 4.

Section 3 (for those using library analysis tools)

1. Please list the software you, or your team, are using to analyze you library.

2. How do you get information about users, user communities, or usage trends of your library from the software listed above? (If you think your software doesn't provide this information, please go to Section 2)

a. User information

b. User communities

c. Usage trends

3. Are you satisfied with your software in getting the information you described in the previous question? Please answer how and why.

a. In getting information about users

b. In getting information about user communities

c. In getting information about usage trends

Please go to Section 4.

Section 4 (Questions about VUDM)

Before answering this section, please read the online VUDM tutorial at http://rocky.dlib.vt.edu/~shkim/VUDM_tutorial/ , download, install, and experiment with it for a while until you feel used to VUDM.

1. How would you rate VUDM for the following functions?
 - a. Providing information about users (0 is worst, 5 is best)

 - b. Providing information about user communities (0 is worst, 5 is best)

 - c. Providing information about usage trends (0 is worst, 5 is best)

2. To provide more details, please rate VUDM, for each of the following five functions?
 - a. Providing information about “trends of information search” (0 is worst, 5 is best)

 - b. Providing information about “distribution of user communities across various topics” (0 is worst, 5 is best)

 - c. Providing information about “user individual characteristics” (0 is worst, 5 is best)

 - d. Providing information about “trends for the near future” (0 is worst, 5 is best)

 - e. Providing information about “drift of search concepts of users, over time” (0 is worst, 5 is best)

3. Do you think VUDM will be useful for your library? Please explain why or why not.

a. In getting information about users (yes/no)

b. In getting information about user communities (yes/no)

c. In getting information about usage trends (yes/no)

4. Would you be willing to use VUDM for your library? (Yes/No)

If you answered “No”, please explain why and list how VUDM needs to be enhanced for your need.

5. Will you willing to use VUDM for your library if VUDM were enhanced as you described above?

6. Please leave any comments.

Appendix B-3: Online Tutorial of VUDM for Library Experts Survey

Contents

- [Introduction](#)
- [System Requirements & Installation](#)
- [Tutorial](#)
 - a. [Download and Installation](#)
 - b. [Start VUDM](#)
 - c. [Data Loading](#)
 - d. [Exploration](#)
- [Release Notes](#)
- [Bug Reports and Feedback](#)
- [Web pages](#)

1. Introduction

This webpage describes VUDM, the Visual User-model Data Mining tool, and provides a guide to its application, i.e., to user model data related to ND LTD, the Networked Digital Library of Theses and Dissertations (<http://www.ndltd.org/>). The goals of VUDM are to provide both an overview and details of users, user communities, and usage trends of digital libraries.

The distinctive approach of this research is that we focus on analysis and visualization of users' implicit rating data, which was generated based on user tracking information, such as sending queries and browsing result sets - rather than focusing on explicit data obtained from a user survey, such as: major, specialties, years of experience, and demographics. The VUDM interface uses spirals to portray virtual interest communities, positioned based on inter-community relationship. Small face icons on the spirals are users in the community. The distances between user icons, or spirals, represent how similar they are. Each spiral has a tag that shows most frequently used items, queries or

anchor texts, accessed by the community in the digital library. Therefore, viewing the overall distribution of spirals and user icons provides insight regarding social networks and their interests as well as usage trends of digital libraries. The visualization strategy of VUDM fully follows Ben Shneiderman's "Information Visualization Mantra: Overview first, then zoom & filter, then details on demand". This tutorial aims to improve the understandability and utility of VUDM.

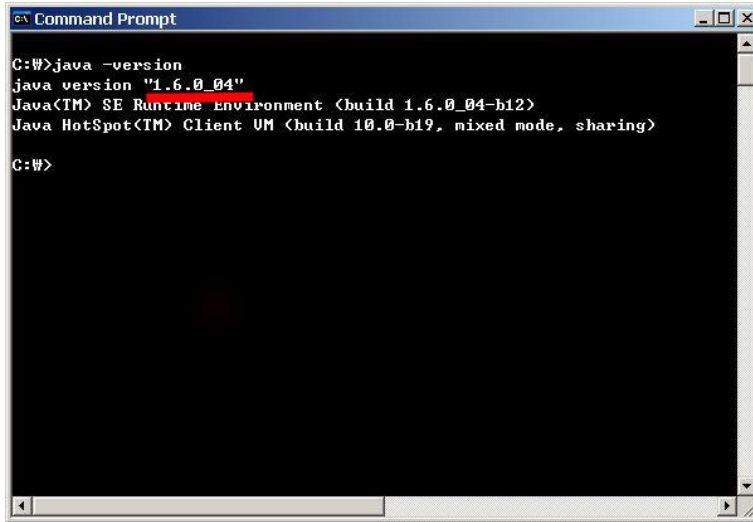
2. System Requirements & Installation

- a. Windows XP, 2000, Server 2003, or Vista
- b. Version 1.6 or later Java Runtime or Java Development Kit
- c. 512MB memory
- d. Mouse
- e. XGA(1024x768) or higher resolution monitor

3. Tutorial

3.1 Download and Installation

1. Check the Java version of your computer. In a Command Prompt window, try "java -version", as shown in Figure 47. If the version is lower than 1.6.x, download a new Java Runtime Environment(JRE) or Java Development Kit(JDK) version of 1.6 or higher from <http://java.sun.com/javase/downloads/index.jsp>. Download "JDK update 4" for JDK or "Java Runtime Environment(JRE) 6 Update 4" for JRE.



```
C:\>java -version
java version "1.6.0_04"
Java(TM) SE Runtime Environment (build 1.6.0_04-b12)
Java HotSpot(TM) Client VM (build 10.0-b19, mixed mode, sharing)
C:\>
```

Figure 47. Java version checking

2. Once a Java Runtime of 1.6 or higher version is installed, download VUDM from http://rocky.dlib.vt.edu/~shkim/VUDM_tutorial/files/VUDM_demo.zip.
3. The zipped file VUDM_demo.zip contains an executable jar file "VUDM_demo.jar" and two data sets "200601" and "200602" in sub-directories. The names of data sets represent Year+Month during which the data were collected. For example, 200602 means that the data was collected in Feb. 2006. The executable jar file, VUDM_demo.jar, is an archive of Java libraries and environment setting files that executes VUDM without tedious configuring of the Java environment of your system.
4. Move the executable jar and data sets into any folder, let's say "C:\VUDM" for this tutorial. Therefore, you will have C:\VUDM\VUDM_demo.jar and C:\VUDM\200601 and C:\VUDM\200602 after successful installation, see Figure 48.

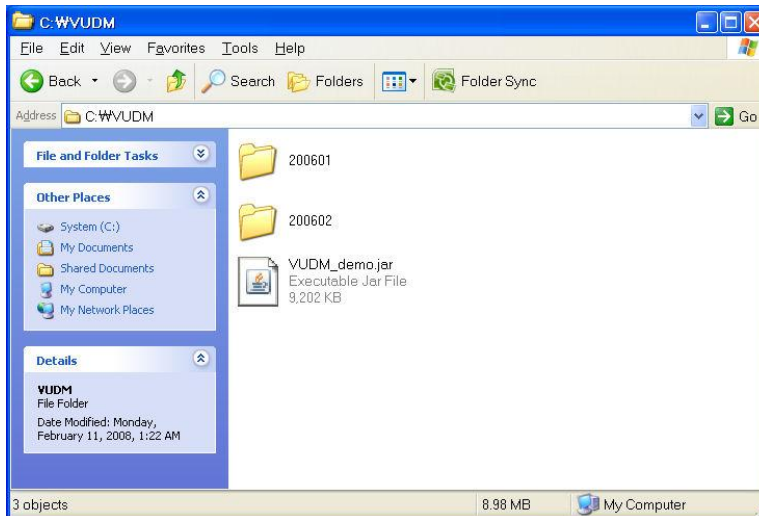


Figure 48. VUDM demo is installed under C:\VUDM

3.2 Start VUDM

1. Open a Command Prompt window and move to the directory by typing in "cd C:\VUDM"
2. Type in "Java -jar -Xmx512m VUDM_demo.jar" to invoke VUDM.

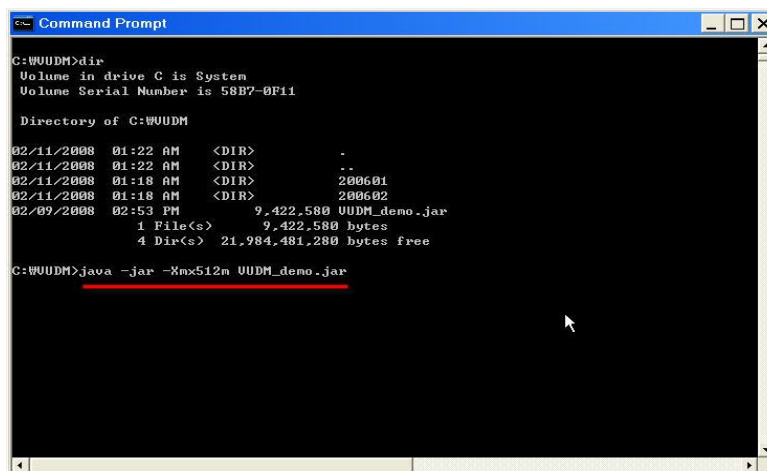


Figure 49. Invoke VUDM

The option "-Xmx512m" secures 512MB memory for running the VUDM. If you experience an "out-of-memory" error during loading a data set, increase the number after "-Xmx". The number should not be greater than the physical memory of your system.

3.3 Data Loading

1. VUDM is a multiple window application, so you can load multiple data set onto multiple windows at a same time. In order to load a data set, open the "File" menu. There are two options, "Open" and "Open in new window" in the "File" menu. Select the "Open" to load a new data set onto the current activated window. If there is no activated window, a new window will be generated, and then data will be loaded.

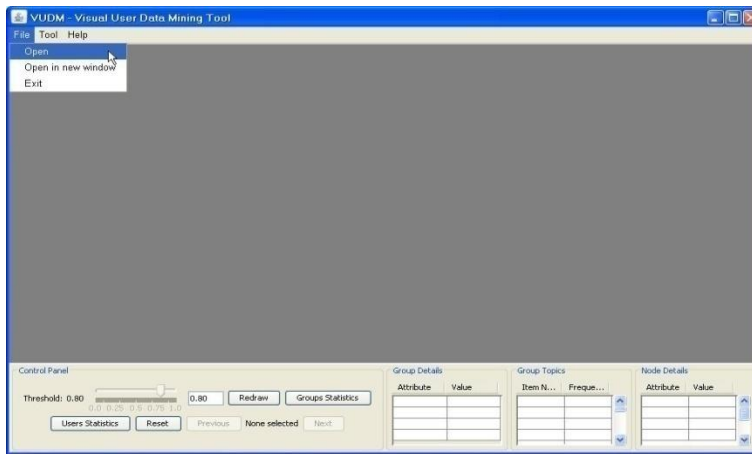


Figure 50. Open new data in current window

2. Select the "Open in new window" to generate a new window and load new data onto the window. This feature is useful when you want to keep current opened windows and load another data set.

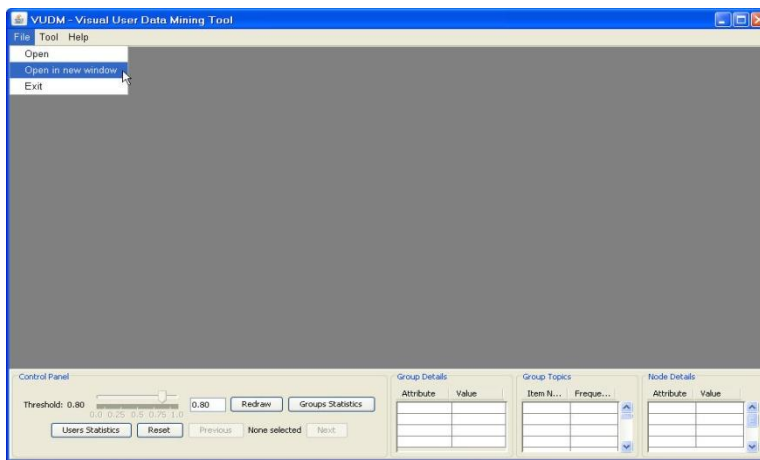


Figure 51. Open new data in a new window

3. Once you select either "Open" or "Open in new window", a file-selection dialog box will be popped up. Browse to "C:\VUDM" to select data folder. Select either "200601" or "200602" then click "Choose folder with data" button. You have to choose a FOLDER, not a single data file in the folders, as shown in Figure 52.

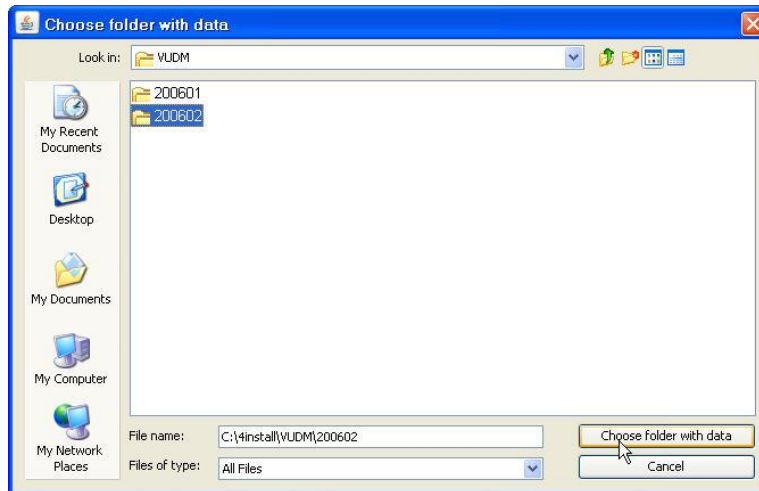


Figure 52. Select a data folder

4. Once you click the "Choose folder with data" button, the data loading process will begin. Progress of loading a data set is visualized by printing "R" in the Command Prompt window as Figure 53 shows.

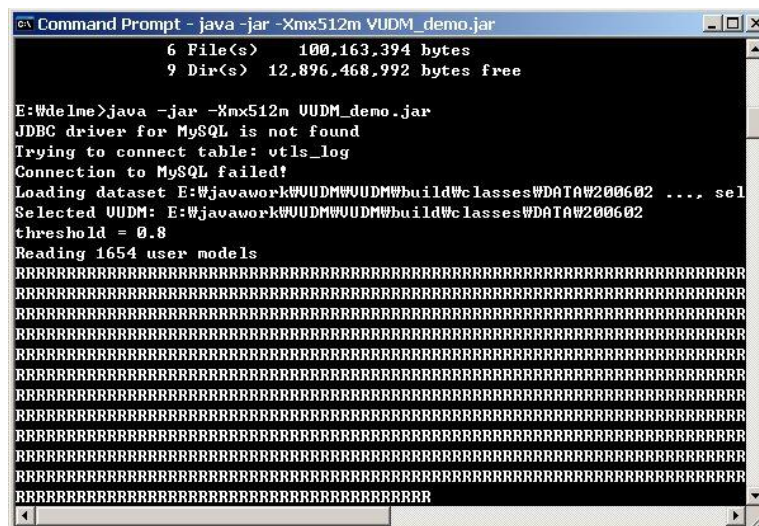


Figure 53. Data loading

5. After data loading is finished, a window with spirals and face icons will be popped up as Figure 54 shows.

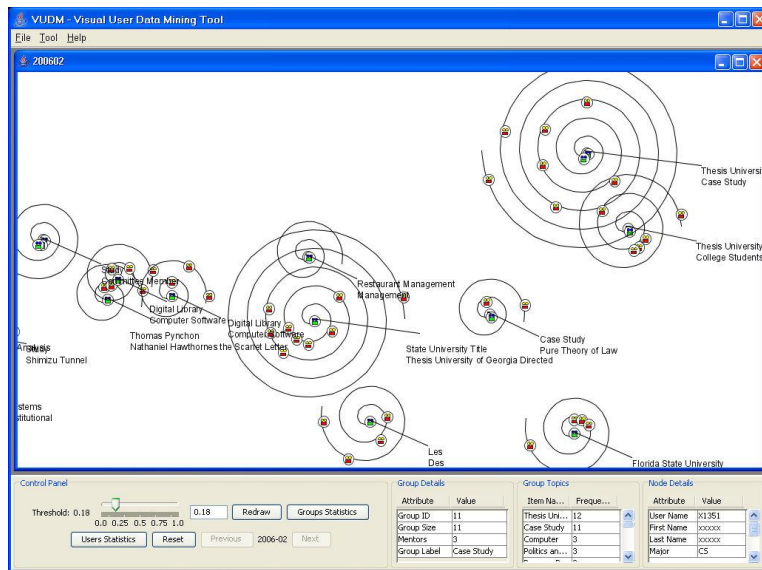


Figure 54. User data of February 2006 is successfully loaded and visualized

3.4 Data Exploration

1. Overview

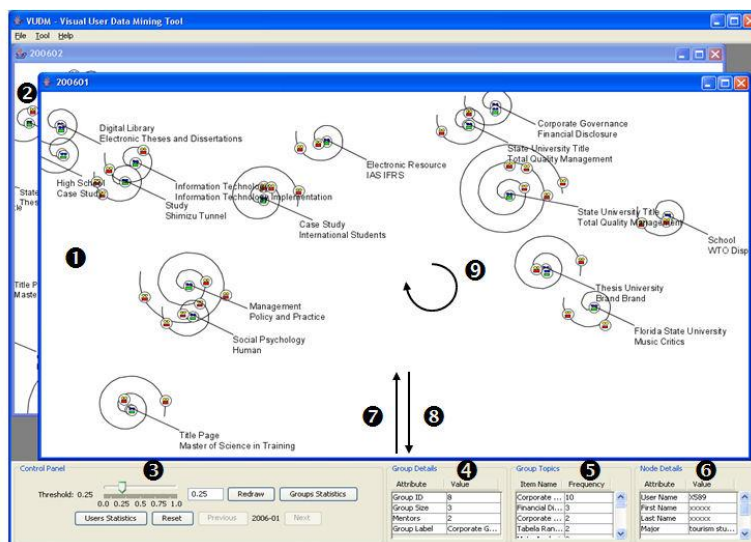


Figure 55. VUDM overview

Once data loading is successful, a new window will be generated. It will have a set of spirals and users; a term 'user' means a patron of a digital library throughout this tutorial; face icons will be visualized in the window in the portion marked '1' in Figure 55. VUDM can visualize multiple data at separate windows at a same time to make it easy to compare them, see the mark '2'. Open two or more different data sets, collected from different times, into separate windows. Through comparing them, we can see how the user communities had been changed during the period. The data name is printed at each window frame. The mark '3' is a filtering device so that you can control the strictness of the user community finding algorithm, which will be explained later in this tutorial. The marks '4', '5', and '6' identify information tables to show more detailed information. That is, '4' shows detailed information about a selected user community. You can click a spiral, which is a user community, with a mouse or just put the mouse cursor on it to select a community. This table contains Group ID, Group Size which is the number of users in the community, number of mentors, and group labels. The information table, marked '5', provides a detailed list of items, which are queries and noun-phrases which are referred by the users in the community you selected, sorted by their frequencies. Information table '6' is for detailed information about a user you selected. Try to move your mouse cursor over user communities and user icons, and see the contents of these three information tables change. The marks '7', '8', and '9' represent mouse interactions for zooming-in, zooming-out, and panning, respectively. For zooming-in and zooming-out, press the RIGHT mouse button and drag forward or backward. For panning, press the LEFT mouse button and drag into your desired direction. Spirals and labels of user communities are clickable and draggable so that you can see hidden objects in a congested area.

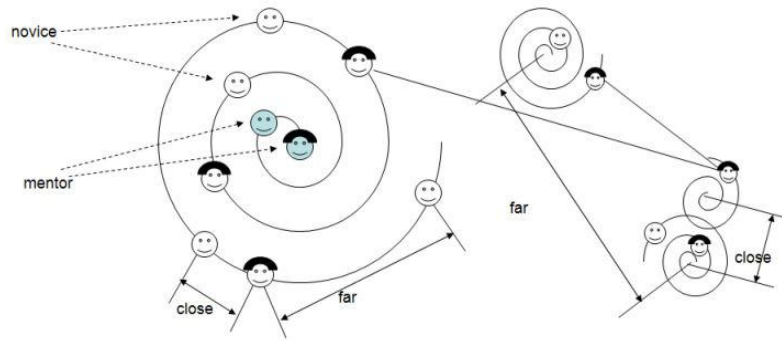


Figure 56. VUDM visualization strategy

Figure 56 briefly describes the visualization strategy of VUDM. The distance between two user communities means the two communities is similar, which means the users of two communities share a lot of research interests or learning topics. A user in a community is ranked according to the similarity between the user and the community she belonged to. More close to the center of a spiral, the more similar and more expert to the topics representing the community. This information is obtained from user registration information, and the similarity between her used items and whole items of the community.

Sometimes, when you load a data set or click the "redraw" button, you may find nothing is displayed in the window. This is because no user community was found in that setting. Two factors affect the result of user community finding. One is the data size and the other is the strictness of community finding algorithm, we will call this value θ throughout this tutorial. The θ is the threshold of similarity between two users that consider the relation between the two users is meaningful for clustering. The other factor is the size of data. If the data set is too small, the probability that any user community is found is very small. If the θ value is too high the probability of finding a user community is also very low. Therefore, if no user communities were discovered, you may add more data or select a more generous θ value to see any user community. An example of using the θ value control slider bar will be provided below in the 'filtering' session.

2. Zoom

After you successfully load a data set, you can zoom any part of users and user communities. Drag the map with the left button of the mouse, so that the area you want to zoom in locates at the center of the window. Click any part of the window with the right button of the mouse, and while holding the right button push or pull the mouse. Pushing the mouse will zoom into the center of the window and pulling the mouse will zoom out from the center of the window.

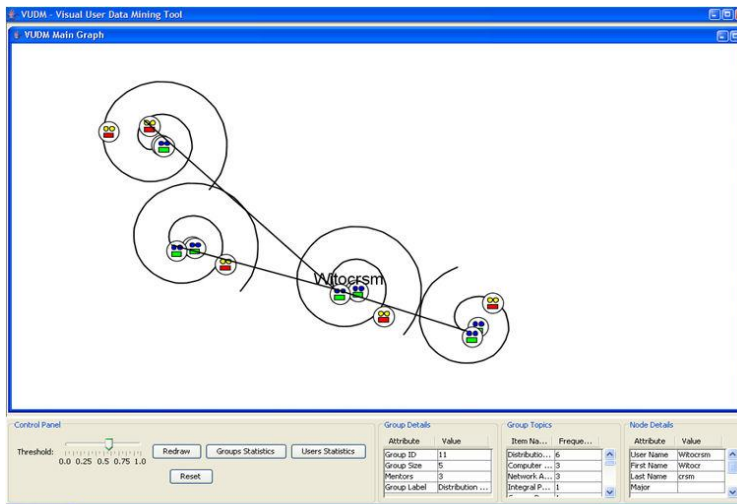


Figure 57. Zoomed image of a community

In VUDM, a user can belong to multiple user communities at a same time. This is reasonable because a normal user has multiple interests and joined in multiple communities. The linked lines among user icons across user communities represent they are an identical user. You can see the linked lines by putting the mouse cursor over the user icon.

3. Filtering

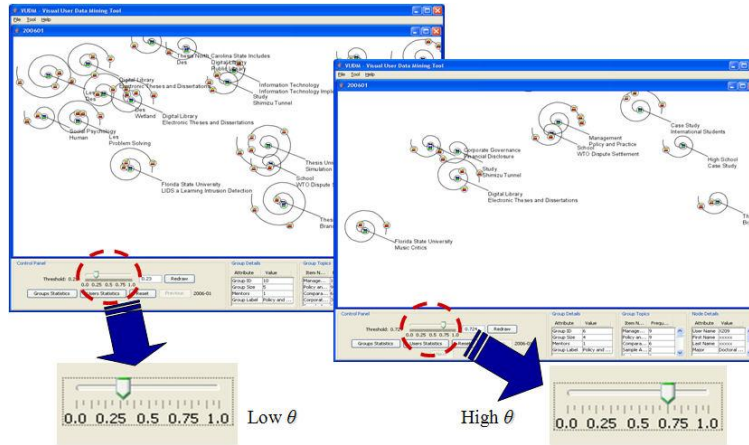


Figure 58. Control the strictness of community finding by adjusting the θ value

VUDM provides a filtering feature, a slider bar for adjusting the θ value, which is a threshold of the strictness of community finding algorithm. The higher the θ the more strict and precise communities will be found. The lower the θ , vice versa, the less precise and the greater number of user communities will be generated, see Figure 58.

In this tutorial, because the demo data size is fixed and cannot be controlled, you can adjust the θ value by sliding the bar, to control the number of user communities. Setting θ to 0 means there is no restriction in making user communities among users. Therefore, the number of found user communities will be $2^n - 1$, which is meaningless, because any combination is allowed. If θ is set to 1, no two users will be grouped together into the same community and all user communities will have only one user, which is also meaningless. VUDM automatically removes too small user communities. Therefore, above a certain θ value, no user communities will be found. For the given data for this demo, the meaningful θ values are between 0 and about 0.2. Finding a meaningful θ is up to the VUDM user, and is dependent on the user data properties.

For examples, open the data set 200601 in a new window and try to change the θ values to 0.2, 0.18, and 0.15 by yourself. Move the slider bar into 0.20 or enter

directly 0.20 at the text input box next to the slider bar, and click the "Redraw" button; then you will get the result shown in Figure 59.

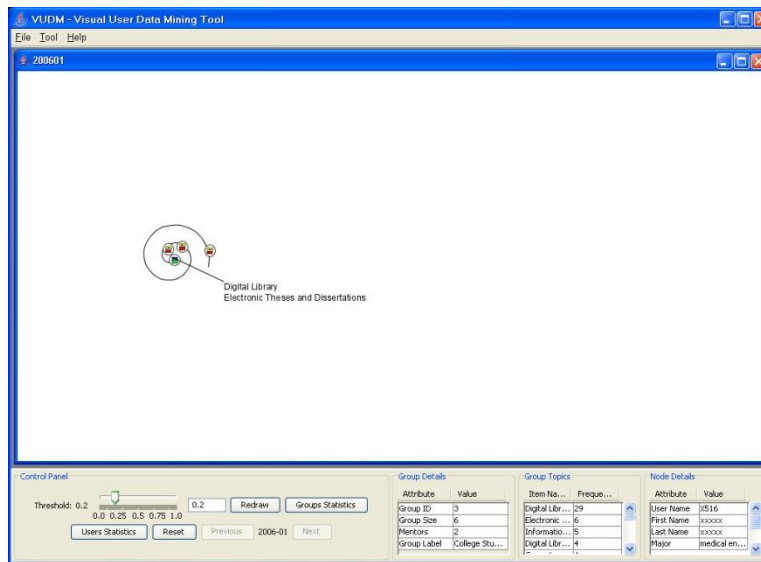


Figure 59. User community of January 2006 ($\theta = 0.2$)

A user community labeled as "Digital Library, Electronic Theses and Dissertations" is found. This label is selected because the items "Digital Library" and "Electronic Theses and Dissertations" are most frequently used by all users in the community. You can check the full list of used items and their frequencies from the second information table located at the bottom right of the VUDM display.

Next, enter 0.18 at the text input box and then click the "Redraw" button. This will present a result as shown in Figure 60.

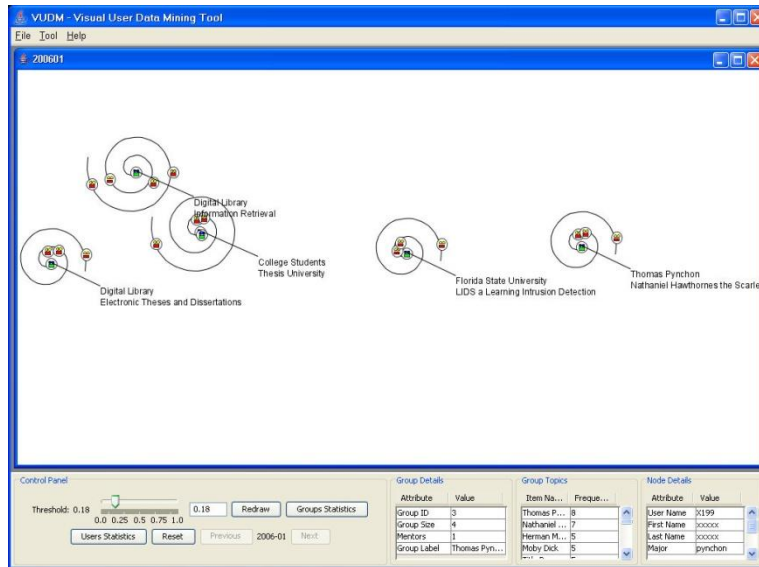


Figure 60. User community of January 2006 ($\theta = 0.18$)

The θ value 0.18, which is lower than the previous 0.20, produced more user communities. A total of five user communities, including the one found when θ is 0.2, were found and visualized. The VUDM also clusters user communities based on their similarities. That is, similar user communities locate close each other while non-similar communities are isolated. In this result, three communities are clustered because they share some topics, such as "Digital Library" and "Electronic Theses and Dissertations". The shared topics may be not fully displayed in their labels. You may have to look into the detailed item tables. The other two communities are isolated because they are not similar to other communities.

Now, for the last example, let's try one more θ value. Enter 0.15 at the text input box and then click the "Redraw" button.

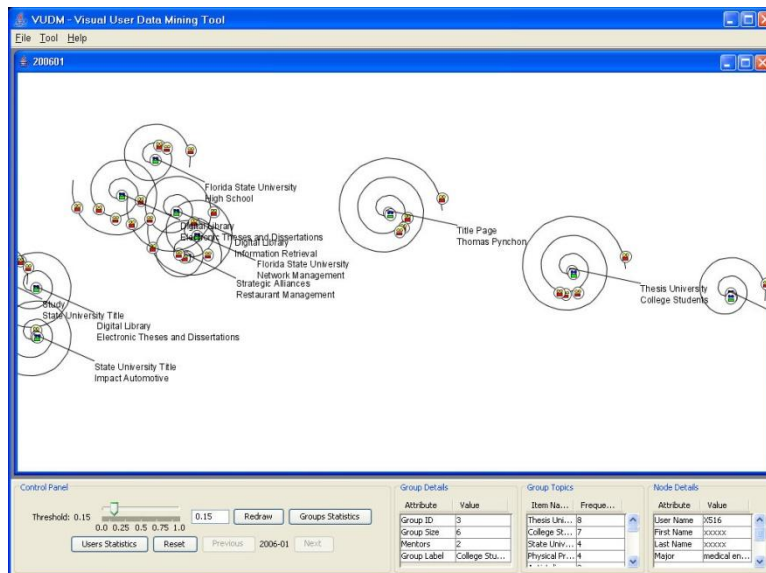


Figure 61. User community of January 2006 ($\theta = 0.15$)

Understandably, more user communities were found because a lower θ value is used. Clusters of user communities are also affected by the low θ value.

Let's compare this data with the data of the following month. Open the data set "200602" in a new window and set the θ to 0.15, the same value used previously, then click "redraw". Now, you will see how the users and user communities have changed between the two months. Also, this provides insight regarding which topics or items attracted more users during the month. It is overview of usage trends of the digital library, NDLTD, at the time the data is collected.

This section described the filtering feature of VUDM. You can control the number of user communities by changing the θ value. You will select a proper θ value according to the size of the data and its properties, to analyze the data.

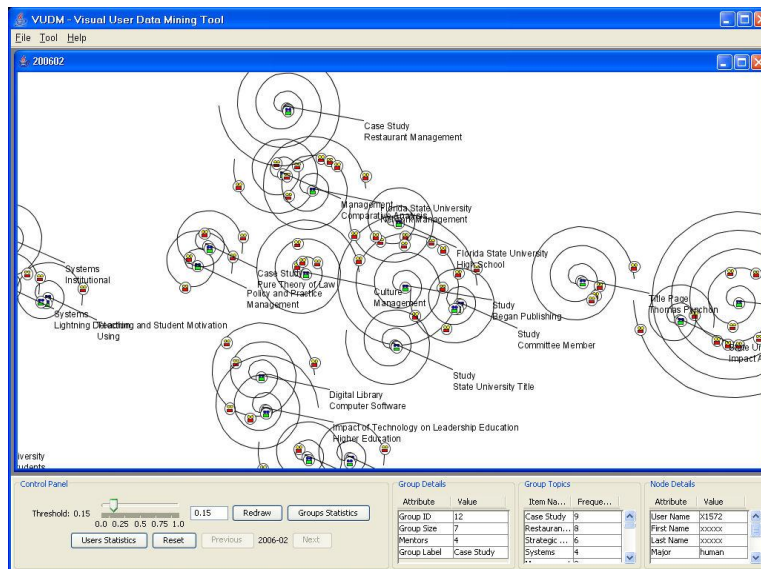


Figure 62. User community of February 2006 ($\theta = 0.15$)

4. Detail on Demand

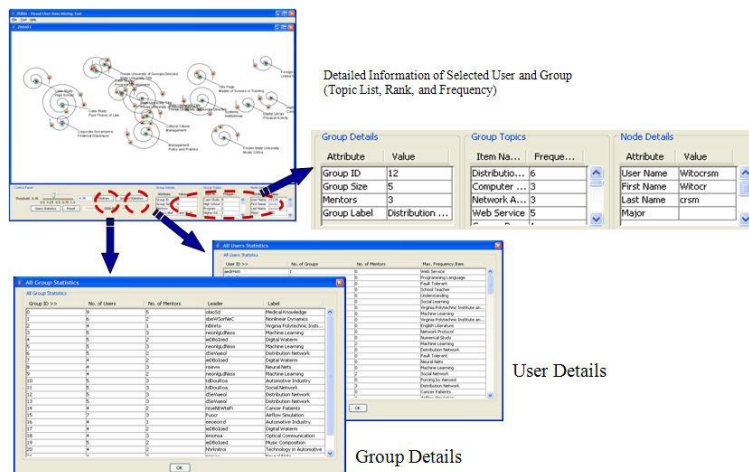


Figure 63. Details on demand

Figure 63 shows how you can get detailed information about users and user communities. Besides the three detailed information tables, which are explained before, there are two buttons labeled as "Groups Statistics" and "Users Statistics" at the bottom left of VUDM main frame. The "Groups Statistics" button invokes a pop-up window that shows a table which contains detailed information about all

user communities in the opened data. This table includes even the information of user communities which are not visualized. The "Users Statistics" button is to invoke a pop-up window that shows a table containing detailed information about all individual users in the data.

4. Release Notes

1st release: Feb. 7th 2008, version 0.9

5. Bug Reports and Feedback

To submit a bug or request a feature, please send an email to shk@vt.edu

6. Web Pages

Related web pages.

Questionnaire for library experts survey on VUDM.

<http://www.dlib.vt.edu>

Virginia Tech Digital Library Research Laboratory

<http://www.ndltd.org>

Networked Digital Library of Theses and Dissertations

<http://www.cs.vt.edu>

Department of Computer Science at Virginia Tech

Appendix B-4: Institutional Review Board Approval Letter

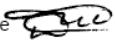
 VirginiaTech		Office of Research Compliance Institutional Review Board 2000 Kraft Drive, Suite 2000 (0497) Blacksburg, Virginia 24061 540/231-4991 Fax 540/231-0959 e-mail moored@vt.edu www.irb.vt.edu FWA00000572(expires 1/20/2010) IRB # is IRB00000667
DATE: February 19, 2008		
MEMORANDUM		
TO:	Edward A. Fox Seonho Kim	Approval date: 2/19/2008 Continuing Review Due Date: 2/4/2009 Expiration Date: 2/17/2009
FROM:	David M. Moore 	
SUBJECT:	IRB Expedited Approval: "Library Expert Survey on VUDM (Visual User Model Mining Tool)", IRB # 08-076	
This memo is regarding the above-mentioned protocol. The proposed research is eligible for expedited review according to the specifications authorized by 45 CFR 46.110 and 21 CFR 56.110. As Chair of the Virginia Tech Institutional Review Board, I have granted approval to the study for a period of 12 months, effective February 19, 2008. As an investigator of human subjects, your responsibilities include the following: 1. Report promptly proposed changes in previously approved human subject research activities to the IRB, including changes to your study forms, procedures and investigators, regardless of how minor. The proposed changes must not be initiated without IRB review and approval, except where necessary to eliminate apparent immediate hazards to the subjects. 2. Report promptly to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others. 3. Report promptly to the IRB of the study's closing (i.e., data collecting and data analysis complete at Virginia Tech). If the study is to continue past the expiration date (listed above), investigators must submit a request for continuing review prior to the continuing review due date (listed above). It is the researcher's responsibility to obtain re-approval from the IRB before the study's expiration date. 4. If re-approval is not obtained (unless the study has been reported to the IRB as closed) prior to the expiration date, all activities involving human subjects and data analysis must cease immediately, except where necessary to eliminate apparent immediate hazards to the subjects. Important: If you are conducting federally funded non-exempt research, this approval letter must state that the IRB has compared the OSP grant application and IRB application and found the documents to be consistent. Otherwise, this approval letter is invalid for OSP to release funds. Visit our website at http://www.irb.vt.edu/pages/newstudy.htm#OSP for further information. cc: File		
<i>Invent the Future</i> VIRGINIA POLYTECHNIC INSTITUTE UNIVERSITY AND STATE UNIVERSITY <i>An equal opportunity, affirmative action institution</i>		

Figure 64. IRB approval letter for the experts review