PERSONALIZATION, COLLABORATION, AND RECOMMENDATION IN THE DIGITAL LIBRARY ENVIRONMENT CYCLADES

Henri Avancini and Umberto Straccia
Istituto di Scienza e Tecnologie dell'Informazione - C.N.R.
Via G. Moruzzi,1 I-56124 Pisa, ITALY.
{Avancini, Straccia}@isti.cnr.it

ABSTRACT

CYCLADES is a system, which combines several technologies from the Information Retrieval and Digital Library area, where users and user communities may deal with a quite large set of heterogeneous digital archives. CYCLADES provides a highly personalized environment where not only users may organize (and search into) the information space according to their individual taste and use, but, and more importantly, also provides advanced features of collaborative work among the users. It is up then to the system to discover interesting properties about the users' interests, relationship between users and user communities, as well as meaningful events that happen in user shared workspaces, and finally to notify the involved users according to their own preferences.

KEYWORDS

Digital library, collaboration, personalization, recommendation.

1. INTRODUCTION

It is widely recognized that *Digital Libraries* (DLs) (Fox01, E. and Marchionini, G. 2001) will play an important role in the next future not merely in terms of the "controlled" digital information they allow access to, but in terms of the *services* they provide to the information society at large. Informally, DLs can be defined as consisting of collections of information (usually, heterogeneous in content and format), which have associated services delivered to users and user communities using a variety of technologies. The services offered on such information can be varied, ranging from content operations to rights management and can be offered to *individuals* as well as to *user communities*. Indeed, an essential technology component of DLs is that they are networked, meaning that access is increasingly becoming *shared* and *collaborative*.

Even though DLs have evolved rapidly over the past decade, typically, DLs still are limited to provide a search facility to the digital society at large. Indeed, they are oriented towards a generic user, as they answer queries crudely rather than, e.g. learn the long-term requirements of a specific user. In practice, users use the same information resource over and over and would benefit from customization: the time consuming effort that the user put in searching documents and possibly downloading them from the DL is often forgotten and lost. This requires a repetition of the manual labor in searching and browsing to find the documents just like the first time. As DLs will become more commonplace and the range of information they provide and the services upon increases, users' expectations will increase and users are expecting more and more sophisticated services from their DLs. A "quick and dirty search" facility is normally an integral part of any digital library, but users' frustrations increase as their demands become more complex and as the volume of information managed by digital libraries increases. There is a need for DLs to move from being *passive* with little adaptation to their users, to being more *proactive* and *personalized* in offering and tailoring information for individual users. Personalization can be defined as the way in which information and services can be tailored in a specific way to match the unique and specific needs of an individual user or a community of users. This is achieved by adapting the presentation and/or the services presented to the user by taking into account the user's task, background, history, device, information needs, location, etc., essentially the user's

context. Personalization can be user-driven which involves a user directly invoking and supporting the personalization process by providing explicit input, or personalization can be completely automatic, where the system observes some user activity and identifies the input used to tailor some aspect of the system in a personalized way. These two examples of user-driven and automatic personalization are at the extreme ends of the spectrum and many personalization tools will have elements of both approaches.

Nowadays, in several DLs some personalization functionalities are provided. Mainly they fall into the category of personalized *alerting services* (see e.g. Bollacker, K. et al. 1999, Faensen, D. et al. 2001, Fernandez, L. et al. 2000, Moukas, A. 1996, Rocha, L. 1999), i.e. services that notify a user (usually, by sending an e-mail), with a list of references to newly available documents in the DLs and deemed as relevant to some of the (manually) user specified topics of interests. Some other DLs, in addition, support the users in being able to organize their information space they are accessing to according to *their own subjective perspective* (see e.g. Fernandez, L. 2000). This is important as not necessarily all the information provided by a DL may be of interest to an user, but just some "slices" of it. Users and communities of users might well profit from being able to organize the information space in a personalized fashion both in terms of restricting the information space in which to search into as well as in terms of organizing it not necessarily in the way a the DL manager thought would be well-suited for anyone.

In this paper we present the CYCLADES system (http://www.ercim.org/cyclades) and stress its "personalization" and alerting features. A major distinction of CYCLADES is the fact that it envisages a DL not only as information space in which individual users may search for and organize the information provided by a DL, but also as a *collaborative meeting place* of people sharing common interests. Indeed, DLs may be viewed as a *common working place* where users may become aware of each other (indeed the system may find out interesting relationships both between users and/or between communities of users and produce the appropriate recommendations), open communication channels, and exchange information and knowledge with each other or with experts. Indeed, usually users and/or communities access a DL in search of some information. This means that it is quite possible that users may have overlapping interests if the information available in a DL matches their expectations, backgrounds, or motivations. Such users might well profit from each other's knowledge by sharing opinions or experiences or offering advice. Some users might enter into long-term relationships and eventually evolve into a community if only they were to become aware of each other. CYCLADES is indeed a DL environment supporting collaboration and personalization at various level, where users and communities may search, share and organize their information space according to their own personal view and where the system generates recommendation of various types based on user and community profiles.

The outline of the paper is as follows. In the next section we will recall the main features of CYCLADES, while in Section 3 we will report some experimental results of the recommendation algorithms adopted within CYCLADES. Section 4 concludes.

2. A PERSONALIZED AND COLLABORATIVE DL

CYCLADES provides an integrated environment for users and groups of users (communities). The logical view of its functionality is depicted in Figure 1.

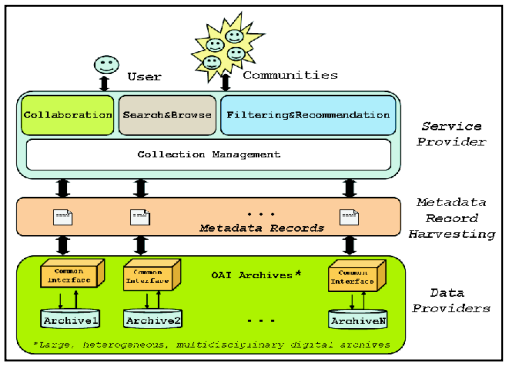

Figure 1. Logical view of CYCLADES

The digital archives to which CYCLADES users have access to are those adhering the *Open Archives Initiative* – OAI (http://www.openarchives.org). Informally, the OAI is an agreement between several digital archives providers in order to provide some minimal level of interoperability between them. In particular, the OAI defines an easy-to-implement gathering protocol over HTTP, which give *data providers* (the individual archives) the possibility to make the documents' metadata in their archives externally available. Indeed, the agreement specifies that each document of an archive should posses a *metadata* record describing the documents properties and content. In particular, the format of the metadata records should be DublinCore (http://dublincore.org). The metadata record consists of several attributes describing author, title, abstract, etc. of documents. The protocol allows then to gather these metadata records (in place of the real documents). A link to the "real" document is also present if the document is accessible. A metadata record may be understood as a statement of existence and short description of a document, which may be then accessible to a user according to the access policies of the archive, which owns the document. To date, there is a wide range of archives available (more than one hundred registered archives) in terms of its content, forming a quite heterogeneous and multidisciplinary information space.

The availability of the metadata records from the OAI compliant archives makes then it possible for *service providers* to build higher levels of functionality. In this sense, CYCLADES allows the access to the metadata provided by these archives, as it gathers these records, and through them provides access to the described documents (if they exist and their access is allowed). On top of them, CYCLADES acts as an OAI service provider providing functionality for (i) advanced search in *large, heterogeneous, multidisciplinary digital archives*; (ii) collaboration; (iii) filtering; (iv) recommendation; and (v) the management of records grouped into *collections*. These functionality are available in several environments described below.

The *Collaborative Work Environment*, which is an extension of the BSCW environment (Basic Support for Collaborative Work) see Bentley, R. 1997, provides a folder-based environment (Figure 2 shows the content of a user folder, in our case the "Physics-Gravity" folder of the community of physicists) for managing e.g. metadata records, queries, collections, external documents, ratings and annotations. In particular, users may organize their own information space according to their own hierarchy of folders. Each folder typically corresponds to one user related subject (or discipline, or field), so that it may be viewed as a thematic and usually semantically related repository of data items. There are two types of folders: (i) *private folders*, i.e. a folder owned by one user only. This kind of folder can only be accessed and manipulated by its owner. They are invisible to other users; and (ii) *community folders*, which can be accessed and manipulated by all members of the community that owns the folder. Community folders are used to share data items with other users and to build up a common folder hierarchy (rating, annotating, downloading and uploading of data items is permitted). Community folders may also contain *discussion forums* where notes may be exchanged in threaded discussions (similar to news groups).

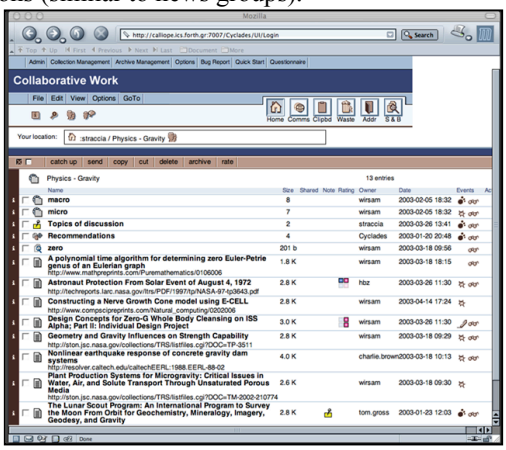

Figure 2. User interface: folder content

In order not to lose shared activity in the collaborative DL environment, mutual awareness can be supported through event icons displayed in the environment. Activity reports that are daily received by email are also be possible. Users can also view the list of all existing communities and can join a community directly if this is allowed by the community's policy, or contact the community administrator in order to be

invited to the community. In the collaborative work environment, the access policies can be set-up, as well as the notification (alerting) modalities.

The *Search and Browse Environment* supports the activity of searching records in the various collections accessible from within CYCLADES as well as to search into the shared folders or private folders a user owns. Users can issue a query and are allowed to store selected records within their folders and community folders they have access to. Essentially, three types of search are supported: (i) in *ad-hoc search* a user specifies a query and the system looks for relevant records within a specified collection; (ii) *filtered search* is like the usual ad-hoc search, except that the user specifies, additionally to a query (e.g. "zero"), also a target folder (e.g. "Physics-Gravity"). The goal of the system consist then to find documents not only relevant to the query, but also relevant to the topic of the target folder (in our example, the request is something like "find records about zero gravity"); and (iii) in *what's new, on-demand*, the user specifies a target folder, without specifying a query, and the goal of the system consists in finding all records, relevant to the target folder, which where become available to the system since the last time the user asked for this request. This corresponds roughly to the functionality provided by alerting services, except that the profile is build implicitly from the folder content, and that records are delivered to the user on-demand. The recommendation environment provides the off-line version.

The *Filtering and Recommendation Environment* supports the personalized search and provides the recommendations functionalities. All recommendations are specific to a given user folder (topic of interest), i.e. they have always to be understood in the context not of the general interests of the user, but of the specific interests (topic) of the user represented by a folder. A user may get recommendations of *metadata records* (suggesting to the user to access to relevant documents), *collections* (suggesting to the user to search within a relevant information space), *users* (suggesting to the user to enter in relationship with a user or give a look to the publicly available documents of the recommended user), and *communities* (suggesting to the user to join the community) issued to users based on user and/or community profiles.

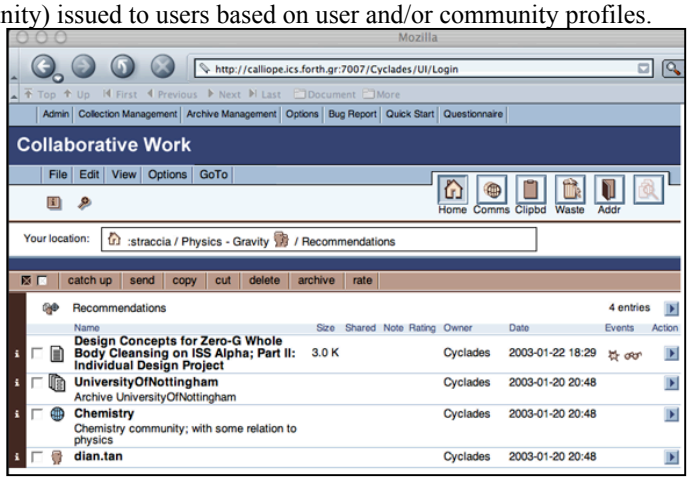

Figure 3. User interface: folder content and recommendation

For instance, Figure 3 shows the recommendations related to the "Physics-Gravity" folder, deemed by the system as relevant to this folder. Finally, the *Collection Management* manages collections (i.e. their definition, creation, and update). Its aim is to allow a dynamic partitioning of the information space according to the users' interests, where to search into. Usually, a collection is meant to reflect a topic of interest of a user or a community, e.g. the collection of records about "Information Retrieval". Informally, a collection specification is the definition of a not materialized view over the information space and it is up to the system to automatically determine the archives in which to search for relevant records (this is accomplished by means of a technique called *automated source selection* see Fuhr, N. 1999).

As pointed out, filtering and recommendation play an important role in making CYCLADES a personalized and collaborative environment. For giving an idea to the reader on how our algorithms work, in the following section, we will detail the algorithm used for metadata record recommendation. We also report some preliminary experimental results of its effectiveness. The recommendation algorithms of users and collections have been sketched out elsewhere (see Renda, M. and Straccia, U. 2002).

3. RECORD RECOMMENDATION & EXPERIMENTAL EVALUATION

In the following, consider a set of users u_k, a set of folders F_i of the users, and a set of available metadata records d_j. For ease, we consider a metadata record as a piece of plain text. Of course, more sophisticated algorithms can be devised by taking into account the metadata structure. Metadata records belong to folders and each user may rate a document within a folder he has access to through the user interface. With r_{ijk} we indicate the rating value given by a user u_k to record d_j, which is stored in folder F_i. We further assume that whenever a data item d_j belongs to a folder F_i of a user u_k, an *implicit* default rating r_{ijk} is assigned. Indeed, a record belonging to a folder of a user is an implicit indicator of being the record relevant to the user folder. Finally, we average the ratings r_{ijk} relative to the same data item--folder pair (i,j) and indicate it as r_{1ij} = mean $_{k>=1}$ $\{r_{ijk}\}$ (see matrix (c) in Table 1).

All records are indexed according to the well-known vector space model (see Salton, G. and McGill, J. 1983). With $d_j = < w_{j1}, \ldots, w_{jm} >$ we indicate its indexed representation, where $0 <= w_{jk} <= 1$ is the "weight" of term t_k in the record d_j (see matrix (a) in Table 1). The *folder profile*, is a machine representation of what a folder is *about* (denoted f_i) for folder F_i is computed as the *centroid*, or average, of the records belonging to F_i, i.e. $f_i = (1/|F_i|) \Sigma_{dj \in Fi} d_j$, thus, it is represented as a vector of weighted terms as well, i.e. $f_i = < w_{i1}, \ldots, w_{im} >$ (see matrix (b) in Table 1).

Table1: (a) The records matrix. (b) The folder profile matrix. (c) The folder-record rating matrix

	…	t_k	…
d_1	…	w_{1k}	…
…	…	…	…
d_j	…	w_{jk}	…
…	…	…	…
d_n	…	w_{nk}	…

(a)

	…	t_k	…
f_1	…	w_{1k}	…
…	…	…	…
f_i	…	w_{ik}	…
…	…	…	…
f_v	…	w_{vk}	…

(b)

	…	d_j	…
F_1	…	r_{1j}	…
…	…	…	…
F_i	…	r_{ij}	…
…	…	…	…
F_n	…	r_{vj}	…

(c)

By relying on matrix (a) of Table 1, the correlation among to rows establishes a similarity between records. Similarly, in matrix (b), the correlation among two rows establishes a correlation among folder profiles. In both cases, the content of records and folders is taken into account only. The measure used for content correlation, denoted CSim(.,.), is the well-know *cosine*}, i.e. the scalar product between two row vectors. By relying on matrix (c), a correlation among folders can be determined by taking ratings issued by users into account only. This similarity is called *rating similarity* of two folders F_i and F_j (denoted RSim(F_i, F_j) and is determined using the *Pearson correlation coefficient* (see Breese, J. et al. 1998), i.e.

$$RSIM(F_1,F_2) = \frac{\sum_{d_j \in D}(r_{1j} - \bar{r}_1).(r_{2j} - \bar{r}_2)}{\sigma_1 \bullet \sigma_2}$$

where $\bar{r}_i$ is the mean of the ratings $r_{i1} \ldots r_{in}$, and is standard deviation is σ_i

3.1 Recommendation algorithm

The objective of the recommendation algorithm is, given a user u and a folder F belonging to u, called the *target folder*, to recommend to F (and, thus, to the user) records relevant to the topic of folder F. Our recommendation algorithm follows a four-step schema: (i) select a set MS(F) of k-most similar folders to F, according to the similarity measures xSim (we can use either Csim, RSim or a combination of both); (ii) determine a pool P_D of candidate records (the number of records is an empirical value), i.e. $F_i \in$ MS(F); (iii) for each of the records $d_j \in P_D$ compute a recommendation score of d_j to F, where $\bar{r}(\bar{r}_i)$ is the mean of the ratings in the target folder F, i.e. the mean of the F row in matrix (c) (mean of $F_i \in$ MS(F) rating row); (iv) Recommend to folder F records having a positive score.

$$s^R(F,d_j) = \bar{r} + \frac{\sum_{F_i \varepsilon MS(F)} (r_{ij} - \bar{r}_i) \bullet RSim(F,F_i)}{\sum_{F_i \varepsilon MS(F)} RSim(F,F_i)}$$

3.2 Experimental evaluation

We tested our recommendation algorithm for effectiveness. Indeed, we first determine a *corpus* of data. From the corpus we select a *test set* of triples (F_i, d_j, r_{ij}, where F_i is the target folder, d_j is a record belonging to F_i and r_{ij} is its average rating. Second, *record recommendations* are given for each of the folders of the test set. Finally, we *analyze the results*.

As to date, neither there is yet a significant corpus within the CYCLADES system, build by real users, nor it was available during the development phase to "tune" our algorithms, nor there exists an available corpus from the literature, which fits to our setting, we build a corpus automatically.

The **corpus** was selected from the *Open Directory Project* hierarchy of (ODP or DMOZ) (http://dmoz.org). ODP is the largest human-edited directory of the Web. The ODP data includes over 3.8 million sites, about 60000 editors and over 460,000 categories. ODP powers the core directory services for the Web's largest search engines and portals, e.g. Google (http://www.google.com). Each category in ODP contains a set of Web documents, which have been evaluated by one or more editors for their relevance to the category. Furthermore, to each document within a category, Google assigns a score (using the PageRank algorithm, see Brin, S. and Page, L. 1998). We construct our corpus as follows. The set of users is the set editors of ODP. The set of records is the set of documents in ODP. The set of folders is the set of categories in ODP. To each record d_j in folder F_i, evaluated by user u_k, we set the rating r_{ijk} equal to the PageRank score s_{ij} assigned to record d_j w.r.t. folder F_i. This means that r_{ij}, the average rating over all users rating records d_j in folder F_i, is indeed s_{ij} (note that all users rate d_j in F_i equally. But this does not matter us, as in the recommendation algorithm just the mean r_{ij} is required). To limit the amount of data, we considered all the categories under Science/Math only, together with the involved records and users. The profiles of the folders has been restricted to the top weighted 100 terms.

To create the **test set**, we considered the set of all records D (corpus), which belong to at least two folders. For each of these records $d_j \in D$ (207 records), we randomly choose a folder F_i in which d_j occurs. The set of chosen folders $\{F_i\}$(195 folders), the records d_j and the relative average rating r_{ij} forms the test set.

Evaluation step is carried out as follows: for each target folder $F_i \in F$, we compute the set of recommended records d_j and consider their recommendation score $s^R(F_i, d_j)$. We measure the *coverage*, which is the percentage of records correctly recommended to the target folders, i.e. the system recommends a record to a folder and in the corpus effectively the record belongs to that folder. We also measure the *accuracy* of the system. Among the many metrics, which have been proposed for this purpose, we consider the widely used *Mean Absolute Error* (MAE) (see Good, N. et al. 1999, Konstan, J. 1997). The MAE is the mean of the absolute error among the prediction $s^R(F_i, d_j)$ of a document $d_j \in D$ and d_j's real rating value $r_{ij} \in F_j$, i.e. the MAE is the mean of $| s^R(F_i, d_j) - r_{ij} |$ over $F_i \in F$, $d_j \in D$. The purpose of the MAE is to determine how far away is the system predicted value from the real rating value of a record.

3.3 Results analysis

The results of our experiments are summarized below. We considered two different parameters: (i) the strategy to select the k-most similar folders w.r.t.. a target folder; and (ii) the impact of varying k. In case (i), we have three different options: either select similar folders by relying on rating similarity only (using RSim), or using content similarity only (using CSim), or to use both. In case (ii), k took one of the values 1,5,10,50,100,150 and 200. Figure 4 reports the coverage in using RSim (top-left) or CSim (top-right) only, for various k. RSim has a coverage about 32%, while CSim reaches about 51%. In the left graph, in each bar, the lighter part indicates the percentage of successfully recommended records using Rsim, which have not been recommended by using CSim (and vice-versa in the right graph). Interestingly, the graph shows that there is a large part of records, that have been recommended using rating similarity, but have not been recommended using content similarity. This may suggest that the combination of both strategies may improve the coverage significantly.

Indeed, Figure 4 (bottom) confirms this, as it shows a significant improvement of both coverage and accuracy. Finally, note that all cases reach a stability at k equals 100 or less. For k=100, combined approach covers almost 70% of test objects, performing 51.4% better than relying on RSim only and 25.2% better than relying on CSim only; on the MAE metric, the combined approach performs 46.2% better than "RSim" and 32.3% better than "CSim".

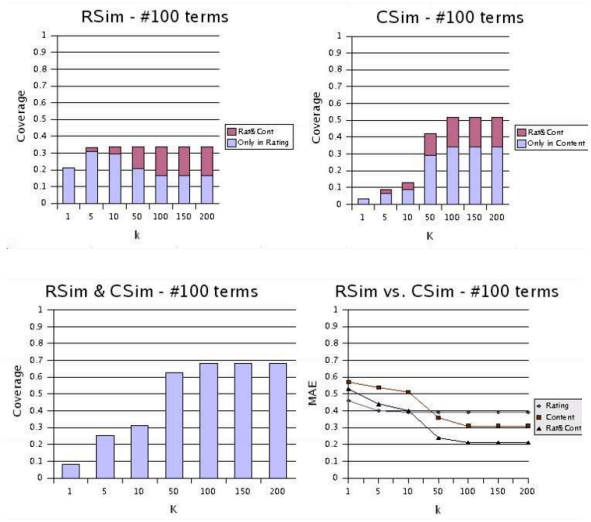

Figure 4. Coverage and Accuracy results

4. CONCLUSION

Every day a huge amount of information is electronically published. In this paper, we describe the Digital Library environment CYCLADES that is not only an information resource where users may submit queries to searching for relevant information, but also a personalized, and more importantly, a collaborative working and meeting space. The functionality provided by the CYCLADES system, can be organized into four categories: users may (i) search for information, not only by means of generic queries, but also by taking into account the learned user topics of interests; (ii) organize the information space according to the folder and personalized collections paradigm, which allow users to personalize the information space made available within CYCLADES; (iii) collaborate, in shared working space, with other users, which may have similar interests or more generally are related according to some purpose; and (iv) get recommendations from the CYCLADES system. CYCLADES not only provides recommendation of records, as it usually happens in personalization system dealing with documents, but by taking advantage of the highly collaborative environment, it may recommend also communities, collections and users as well. Particular attention has been paid to the recommendation part and some experiments showing its effectiveness.

ACKNOWLEDGEMENT

This work is funded by the European Community in the context of the CYCLADES project IST-2000-25456, under the Information Societies Technology programme.

REFERENCES

Bentley, R. et al. 1997. Basic support for cooperative work on the world wide web. *International Journal of Human Computer Studies*, (46):827—846.

Bollacker, K. Et al. 1999. A system for automatic personalized tracking of scientific literature on the web. *Proceedings of the Fourth ACM Conference on Digital Libraries*, pages 105--113, New York, ACM Press.

Breese, J. et al. 1998. Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 43—52, Madison, Wisconsin, USA.

Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems}*, 30(1-7):107—117.

Faensen, D. et al. 2001. Hermes: a notification service for digital libraries. *Proceedings of the ACM - IEEE Joint Conference on Digital Libraries*, pages 373—380.

Fernandez, L. et at. 2000. Mibiblio: personal spaces in a digital library universe. *ACM Digital Libraries*, pages 232—233.

Fox, E. and Marchionini, G. 2001. Digital libraries: Introduction. *Communications of the ACM*, 44(5):30—32.

Fuhr, N. 1999. A decision-theoretic approach to database selection in networked IR. *ACM Transactions on Information Systems*, 17:229—249.

Good, N. et al. 1999. Combining collaborative filtering with personal agents for better recommendations. *Proceedings of the American Association of Artificial Intelligence Conference*, pages 439—446.

Konstan, J. 1997. Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77—87.

Moukas, A. 1996. Amalthaea: Information discovery and filtering using a multiagent evolving ecosystem. *Proceedings of the Practical Applications of Agents and Multiagent Technology*, London, GB.

Renda, M. and Straccia, U. 2002. A personalized collaborative digital library environment. *Proceedings of the 5th International Conference on Asian Digital Libraries*, Number 2555 in Lecture Notes in Computer Science, pages 262—274, Singapore, Republic of Singapore. Springer-Verlag.

Rocha, L. 1999. Talkmine and the adaptive recommendation project. *ACM Digital Libraries*, pages 242—243.

Salton, G. and McGill, J. 1983. *Introduction to Modern Information Retrieval*. Addison Wesley Publ. co., Reading, Massachussetts.