Visualizing User Communities and Usage Trends of Digital Libraries Based on User Tracking Information

Seonho Kim, Subodh Lele, Sreeram Ramalingam, and Edward A. Fox

Department of Computer Science
Virginia Tech
Blacksburg, Virginia 24061 USA
{shk, subodhl, sreeram, fox}@vt.edu

Abstract. We describe VUDM, our Visual User-model Data Mining tool, and its application to data logged regarding interactions of 1,200 users of the Networked Digital Library of Theses and Dissertations (NDLTD). The goals of VUDM are to visualize social networks, patrons' distributions, and usage trends of NDLTD. The distinctive approach of this research is that we focus on analysis and visualization of users' implicit rating data, which was generated based on user tracking information, such as sending queries and browsing result sets – rather than focusing on explicit data obtained from a user survey, such as major, specialties, years of experience, and demographics. The VUDM interface uses spirals to portray virtual interest groups, positioned based on inter-group relationships. VUDM facilitates identifying trends related to changes in interest, as well as concept drift. A formative evaluation found that VUDM is perceived to be effective for five types of tasks. Future work will aim to improve the understandability and utility of VUDM.

1 Introduction

Digital libraries (DLs) support diverse users, but might do so even better if available data about those users could be employed to facilitate personalization. Work toward such a goal is in keeping with new trends to improve the WWW, such as the move toward "Web 2.0", where web applications become more flexible, and evolve with the collaboration of users. Fortunately, we may benefit from data mining and unsupervised learning techniques applied to the large volume of usage data from community-driven information systems like blogs [1], wikis, and other types of online journals. Thus, we can go beyond what is possible if only examining data from OLAP systems [2]. We begin to address challenging research questions about a DL such as:

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

S. Sugimoto et al. (Eds.): ICADL 2006, LNCS 4312, pp. 111 – 120, 2006.

Fortunately, visualization supports direct involvement of users in exploration and data mining, so they can utilize their creativity, flexibility, and general knowledge [3].

There has been a great deal of prior work that relates to our research, so we only can touch on a small sample of papers touching on particular aspects of our approach.

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 [4]. Boyd, working with Social Network Fragments [5], visualized clusters of contacts derived from the *to* and *cc* lists in email archives. Heer, in Vizster, visualized relationships between members in an online date site Friendster [6], SPIRE Themescape [7] 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 [8, 9]. Tang utilized users' browsing patterns for collaborative filtering [10]. Webb examined challenging user modeling approaches like data rating, concept drift, data sparseness, and computational complexity [11].

In Section 2 we describe the DL context and data preprocessing aspects of our study. Section 3 introduces VUDM and our approach to visualization. Section 4 gives details about the visualization and illustrates its use for key tasks. Section 5 summarizes our pilot user study, and identifies important areas for future work, while Section 6 presents conclusions.

2 Data Description and Preprocessing

The Networked Digital Library of Theses and Dissertations (NDLTD) union catalog [12] describes a collection of over 240,000 electronic theses or dissertations (ETDs) from more than 325 member institutions, such as universities and libraries. Our data set consists of 1,200 user models, describing those who registered to use our service 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. Explicit data is easy to analyze with normal analysis tools. However, such data is insufficient when addressing the comprehensive questions listed in Section 1. Further, user interests and behavior change over time, so it is important to enhance user models with data from a user tracking system [13], i.e., "implicit (rating) data" (so-called because the data was not entered explicitly in answer to questions). Our implicit data consists of "queries" 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" [14]. Our 1,200 user models contain both explicit data and implicit rating data that grow with 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. The next section explains our visualization interface (VUDM) that supports all this.

3 VUDM and Visualization Strategies

Our Visual User model Data Mining (VUDM) tool transforms available data into a set of windows as illustrated in Figure 1. The main window presents an overview of all users (shown as icons) and communities (i.e., groups, shown as spirals). The presentation is controlled by a slider, that specifies a user correlation threshold (θ, which will be explained later in this section), in order to determine if users should be in the same group. Another control determines whether an overview is shown, or if one should zoom into a region of particular interest. In addition, all user icons and group spirals can be dragged with the mouse, e.g., to examine a congested area.

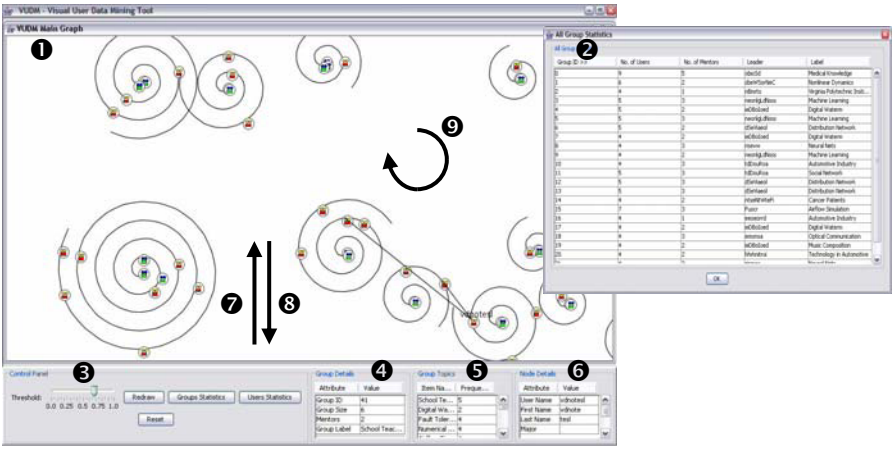

Fig. 1. The main window, 1, displays an overview of users, virtual interest groups, and their relationships. The statistics window, 2, presents detailed information, either about all users or about all groups in the system. The slide bar, 3, controls the correlation threshold (θ). The small tables at 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.

On demand, a small pop up window appears – see Figure 1 on the right (2). It provides detailed information about users or groups. It supports basic OLAP functions, such as sorting and listing. This and the main window are linked and synchronized. Thus, VUDM services combine strengths of graphical and text-oriented presentations.

The visualization of users and topic-based groups aims to summarize high dimensionality data, in order to support key tasks (see Section 4). Three degrees of freedom (three dimensions) are shown, since one can vary the position (x, y coordinates) of a spiral center, as well as the distance (of a user icon) from the center.

The positions of spirals (groups) are not controlled absolutely because the dimensionality of data is too high. It is only important to maintain relative distances among spirals (interest groups). For laying out the spirals, a "grid layout" method [15] 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.

Regarding classifying users into virtual interest groups and finding "groups of similar groups", we use the same algorithm. Because any statistical information about distribution and underlying densities of patrons, such as sample mean and standard deviation, are not known, nonparametric classification techniques, such as Parzen Windows and *k*-Nearest-Neighbor (*k*NN), 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 sample. For these reasons we devised a modified *k*NN algorithm: "fixed-size window multi-classification" (FSWMC) algorithm. Figure 2 illustrates the difference between *k*NN and FSWMC. Distances between samples (the spots in the hyperspace) are calculated using Formula (1) in Section 4.1. While the window size, *r*, of the *k*NN is dependent on 'n', the total number of samples, the window size of FSWMC is fixed to the correlation threshold θ. The θ value is entered from the user interface. In this algorithm, a test sample will be assigned to 0 or more classes, depending on the number of neighbors within the distance θ. Theoretically a maximum of 'n' classes, one class for each sample, can be found. However, we reduce the number by the "removing subclass rule": a class whose elements are all elements of another class can be removed to ensure there are no hierarchical relationships among classes. Also, we remove trivial classes, where the number of elements is smaller than a specified value. Even though Parzen Windows also uses a fixed-size window, our algorithm is more similar to *k*NN

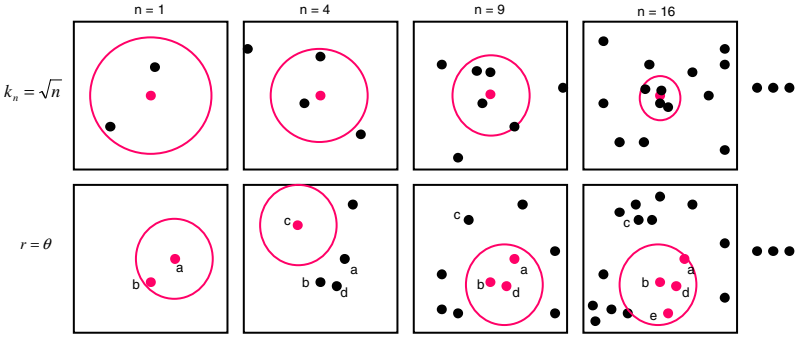

Fig. 2. Top Row: The *k*NN rule starts at the test point, red spot, among classified samples, and grows the surrounding circle until it contains '*k*' samples. Then, it classifies the test point into the most dominant class in the circle. Bottom Row: The fixed-window multi-classification rule classifies all samples enclosed by the fixed sized, $r=\theta$, 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, two classes, {c} and {a,b,d,e}, are found up to stage n=16.

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".

4 Support for Knowledge Finding Tasks

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.

4.1 User Characteristics and Relations

User characteristics are the most important information for personalization. Many commercial online shopping malls, such as amazon.com and ebay.com, are already utilizing user characteristics for personalized services. VUDM visualizes each user's interest topics and expertise level by putting his icon on spirals in a 2D user space (see Figure 3 left). Each spiral represents a set of closely related topics shared by the users placed on the spiral. 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 1).

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

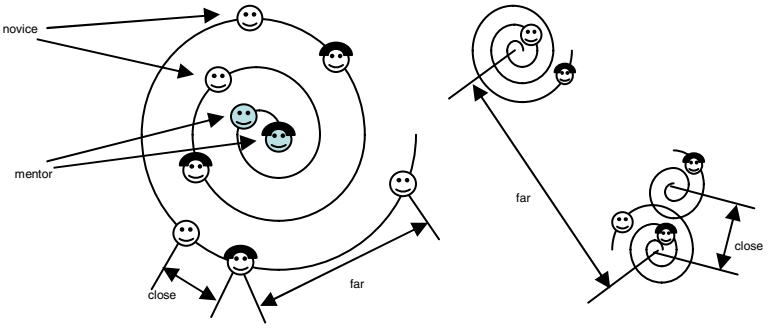

Fig. 3. Left: Each small face icon on the spiral is a user. A spiral represents a set of closely related topics and, thus, forms a virtual interest group with the users on the spiral who share the topics. The size of a spiral is proportional to the size of the group. Distance between user icons within a group reflects their similarity with regard to topics. Right: Distance between two spirals reflects the similarity between the two groups.

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 formulas (1) and (2). We used mainly implicit data rather than explicit data, because collecting implicit data is more practical than collecting explicit data, and it helps us avoid terminology issues (e.g., ambiguity) which are common in information systems [14].

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

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

(1) represents the correlation of users 'a' and 'b'. 'v_{aj}' is the rating value of item 'j' of user 'a' which means the number of positive ratings on 'j' made by 'a'. 'j' represents common topics or research interests which are rated by users 'a' and 'b'. '$\bar{v}_a$' is the average probability of positive rating of the user, as obtained by (2) [16].

4.2 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 measure the degree of relationship between two groups either by calculating the vector similarity between the two group representatives (a union of the model data for all members), using Formula (3), or by calculating the Tanimoto Metric (4) which uses the number of members in common [17]. Compared to vector similarity, the Tanimoto Metric has lower computational cost but still is effective.

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

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

(3) represents the group similarity between two virtual interest groups 'A' and 'B'. '$v_{A,j}$' is the sum of the frequencies of positive rating on topic 'i' made by all users in

group ‘A’. ‘T’ is the set of all topics in the system that are rated positive at least once. (4) represents the similarity distance between two groups ‘A’ and ‘B’. ‘n_A’ and ‘n_B’ are the number of users in A and B, respectively. ‘n_{AB}’ is the number of users in both groups A and B.

4.3 Usage Trends

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 4 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.

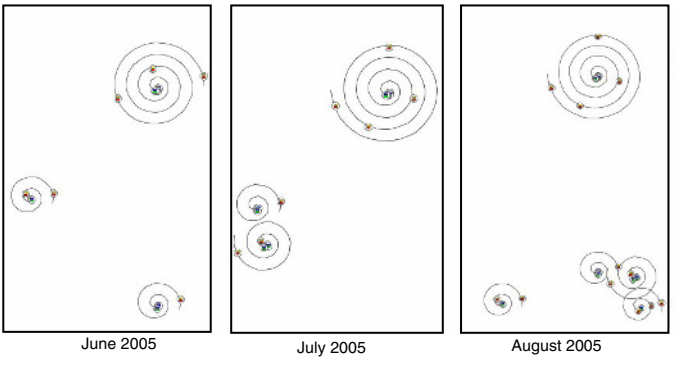

Fig. 4. Visualizing user space at different times makes it possible to figure out and predict retrieval trends, emerging attractive topics, drifts of concepts, etc.

VUDM also can help digital librarians visualize *concept drift*, which is a well known problem in the machine learning area [11]. The real attributes of a user are likely to change over time [18]. In recommender systems, detecting the concept drift of a user allows making more timely recommendations (see Figure 5).

As a virtual interest group spiral represents a set of closely related topics and interests, it also can be regarded as a concept for each user who belongs in the spiral. If a concept of a user drifts to a new concept, a clone of his icon appears on the new spiral and a connection line links the new icon together with the previously existing icons to indicate that they are for a single person. Therefore, by tracing connection lines and spirals over time, it is possible to detect occurrences of concept drift.

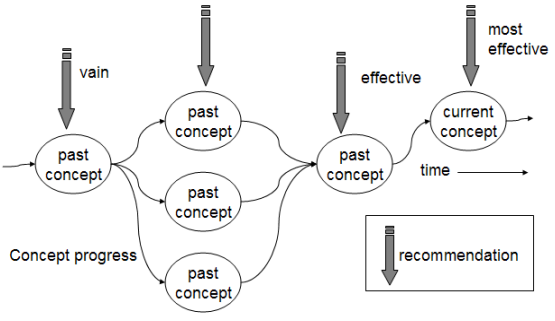

Fig. 5. Detecting drift of concepts makes it possible to provide timely recommendation

5 Evaluation and Future Work

It is difficult to evaluate a visualization tool objectively. Therefore, we conducted an analytic formative evaluation based on user interviews. Unlike a summative evaluation, whose goal is to prove the effectiveness of software statistically with many random participants, formative evaluation aims to collect professional suggestions from several domain-knowledgeable participants, as the system is developed [19]. Eight Ph.D students majoring in computer science were recruited, who have basic knowledge about Digital Library, Data Mining, and information visualization. Participants were given 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:

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

For each question, participants could answer either 'negative' or 'positive'. If they selected 'positive', they were asked to select the degree of agreement from 1 to 10. All participants answered positively for all questions, except two questions were answered negatively by one participant (see below). The average (non-negative) scores for each question were 89%, 85.5%, 86.2%, 75.8%, and 69%, respectively.

During the interview sessions, participants were asked to comment on problems with VUDM and to make suggestions. Most participants had difficulty 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 users and groups based on sets of topics (the user and group are involved with), and provides topic tables which consist of hundreds of topics ordered by frequencies, instead of labels.

One negative answer was about question 'c', using the topic tables. The participant commented that the topic tables don't work with visualization because they contain

too much detail information. The other negative answer was about question 'd'. It is difficult for VUDM users to spot changes in usage trends since they must see multiple pictures about usage trends for the past several months to predict the next month. The participant commented that VUDM should provide better visualization for this task, such as animation or colored traces showing changes. Since our approach is new, it is not surprising that some users were confused about the novel features of VUDM.

Further testing, with more time allowed for users to become familiar with our approach, is needed. Another problem we identified is that our user model data is just cumulative. It is not easy to determine if and when a topic goes out of favor. If we worked with sliding windows covering different time periods, we might solve such problems. Also, because the NDLTD union catalog covers all scholarly fields, and we only had 1,200 registered users, finding virtual interest groups was hard. Adding more user data or applying VUDM to subject-specific DLs, like CITIDEL [20] or ETANA-DL [21], should solve this problem. Finally, privacy issues were identified. Devices and modifications were requested to secure sensitive information, such as user IDs.

6 Conclusions

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, 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.

Acknowledgements

We thank the: people and organizations working on NDLTD, students participating in formative evaluation interviews, developers 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.

References

1. Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, Andrew Tomkins: Structure and Evolution of Blogspace. In Communications of the ACM, Vol. 47, No. 12 (2004) 35-39
2. Tom Soukup, Ian Davidson: Visual Data Mining: Techniques and Tools for Data Visualization and Mining. Wiley Computer Publishing, (2002)
3. Daniel A. Keim: Information Visualization and Visual Data Mining. IEEE Transaction on Visualization and Computer Graphics, Vol 8, No. 1 (2002) 1-8

4. Jennifer Xu, Hsinchun Chen: Criminal Network Analysis and Visualization. Communications of the ACM, Vol. 48, No. 6 (2005) 101-107
5. Danah Boyd, Jeffrey Potter: Social Network Fragments: An Interactive Tool for Exploring Digital Social Connections. In Proceedings of International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 2003) (2003) 1
6. Jeffrey Heer, Danah Boyd: Vizster: Visualizing Online Social Networks. In Proceeding of the 2005 IEEE Symposium on Information Visualization (INFOVIS'05) (2005) 5
7. James A. Wise, James J. Thomas, Kelly Pen-nock, David Lantrip, Marc Pottier, Anne Schur: Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In Proc. of the Information Visualization Symposium, IEEE Computer Society Press, (1995) 51-58
8. Eren Manavoglu, Dmitry Pavlov, C. Lee Giles: Probabilistic User Behavior Models. In Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03), (2003), 203-210
9. Michael Pazzani, Daniel Billsus: Learning and Revising User Profiles: The Identification of Interesting Web Sites. Machine Learning, Kluwer Academic Publishers, Vol 27 (1997) 313-331
10. Tiffany Ya Tang, Gordon McCalla: Mining Implicit Ratings for Focused Collaborative Filtering for Paper Recommendations. In Online Proceedings of Workshop on User and Group models for Web-based Adaptive Collaborative Environments (UM'03). Available at http://www.ia.uned.es/~elena/um03-ws/ (2006)
11. Geoffrey I. Webb, Michael J. Pazzani, Daniel Billsus: Machine Learning for User Modeling, User Modeling and User-Adapted Interaction. Kluwer Academic Publisher, Vol 11 (2001) 19-29
12. NDLTD, Networked Digital Library of Theses and Dissertations, available at http://www.ndltd.org (2006)
13. Uma Murthy, Sandi Vasile, Kapil Ahuja: Virginia Tech CS class project report. Available at http://collab. dlib.vt.edu/runwiki/wiki.pl?IsRproj_UserMod_Con (2006)
14. Seonho Kim, Uma Murthy, Kapil Ahuja, Sandi Vasile, Edward A. Fox: Effectiveness of Implicit Rating Data on Characterizing Users in Complex Information Systems. 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'05), LNCS 3652, Springer-Verlag, Berlin Heidelberg New York (2005) 186-194
15. Ivan Herman, Guy Melançon, M. Scott Marshall: Graph Visualization and Navigation in Information Visualization: A Survey. IEEE Transactions on Visualization and Computer Graphics, Vol 6, Issue 1 (2000) 24-43
16. Seonho Kim, Edward A. Fox: Interest-based User Grouping Model for Collaborative Filtering in Digital Libraries. 7th International Conference of Asian Digital Libraries (ICADL'04), LNCS 3334, Springer-Verlag, Berlin Heidelberg New York (2004) 533-542
17. Richard O. Duda, Peter E. Hart, David G. Stork: Pattern Classification. A Wiley-Interscience Publication, (2000)
18. Gerhard Widmer, Miroslav Kubat: Learning in the Presence of Concept Drift and Hidden Contexts. Machine Learning. Kluwer Academic Publishers, Vol. 23 (1996) 69-101
19. Deborah Hix, H. Rex Hartson: Developing User Interfaces: Ensuring Usability Through Product & Process. Wiley Professional Computing, (1993)
20. CITIDEL, Computing and Information Technology Interactive Digital Educational Library. Available at http://www.citidel.org (2006)
21. ETANA-DL, Managing complex information applications: An archaeology digital library. Available at http://etana.dlib.vt.edu (2006)