

Interest-Based User Grouping Model for Collaborative Filtering in Digital Libraries

Seonho Kim and Edward A. Fox

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

Abstract. Research in 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 recommendations 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 for recommender systems 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 paper, we present an interest-based user grouping model for a collaborative recommender system for Digital Libraries. Also, we present several user interfaces that obtain implicit user rating data. Our model uses a high performance document clustering algorithm, LINGO, to extract document topics and user interests from documents users access in a Digital Library. This model is better suited to Digital Libraries than traditional recommender systems because it focuses more on users than items and because it utilizes implicit rating data.

1 Introduction

One of the most anticipated roles of a Digital Library (DL) is to support a variety of user classes, such as researchers and learners, who are interested in similar research and learning topics. Another expected role for DLs is the active support of social communities. In order to implement these complex functions of DLs, more research should be carried out on the users rather than the system. Many studies of collaborative filtering (CF) and adaptive techniques for recommender systems have been conducted for online shopping malls and personalized websites [1, 4]. However, those techniques focus on exploring learning methods to improve accuracy instead of focusing on 1) scalability, 2) accommodating new data and 3) comprehensibility - three important aspects of recommender systems in DLs [5]. We propose a user grouping technique to improve these aspects by concentrating on user interests.

2 Previous Studies

Recommending has been studied with regard to Ungar's system [4] for online media shopping malls; GroupLens [13] for Usenet netnews systems; and for DLs: PASS [2],

Renda's system [3], and DEBORA [15]. The main idea of these studies is the sharing of user profile and preference information, which was entered explicitly by users, with other users within the system, to provide them with more productive service. PASS provided personalized service by computing similarity between user profiles and documents in pre-classified research domains. Renda's system emphasized collaborative functions of DLs by grouping users based on their profiles.

Unlike other systems, GroupLens employs a time consuming factor, which is a type of implicit rating data, together with explicit rating data. Nichols's research [11] emphasized the potential of the implicit rating technique and suggested the use of implicit data as a check on explicit ratings. Gonçalves' research defined an XML based log standard for DLs [16] which paved the way for the study of implicit ratings in DLs. This research emphasized the interoperability, reusability and completeness of the log system. However, without extension, this log suggestion is not suitable for recommender systems, that use the standard HTTP log system, because it initially focused on extracting system rather than content information.

3 Recommending Through Interest-Based User Grouping

In our system, users that share a research or learning interest are placed in the same group. Users' interests are collected from their implicit ratings, based on browsing and selection of clustered document sets, rather than explicit ratings such as from questionnaire responses. One hypothesis of our model is that high performance document clustering algorithms, such as LINGO [7], can be used to extract topics from documents. We also can use document clustering algorithms to overcome data sparseness problems. Document topics are stored in the user model and are used for grouping the users and making recommendations. Since we integrate item information into the user model, we can concentrate on the user side and make it easier to formulate recommendations. Because of these features, our model is especially suitable for DLs.

Our system employs LINGO to extract topics from documents. Document clustering also is used to overcome the bias caused by sparse ratings in most research areas and highly-duplicated ratings in very narrow areas. LINGO was selected because of its ability to discover descriptive names for clusters. Unlike most other algorithms, LINGO first attempts to find descriptive names for future clusters and only then proceeds to assign matching documents to clusters [7]. The names found by LINGO are treated as "document topics". When a user performs a search and gets a set of result documents, a selection of topics is provided to the user. Topics which are rated positively are treated as user "interests". Thus, user ratings are made implicitly.

Our system uses implicit user rating information collected by tracking user interactions with the interface: sending a query, expanding a dynamic tree, selecting a document cluster, and selecting a document to read or download. When a user rates document topics in a DL, the document topics are saved in temporary storage by the client. The saved user interests are sent to the server and stored in the user profile database for use by the recommender system when the user returns to the DL. Unlike GroupLens [12], our system doesn't use a time-consuming factor because it causes so much noise in real world use and distracts experiment participants.

Our recommender system is based on collaborative filtering; to make a user recommendation, the system refers to other users that have similar interests. User

grouping is based on similarity. In the field of information retrieval, there are two general methods (which we employ) to calculate similarity between documents: correlation and vector similarity. Thus we compute:

$$w(a,i) = \frac{\sum_j (v_{a,j} - \bar{v}_a)(v_{i,j} - \bar{v}_i)}{\sqrt{\sum_j (v_{a,j} - \bar{v}_a)^2 \sum_j (v_{i,j} - \bar{v}_i)^2}} \quad (1)$$

$$w(a,i) = \sum_j \frac{v_{a,j}}{\sqrt{\sum_{k \in I_a} v_{a,k}^2}} \frac{v_{i,j}}{\sqrt{\sum_{k \in I_i} v_{i,k}^2}} \quad (2)$$

$$\bar{v} = \frac{\text{total number of topics selected by the user}}{\text{total number of topics proposed to the user by the system}} \quad (3)$$

(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 user ‘a’ and ‘i’. ‘ $\bar{v}$ ’ is the average probability of positive rating of the user which is obtained by (3). (2) represents vector similarity between two users. ‘ I_a ’ is a set of interests of user ‘a’, and ‘k’ is an element of the interests set. Either formula can be used to calculate the similarity between two users. 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”.

A recommendation is triggered either when an item in a DL attains a rating greater than a certain value defined by the system manager, or when a newly arrived item gets a positive rating, after reading or downloading, from any user that has mentor level status in his/her group. The recommendation process of our model consists of two phases. In phase one, the recommender system decides which user groups will be recommended by a new positive rating. These groups are determined based on the probabilities that each of its users makes the same rating. R_k is the probability that user group ‘k’ is affected by the rating, which is made by a user ‘a’ to an item ‘j’.

$$R_k = P_k \frac{1}{N} \sum_{i: C_i=k} v_{i,j} (1 - P_k)^{\frac{1}{T-N} \sum_{i: C_i \neq k} (1 - v_{i,j})} \quad (4)$$

where T is the total number of users registered in the system, C_i is the group that user ‘i’ belongs to, v_{ij} is the probability that user ‘i’ rates item ‘j’, N is the total number of users in group ‘k’ and P_k is the base rate of group ‘k’ observed from the database (which is calculated by dividing the number of users in group ‘k’ by ‘T’, the total number of registered users in the system).

The second phase of the recommendation process is individual recommendation. Once the groups are selected for a new rating, individual users in highly ranked groups are checked for eligibility to get a recommendation for the rating that evoked the recommendation. We compute $P_{a,j}$, the probability that a user ‘a’ in group ‘k’ likes item ‘j’.

$$P_{a,j} = \bar{v}_a + \kappa \sum_{i=1}^n w(a,i)(v_{i,j} - \bar{v}_i) \quad (5)$$

where ‘n’ is the number of users in the group. This formula is derived from the Memory-Based Algorithm [14] but we use average probability of positive rating of

user ‘a’ for $\overline{v_a}$ instead of the mean rate value of the user. By only calculating this formula for users in the selected groups we decrease the cost of the recommendation.

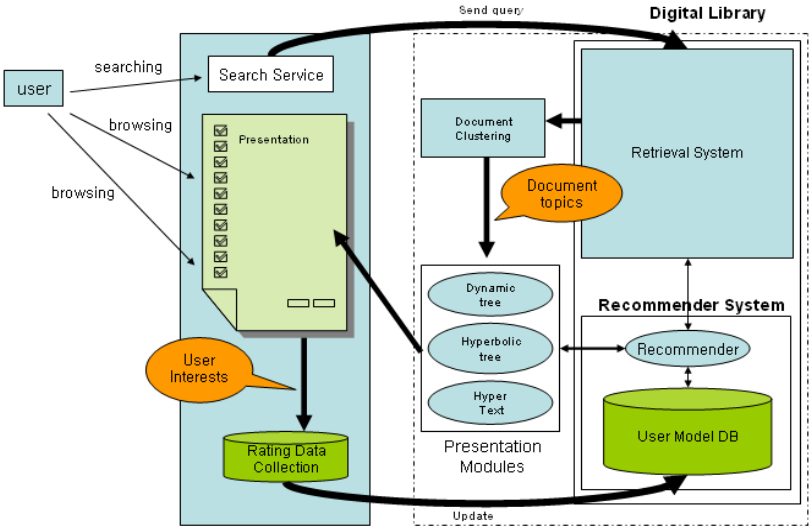


Fig. 1. A schematic drawing of the components in the interest-based recommender system for DLs. Document topics are extracted by the naming algorithm of the document clustering system, LINGO, and are presented to the user to be rated by way of hyperbolic tree, dynamic tree, and normal HTML pages. While using the DL, the user gives implicit ratings, such as by sending queries, browsing result documents, and reading some documents. User interests are gathered from these implicit ratings and stored into a “rating data collection” which will affect the user’s profile that is stored in the “user model DB” in the DL

Generally, user models are implemented with user profile data such as name, ID, sex, major, interests, position, hobby, etc. Our system, illustrated in Figure 1, which is based on the Carrot2 system [6], 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 recommender and added to the user model database as XML. Unfortunately, the standard HTTP log protocol cannot extract the title of an anchor, which is treated as a document topic in our system. Specially designed user interfaces implemented in JavaScript (see Figure 2) or Java (see Figure 3) [9, 10, 17], are required. 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. The interests of the current user are stored in a cookie at the client and are transferred to the server the next time the user accesses the DL. The cookie size limit of 4000 bytes is large enough to store the user’s behavior for a single login session.

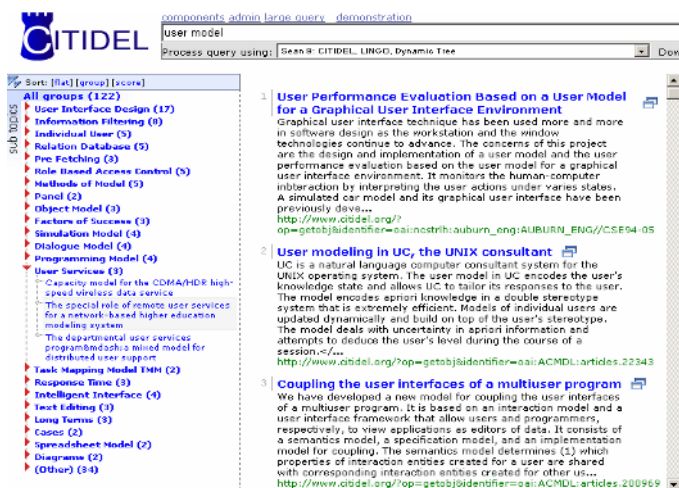


Fig. 2. A JavaScript based user interface. The left (dynamic tree) and right (HTML page) frames present a clustered result set efficiently. Data on user interactions such as opening clusters or selecting a document are stored temporarily in a cookie

4 Hypotheses

With our system, which extracts user interests for user grouping, we may test three hypotheses about proper document clustering algorithm behavior:

1. For any serious user who has his own research interests and topics, show consistent output for the document collections referred to by the user.

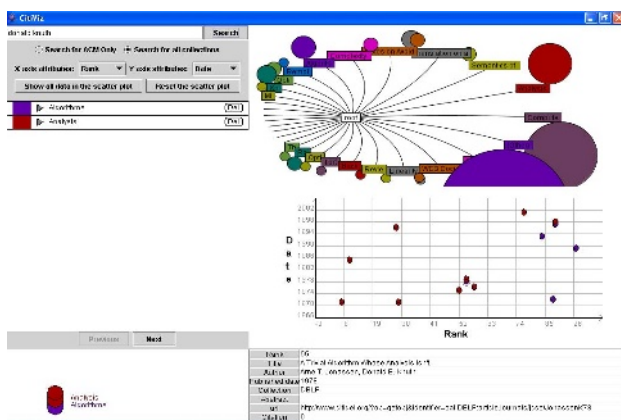


Fig. 3. CITIVIZ: An Interactive Visualization Interface for CITIDEL. Search result clusters are presented as nodes in a hyperbolic tree. Users can select relevant nodes to see detailed information of the document in clusters [17]

- 2. For serious users who share common research interests and topics, show overlapped output for the document collections referred to by them.
- 3. For serious users who don't share any research interests and topics, show different output for the document collections referred to by them.

5 Experiments: Collecting User Interests and User Grouping

Our experiments test the hypotheses. Participants completed a questionnaire with a question asking about their research interests. Then, participants were asked to list 10 specialized and complex queries in their research field to input into the DL for retrieval. In answering this difficult question, the participants were allowed to refer to any web site, document, or printed paper to get ideas. After completing the questionnaire, participants were instructed to use our JavaScript-based experimental interface to CITIDEL [8] (see Figure 2), to search for documents with the queries listed in the questionnaire.

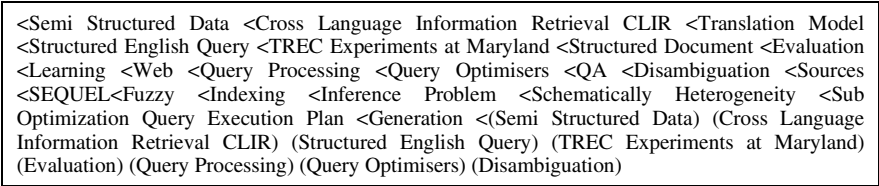


Fig. 4. Implicit rating data collected from one user

After CITIDEL displayed the search results, the participants were asked to browse through them. Implicit rating data was collected because participants were supposed to open and read relevant clusters and documents. Figure 4 shows an example of the implicit rating data collected during one participant's interaction with our system. This data 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). The parenthesized topics mean they are rated positively by the user.

6 Results

We collected user interests 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 at Virginia Tech were analyzed for this study. For convenience, we named each subject according to their research interests, which were obtained from the questionnaires, as shown in Table 1.

User grouping was based on vector similarity as in (2). User similarities among all subjects are shown in Figure 5. A longer column in the graph represents greater similarity. Columns are either high or very low. This means the document clustering algorithm is able to distinguish a user from others who have different interests.

Table 1. Symbols of participants of this experiment and their profiles. They were obtained from explicit ratings made by their answering the questionnaire

	User Symbols	User profiles collected from questionnaire
1	dlmember	The one who belonged to the Digital Library Research Laboratory
2	softeng	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	person_1	The one who has an interest in Personalization
16	person_2	The one who has an interest in Personalization
17	se_me	The one who has an interest in Software Engineering
18	fuzzy	The one who has an interest in Fuzzy Theory

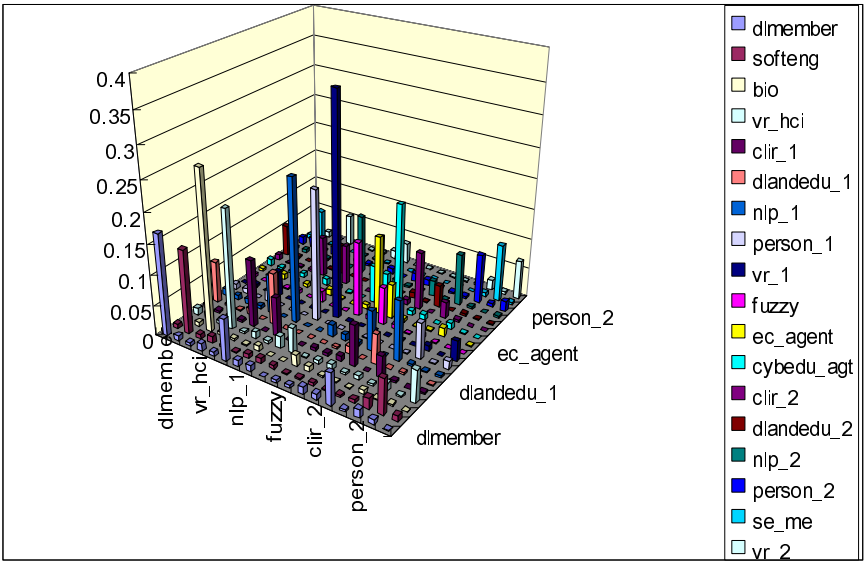


Fig. 5. User similarities among the users

Hypothesis 1 is supported by the results of Figure 5, which shows that similarities among users are distinctive according to their research topics. Hypotheses 2 and 3 are supported in Table 2. Several pairs of users with the same interests, such as “person_1 and person_2”, “nlp_1 and nlp_2”, “clir_1 and clir_2”, “dlandedu_1 and dlandedu_2”, and “softeng and se_me” show high similarities, which indicates that the document clustering algorithms generated highly overlapped outputs. Further support and clarification is given in Table 3, which shows nine user groups from eighteen participants in the experiment, based on our interest-based user grouping model.

Table 2. Similarities levels, sorted & grouped between each user & the other (Level 5) users

User ID	Level 1	Level 2	Level 3	Level 4
dlmember			dlandedu_1, dlandedu_2	
softeng			se_me	person_2
bio				
vr_hci			vr_2, vr_1	person_1, person_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	person_1, person_2
EC_agent		CybEdu_agt		person_2
CybEdu_agt		EC_agent	fuzzy	
dlandedu_1	dlmember	dlandedu_2		vr_hci, CybEdu_agt
dlandedu_2	dlmember	dlandedu_1		CybEdu_agt, vr_hci
person_1			person_2	
person_2			person_1	
se_me		softeng		
fuzzy			CybEdu_agt	bio, nlp_1

Table 3. User groups generated by merging a user and other members with the closest similarity level from Table 2

User Group ID	Members
A	dlmember, dlandedu_1, dlandedu_2
B	softeng, se_me,
C	vr_hci, vr_1, vr_2
D	clir_1, nlp_1, clir_2
E	nlp_1, nlp_2
F	person_1, person_2
G	fuzzy, cybedu_agt
H	EC_agent, cybedu_agt, fuzzy
I	Bio

7 Conclusions and Future Work

We proposed an interest-based user grouping model for collaborative filtering recommender systems. These are suitable for systems such as DLs which have sparse rating data and dynamically increasing contents. This model focuses more on users than items. Also, it actively finds user communities. A document clustering algorithm and specially designed user interfaces are employed to extract document topics and user interests by way of implicit ratings. Results of experiments to test the hypotheses support our approach. Future work will include detailed analyses of this model for accuracy, scalability, and efficiency. Other work will seek further confirmation of our hypotheses by way of wider deployment of the visualization front-ends for CITIDEL.

Acknowledgements

We thank the: people and organizations working on CITIDEL, student participants in our experiments, developer of the LINGO algorithm, and developers of the CITIVIZ visualization tool. Thanks go to the National Science Foundation for its support of grants NSF DUE-0121679, DUE-0121741, IIS-0307867, and NSF IIS-0325579.

References

1. Georgios Paliouras, Vangelis Karkaletsis, Christos Papatheodorou, Constantine D. Spyropoulos: Exploiting Learning Techniques for the Acquisition of User Stereotypes and Communities. In Proc. Int. Conference on User Modelling, UM '99 (1999) 169-178.
2. Chun Zeng, Xiaohui Zheng, Chunxiao Xing, Lizhu Zhou : Personalized Services for Digital Library. In Proc. 5th Int. Conf. on Asian Digital Libraries, ICADL (2002) 252-253.
3. M. Elena Renda, Umberto Straccia: A Personalized Collaborative Digital Library Environment. In Proc. 5th Int. Conf. on Asian Digital Libraries, ICADL (2002) 262-274
4. Lyle H. Ungar and Dean P. Foster: A Formal Statistical Approach to Collaborative Filtering. CONALD '98, Carnegie Mellon U., Pittsburgh, PA (1998).
5. John S. Breese, David Heckerman and Carl Kadie: Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (1997) 43-52.
6. Carrot² Project, A Research Framework for experimenting with automated querying of various data sources, processing search results and visualization, <http://www.cs.put.poznan.pl/dweiss/carrot/> (2004)
7. Stanislaw Osinski and Dawid Weiss: Conceptual Clustering Using Lingo Algorithm: Evaluation on Open Directory Project Data, Advanced in Soft Computing, Intelligent Information Processing and Web Mining, Proceedings of the International IIS: IIPWM'04 Conference, Zakopane, Poland (2004) 369-378
8. CITIDEL project, Computing and Information Technology Interactive Digital Educational Library, <http://www.citidel.org/> (2004)
9. Abdelmoumin Ghada, Alafaliq A. Ahmad, Nithiwat Kampanya, Seonho Kim: Virginia Tech report, at <http://infovis.cs.vt.edu/cs5764/Fall2003/projects.html> (2003)

10. Saverio Perugini, Kathleen McDevitt, Ryan Richardson, Manuel Perez-Quinones, Rao Shen, Naren Ramakrishnan, Chris Williams, Edward A. Fox: Enhancing Usability in CITIDEL: Multimodal, Multilingual and Interactive Visualization Interfaces, in Proc. of the Fourth ACM/IEEE Joint Conference on Digital Libraries (2004) 315-324
11. David M. Nichols: Implicit Rating and Filtering. In Proceedings of 5th DELOS Workshop on Filtering and Collaborative Filtering (1997) 31-36
12. Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, John Riedl: GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In Proceedings of Conference on Computer Supported Cooperative Work (1994) 175-186
13. Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl: GroupLens: Applying Collaborative Filtering to Usenet News. Communications of the ACM, Vol. 40 (1997) 77-87
14. Kai Yu, Anton Schwaighofer, Volker Tresp, Xiaowei Xu, Hans-Peter Kriegel: Probabilistic Memory-based Collaborative Filtering. IEEE Transactions on Knowledge and Data Engineering (2003) 56-69
15. David M Nichols, Duncan Pemberton, Salah Dalhoumi, Omar Larouk, Clair Belisle, Michael B. Twidale: DEBORA: Developing an Interface to Support Collaboration in a Digital Library. In Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries (2000) 239-248
16. Marcos A. Gonçalves, Ming Luo, Rao Shen, Mir Farooq, and Edward A. Fox: An XML Log Standard and Tools for Digital Library Logging Analysis, in Proc. of Sixth European Conference on Research and Advanced Technology for Digital Libraries (2002) 16-18
17. Nithiwat Kampanya, Rao Shen, Seonho Kim, Chris North, Edward A. Fox: Citiviz: A Visual User Interface to the CITIDEL System, in Proceedings of Fourth European Conference on Digital Libraries, ECDL 2004 (2004)