Department of Computer Science	660 McBryde Hall
Virginia Tech	(0106), Virginia Tech,
Technical Report Series	Blacksburg, VA 24061

Design and Evaluation of Techniques to Utilize Implicit Rating Data in Complex Information Systems

Seonho Kim[†], Edward A. Fox[†], Weiguo Fan[†], Chris North[†],
Deborah Tatar[†], Ricardo da Silva Torres[‡]

shk, fox, wfan, north, tatar@vt.edu[†], rtorres@ic.unicamp.br[‡]

[†] Department of Computer Science at Virginia Tech
[‡] Institute of Computing at State University of Campinas, Brazil

January 1, 2007

Table of Contents

Abstract

Research in 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 - where it is possible to classify items involved and to make personalization 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. In this report, we present techniques for collecting, storing, processing, and utilizing implicit rating data of Digital Libraries for analysis and decision support. We present our pilot study to find virtual user groups using implicit rating data. We demonstrate the effectiveness of implicit rating data for characterizing users and finding virtual user communities, through statistical hypothesis testing. Further, we describe a visual data mining tool named VUDM (Visual User model Data Mining tool) that utilizes implicit rating data. We provide the results of formative evaluation of VUDM and discuss the problems raised and plans for further studies.

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Problem Statement and Related Work

As two-way World Wide Web services such as blogs, wikis, online journals, online forums, etc. became popular, more people were able to express themselves and play more active roles in online societies [22, 34]. This trend changed WWW users from passive anonymous observers to visible individuals with personalities. Such users, in increasing numbers, are patrons of digital libraries (DLs), e.g., researchers and distance learners. Studying users of DLs is providing opportunities for research on collaborative filtering, personalization, user modeling, and recommender systems [40]. Most such studies consider users' explicit ratings on the information they select, as well as users' preferences – e.g., on research areas, majors, learning topics, or publications – which are entered explicitly [23, 26, 40]. However, as Table 1 presents, obtaining explicit rating data is difficult and expensive. Also, the amount of information is small and fixed by the questions to which the answers are given, and the possible questions should be limited. Further, terminology associated with the broad topical coverage of most DLs 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. Therefore, we are trying to utilize implicit rating data (so-called because the data was not entered explicitly in answer to questions) which is easy to obtain and suffers less from terminology issues. Table 1 presents a comparison between implicit rating data and explicit rating data.

Table 1. Comparison of explicit and implicit rating data in digital libraries.

	Explicit Rating Data	**Implicit Rating Data**
Source	User questionnaire, Online survey, Offline survey, User review, etc.	User activities, System states and variables, Consumed time, etc.
Cost to collect	Expensive	Cheap

Information	Correct and specific	Needs to be analyzed, contains potential knowledge
Amount of Information	Fixed	Depends on analysis methods and applications
Possible Questions	Limited	Unlimited
Problems in applying to DLs	Terminology problem, Interrupt user tasks	Technologies to collect, store, and process are under-developed

In collecting implicit rating data, we assume the user is engaged in tasks to achieve her goals, such as finding books or documents.

Many previous research studies utilized implicit rating data for various purposes in complex information systems. Nichols [25] and the GroupLens team [21] showed the great potential of implicit rating data when it is combined with existing systems to form a hybrid system. Further, we utilized users' implicit rating data in visualizing users, user communities and usage trends of DLs [19], and proposed collaborative filtering techniques for personalization [18]. Visualization can help us to answer complex and comprehensive questions on DLs by supporting direct involvement of users in exploration and data mining, so they can utilize their creativity, flexibility, and general knowledge [15]. Some of the broad areas of related work include: visualization of social networks, visualization of documents and topics, learning about users, and user modeling. For example, visualization of networks of criminals and criminal events can help unearth hidden patterns in crime data as well as detect terrorist threats [39]. Boyd, working with Social Network Fragments [7], visualized clusters of contacts derived from the *to* and *cc* lists in email archives. Heer, in Vizster [11], visualized relationships between members in an online date site Friendster [4]. Wise's SPIRE Themescape [38] facilitates visualization of the topic distribution in a large document space. Probabilistic approaches to user modeling have made it possible to learn about user profiles, as well as to revise them based on additional data [24, 31]. Tang utilized users' browsing patterns for collaborative filtering [35]. Webb examined challenging user modeling approaches like data rating, concept drift, data sparseness, and computational complexity [36].

1.2 Organization of this Technical Report

This report consists of six chapters and one appendix. In this chapter, Introduction, we have stated the current trends of web technologies and problems that motivated this research, and described some related works.

Chapter 2, *Data Collection and User Modeling*, presents a detailed look at two sources of data, a user tracking system and an online user survey, and methods to store and process the data.

Chapter 3, *Preliminary Experiment: Characterizing Users with Implicit Rating Data and Verification with Explicit Rating Data*, presents a result of an experiment that tested the effectiveness of implicit rating data on user characterization and community finding. Explicit rating data also was used for evaluation of the result.

Chapter 4, *Hypotheses Testing: Effectiveness of Implicit Rating Data in Characterizing Users and User Communities*, presents results of hypothesis tests to support the result of Chapter 3.

Chapter 5, *Effectiveness of Four Different Data Types in Community Finding*, presents results of a study on comparing the effect of different types of data on the performance of user community finding.

Chapter 6, *Supply / Demand Analysis in NDLTD (Networked Digital Library of Theses and Dissertations): Using Implicit Rating Data*, demonstrates how we could utilize implicit rating data to analyze NDLTD [5] by measuring the information supply and demand in the system. This experiment shows that analyzing implicit rating data provides particular information, such as the amount of information demands, which is hard to obtain from analyzing explicit rating data.

Chapter 7, *VUDM: A Visual Data Mining Tool Utilizing Implicit Rating Data*, describes a Visual Data Mining Tool, which is developed to visualize users and user communities in Digital Libraries using implicit rating data, and a result of formative evaluation.

Chapter 8, *Conclusions*, discusses the results of experiments, hypothesis tests, the formative evaluation, and future work.

Finally, the *Appendix* includes the Institutional Review Board (IRB) documents we prepared for our experiments.

Chapter 2. Data Collection and User Modeling

In this research, we employ user modeling techniques to represent characteristics of users. Each user in DLs has her own user model which is a data structure that contains information that reflects her research interests. User model data is a realization of a user model which is represented using XML and contains demographic information, explicit rating data, and implicit rating data. Explicit rating data, including demographic information, is collected from online surveys and questionnaires. Implicit rating data is collected by the user tracking system which is embedded in a web user interface of the DL. This chapter explains two sources of data, the user tracking system and the online survey, and methods to store and process the user model data.

2.1 User Tracking System: Source of Implicit Rating Data

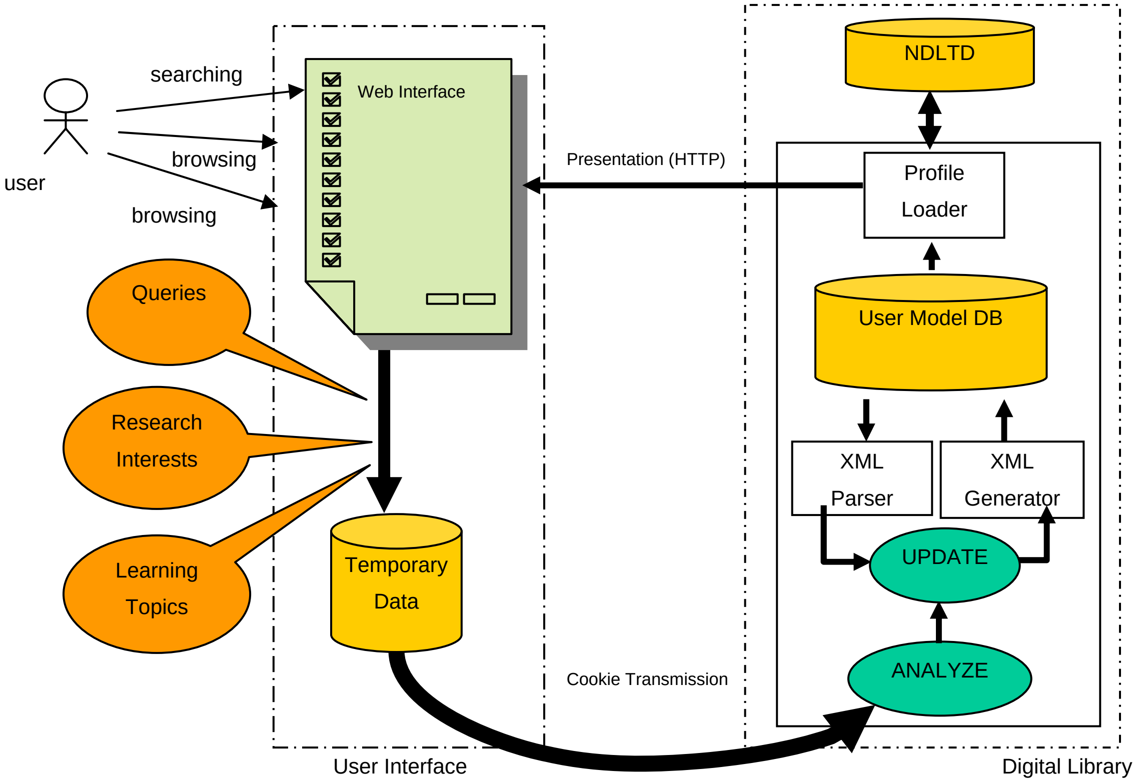
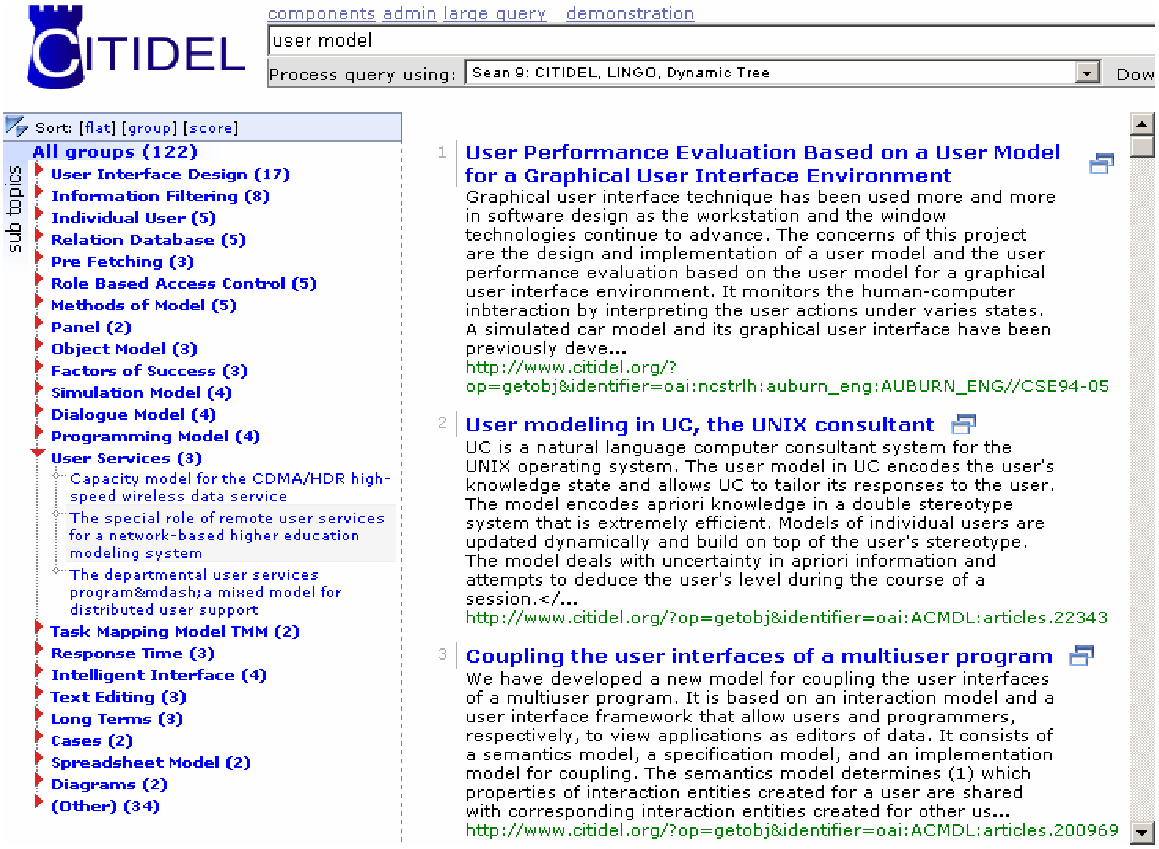


Figure 1. System diagram of a user tracking system.

The standard HTTP log protocol, unfortunately, cannot extract 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 need 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 embedded a user tracking system. Figure 1 is a overall system diagram of our user tracking system. It consists of mainly two parts, one is user tracking part in user interface at client-side, and the other is data updating part at server-side. At client-side, user behavior, such as entering queries, clicking hyperlinks, and browsing hyperbolic tree are captured by JavaScript embedded user web interface. The captured information is stored in local disk by using cookie technique. The cookie is transferred to server-side when the user ends login session or terminates the internet explorer. At server-side, the transferred cookies is analyzed and transformed into XML format by “analyzer”. Therefore, this temporary XML file includes user tracking information for one login session. The “update” module checks whether the user model of current user already exists in the database. If the user’s user model does not exist in database, it is entered into the database as a new record. If the user’ user model already exists in the database, the temporary data is merged with the previous data and entered back into database. User model database is also accessed when a returning user login the digital library to load her profile.

Our system is embedded in the web search interface of CITIDEL [2] and NDLTD [5]. When a user logs in these DLs, her profile information is loaded, and the user tracking system starts to collect data about the user’s activities. Besides queries, it collects information from the search result documents presented to the user, see Figure 2 [16]. The left (dynamic tree) and right (HTML page) frames present a clustered result set efficiently. Data on user interactions such as opening clusters or selecting a document are stored temporarily in a cookie. Names of document clusters in the left frame of this interface were generated by finding the most representative noun phrase among the documents in the cluster. We assume that the user will browse the dynamic tree based on her judgment whether the name of cluster is relevant to her query or not; also this judgment is closely related with her research interests or learning topic. Therefore, we collect the names of clusters and information about which clusters were examined and which clusters were not examined; we treat this as implicit rating data. 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 to be transferred to the server-side process module, analyzer, for updating the user data when the session is closed.

Once the temporary rating data is transferred to the DL 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.


CITIDEL
components admin large query demonstration
Dow
Sort: [flat] [group] [score]
sub topics

user model

Process query using: Sean 9: CITIDEL, LINGO, Dynamic Tree

All groups (122)
- User Interface Design (17)
- Information Filtering (8)
- Individual User (5)
- Relation Database (5)
- Pre Fetching (3)
- Role Based Access Control (5)
- Methods of Model (5)
- Panel (2)
- Object Model (3)
- Factors of Success (3)
- Simulation Model (4)
- Dialogue Model (4)
- Programming Model (4)
- User Services (3)
  - Capacity model for the CDMA/HDR high-speed wireless data service
  - The special role of remote user services for a network-based higher education modeling system
  - The departmental user services program&mdash;a mixed model for distributed user support
- Task Mapping Model TMM (2)
- Response Time (3)
- Intelligent Interface (4)
- Text Editing (3)
- Long Terms (3)
- Cases (2)
- Spreadsheet Model (2)
- Diagrams (2)
- (Other) (34)

1 **User Performance Evaluation Based on a User Model for a Graphical User Interface Environment**
Graphical user interface technique has been used more and more in software design as the workstation and the window technologies continue to advance. The concerns of this project are the design and implementation of a user model and the user performance evaluation based on the user model for a graphical user interface environment. It monitors the human-computer inbteraction by interpreting the user actions under varies states. A simulated car model and its graphical user interface have been previously deve...
http://www.citidel.org/?op=getobj&identifier=oai:ncstrlh:auburn_eng:AUBURN_ENG//CSE94-05

2 **User modeling in UC, the UNIX consultant**
UC is a natural language computer consultant system for the UNIX operating system. The user model in UC encodes the user's knowledge state and allows UC to tailor its responses to the user. The model encodes apriori knowledge in a double stereotype system that is extremely efficient. Models of individual users are updated dynamically and build on top of the user's stereotype. The model deals with uncertainty in apriori information and attempts to deduce the user's level during the course of a session.</...
http://www.citidel.org/?op=getobj&identifier=oai:ACMDL:articles.22343

3 **Coupling the user interfaces of a multiuser program**
We have developed a new model for coupling the user interfaces of a multiuser program. It is based on an interaction model and a user interface framework that allow users and programmers, respectively, to view applications as editors of data. It consists of a semantics model, a specification model, and an implementation model for coupling. The semantics model determines (1) which properties of interaction entities created for a user are shared with corresponding interaction entities created for other us...
http://www.citidel.org/?op=getobj&identifier=oai:ACMDL:articles.200969

Figure 2. A JavaScript based user interface.

Figure 3 shows an example of the user tracking data collected during one login session of a participant in one of our experiments using our user tracking system. Information about whether it was a query or a cluster, examined or not-examined is tagged by special characters like %28, %29, and %3C. This part corresponds to the browsing of the result set of a single query. This participant explicitly answered in the questionnaire that he has an interest in Cross Language Information Retrieval (CLIR).

```
 %3CExample%20Based%20Machine%3CEnglish%20Japanese%3CMachine%20Translation%3C
 Approach%20to%20Machine%3CMachine%20Adaptable%20Dynamic%20Binary%3CStatistical
 %20Machine%3CFuture%20of%20Machine%20Translation%3CCross%20Language%20Informa
 tion%20Retrieval%3CModel%3CKnowledge%3CNatural%3CLinguistic%3CApplication%3CMul
 ti%3CTechnology%3CSyntax%20Directed%20Transduction%3CChinese%20Machine%3CInterlin
 gual%3CUnderstanding%3CLexical%20Conceptual%3CMachine%3CIntegrated%3CLexicon%3C
 %20Language%3CPhrase%3CRecombination%20of%20Genes%3CCross%20Language%20Information
%20Access%3CUser%3CGraphical%3CBaseline%3CDisambiguation%3CRouting%3C Indexin
 g%3CMorphology%3CExploiting%3C%28Other%29%28%3CCross%20Language%20Information
 %20Retrieval%29%28%3CDictionary%20Based%29%28%3CCross%20Language%20Information
 %20Retrieval%20CLIR%20Track%29%28%3CEnglish%20Chinese%29%28%3CResolving%20Am
 biguity%20for%20Cross%20Language%29%28%3CCross%20Language%20Text%29%28%3CTran
 slation%20Resources%29%28%3CCross%20Language%20Evaluation%20Forum%20CLEF%29%2
 8%3CTREC%20Cross%20Language%29%28%3CCross%20Language%20Information%20Access%
 29%3CCross%20Language%20Information%20Retrieval%20CLIR%3CTREC%20Experiments%20
 at%20Maryland%3CCLIR%20Using%20a%20Probabilistic%20Translation
```

Figure 3. Captured temporary data in cookie file.

Figure 4 is our Domain Generalization Graph (DGG) for the user activity attribute in our model; DGGs are more commonly used in connection with data mining targeted on sales or transaction data [9] to represent the comprehensive relations between attributes. Each node in the graph represents a partition of the values that can be used to describe the attributes. The higher nodes, destination of arrows, means the attributes represented by the node is more general attributes than their lower nodes, which are at starting of arrows. The “ANY” node means, therefore, the most general attribute, which has no specific characters, and every attribute relation arrows end at “ANY”. The discrete attribute “frequency” is independent of other attributes of user activity. Edges between adjacent nodes describe the generalization relations between the nodes. Each user activity has a direction, where:

- “rating” means the user gives some feedback to the system;
- “perceiving” means the user doesn’t give feedback to the system; and
- regarding intention, “implicit” means the user gives feedback implicitly while “explicit” means feedback is given explicitly.

Thus, sending a query and reading a title are not “rating,” since we don’t give any feedback. However, expanding and skipping a cluster are “rating” – by which we indicate whether the cluster is interesting or not. For an example of intention, note that entering a query is “implicit,” because our purpose is not to characterize ourselves. However, entering user information or preferences is “explicit.”

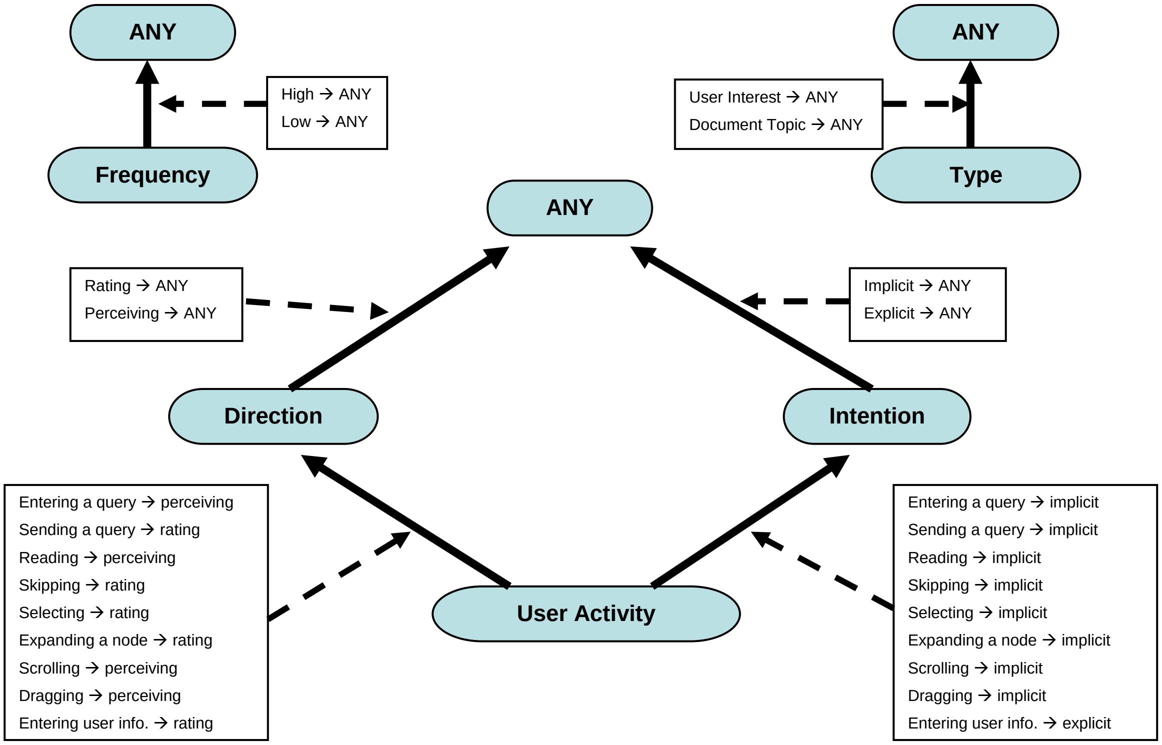
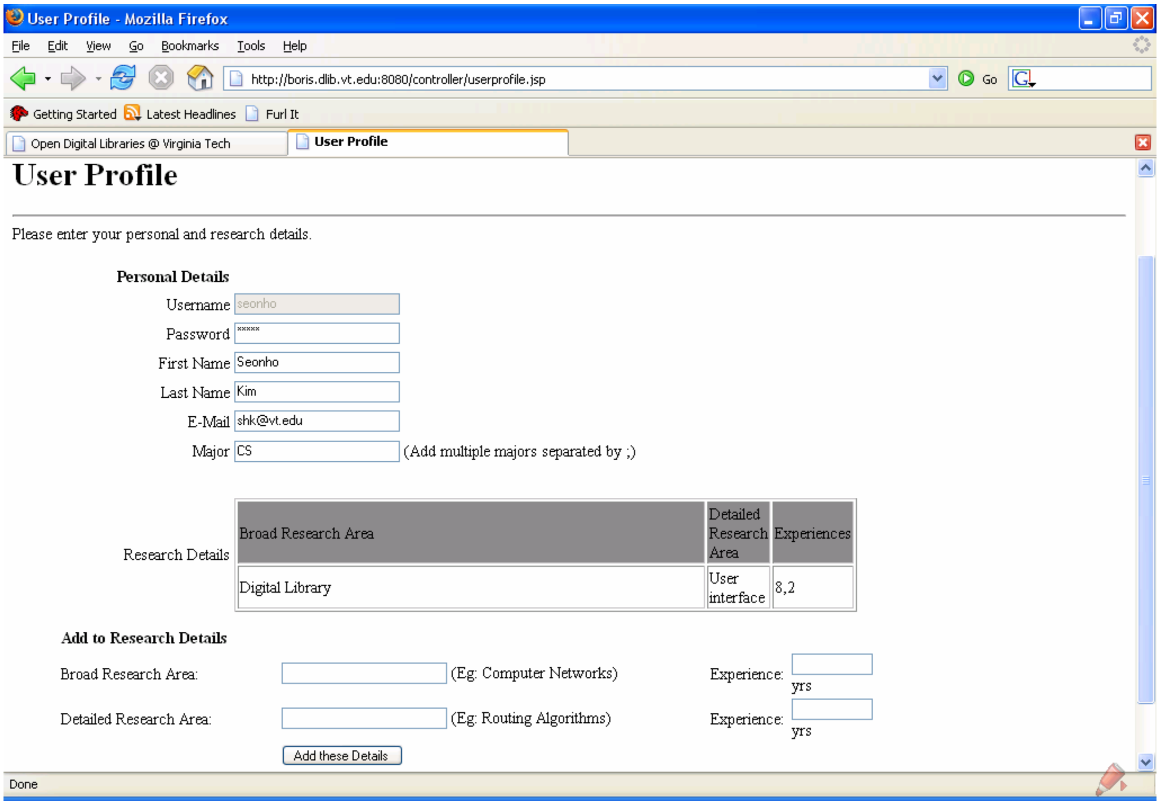


Figure 4. Domain Generalization Graph (DGG) for “user activity” attributes.

2.2 Online User Survey: Source of Explicit Rating Data

Even though our focus in this doctoral research work is proving the effectiveness of implicit rating data and utilizing it, we also collect explicit rating data about user’s research interests, learning topics, and basic demographic information for evaluation purposes. Some explicit rating data will be used in our visual data mining tool (VUDM), see Chapter 7, for labeling user icons and estimating a user's level of expertise. Figure 5 is a screen shot of an online survey, which is a part of the user registration for the NDLTD search result document clustering service. Using this survey, we collect explicit rating data, such as user's name, email, major, broad research interest, detailed research interests, and experience years for each research topic. Explicit rating data collected from this survey also is used to initialize the user's user model data, which will be described in the next section.


User Profile - Mozilla Firefox
File
Edit
View
Go
Bookmarks
Tools
Help
http://boris.dlib.vt.edu:8080/controller/userprofile.jsp
Go
Getting Started
Latest Headlines
Furl It
Open Digital Libraries @ Virginia Tech
User Profile
Add these Details
Done

User Profile

Please enter your personal and research details.

Personal Details

Username	seonho
Password	*****
First Name	Seonho
Last Name	Kim
E-Mail	shk@vt.edu
Major	CS (Add multiple majors separated by ;)

Research Details

Broad Research Area	Detailed Research Area	Experiences
Digital Library	User interface	8,2

Add to Research Details

Broad Research Area: (Eg: Computer Networks) Experience: yrs

Detailed Research Area: (Eg: Routing Algorithms) Experience: yrs

Figure 5. Online survey as a part of registration process: collecting explicit rating data.

2.3 User Model Data: Structure for Storing Explicit and Implicit Rating Data

Our user tracking system employs user modeling techniques. User models are implemented with user profile data such as name, ID, sex, major, interests, position, hobby, etc. Our system, illustrated in Figure 1, uses XML formatting for all messages exchanged and stored among the DL components. Once the temporary user rating data is transferred to the DL server, it is processed by the *analyzer* and added to the user model database in XML form. The XML SAX parser library [30] is used to analyze and update user model data files.

Figure 6 shows the XML schema of our user model. The information in the user model can be broken into two categories: personal information and search history information. The *proposed* element contains set of terms and their frequencies which are identified by the search engine and document clustering system of digital library, and “proposed” to the user through user interface to select if she is interested in the terms. Also, the *selected* element contains set of terms and their frequencies, which are actually “selected” by the user to obtain more information by updating current page or moving to 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"/>
<element name="name">
<complexType>
<sequence>
            <element name="first" type="string"/>
            <element name="last" 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>
<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 6. Structure of user model data.

The following tags fall under the category of personal information:

- <UserID> – holds the unique username (also the name of the file)
- <email> – holds the user's email address
- <name> – parent element
  - <first> – child element, holds first name of user
  - <last> – child element, holds last name of user
- <major> – holds user's major
- <research> – parent element, holds a research area entered by the user (can have more than one)
  - <interest> – child element, holds an interest within the specified area
  - <experience> – child element, holds the number of years involved in the research interest
- <group> – group element, hold a list of interest groups this user belonged and their probabilities. This information is generated by analyzer after user model is loaded.

The following tags make up the search history (and other tags used by the analyzer):

- <query> – holds the query text in <item> tags. Also keeps track of how many times the user has entered that query string (by using a freq (frequency) attribute in the item tag.
- <proposed> – holds all cluster titles that have been proposed to the user (and the frequency)

- <selected> – holds all cluster titles that the user has selected (and the frequency)

Figure 7 presents a logical structural overview of user model data. “User Description” is entered by the user through the online survey. “Groups” is generated by the analyzer. “User tracking” is collected by the tracking system. Based on the source of information, user description, group information, and user tracking information, which are from the survey, analyzer, and tracking system respectively, comprise the user model data. Items in piled rectangles in the figure mean that multiple items can be added.

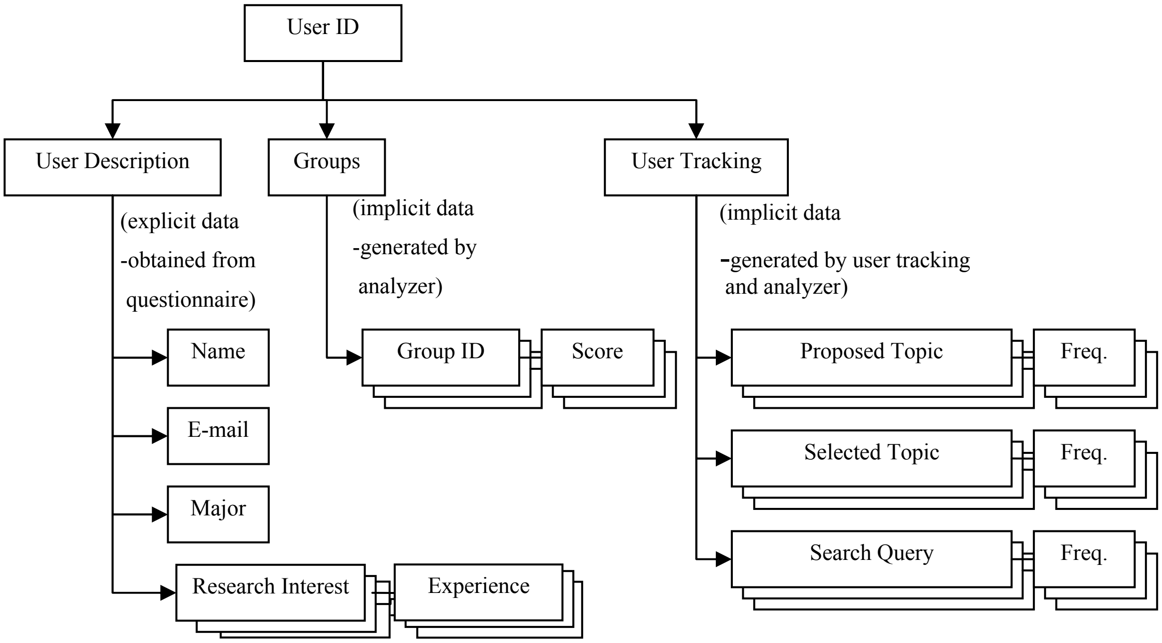


Figure 7. Structure of user model.

Tables 2 through 6 describe detailed technical information of each item in the user model data such as data type, generation source, and valid value ranges,

Table 2. Data structure of ResearchInterest in user model data.

	Source	Data Structure	Range
Research Interest	Online Survey	String	N/A
Experience	Online Survey	Real	0.0 – 60.0

Table 3. Data structure of ResearchLearningTopic in user model data.

	Source	**Data Structure**	**Range**
Topic	User Tracking	String	N/A
Frequency	User Data Analyzer	Integer	0 - MaxInt

Table 4. Data structure of SearchQuery in user model data.

	Source	**Data Structure**	**Range**
Query	User Tracking	String	N/A
Frequency	User Data Analyzer	Integer	0 - MaxInt

Table 5. Data structure of Groups in user model data.

	Source	**Data Structure**	**Range**
GroupID	Grouping Algorithm	Integer	Unique Identification Number
Score	Grouping Algorithm	Double	0.0 – 1.0

Table 6. User model data description.

	Type (Implicit/Explicit)	**Source**	**Data Structure**	**Range**
ID	Explicit	User Entered	String	N/A
FirstName	Explicit	User Entered	String	N/A
LastName	Explicit	User Entered	String	N/A
Email	Explicit	User Entered	String	N/A
Major	Explicit	User Entered	String	N/A
ResearchInterest List	Explicit	User Entered	List of Strings	3
GroupList	Implicit	User Data Analyzer	List of Communities	200 maximum
SearchQueryList	Implicit	User	List of	500 maximum

		Tracking	SearchQueries	
ProposedList	Implicit	User Tracking	List of ResearchLearningTopics	500 maximum
SelectedList	Implicit	User Tracking	List of ResearchLearningTopics	500 maximum

Generally, user models are implemented with shared feature equations and parameters for the equations that represent the users' characteristics. However, our user models consist of common feature equations, raw data items, and statistics collected by a user tracking system. Parameters for feature equations, such as probability of belonging to certain user groups, similarity with certain users, and amount of information demands can be calculated from the raw data and statistics when they are needed. Therefore, our user models are more interoperable and transferable, also, containing more potential knowledge.

```
<?xml version="1.0" ?>
- <user>
  <userID>seonho</userID>
+ <userInfo>                                          (1)
+ <userInterests>
- <community>                                         (2)
     <member score="0.743">001</item>
     <member score="0.183">003</item>
  </community>
- <query>
     <item freq="3">Educational Library</item>
     <item freq="2">User modeling</item>
     <item freq="1">Log System</item>
  </query>
- <proposed>                                          (3)
     <item freq="3">Curriculum in Computer</item>
     <item freq="3">Distance learning</item>
```

```
      <item freq="2">Computer Communication</item>
      <item freq="2">Computer and Computer Education</item>
      <item freq="1">Computer Security</item>
      <item freq="1">Computer Integrated Manufacturing</item>
      <item freq="1">Computer and Public</item>
      <item freq="1">Computer Anxiety</item>
     <item freq="1">Data Parallel</item>
     <item freq="1">IEEE Computer Society</item>
  </proposed>
- <selected>                                                    (4)
     <item freq="3">Curriculum in Computer</item>
     <item freq="2">Distance learning</item>
     <item freq="2">Computer and Computer Education</item>
     <item freq="1">Computer and Public</item>
     <item freq="1">IEEE Computer Society</item>
 </selected>
  </user>
```

Figure 8. An example of user data: consists of both explicit and implicit data.

Figure 8 shows an example of a user model. This model consists of four highest level elements (in addition to a log of queries submitted): 1) “userInfo” and “userInterests” (not expanded) are for explicit answers to a questionnaire, 2) “community” is for the communities of the user found by the recommender, 3) “proposed” is for document topics which are shown to the user and skipped, and 4) “selected” is for document topics which are selected or expanded by the user. Therefore, (1) is explicit rating data, (2) reflects computer inference, and (3) and (4) are implicit rating data. Each entry has accompanying statistics (e.g., frequencies, probabilities).

2.4 Analyzer: User Model Data Processor

This section describes how the user tracking data is processed and merged into user model data which is stored in the user model database. The *analyzer* module, which is server-side, see at the bottom of the right rectangle of the system diagram in Figure 1, is in charge of three functions, namely 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, while four classes, Item, GroupItem, Research, and UserModel are for analyzing and representing the User Model.

class TestSAX

Description:

This class is an implementation of a SAX parser that modifies a UserModel object directly as the XML file is parsed. The SAX parser requires a UserModel parameter to the constructor – this is the UserModel that will be modified. There is no main method. The UserModel class calls the parse method directly. This implementation works by keeping track of which element it is currently reading and modifying the UserModel through calls to appropriate modification methods.

Methods:

Not listed since they are standard to a SAX Parser implementation.

class Item

Description:

An Item object consists of a name and a frequency that represents the data stored in an item tag in the XML User Model.

Methods:

1. Item() – Default constructor, sets the name to null and sets the frequency to 0.
2. Item(String) – Constructor, sets the name to the parameter and sets the frequency to 1.
3. Item(String, int) – Constructor, sets the name and the frequency by the value of the parameters.

4. int getFreq – Accessor method, returns the frequency of the Item.
5. String getName() – Accessor method, returns the name of the Item.
6. void increment() – Increments the frequency of the Item.
7. void setName(String) – Sets the name of the Item.

class GroupItem

Description:

The GroupItem object represents the data stored in the groupID tag in the user model data. A GroupItem consists of a group ID (int) and a user's probability of belonging to the group (double).

Methods:

1. GroupItem() – Default constructor, sets the group ID to 0 and the probability to 0.0.
2. GroupItem(int groupID) – Constructor, sets the group ID to the parameter and the probability to 0.0.
3. GroupItem(int groupID, double sim) – Constructor, sets the group ID and the probability by the values of the parameters.
4. int getID() – Accessor method, returns the group's ID.
5. void setID(int) – Mutator method, sets the group's ID to the value of the parameter.
6. void setSimilarity(double) – Mutator method, sets the probability to the value of the parameter.
7. double getSimilarity() – Accessor method, returns the probability of belonging to this group.

class Research

Description:

The Research class consists of research topics and expertise year for each topic. The Research class represents the data held in a research tag in the XML User Model.

Methods:

1. Research() – Default constructor, sets the topic to null
2. void addInterest (String, double) – Mutator method, adds the String parameter to the list of research interests.
3. String getInterest (int index) – Accessor method, returns the interest at the specified index in the List of interests.
4. Double getExpyear (int index) – Accessor method, returns the number of years at the specified index in the List of interests.
5. int getNumInterests() – Accessor method for the number of research interests of this user.

class UserModel

Description:

The UserModel class holds the data from the XML formatted tracking data file. For ease of integration, it contains all methods needed for access to the other classes, including the SAX parser (TestSAX) and the XML Writer class (WriteXML). The UserModel represents all the data that is stored in the XML document. The class consists of Strings to hold the user's first name, last name, user ID, password, and email and a List to hold the user's research interests (List of Research objects), groups (List of GroupItem objects), and majors (List of Strings), and the user's history of activities on the web user interface of the Digital Library, including the queries that the user has entered, the text of HTTP links that have been shown to the user, and the text of HTTP links that the user has selected. It also includes instances of the TestSAX and WriteXML classes. The UserModel class includes a total of 61 methods for analysis and later grouping of users.

Methods (Selected):

1. UserModel() – Constructor
2. void addFromCookie(String) – parses the cookie string into the query string, proposed topics, and selected topics, and adds the data to the appropriate fields.

3. void addGroup(int, double) – Mutator method, adds a group with the ID and probability passed to the method to the user's list of groups.

4. void addMajor(String) – Mutator method, adds the String parameter to the user's list of majors.

5. void addProposed(String, int) – Mutator method, adds an Item with the name and frequency passed to the method to the user's list of proposed clusters. (Only for use by the TestSAX class.)

6. void addQuery(String, int) – Mutator method, adds an Item with the name and frequency passed to the method to the user's list of queries. (Only for use by the TestSAX class.)

7. void addResearchInterest(String interest) – Mutator method, adds an interest with the name given to the Research object and adds it to the list.

8. void addSelected(String, int) – Mutator method, adds an Item with the name and frequency passed to the method to the user's list of selected topic. (Only for use by the TestSAX class.)

9. void loadProfile() – Parses the XML document and updates the data in the UserModel class based on the user's ID. This method is called by the setUserID method if the user is not marked as a newUser (see method below).

10. void writeItems(WriteXML, List) – Helper method for write(), writes a List of Items to XML using the WriteXML parameter.

Chapter 3. Preliminary Experiment: Characterizing Users with Implicit Rating Data and Verification with Explicit Rating Data

3.1 Experimental Design

Our preliminary experiment described in this chapter tests the effectiveness of implicit rating data on characterizing users. We will find virtual interest groups by using only implicit rating data and evaluate the result with the users' research interests, which are entered explicitly through an online survey. The experimental environment was created to replicate serious patrons' real use of our Digital Library. Participants of this experiment completed a general questionnaire for normal demographic information along with questions asking about their research interests. After completing the questionnaire, participants were instructed to use our JavaScript-based experimental interface to CITIDEL [2] (see Figure 2), to search for documents with the queries in their research and learning interests.

3.2 Data Description and Preprocessing

In order to collect implicit rating data, we developed a special interface for the CITIDEL system [2], part of the NSF-funded National Science Digital Library. Our interface was based on Carrot^2 [1], coupled with our user tracking system, which together support and record selection of clusters (i.e., the output of the system) [19, 20].

We collected data from 22 graduate students at both the Ph.D. and Master's level. Data sets from four graduate students were excluded as their research domains are outside the field of computer science and the CITIDEL system only contains documents in the "computing" field. Therefore, data from 18 selected graduate students in the Department of Computer Science were analyzed for this study. Table 7 describes the data used for this experiment.

Table 7. Data set: collected from 18 Ph.D. and M.S. students in computer science.

Data Type	Data Source	Quantity		Description
		Number of Tasks per Participant	Number of Records per Task	
Implicit rating data	User Tracking Interface in CITIDEL	Each participant conducted 10 searches in their specialties.	Average 28 research and learning topics	Selected topics by the user are tagged "positive" and not-selected topics are tagged "negative". Frequencies of all topics are counted.
Explicit rating data	Online User Survey during the registration	Each participant listed their research areas, with a maximum of 3 areas.	3 levels	Research area is described in 3 different levels, e.g., Computer Science>Data and Information>Digital Library.

For convenience of observation, we named each subject according to their research interests, using explicit rating data, which were obtained from the questionnaires, as shown in Table 8.

Table 8. Symbols of participants and their profiles.

	User Symbols	User profiles collected from questionnaire
1	DLmember	The one who belonged to the Digital Library Research Laboratory
2	SW_eng1	The one who has an interest in Software Engineering
3	Bio	The one who has an interest in Bioinformatics
4	VR_hci	The one who has an interest in Virtual Reality and Human Computer Interaction
5	CLIR_1	The one who has an interest in Cross Language Information Retrieval

6	CLIR_2	The one who has an interest in Cross Language Information Retrieval
7	NLP_1	The one who has an interest in Natural Language Processing
8	NLP_2	The one who has an interest in Natural Language Processing
9	VR_1	The one who has an interest in Virtual Reality
10	VR_2	The one who has an interest in Virtual Reality
11	EC_agent	The one who has an interest in E-Commerce and Agent
12	CybEdu_agt	The one who has an interest in Cyber Education and Agent
13	DLandEDU_1	The one who has an interest in Digital Library and Education
14	DLandEDU_2	The one who has an interest in Digital Library and Education
15	Personal_1	The one who has an interest in Personalization
16	Personal_2	The one who has an interest in Personalization
17	SW_eng2	The one who has an interest in Software Engineering
18	Fuzzy	The one who has an interest in Fuzzy Theory

3.3 Searching for Virtual Interest Group by Using Implicit Rating Data

User grouping was based on calculating user similarity, based on a correlation function, as in equation (1). User similarities among all subjects are shown in Figure 9. A longer column in the graph represents greater similarity. Columns are either high or very low. This means our user similarity equation (1), using only implicit rating data, is able to distinguish a user from others who have different interests.

$$correlation(a,i) = \frac{\sum_j (v_{a,j} - \overline{v_a})(v_{i,j} - \bar{v}_i)}{\sqrt{\sum_j (v_{a,j} - \bar{v}_a)^2 \sum_j (v_{i,j} - \bar{v}_i)^2}} \qquad \textbf{(1)}$$

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

(1) represents the correlation of user ‘a’ and user ‘i’. ‘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 items which are rated by users ‘a’ and ‘i’. ‘$\bar{v}_a$’ is the average probability of positive rating of the user which is obtained by (2) [17]. In these calculations, we treat user interests as atomic terms that are strings. Thus, the strings “digital library” and

"digital camera" are not considered similar even though they share the common term "digital".

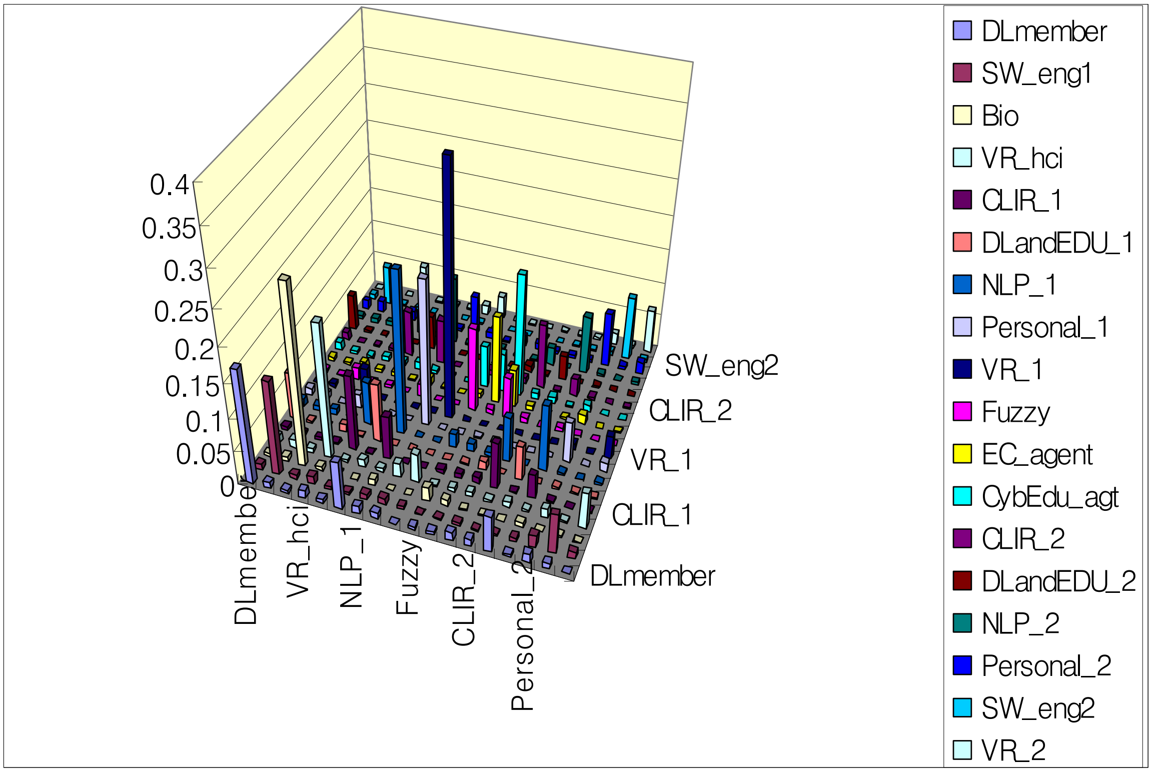


Figure 9. User similarities among the users.

After calculating similarities among all participants, we allocated their IDs according to their similarities to others. Table 9 shows the result of allocation after less similar participants were truncated. The high similarity range, above 0.075, was divided into four levels. That is, the range between max similarity and 0.075 is divided into 4 levels. Then, similar participants were allocated according to their similarity with the participant of the row. Because we named participants with their research interests, entered explicitly, it is easy to see that participants with similar research interests have greater similarity than other users. For example, DLmember was more similar to DLandEDU_1 and DLandEDU_2 than others. Bio does not have a similar participant.

Table 9. Similarities levels, sorted & grouped for each user.

User ID	Level 1	Level 2	Level 3	Level 4
DLmember			DLandEDU_1, DLandEDU_2	
SW_eng1			SW_eng2	Personal_2
Bio				
VR_hci			VR_2, VR_1	Personal_1, Personal_2
CLIR_1	NLP_1, CLIR_2		NLP_2	
CLIR_2	CLIR_1, NLP_1		NLP_2	
NLP_1	NLP_2		CLIR_1, CLIR_2	
NLP_2	NLP_1		CLIR_1, CLIR_2	
VR_1				VR_2, VR_hci
VR_2		VR_hci	VR_1	Personal_1, Personal_2
EC_agent		CybEdu_agt		Personal_2
CybEdu_agt		EC_agent	Fuzzy	
DLandEDU_1	DLmember	DLandEDU_2		VR_hci, CybEdu_agt
DLandEDU_2	DLmember	DLandEDU_1		CybEdu_agt, VR_hci
Personal_1			Personal_2	
Personal_2			Personal_1	
SW_eng2		SW_eng1		
Fuzzy			CybEdu_agt	Bio, NLP_1

Deciding upon virtual interest groups is achieved by merging the participants of the row and the participants in the level with the closest similarity. This primitive grouping algorithm is simple and fast but has a problem with grouping not-similar users into a group. We propose an improved grouping algorithm, "fixed-size window multi-classification" (FSWMC), a modified *k*NN algorithm, in Section 7.2.

In this experiment, eight virtual interest groups were found, as Table 10 shows. These groups are found by merging a user and other members with the closest similarity level from Table 9. Three participants have research interests in Digital Library, two participants in Software Engineering, three participants in Virtual Reality and Human Computer Interaction, three participants in Natural Language Processing and Cross Language Information Retrieval, two participants in Natural Languages, two participants in Personalization, and three participants in fuzzy theory and agents for E-Commerce and Cyber-Education. A participant interested in Bio is grouped alone because no participant was close to her.

Table 10. Result of interest group finding.

User Group ID	Members
A	DLmember, DLandEDU_1, DLandEDU_2
B	SW_eng1, SW_eng2,
C	VR_hci, VR_1, VR_2
D	CLIR_1, NLP_1, CLIR_2
E	NLP_1, NLP_2
F	Personal_1, Personal_2
G	EC_agent, CybEdu_agt, Fuzzy
H	Bio

3.4 Conclusion of Experiment

In this preliminary experiment, we demonstrated how we could use the implicit rating data in characterizing and finding virtual interest groups in a Digital Library to show that implicit rating data by itself, without mixing in explicit rating data, is useful information for characterizing users. We performed user clustering according to their research interests by using implicit rating data, which is user tracking data. The evaluation of user clustering is performed by comparing explicitly entered research interests among users in each user cluster. Table 8 and Table 10 show our user clustering was effective.

This result is meaningful to help us in moving forward beyond the previous general belief, namely that implicit rating data is just auxiliary information for supporting explicit rating data, because collecting explicit rating data in a complex information system like a Digital Library is very expensive and difficult. In the next chapter, we will continue this study with statistical methods through hypothesis testing.

Also, in this experiment, we found there is no known method to evaluate the correctness of user clustering based on their shared interests. A study to develop the method will be useful.

Chapter 4. Hypotheses Testing: Effectiveness of Implicit Rating Data in Characterizing Users and User Communities

In the previous chapter, we described an experiment to show that implicit rating data is effective in characterizing users and user communities. Explicit rating data is used to validate that user communities found in the experiment were correct. However, using explicit rating data for evaluation still has problems that caused by terminological issues as we mentioned in Section 1.1. In this chapter, to avoid this, we make two hypothesis tests to validate our hypotheses objectively.

4.1 General Principles of Hypothesis Testing

Hypothesis tests are procedures for making rational decisions about the reality of effects. There are two cases when we say an *effect* is present. The first case is when a change in one thing is associated with a change in another. For example, if changes of salt intake are related to the chance of heart failure, we say an effect exists. Another case is when a difference in distribution exists between two aspects. For example, if the distribution of political party preference (Republicans, Democrats, or Independents) differs for sex, then an effect is present for the parameter 'sex' [32]. Once a parameter is proven to have an effect, it is possible and meaningful to analyze the data according to the parameter.

All hypothesis tests conform to similar principles and procedures as listed below [32].

Step 1: A model of the world is created in which there are no effects. That is, a *null hypothesis* is generated.

Step 2: The experiment is then repeated an infinite number of times.

Step 3: If the results of the experiment are unlikely in the model generated in step one, then the model is rejected and we accept the effects as real. If the results of the experiment could be explained by the model, retain the model in step one and no decision can be made about the reality of effects.

4.2 Hypotheses

For the hypotheses tests in this chapter, we use the same data we used in the preliminary experiment presented in Chapter 3. Since our user tracking system will collect the names of document clusters, as implicit rating data, while the users browse the search result pages, we test three hypotheses about proper human-computer interaction and document clustering. As implicit rating data in DLs is generated by users who have their goals in mind, such as finding some documents or books, and use DLs to achieve their goals, we call these users *serious users*; our hypotheses assume all the participants were serious users.

In our two hypotheses tests in this chapter, our goals are to 1) show that the effect of implicit rating data of DL users is real by showing that the distribution of implicit rating data is different from that of *non-implicit rating data*, which is "un-rated" user tracking data, and to 2) show that a difference in distribution exists between two implicit rating data sets, one from an interest-sharing user group and the other from an interest-exclusive user group.

Three hypotheses are:

1. H_1: For any serious user with their own research interests and topics, show consistent output for the document collections referred to by the user.
2. H_2: For serious users who share common research interests and topics, show overlapped output for the document collections referred to by them.
3. H_3: For serious users who don't share any research interests and topics, show different output for the document collections referred to by them.

The first goal is represented by $H_{1,}$ which means that unlike the *non-implicit rating data* collected from a non-serious user, the deviation of implicit rating data occurrence tends to converge to some value as the user uses the Digital Library an infinite number of times. The second goal is represented by H_2 and $H_{3,}$ which mean the deviation of overlapped implicit rating data occurrence within the interest-sharing user group tends to converge to some value while that of the interest-exclusive user group will diverge as the users use the Digital Library an infinite number of time. Because H_3 is the contrapositive H_2, we won't perform a separate experiment for H_3.

4.3 Data Set and Hypotheses Test Procedures

For this hypotheses test, we used the data set collected during the experiments presented in Chapter 3. Table 7 presents detailed information about the data set.

We performed hypothesis testing [29] as follows. Because the data collected from the user tracking system is independent and identically distributed (i.i.d.), we use inference processes to verify hypotheses and estimate properties, starting with HT1.

HT1: Hypothesis testing and confidence intervals for H_1.

1. H_0 (Null hypothesis of H_1): Mean values (μ) of the frequency of document topics proposed by the Document Clustering Algorithm are NOT consistent ($\mu_0 = 1$) for a user.

   Hypothesis Testing about $\underline{H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu > \mu_0}$

2. Conditions: 95% confidence (test size = 0.05), sample size 'n' < 25, unknown standard deviation 'σ', i.i.d. random sample from normal distribution, → estimated z-score t-test.

3. Test statistics: sample mean '$\bar{y}$' = 1.1429, sample standard deviation 's' = 0.2277 are observed from the experiment.

4. Rejection Rule is to reject H_0 if $\bar{y} > \mu_0 + z_{\alpha/2}\, \sigma/\sqrt{n}$

5. From the experiment, $\bar{y} = 1.1429 > \mu_0 + z_{\alpha/2}\, \sigma/\sqrt{n} = 1.0934$

6. Therefore decision is to Reject H_0 and accept H_1, 95% Confidence Interval for μ is $\underline{1.0297 \le \mu \le 1.2561}$, and P-value = 0.0039

HT2: Hypothesis testing and confidence intervals for H_2.

1. H_0 (Null hypothesis of H_2): A user's average ratio of overlapped topics with other persons in her groups over her total topics which have been referred, μ_1, is the same as the average ratio of overlapped topics with other persons out of her groups over her total topics which have been referred, $\mu_{2.}$

Hypothesis Testing about $\underline{H_0 : \mu_1 = \mu_2 \text{ vs. } H_2 : \mu_1 > \mu_2}$

Because a user can belong to multiple groups, population means μ_1 and μ_2 are calculated as in the formulas below, (3) and (4), respectively,

$$\mu_1 = \frac{\sum_{k=1}^{G} \sum_{i=1}^{n_K} \sum_{j=1, j \neq i}^{n_K} O_{i,j}}{\sum_{k=1}^{G} n_K (n_K - 1)} \qquad (3)$$

$$\mu_2 = \frac{\sum_{k=1}^{G} \sum_{i=1}^{n_K} \sum_{j=1, j \notin K}^{N} O_{i,j}}{\sum_{k=1}^{G} n_K (N - n_K)} \qquad (4)$$

where $O_{i,j}$ is user i's topic ratio overlapped with user j's topics over i's total topics, G is the total number of user groups in the system, n_K is the total number of users in group K, and N is the total number of users in the system. One instance of random variables in this testing, one user's overlapped topic ratio with other persons in her group, in-group overlapping ratio, and overlapped topic ratio with other persons out of her group, out-group overlapping ratio, is illustrated in Figure 10. In this figure, all overlapping ratios are directed. $\overrightarrow{ab}$ means the overlapping ratio from user 'a' to user 'b'. Because the ratio is the number of topics overlapped over the total number of topics in her user model, $\overrightarrow{ab} \neq \overrightarrow{ba}$. In this case, the in-group overlapping ratio of user 'a' is the average of $\overrightarrow{ab}$, $\overrightarrow{ac}$, and $\overrightarrow{ad}$, and the out-group overlapping ratio is the average of $\overrightarrow{ae}$ and $\overrightarrow{af}$.

2. Conditions: 95% confidence (test size $\alpha = 0.05$), two i.i.d. random samples from normal distribution, for two sample sizes n_1 and $n_{2,}$ $n_1 = n_2 < 25$, standard deviations of each sample σ_1 and σ_2 are unknown → two-sample Welch t-test.

3. Test statistics: Welch score ‘w0’ = $(\overline{y_1} - \overline{y_2}) / \sqrt{\frac{s_1^{\ 2}}{n_1} + \frac{s_2^{\ 2}}{n_2}}$, where $\bar{y}_1$, $\bar{y}_2$ are the sample means of each sample and s_1, s_2 are the sample standard deviations of each sample.

4. Rejection Rule is to reject H_0 if the $w_0 > t_{df_s,\alpha}$ where *t* refers to the t-cutoff of the t-distribution table, and df_s is the Satterthwaite’s degree of freedom approximation [44] which is calculated by

$$df_S = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}} \qquad (5)$$

5. From the experiment, $\bar{y}_1 = 0.103$, $\bar{y}_2 = 0.0215$, $df_s = 16.2$ and $w_0 = 4.64 > t_{16.2,\ 0.05} = 1.745$

6. Therefore the decision is to Reject H_0 and accept H_2. 95% Confidence Intervals for μ_1, μ_2 and $\mu_1 - \mu_2$ are $0.0659 \le \mu_1 \le 0.1402$, $0.0183 \le \mu_2 \le 0.0247$ and $0.0468 \le \mu_1 - \mu_2 \le 0.1163$, respectively, and P-value = 0.0003

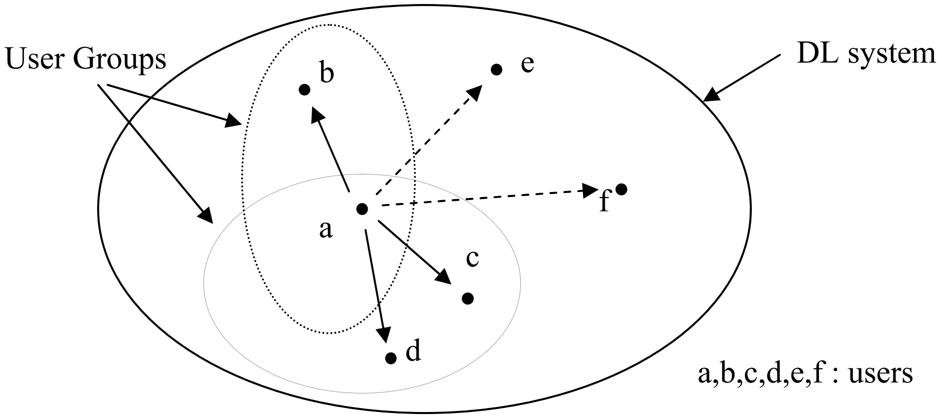


Figure 10. User *a*’s in-group overlapping and out-group overlapping.

4.4 Conclusions of Hypotheses Testing

We performed two hypotheses tests to prove the effectiveness of implicit rating data in characterizing users. Our user tracking system, the source of implicit rating data, collects information about document clusters while a user browses the result set in CITIDEL. Thus, we tested how the results of document clustering reflect users' characteristics, such as research interests, learning topics, and preferences while users use DLs in a serious manner. The test results support the claim that implicit ratings are meaningful information for studies on user analysis, personalization, collaborative filtering, and recommending. These results are more meaningful in complex information systems, like DLs, because such systems have dynamic contents and sparse rating data, and thus implicit rating data is more feasible to collect than explicit rating data.

This experiment was a closed experiment in a designed environment. This was because the cost of experimenting in an open environment with an unlimited number of participants was so high. In our future work, we plan to conduct this experiment again as an open experiment in a real DL environment, NDLTD, with more data collected from thousands of users.

Chapter 5. Effectiveness of Four Different Data Types in Community Finding

5.1 Purpose

Studies on the effect of different types of data on the performance of user cluster mining have highlighted a basic problem caused by the variety of academic terms, as we mentioned in Section 1.1. However, we can explore user cluster mining more objectively, because we can obtain user groups without depending on user's subjective answers to questionnaires about their research interests or preferences. We conducted an ANOVA test to compare the effectiveness of four different user rating data types on the performance of user cluster mining by using implicit rating data and user groups collected from experiments [20].

5.2 Data Set

Table 11. Data set: Four types of implicit rating data collected from CITIDEL.

Quantity			**Description**
Number of Participants	**Number of Tasks per Participant**	**Number of Records**	
18 Ph.D. and MS. Students in Computer Science Major	Each participant conducted 10 searches in their specialties and browsed the results to find documents.	Average = 28 research and learning topics for each search. Therefore, each participant provided an average of 280 topics.	Topics are tagged either "positive" if the user had browsed it or "negative" if the user hadn't browsed it. At the same time, each topic is considered either as a set of words or as an atomic term.

Table 11 describes the data set we used for this experiment. Collected topics are tagged either "positive" or "negative" by the user tracking system. Also, because the topics are in the form of a noun phrase, each topic could be considered either as a set of words or as one atomic term. Therefore, we have four different data types to compare regarding their effectiveness in characterizing users:

1. Selected Topics: Set of noun phrases that are displayed on the screen and selected for browsing by the user
2. Proposed Topics: Set of noun phrases that are displayed on the screen but not selected by the user
3. Selected Terms: Set of words that are displayed on the screen as a part of "Topics" and selected for browsing by the user
4. Proposed Terms: Set of words that are displayed on the screen as a part of "Topics" but not selected by the user

5.3 ANOVA Test

Figure 11 shows the result; ANOVA statistics F(3, 64) = 4.86, p-value = 0.0042, and the least significant difference (LSD) = 1.7531. The object of this ANOVA test is to see how the performance of user cluster mining is affected by four different data types, such as selected topics, proposed topics, selected terms, and proposed terms. Topics mean noun phrases generated by LINGO (according to the *Collins English Dictionary*, lingo is "a range of words or a style of language which is used in a particular situation or by a particular group of people") [27, 28]. Terms indicate single nouns contained in the original documents, queries, and topics. Although we gained a relatively large LSD because of the small number of participants, we still found statistical significance in this test. This figure also shows that the test using proposed terms performs significantly worse. Except for the test using the proposed terms, the other three tests that use selected topics, proposed topics, and selected terms don't show statistically significant differences from each other, even though the test using proposed document topics shows slightly higher performance. We believe that this is because using proposed terms causes too sensitive overlapping both in the in-group testing and out-group testing (to distinguish proper relations between users). This leads us to conclude that term-frequency based approaches, to user cluster mining are not as efficient as document-topic based approaches using user rating and document clustering, because using proposed terms performed poorest in this test.

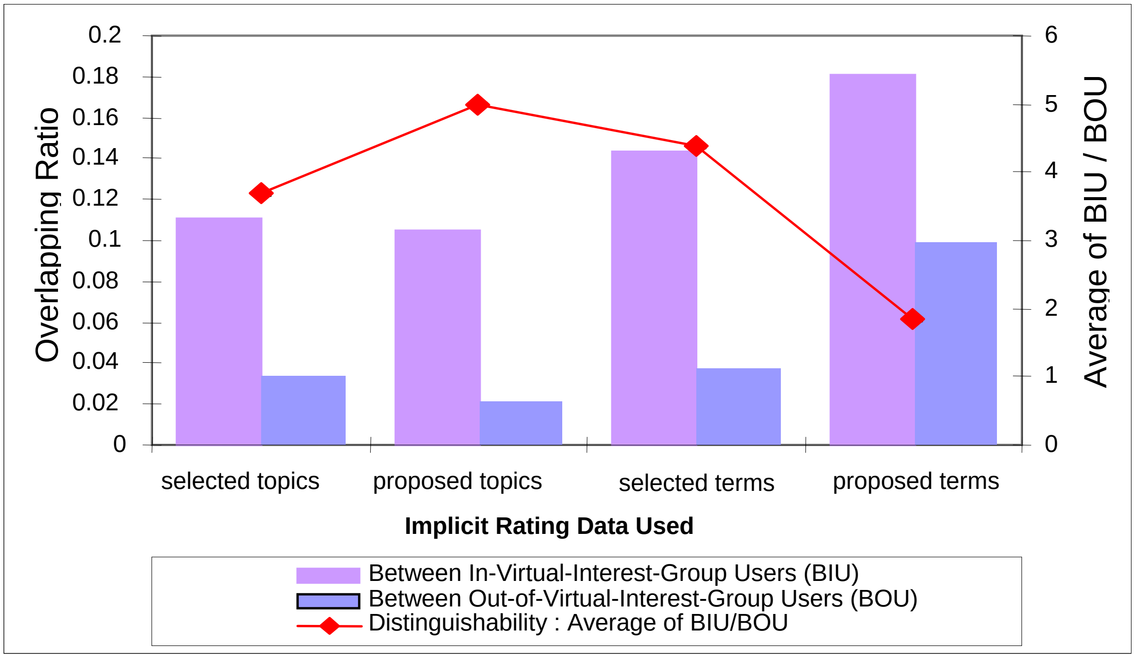


Figure 11. Effects of four different implicit rating data type used.

5.4 Conclusions

We tested the effect of different types of implicit rating data on the performance of user community mining, and found that using proposed terms performed worst. We explain this be because of the sensitive overlapping ratio of appearance on the screens among participants. Using the proposed topics showed higher performance than using selected topics, or selected terms in our experiment. However, the difference was not statistically significant. Because of small data, we only can say that using proposed terms performed significantly poorer than did the other three data types. For further study, in order to find most effective data type, we will perform this experiment again with large amounts of real data.

Chapter 6. Supply / Demand Analysis in NDLTD: Using Implicit Rating Data

Analyzing implicit rating data provides particular information, which is hard to obtain from analyzing explicit rating data. This experiment demonstrates how we could utilize implicit rating data to analyze NDLTD by measuring the amount of information supply and demand in NDLTD. The goal of this experiment is to reveal how well the Electronic Theses and Dissertations (ETDs) in NDLTD match with the information demands of users in each scholarly field.

6.1 Data Set and Preprocessing

For this experiment, we employ a user interface embedded user tracking system to collect 1,100 users' implicit rating data, using query logs and analyzing browsing activities. Table 12 describes the type, source, and quantity of data used for this experiment.

Table 12. Data set used for supply / demand analysis of NDLTD.

Type	**Source**		**Number of Record**	**Description**
Supply Analysis	Electronic Thesis and Dissertation (ETD)		242,688 ETDs	Harvested from union catalog at Online Computer Library Center (OCLC) using "OAI/ODL Harvester" [33]. Contains ETDs until Fall 2005 and part of Spring 2006 graduation.
Demand Analysis	Explicit Data	User Survey	From 1,100 users	Online user survey conducted from August 2005 to April 2006 as part of User Modeling Study [17]. Contains demographic information, major, research fields, and expertise years in the fields for each user.
	Implicit	Query Log		Collected by User Tracking System [16]

	Data	Browsing Activities		as part of User Modeling Study. Consists of queries and their frequencies for each user.

6.2 Classification of ETDs and Users

The goal of this study is to figure out how well the ETDs in NDLTD match with the information demands of users in each scholarly field. Our approach is based on classifying both the ETDs and user data into the same scholarly classes with the same criteria to see their distributions. Then we compare these two distributions with each other. Classification of ETD and user data was done by examining “key fields”, which are fields used for classification, such as “subject” fields in ETD metadata, and the “major”, “broadresearch”, and “specific” fields in user data. We built a common matching table that consisted of identification string patterns for 77 subcategories. After that, we grouped the subcategories into 7 higher level categories as shown in Table 13. These categories were created based on the faculty/college systems of five universities in Virginia.

Table 13. Seven categories and 77 subcategories.

	7 categories	77 subcategories
1	Architecture and Design	ArchitectureConstruction, LandscapeArchitecture
2	Law	Law
3	Medicine, Nursing and Veterinary Medicine	Dentistry, Medicine, Nursing, Pharmacy, Veterinary
4	Arts and Science	Agriculture, AnimalPoultry, Anthropology, ApparelHousing, Archaeology, Art, Astronomy, Biochemistry, Biology, Botany, Chemistry, Communication, CropSoilEnvSciences, DairyScience, Ecology, EngineeringScience, English, Entomology, Family, Food, ForeignLanguageLiterature, Forestry, Geography, Geology, GovernmentInternationalAffair, History, Horticulture, HospitalityTourism, HumanDevelopment,

		HumanNutritionExercise, Informatics, Interdisciplinary, LibraryScience, Linguistics, Literature, Meteorology, Mathematics, Music Naval, Philosophy, Physics, Plant, Politics, Psychology, PublicAdministrationPolicy, PublicAffair, Sociology, Statistics, UrbanPlanning, Wildlife, Wood, Zoology
5	Engineering and Applied Science	Aerospace, BiologicalEnginerring, Chemical, ComputerScience, Electronics, Environment, Industrial, Materials, Mechanics, MiningMineral, Nuclear, OceanEngineering
6	Business and Commerce	AccountingFinance, Business, Economics, Management
7	Education	Education
8	Others (unclassifiable)	(Unclassifiable)

6.3 Measurement of Supply and Demand

The supply of NDLTD is measured from the contents of NDLTD, especially ETD metadata, harvested from the union catalog run by the Online Computer Library Center (OCLC) [6] using an OAI/ODL Harvester [33]. Supply of a certain category is measured by classifying each ETD into one of 77 subcategories based on the key field, “subject”, and then results are counted.

Regarding measuring demand, we assume that the amount of information demand is proportional to the number of queries sent for search, plus the user browsing activities within the search result set. That is, the more queries users send for search and then examine the returned results, the more the information has been demanded. Further, from our assumption about measuring demand, we also assumed the measured demand for a certain category is proportional to the sum of query numbers and the amount of browsing activities of all users in the category as represented by Equation 6.

$$Demand\ of\ a\ Category \propto \sum_{user\ \in\ category} number\ of\ queries + number\ of\ browse\ activities \qquad (6)$$

6.4 Analysis of Summary Statistics

Supply-Demand comparison in each category tells how well the supplies of ETDs in NDLTD are matching with the demands of users in each category. Figures 12 and 13 show the results of comparisons in each of the 77 subcategories. These two figures show that the supplies in “Business” and “Economics” are relatively insufficient relative to that for other fields. Several engineering areas, such as “Computer Science” and “Electronics”, are also in the same situation. Figure 14 is a summary of the results shown in Figures 12 and 13, based on merging subcategories into 7 higher level categories.

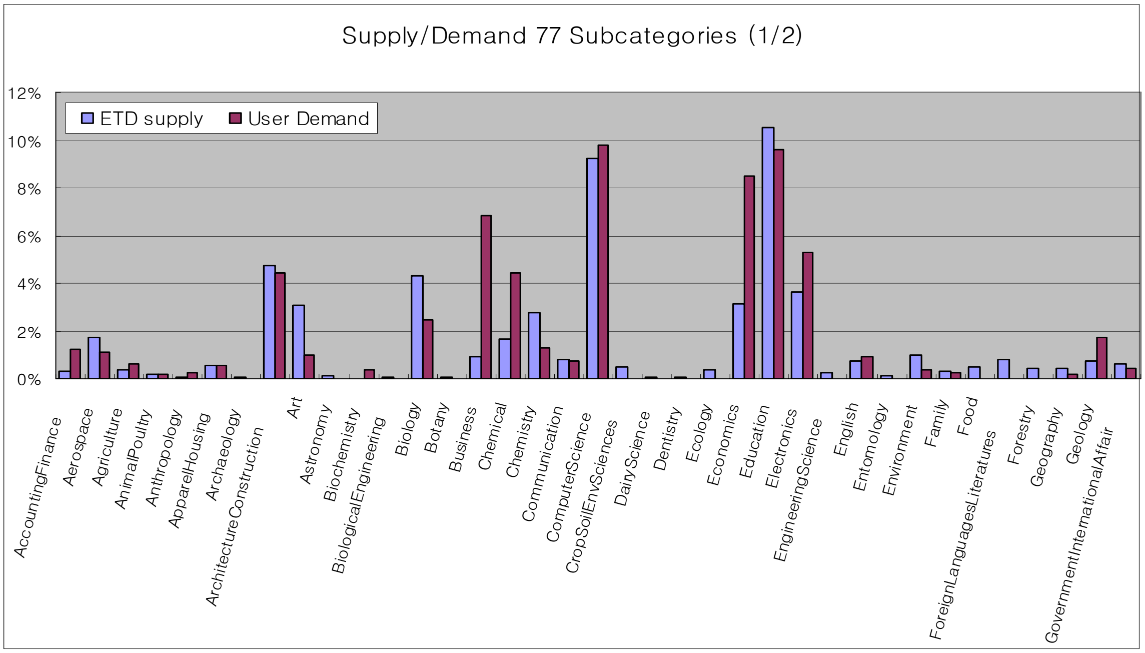


Figure 12. Supply-demand comparison in 77 subcategories (part 1 of 2).

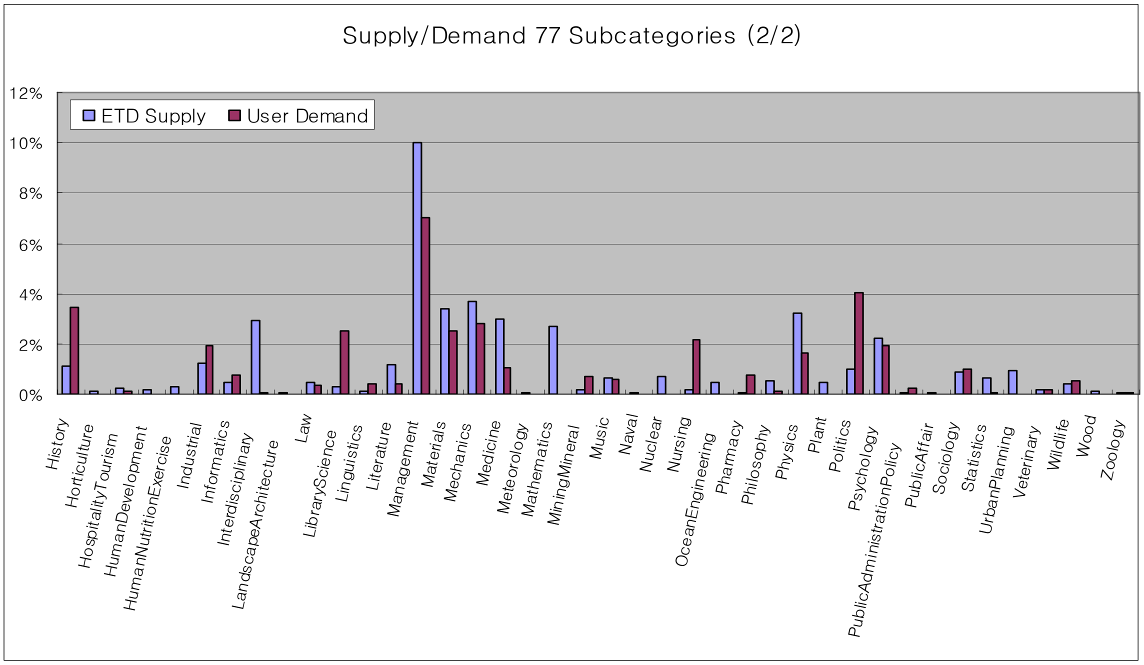


Figure 13. Supply-demand comparison in 77 subcategories (part 2 of 2).

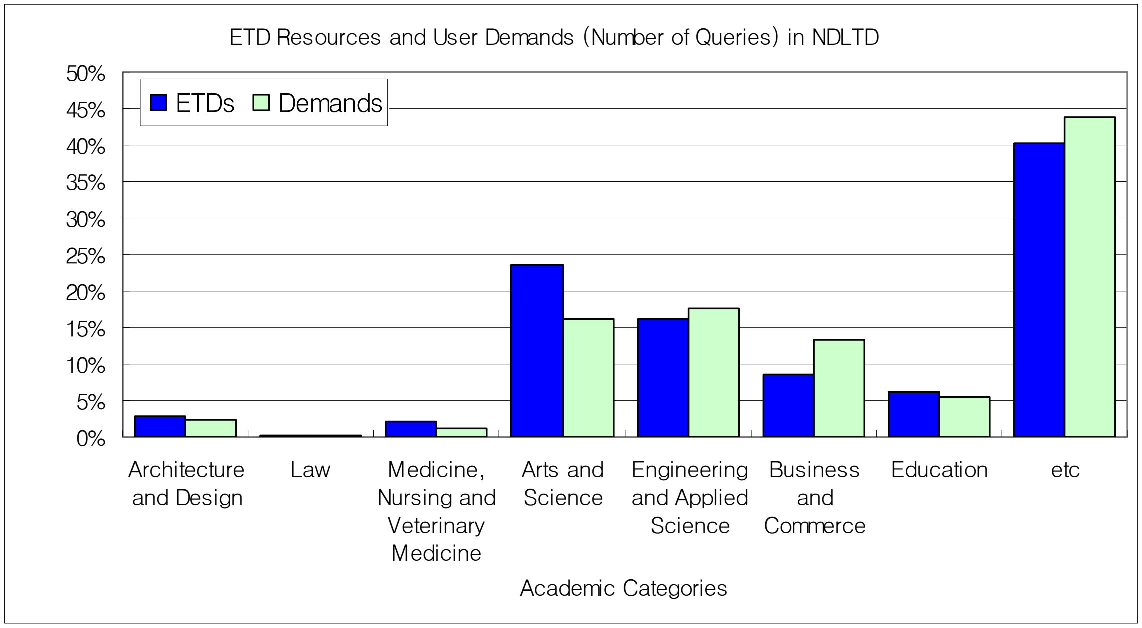


Figure 14. Supply-demand comparison in 7 categories.

From these charts, we can tell that NDLTD is supplying enough ETDs in “Architecture and Design”, “Medicine, Nursing and Veterinary Medicine”, “Arts and Science” and

“Education”, fields, while it may not be in “Engineering and Applied Science” and “Business and Commerce”.

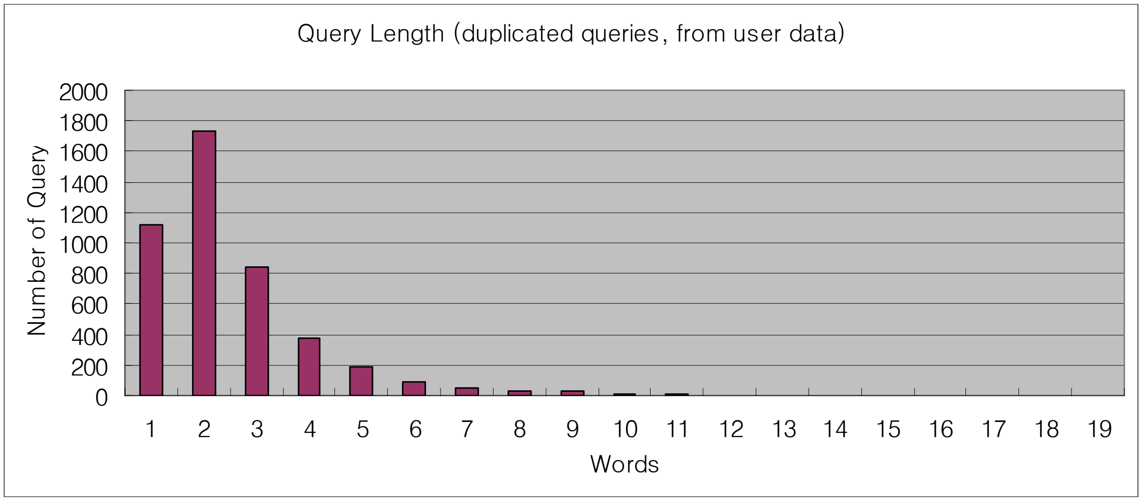


Figure 15. Distribution of query length.

Analyzing query length also provides good information for information system research. We can collect this information from user data, as is shown in Figure 15. According to this figure, most common length of queries was two words, and 81.9% of total queries were shorter or equal to three words.

6.5 Conclusions of Experiment

We analyzed ETDs and users in NDLTD to understand how well NDLTD supplies ETDs for users in each scholarly area. We measured the supply-demand by classifying ETDs and user data into the same 7 high-level categories, and also into 77 lower-level subcategories. We measured information demand by analyzing implicit rating data. This showed implicit rating data’s potential (in providing important information which is not obtainable from explicit rating data).

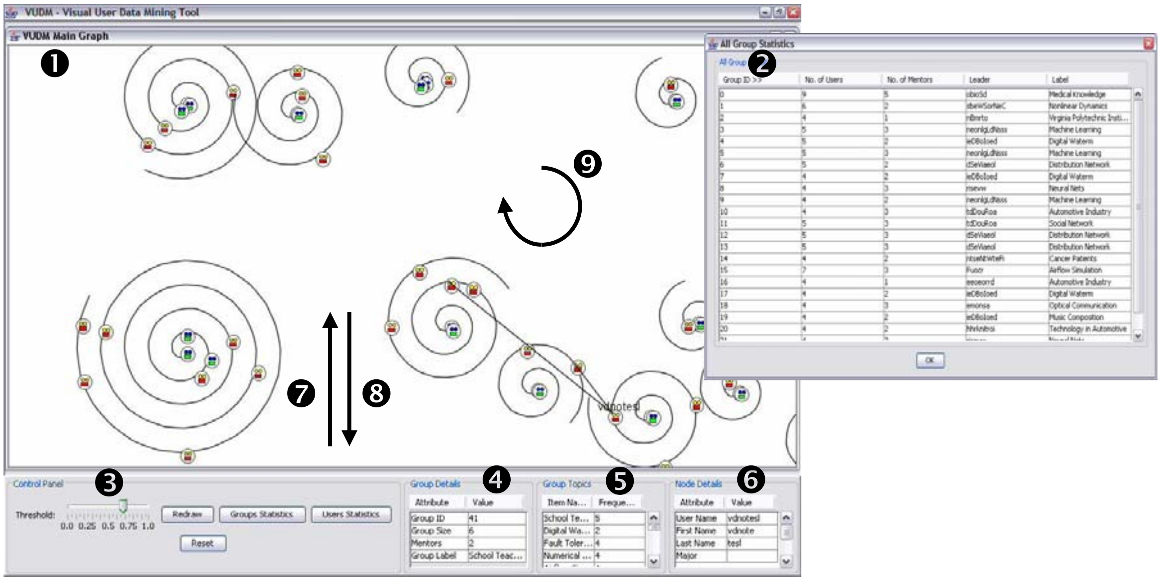
Chapter 7. VUDM: A Visual Data Mining Tool Utilizing Implicit Rating Data

7.1 Visualization Strategies

The visualization strategies of our Visual User model Data Mining (VUDM) tool fully follow the Ben Shneiderman information visualization mantra [8], "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 16. In this figure, 1, displays an overview of users, virtual interest groups, and their relationships. The statistics window, 2, presents detailed information, either about all users or about all groups in the system. 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 7.3). 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 [12] 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.



Group ID >>	No. of Users	No. of Mentors	Leader	Label
0	9	5	obioSd	Medical Knowledge
1	6	2	sbeWSorNeC	Nonlinear Dynamics
2	4	1	nBmrto	Virginia Polytechnic Insti...
3	5	3	neonlgLdNass	Machine Learning
4	5	2	ieOBoloed	Digital Waterm
5	5	3	neonlgLdNass	Machine Learning
6	5	2	dSeVlaeol	Distribution Network
7	4	2	ieOBoloed	Digital Waterm
8	4	3	rsevw	Neural Nets
9	4	2	neonlgLdNass	Machine Learning
10	4	3	tdDouRoa	Automotive Industry
11	5	3	tdDouRoa	Social Network
12	5	3	dSeVlaeol	Distribution Network
13	5	3	dSeVlaeol	Distribution Network
14	4	2	ntseMtWteR	Cancer Patients
15	7	3	Fusrr	Airflow Simulation
16	4	1	eeoeorrd	Automotive Industry
17	4	2	ieOBoloed	Digital Waterm
18	4	3	emonsa	Optical Communication
19	4	2	ieOBoloed	Music Composition
20	4	2	Mrknlrsi	Technology in Automotive

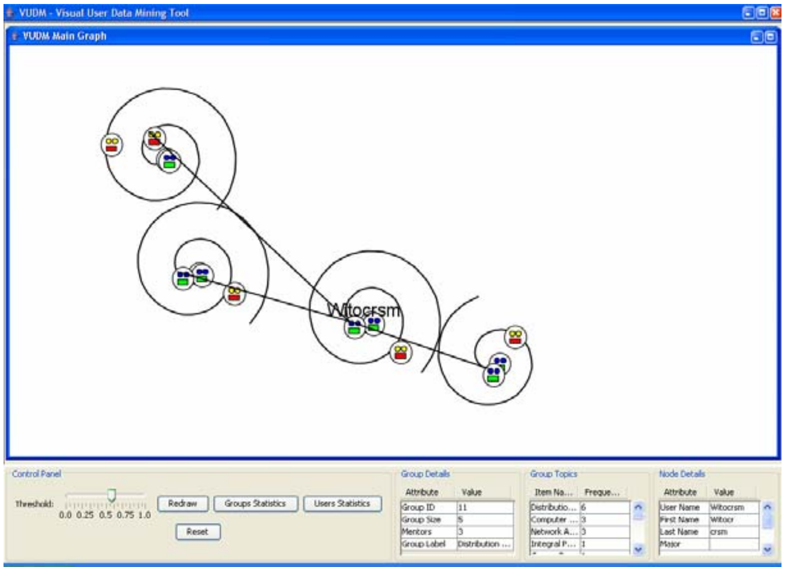
Attribute	Value
Group ID	41
Group Size	6
Mentors	2
Group Label	School Teac...

Item Na...	Freque...
School Te...	5
Digital Wa...	2
Fault Toler...	4
Numerical ...	4

Attribute	Value
User Name	vdnotesl
First Name	vdnote
Last Name	tesl
Major	

Figure 16. The main window of VUMD shows an overview of virtual user communities.

Figure 17 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 the desired area to make it easy to distinguish, locate, and select users and user communities. In zoomed user space, it is easier to locate and select a user or a community spiral. 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.



Attribute	Value
Group ID	11
Group Size	5
Mentors	3
Group Label	Distribution ...

Item Na...	Freque...
Distributio...	6
Computer ...	3
Network A...	3
Integral P...	1

Attribute	Value
User Name	Witocrsm
First Name	Witocr
Last Name	crsm
Major	

Figure 17. Zoomed user space.

VUDM also provides a filtering feature. Filtering is achieved by moving a sliding bar which is associated with the user correlation threshold θ, which will be explained later in chapter, on the user interface. With higher θ, stricter community finding is performed, so that less probable user communities are filtered out, therefore, finding user communities becomes strict too. Figure 18 shows how the correlation threshold θ influences the finding of 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 the high correlation threshold. In addition, all user icons and group spirals can be dragged with the mouse, e.g., to examine a congested area.

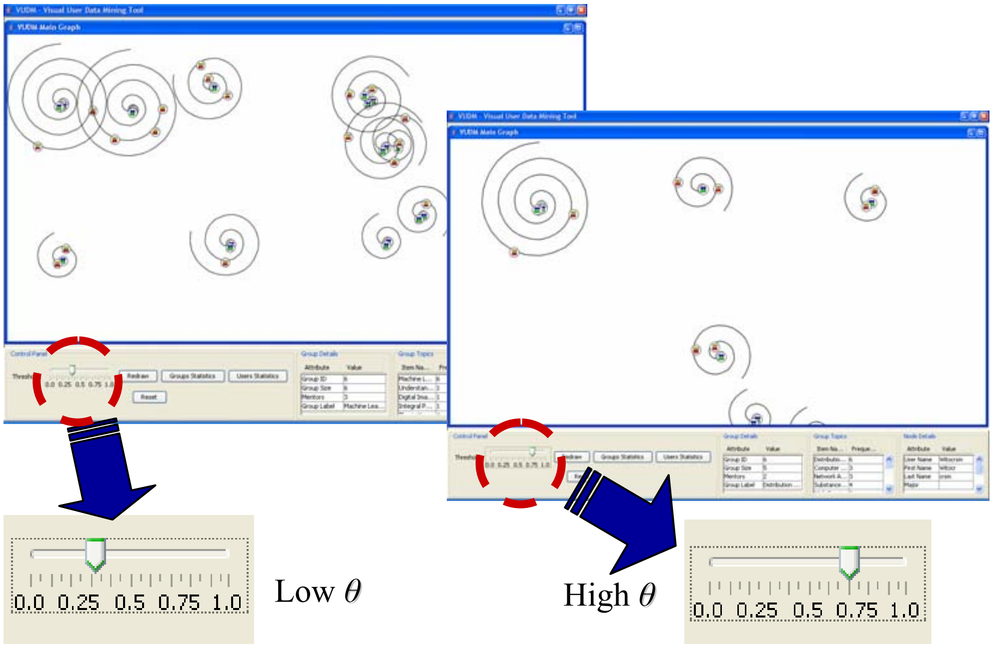
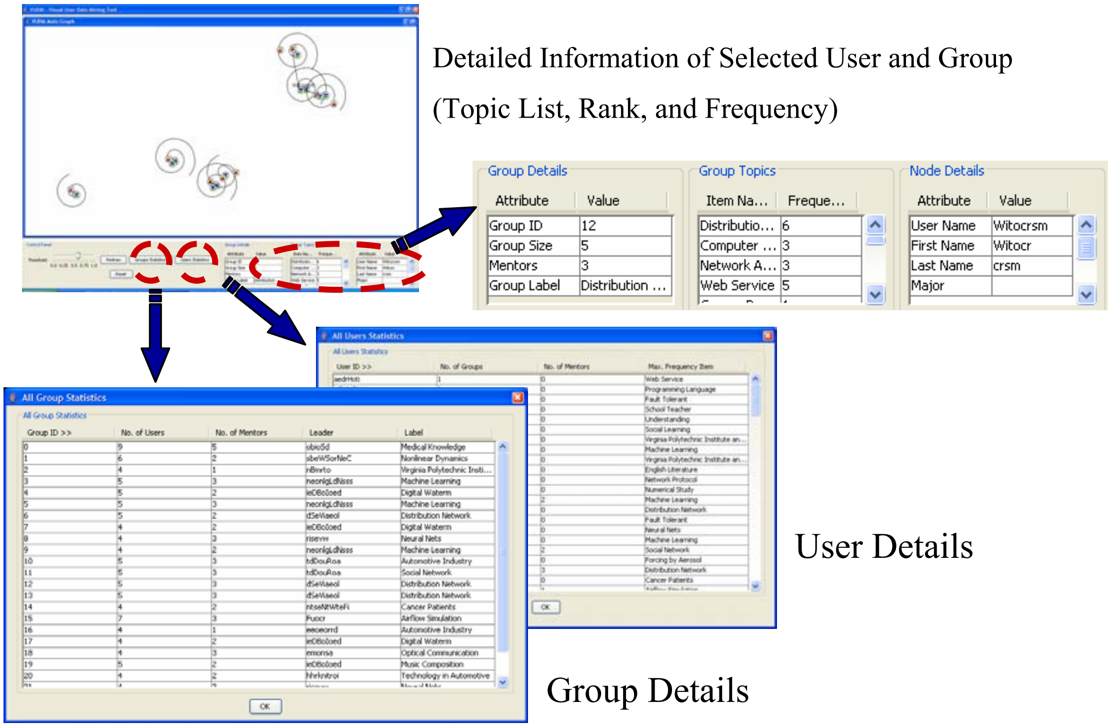


Figure 18. Filtering user space.

VUDM provides detailed information about users and user communities on demand – see the two windows at the bottom of Figure 19. 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 her research interests are listed. In the window for group details, the group's ID, number of users it contains, and a list of topics of the group is 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.


Detailed Information of Selected User and Group
(Topic List, Rank, and Frequency)
User Details
Group Details
All Users Statistics
All Group Statistics
OK

Group Details

Attribute	Value
Group ID	12
Group Size	5
Mentors	3
Group Label	Distribution ...

Group Topics

Item Na...	Freque...
Distributio...	6
Computer ...	3
Network A...	3
Web Service	5

Node Details

Attribute	Value
User Name	Witocrsm
First Name	Witocr
Last Name	crsm
Major	

All Users Statistics

User ID >>	No. of Groups	No. of Mentors	Max. Frequency Item
aedrHotI	1	0	Web Service
		0	Programming Language
		0	Fault Tolerant
		0	School Teacher
		0	Understanding
		0	Social Learning
		0	Virginia Polytechnic Institute an...
		0	Machine Learning
		0	Virginia Polytechnic Institute an...
		0	English Literature
		0	Network Protocol
		0	Numerical Study
		2	Machine Learning
		0	Distribution Network
		0	Fault Tolerant
		0	Neural Nets
		0	Machine Learning
		2	Social Network
		0	Forcing by Aerosol
		3	Distribution Network
		0	Cancer Patients

All Group Statistics

Group ID >>	No. of Users	No. of Mentors	Leader	Label
0	9	5	obioSd	Medical Knowledge
1	6	2	sbeWSorNeC	Nonlinear Dynamics
2	4	1	nBmrto	Virginia Polytechnic Insti...
3	5	3	neonlgLdNsss	Machine Learning
4	5	2	ieDBoIoed	Digital Watern
5	5	3	neonlgLdNsss	Machine Learning
6	5	2	dSeVlaeol	Distribution Network
7	4	2	ieDBoIoed	Digital Watern
8	4	3	riservr	Neural Nets
9	4	2	neonlgLdNsss	Machine Learning
10	5	3	tdDouRoa	Automotive Industry
11	5	3	tdDouRoa	Social Network
12	5	3	dSeVlaeol	Distribution Network
13	5	3	dSeVlaeol	Distribution Network
14	4	2	ntseNtWteFi	Cancer Patients
15	7	3	Fuocr	Airflow Simulation
16	4	1	eeoeorrd	Automotive Industry
17	4	2	ieDBoIoed	Digital Watern
18	4	3	emonsa	Optical Communication
19	5	2	ieDBoIoed	Music Composition
20	4	2	Nhrlnitroi	Technology in Automotive

Figure 19. Detailed information on demand.

7.2 Loose Grouping Algorithm

For classifying users into virtual interest groups and finding "groups of similar groups", we use the same, FSWMC, 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), 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 [10]. For these reasons we devised a modified *k*NN algorithm: "fixed-size window multi-classification" (FSWMC) algorithm. Figure 20 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. 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*=*θ*, circle, surrounding the test point, into a new class. If this new class is a sub- or super-class of an already found class, remove the redundant sub-class. In this figure, first stage shows a new group [a, c] is found by grouping users around ‘a’. Second stage shows a new group [b] is found by grouping users around ‘b’. Third stage shows a new group [a, c, d] is found by grouping users around ‘c’. And now, this group is super-class of a previously identified group [a, c], thus, discard the sub-class group [a, c]. In the next stage, a new group [a, c, d, e] is found by grouping users around ‘d’, and discard a previously identified group [a, c, d] because this new group is super-class of the group. Therefore, two classes, [b] and [a,c,d,e] are found up to stage n=16.

Distances between samples (the spots in the hyperspace) are calculated using Formula 7 in Section 7.3.1. 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 posterior” probabilities, P(class|feature), while the Parzen Windows estimates the density function p(feature|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”.

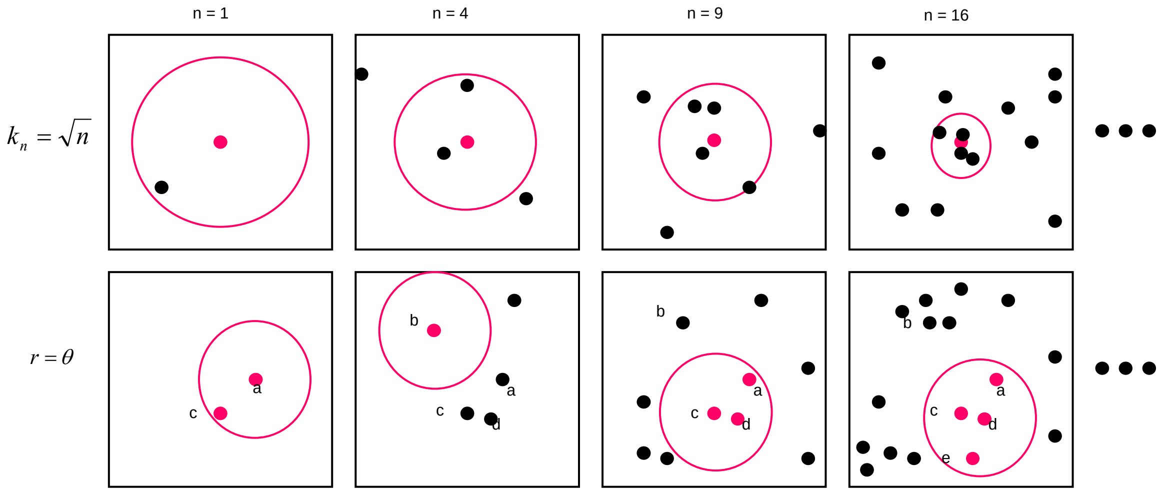


Figure 20. Illustrated kNN and FSWMC algorithm.

The pseudo code for our fixed-window multi-classification algorithm is as below.

```
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 $c$;
        }
}
```

7.3 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.

7.3.1 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 her icon on spirals in a 2D user space (see Figure 21 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 21 right). Distance between two spirals reflects the similarity between the two groups. By using JUNG Network/Graph library [14] the relative distances between groups are maintained even while the whole user space was zoomed or un-zoomed.

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, such as total number of queries and topics collected in user model data. High-ranked persons in a group are colored differently, and are classified as mentors; novice users may be encouraged to collaborate with them.

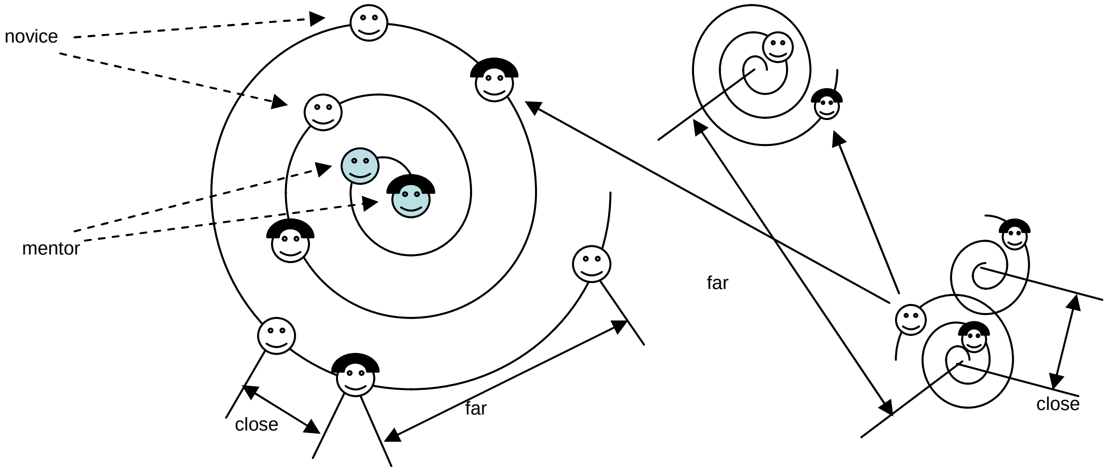


Figure 21. User and user community characteristics and relations.

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 equation (1) and (2) in Section 3.3. 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 [20].

Equation (7) 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 (8) [18].

7.3.2 Virtual Interest Group and Relations

Virtual Interest Groups are virtual clusters of DL users who share specific research interests and topics. Visualizing virtual interest groups helps us understand the characteristics of DL 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 used two methods to compute similarity to measure the degree of relation between two groups. One of those is cosine similarity which is computing vector similarity between the two group representatives (a union of the model data for all members), using equation (9). Another is Tanimoto Metric which is computing normalized overlapping ratio of members between two groups, using equation (10). Compared to cosine similarity, the Tanimoto Metric has lower computational cost but still is effective.

$$groupsim(A,B) = \sum_{i \in T} \frac{v_{A,i}}{\sqrt{\sum_{i \in T} v_{A,i}^2}} \frac{v_{B,i}}{\sqrt{\sum_{i \in T} v_{B,i}^2}} \qquad \textbf{(9)}$$

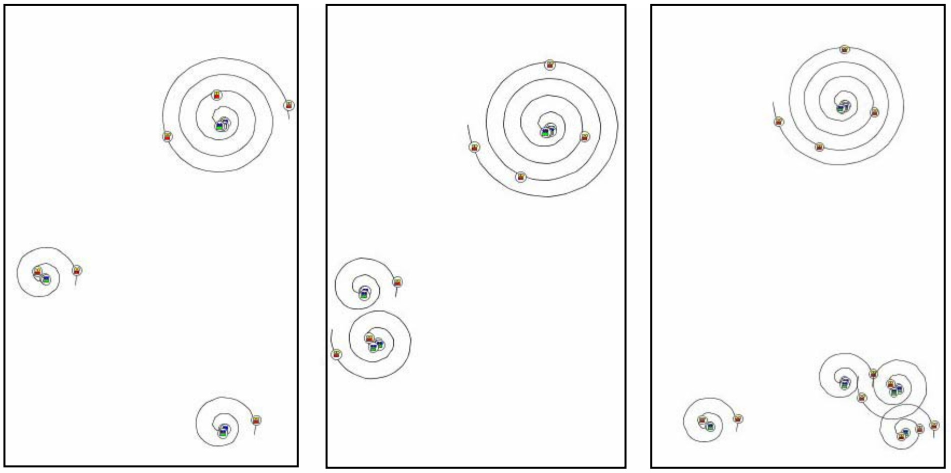
$$D_{Tanimoto}(A,B) = \frac{n_A + n_B - 2n_{AB}}{n_A + n_B - n_{AB}} \qquad \textbf{(10)}$$

(9) represents the group similarity between two virtual interest groups ‘*A*’ and ‘*B*’. ‘$v_{A,j}$’ is the sum of the frequencies of positive ratings on topic ‘*i*’ made by all users in group ‘*A*’. ‘*T*’ is the set of all topics in the collecting that are rated positively at least once. (10) represents the similarity distance between two groups ‘*A*’ and ‘*B*’. ‘n_A’ and ‘n_B’ are the numbers of users in *A* and *B*, respectively. ‘n_{AB}’ is the number of users in both groups *A* and *B*.

7.3.3 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 22 shows VUDM results for three months. 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.



June 2005 July 2005 August 2005

Figure 22. Visualizing usage trends of digital library.

7.3.4 Concept Drift

Detecting concept drifts is a well known problem in the machine learning area, that involves user models dynamically adjusting to user's changes quickly, as the real attributes of a user are likely to change over time [36, 37].

In recommender systems, detecting the concept drift of a user allows making recommendations at the proper times, as is illustrated in Figure 23. A User's concept of information search changes as time pass by. Sometimes a concept divides into multiple concepts, which will be visualized in VUDM by locating the user's icons on multiple spirals and linked with a line. Sometimes concepts disappear or merge into other concepts, which will be visualized in VUDM by removing some of the user's icons from spirals. VUDM is able to visualize states of user model data of different time, and HTTP log data, normally including time stamp, can be converted to user model data. With the time information, VUDM provide multiple time series image of user space. Figure 23 shows a progress of a user's concepts. Four different stages of her concepts are identified. One single concept, at the first stage, is divided into two and got another new concept at second stage. At the third stage, her two concepts merged into one, and lost interests for the other concept. At the last stage, her concept evolved into another a new concept. Detecting the last concept and making a recommendation for the concept is necessary for timely recommendation.

As a spiral in VUDM represents a set of closely related topics and interests, it also can be regarded as a concept 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 concept.

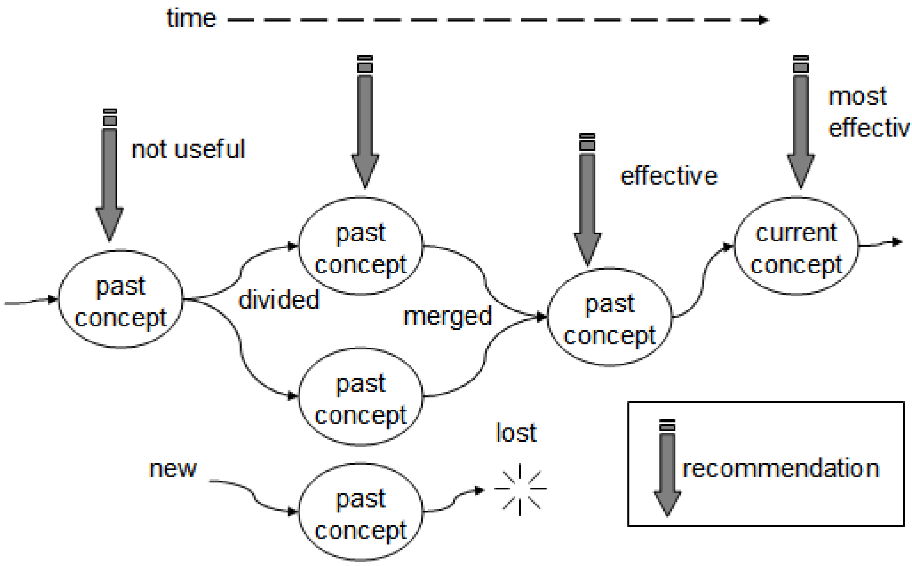


Figure 23. Detecting drift of concepts and timely recommendation.

7.4 Formative Evaluation of VUDM

7.4.1 Experiment Data

Our data set consists of 1,200 user models, describing those who registered to use our search result clustering service in NDLTD 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 14 describes the data set used for VUDM.

Explicit data is easy to analyze with normal analysis tools. However, such data is insufficient when addressing some comprehensive questions [19]. Further, user interests and behavior change over time, so it is important to enhance user models with implicit rating data. 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” [20]. Our 1,200 user models contain both explicit data and implicit rating data, as described in Table 14, that grows 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.

Table 14. Data Set: User data for VUDM.

	Data Type	**Number of Records**	**Description**
Implicit Data	Query Log	From 1200 Users	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		N/A

	Experience		
	Demographic Information		N/A

7.4.2 Evaluation Design

It is difficult to evaluate a visualization tool objectively, and VUDM is a data mining tool to support mainly Digital Library administrators or decision makers rather than normal Digital Library customers. Therefore, we conducted an analytic formative evaluation, which goal is to collect professional suggestions from several domain-knowledgeable participants, including user interviews as the system is developed [13]. Our evaluation consists of two sessions, answering sessions and interview sessions. 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 V isualization. Participants were given enough time to become familiar with 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. The questionnaire used for this experiment is included in the appendix at the end of this report.

7.4.3 Results of Evaluation

All participants answered positively for all questions, except two questions were answered negatively by one participant (see below). Table 15 shows the result of the answering session.

Table 15. Average (non-negative) scores for each question in answering session.

Question	Score
Information seeking trends	89
Virtual interest group distributions	85.5
User characteristics	86.2
Trends in the near future	75.8
Drift of concepts	69

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 sets 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 detailed 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 DLs, like CITIDEL [2] or ETANA-DL [3], should solve this problem.

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

7.4.4 Conclusion of Formative Evaluation

We developed a visualization tool, VUDM, to support knowledge finding and decision making in personalization. VUDM visualizes user communities and usage trends. VUDM makes use of unsupervised learning methods for grouping, labeling, and arranging a presentation in a 2-dimensional space. For this, a modified *k*NN neighboring algorithm, our fixed-size window multi-classification algorithm, was devised, which is suitable for flexible classification of users and user groups. Also, we categorized the knowledge needs required for personalization into three subcategories: user characteristics and relationships, virtual interest group characteristics and relationships, and usage trends. We showed how each of these can be addressed. We applied VUDM to NDLTD, analyzing 1,200 user models which are largely based on implicit ratings collected by a user tracking system. Through a formative evaluation, we found that VUDM is positively viewed with regard to the three categories.

Chapter 8. Conclusions

A new trend is for the WWW to be considered as a platform for dynamic and flexible web applications which is driven by, and evolves through, users' cooperation. Complex information systems, such as DLs, also are following this trend as they improve to provide more personalized and interactive services. Thus, user analysis and user-centered DL evaluation is being considered more important than before. In this technical report, we proposed several techniques to analyze users and DLs through utilizing user-providing information, implicit rating data, to enhance DL services. Implicit rating data is more important in complex information systems because it is more feasible to collect and utilize than explicit rating data. We showed how implicit rating data can be collected, stored, and processed. User tracking and user modeling techniques are proposed and implemented for this purpose. Also, we showed that implicit rating data can be used effectively to characterize users and find user communities in DLs, experimentally. Further, we provided results of hypothesis tests to support the potential of implicit rating data statistically, and an example of utilizing implicit rating data, in analyzing NDLTD usage, to obtain specific knowledge which is hard to get from other methods. Finally, we developed a visualization tool for analyzing users, user communities, and usage trends of DL by using implicit rating data. A conclusion of our study is that implicit rating data is effective for charactering users, user communities, and usage trends. We observe that it is meaningful to move forward from the previous generally held belief that implicit rating data is just auxiliary information for supporting explicit rating data, because collecting explicit rating data in DLs is expensive and has many problems.

Acknowledgements

We thank the: people and organizations working on CITIDEL and NDLTD, students participating in experiment and formative evaluation interviews, developers of the Carrot system, creators and supporters of the Piccolo and Jung Java libraries, and NDLTD users who answered our special survey. Thanks also go to NSF for support of grants DUE-0121679, DUE-0121741, IIS-0307867, and IIS-0325579.

Appendix: Institutional Review Board (IRB) Documents

Informed Consent Form

INFORMED CONSENT[1] FOR PARTICIPANTS OF INVESTIGATIVE PROJECTS
TO BE USED WITH THE QUESTIONNAIRE

TITLE OF PROJECT: GrapeZone: Document Clustering Techniques for Digital Libraries.

SUB TITLE : User Model Construction By Using Contents Clustering

INVESTIGATORS: Seonho Kim.

I. THE PURPOSE OF THIS RESEARCH

You are invited to participate in a study concerning the evaluation of document clustering techniques for use with digital libraries. This part of the study involves evaluating the effectiveness and the quality of three document clustering techniques using task-oriented evaluation methodology.

II. PROCEDURES

To accomplish the goals of this user study, you will be asked to perform a set of searching tasks using document clustering techniques to search a digital library for a particular document(s) and save your search results, complete both a paper-and-pencil task-questionnaire and a post-questionnaire, relating to previous experience using portals and search engines, as well as your recent experience using the document clustering techniques under investigation. Participation in this study will require approximately 1 hr of your time, and in order to participate, you must be at least 18 years old.

III. RISKS

There are no apparent risks involved with participation in this study.

IV. BENEFITS OF THIS PROJECT

[1] This informed consent is based on an approved previous informed consent, to be found at http://ei.cs.vt.edu/~cs5724/projects97f/cs3604www/icf.html, which has been modified to suit our project purpose.

A general benefit of this project is the opportunity to provide information which may ultimately lead to the improvement of digital library search results and clustering techniques, for the purpose of providing a more satisfying experience for digital library (such as CITIDEL) patrons.

No guarantee of direct benefits has been made to encourage you to participate.

If you would like to view a summary of this research when it is completed, please check the following web address for a link related to this project on or after December 25, 2003:
http://thorn.dlib.vt.edu:8080/controller/index.jsp

V. EXTENT OF ANONYMITY AND CONFIDENTIALITY

The written responses collected in this study will be kept strictly confidential. At no time will the researchers release an individual participant's responses. The information you provide will be identified through the use of a randomly assigned participant number. Only this number (not your name) will be used during data analyses and in any reports of this research.

VI. COMPENSATION

No financial compensation will be offered to you for participation in this project.

VII. FREEDOM TO WITHDRAW

You are free to withdraw at any time without penalty.

VIII. APPROVAL OF RESEARCH

Virginia Polytechnic Institute and State University’s Department of Computer Science (IRB # 97-255) have approved this research project, as required, by the Institutional Review Board (IRB) for Research Involving Human Subjects at Virginia Polytechnic Institute and State University and.

IX. PARTICIPANT'S RESPONSIBILITIES

I voluntarily agree to participate in this study. I understand that I have the following responsibilities:

1) To read all of the questionnaire's instructions.
2) To provide a written response for each of the questionnaire's items

X. PARTICIPANT'S PERMISSION

I have read and understand the informed consent and conditions of this project. I have had all my questions answered. I hereby acknowledge the above and give my voluntary consent for participation in this project.

If I participate, I may withdraw at any time without penalty. I agree to abide by the rules of this project.

____

Participant's Signature and Date

Should I have any pertinent questions about this research or its conduct, I may contact:

(Investigator) Seonho Kim, shk@vt.edu

(Faculty Advisor) Edward A. Fox- fox@vt.edu

Advertisements and Recruitment

To: gradstudents@cs.vt.edu

Hello everyone. I am a Ph.D. graduate student in department of Computer Science, working in the Digital Library Research Lab. I am studying user modeling techniques for Digital Libraries and looking for participants of an experiment for my study.

Your role in this test is playing a serious Digital Library user, using a search service of the Digital Library, and answering to a questionnaire.

This experiment will take about 40 minutes and you can take this experiment more than once, if you agree, under some conditions. The longer you take this test, the more data I can gain, so I will thank for your long participation.

If you are a student of Dept. of CS, ECE or anything related with computer, you can be a good helper for this experiment.

If you are taking usability engineering class, you can get some participation points.

If you need this, download, print and bring below form to the experiment.

http://courses.cs.vt.edu/~cs5714/spring2004/Participation/certification%20form.doc

I will appreciate it if you select your convenient time from following time table and let me know via email.

Experiment schedule time table = http://csgrad.cs.vt.edu/~haebang/timeslot.html

Contact: Seonho Kim (shk@vt.edu), Torgersen hall 2030.

Experiment Place: CS Grad Lab. (McBryde 659A)

Date: From 8th April

Questionnaire for User Tracking

Section 1: Back-Ground Questionnaire

1. User number [ ]

2. Age:

   ___ Under 18

   ___ 18 – 24

   ___ 25 – 34

   ___ 35+

3. Gender: ___ Male ___ Female

4. Profession:

   ___ Undergraduate student

   ___ Master student

   ___ Ph.D student

   ___ Post Ph.D

   ___ Researcher

   ___ Faculty

   ___ Others:_____________

5. What is your major? ________________________

6. If your major is(was) not CS/CE, list CS courses you have taken so far.

Under Level :

Grad Level :

7. What is your research interest topic? Please describe into 3 stage detail levels from general to specific topic.

e.g. : Computer Science > Digital Library > Document Clustering

e.g. : Computer Science > Artificial Intelligence > Machine Learning

_______________ > _______________ > _______________

If you have more than one topic, list them and please contact facilitator.

_______________ > _______________ > _______________

_______________ > _______________ > _______________

8. How long have you been studied in your major ? ___ years

9. What tools do you normally use to search for published papers?

___ CiteSeer.com / ResearchIndex.com

___ Search engine (i.e. Google, Yahoo, MSN)

___ ACM Portal

___ Citidel.org

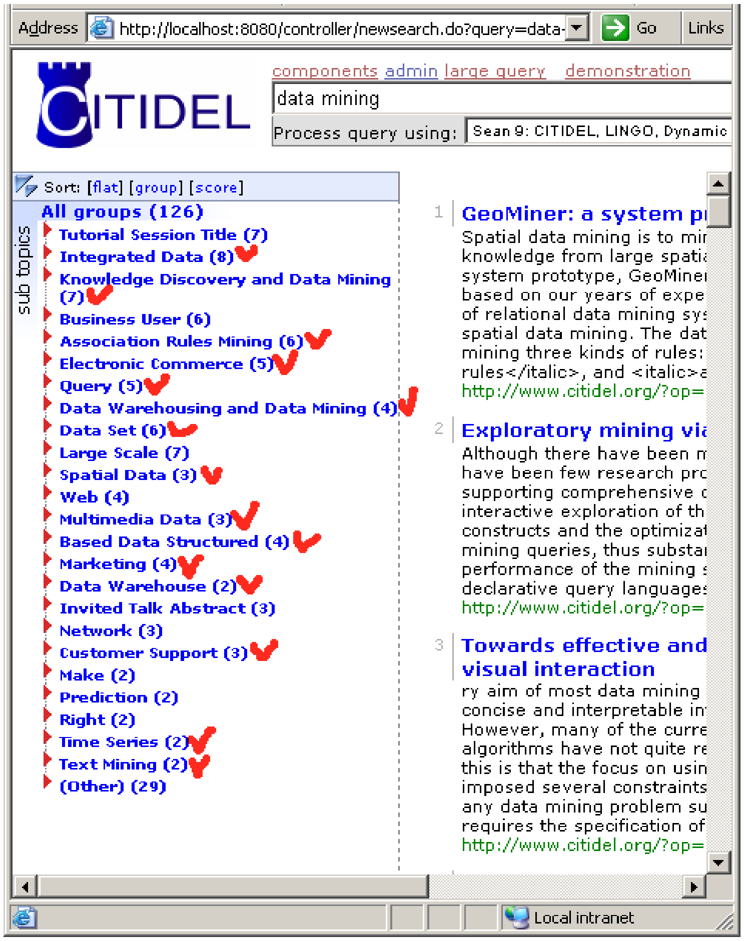
___ Printed documents

___ Traditional Library

___ Others:_____________________

Section 2: User Tasks

Your role of this experiment is playing a “serious” user of a Digital Library and using a search service. A serious user has **goals** in using a Digital Library, such as finding a paper for his/her research or finding a book on a specific topic for a class presentation. The user has enough knowledge to tell whether a document from the search result is relevant to his query or not. A serious user will try to use sophisticate queries to find relevant document fast and will not select queries randomly or just for fun.



Address
http://localhost:8080/controller/newsearch.do?query=data-
Go
Links
CITIDEL
components admin large query demonstration

data mining

Process query using: Sean 9: CITIDEL, LINGO, Dynamic

Sort: [flat] [group] [score]
sub topics

All groups (126)
- Tutorial Session Title (7)
- Integrated Data (8) ✓
- Knowledge Discovery and Data Mining (7) ✓
- Business User (6)
- Association Rules Mining (6) ✓
- Electronic Commerce (5) ✓
- Query (5) ✓
- Data Warehousing and Data Mining (4) ✓
- Data Set (6) ✓
- Large Scale (7)
- Spatial Data (3) ✓
- Web (4)
- Multimedia Data (3) ✓
- Based Data Structured (4) ✓
- Marketing (4) ✓
- Data Warehouse (2) ✓
- Invited Talk Abstract (3)
- Network (3)
- Customer Support (3) ✓
- Make (2)
- Prediction (2)
- Right (2)
- Time Series (2) ✓
- Text Mining (2) ✓
- (Other) (29)

1 GeoMiner: a system pı
Spatial data mining is to mir
knowledge from large spatiz
system prototype, GeoMinei
based on our years of expe
of relational data mining sys
spatial data mining. The dat
mining three kinds of rules:
rules</italic>, and <italic>a
http://www.citidel.org/?op=

2 Exploratory mining via
Although there have been n
have been few research prc
supporting comprehensive c
interactive exploration of th
constructs and the optimizat
mining queries, thus substaı
performance of the mining s
declarative query languages
http://www.citidel.org/?op=

3 Towards effective and visual interaction
ry aim of most data mining
concise and interpretable in
However, many of the curre
algorithms have not quite re
this is that the focus on usin
imposed several constraints
any data mining problem su
requires the specification of
http://www.citidel.org/?op=

Figure 24 : A result for query “data mining”.

Our user interface is featured with document clustering on the search result document set, which is grouping similar documents and naming the groups. Your task is using our user interface, such as sending queries, browsing the clusters and opening some documents you think relevant to your queries.

For example, see the figure 18, assume you have interests in “data mining”, you may throw a series of queries related with “data mining”, such as “data mining”, “recommender system”, “knowledge discovery”, “Bayesian network”, etc. After sending the query “data mining”, you will see this result screen. In the left frame, there are many clusters for the result documents. In the right frame, a list of documents in current cluster is shown. Clicking a cluster name of the left will replace the right frame with the information of documents in the selected cluster.

For example, see the figure, the user with interests in “data mining” may click the clusters which are marked to browse and examine the documents, and may not click rest of clusters because she think they may not contain any interesting documents. Your task in this experiment is using this interface this way for ten queries. Because there is no correct answer for this task, don’t be afraid of selecting wrong clusters.

Section 3: Post-Test Questionnaire

1. Do you think most clusters in the left frame were relevant to your query?

Never always

0 1 2 3 4 5 6 7 8 9 10

2. What kind of assistances are you expecting from the Digital Libraries for the future?

___ Material recommendation

___ Query recommendation

___ Conference news

___ Call for Paper

___ Push service (via email)

___ New material arrival

___ Search result clustering

___ Document summarization

___ Community recommendation

___ Web site recommendation

___ Course recommendation

List more services in your mind. Please stimulate your imagination!

____

____

____

____

3. Some web sites are gathering information about you, whether explicitly or implicitly, to provide more intelligent and personalized user interface and recommendation to users. What kind of information would you willing to provide

to the web sites for these purposes? Assume those information will be used only by the system for analysis purpose.

Agree Not agree

My gender 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

My age 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

My major 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

My company/school 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

My login time 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

My hobby 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

My research area 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

Courses you're taking 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

Courses you've been taken 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

Queries I used 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

Academic associations 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

Websites I visited 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

Websites in my favorite list 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

Documents I've accessed 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

Document I downloaded 0 1 2 3 4 5 6 7 8 9 10

Agree Not agree

My published papers 0 1 2 3 4 5 6 7 8 9 10

4. Any comment?

Questionnaire for Formative Evaluation of VUDM

Section 1: Back-Ground Questionnaire

1. User number [ ]

2. Age:
   ___ Under 18
   ___ 18 – 24
   ___ 25 – 34
   ___ 35+

3. Gender: ___ Male ___ Female

4. Profession:
   ___ Undergraduate student
   ___ Master student
   ___ Ph.D. student
   ___ Post Ph.D.
   ___ Researcher/Instructor
   ___ Faculty
   ___ Others:_____________

5. What is your major? ________________________

6. If your major is(was) not CS/CE, list CS/CE courses you have taken so far.

Undergraduate Course:

Graduate Course:

7. What are your research interests? Please describe your research interests in three detail levels of from general to specific.

e.g.: Computer Science > Digital Library > Document Clustering

e.g.: Computer Science > Artificial Intelligence > Machine Learning

________________ > ________________ > _______________

If you have more than one research interest, list them and please contact facilitator.

________________ > ________________ > _______________

________________ > ________________ > _______________

8. How long have you been studied the topics described at question 7?
_____ years

9. Which tools do you normally use to search documents for your research?

___ CiteSeer.com / ResearchIndex.com

___ Search engine (i.e. Google, Yahoo, MSN)

___ ACM Portal

___ Citidel.org

___ Printed documents

___ Traditional Library

___ Others:______________________

Section 2: User Tasks

After trying to be familiar with this visualization tool, answer the questions below.

1. This tool is able to show the "user space" at last three different months in timely order. Does this visualize how information search trend has been changed? (Yes / No)

If you are positive please rate your agreement:

Little Very much

1 2 3 4 5 6 7 8 9 10

If you are negative, could you explain why?

2. Does this tool show how users' research interests have been distributed? (Yes / No)

If you are positive please rate your agreement:

Little Very much

1 2 3 4 5 6 7 8 9 10

If you are negative, could you explain why?

3. Does this tool show how peoples are similar to each other? (Yes / No)

If you are positive please rate your agreement:

Little Very much

1 2 3 4 5 6 7 8 9 10

If you are negative, could you explain why?

4. Can you roughly predict attractive topics for next month? (Yes / No)

If you are positive please rate your agreement:

Little Very much

1 2 3 4 5 6 7 8 9 10

If you are negative, could you explain why?

5. Can you trace how a user's retrieval focus has been drifted?

If you are positive please rate your agreement:

Little Very much

1 2 3 4 5 6 7 8 9 10

If you are negative, could you explain why?

Section 3: Post-Test Questionnaire

1. Do you think it will be useful to you if Digital Libraries recommends some documents that you may be interested in?

Not at all Absolutely

0 1 2 3 4 5 6 7 8 9 10

2. Do you mind if your usage history in a Digital Library is stored somewhere for purposes of research and service improvement? (assume that your privacy will be secured)

Mind very much Don’t mind

0 1 2 3 4 5 6 7 8 9 10

3. Do you mind if an intelligent software analyzes the log data of all users, including yours, in Digital Library to develop better service? (assume that your privacy will be secured)

Mind very much Don’t mind

0 1 2 3 4 5 6 7 8 9 10

4. What kind of digital material types do you prefer when you don’t know what kind of knowledge you were supposed to find from it?

text graphic

0 1 2 3 4 5 6 7 8 9 10

5. Any comment or question?

References

1. Carrot2 Project, A research framework for experimenting with automated querying of various data sources, processing search results and visualization, Available at http://www.cs.put.poznan/pl/dweiss/carrot/, 2006
2. CITIDEL, Computing and Information Technology Interactive Digital Education Library, Available at http://www.citidel.org, 2006
3. ETANA-DL, Managing Complex Information Application: An Archaeology Digital Library, Available at http://etana.dlib.vt.edu, 2006
4. Friendster, Online Social Network Portal, Available at http://www.friendster.com/, 2007
5. NDLTD, Networked Digital Library of Theses and Dissertations, Available at http://www.ndltd.org, 2006
6. OCLC, Online Computer Library Center, Available at http://www.oclc.org/, 2006
7. 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.
8. Stuart K. Card, Jock D. Mackinlay and Ben Shneiderman, *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers, San Francisco, 1999.
9. Aaron Ceglar, John Roddick and Paul Calder, Guiding Knowledge Discovery Through Interactive Data Mining. *Managing Data Mining Technologies in Organizations: Techniques and Applications*, Idea Group Publishing, 2003, 45-87.
10. Richard O. Duda, Peter E. Hart and David G. Stork, *Pattern Classification*. A Wiley-Interscience Publication, 2000.
11. 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.
12. 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.
13. Deborah Hix and H. Rex Hartson, *Developing User Interfaces: Ensuring Usability Through Product & Process*. Wiley Professional Computing, 1993.
14. JUNG, Java Universal Network/Graph Framework, Available at http://jung.sourceforce.net/, 2007
15. Daniel A. Keim, Information Visualization and Visual Data Mining. In *IEEE Transactions on Visualization and Computer Graphics*, *7* (1), 2002, 100-107.
16. Seonho Kim, NDLTD, Search Interface Embedded User Tracking System, Available at http://boris.dlib.vt.edu:8080/controller/index.jsp, 2006
17. Seonho Kim, User Modeling for Educational Digital Libraries, Coursework Project Contract Page, Available at Available at http://collab.dlib.vt.edu/runwiki/wiki.pl?IsRproj_UserMod_Con, 2004

18. 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.
19. 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.
20. 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.
21. 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.
22. Ravi Kumar, Jasmine Novak, Prabhakar Raghavan and Andrew Tomkins, Structure and Evolution of Blogspace. In *Communications of the ACM*, *47* (12), 2004, 35-39.
23. 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.
24. 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.
25. David M. Nichols, Implicit Rating and Filtering. In *Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering*, Budapest, Hungary, 1997, 31-36.
26. 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.
27. Stanislaw Osinski and Dawid Weiss, A Concept-Driven Algorithm for Clustering Search Results. In *IEEE Intelligent Systems*, *20* (3), 2005, 48-54.
28. Stanisław Osiński and Dawid Weiss, Conceptual Clustering Using Lingo Algorithm: Evaluation on Open Directory Project Data, Advanced in Soft Computing, Intelligent Information Processing and Web Mining. In *Proceedings of the the International IIS: IIPWM'04 Conference*, Zakopane Poland, 2004, 369-378.
29. R. Lyman Ott and Michael Longnecker, *An Introduction to Statistical Methods and Data Analysis*. Wadsworth Group, 2001.
30. SAX Parser, Simple API for XML, Available at http://www.saxproject.org/, 2007

31. Michael Pazzani and Daniel Billsus, Learning and Revising User Profiles: The Identification of Interesting Web Sites. *Machine Learning*, Kluwer Academic Publishers, 1997, 313-331.

32. David W. Stockburger, Introductory Statistics: Concepts, Models, and Applications. Online book avaliable at http://www.psychstat.missouristate.edu/ introbook/sbk00.htm, 2007.

33. 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)*, 2004 414.

34. Cass R. Sunstein, Democracy and Filtering. In *Communications of the ACM*, *47* (12), 2004, 57-59.

35. 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.

36. 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.

37. Gerhard Widmer and Miroslav Kubat, Learning in the Presence of Concept Drift and Hidden Contexts. *Machine Learning*, Kluwer Academic Publishers, 1996 69-101.

38. 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.

39. Jennifer Xu and Hsinchun Chen, Criminal Network Analysis and Visualization. In *Communications of the ACM*, *48* (6), 2005, 101-107.

40. 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.