

User Awareness in CYCLADES: a Collaborative and Personalized Environment for Open Digital Archives

Henri Avancini and Umberto Straccia

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

Abstract

Every day a huge amount of newly created information is electronically published in Digital Libraries. Complementary to the usual vision, we envisage a Digital Library not only as an information source where users may submit queries to satisfy their daily information need, but also as a collaborative working and meeting space of people sharing common interests. CYCLADES is a system, which combines several technologies from the Information Retrieval and Digital Library area, providing a highly personalized environment where not only users may organize (and search into) the information space according to an individual flavour, but, and more importantly, also provides support for advanced collaborative work. In particular, users may build communities, may become aware of each other, exchange information and knowledge with other users, and may get recommendations from the system based on preference patterns of users.

1 Introduction

A characteristic of the digital information age is that the amount of information published in electronic format, the services provided on it and the number of users accessing to it to satisfy their daily information need is growing at a tremendous rate. In this scenario *Digital Libraries* (DLs) [FM01] will play an important role in the next future not merely in terms of the 'controlled' digital information (the *content* of DLs) they allow access to, but especially in terms of the *services* they provide. Informally, DLs can be defined as consisting of collections of information, which have associated services delivered to users and user communities using a variety of technologies. The information accessible from DLs are usually heterogeneous both in content (addressing many aspects of human knowledge) and format, and can be represented as digital text, image, audio, video, or other media. This information can be digitized paper or born digital material. 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*.

Without doubt, DLs have evolved rapidly over the past decade and as DLs become more commonplace and the range of information they provide services upon increases, users' are expecting more and more sophisticated services from their DLs. In fact, more and more

DLs, rather than to provide a search facility to the digital society at large ¹, are going 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. For instance, towards this direction fall those DLs, which offer so-called *personalized alerting services* (see e.g. [BLG99, FFS⁺01, FSG00, Mou96, Roc99]), 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. [FSG00]). In particular, they usually allow to organize the data retrieved in a DL into thematic folders, like computer users do within their own computer. This is important as 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 the DL manager thought would be well suited for anyone.

Our vision is that, DLs can also be considered as *collaborative meeting place* of people sharing common interests. Indeed, DLs may be viewed as a *common working places* where users may become aware of each other (indeed a DL may find out interesting relationships both between users and/or between communities of users and produce the appropriate recommendations/advice), open communication channels, and exchange information and knowledge with each other or with experts. In fact, 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. Such a service might be important for a DL as it supplies very focused information. Hence, we are moving from services supporting an individual user towards services supporting *groups* (or a *community*) of users: thus, move from the study of individual human behaviour towards the discipline concerned with the study of human behaviour in groups and the technical support thereof. More fundamentally, we make a conceptual shift in our understanding of DLs: whereas the classical view of DLs was manipulation of data by isolated individuals, our view of DLs is manipulation and exchange of data and information as well as cooperation by individuals aware of their environment as well as other users.

We have developed a system, named CYCLADES ², whose aim is to provide advanced services for both personalization and collaborative work. The description of its features are one of the topics of this paper. Besides, we will present some experimental results showing the effectiveness of CYCLADES in relating users with similar interests to each other. This is important, as described above, as it may help users to get aware of other users sharing similar interest and to enter maybe into long-term relationships.

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 user recommendation algorithms adopted within CYCLADES. Section 4 concludes.

¹They are oriented towards a generic user, as they answer queries crudely rather than learn the long-term requirements of a specific user.

²<http://www.ercim.org/cyclades>

2 CYCLADES

The CYCLADES system offers a broad range of functionality for both individual scholars, who want to search and browse in digital archives, as well as for communities of scholars who want to share and exchange information. A multi-disciplinary and distributed team designed this functionality with backgrounds in digital libraries, databases, information retrieval, Web-based systems, as well as computer-supported cooperative work and virtual communities. After a first suggestion for the basic functionality of the system was finished, this functionality was presented to prospective users, and their feedback was captured via a Web-based questionnaire.

The CYCLADES system integrates a set of functionalities that support the user when accessing very large virtual e-print archives with: functionality for efficient and effective retrieval of relevant information from many large, distributed and multi-disciplinary digital archives; feedback on the degree of relevance of the retrieved information; regular information about new publications in the archive environment which are relevant to the users' interests; automatic retrieval of users' long-term information needs; as well as rapid dissemination of the search results world-wide. A special set of features provides communities of scholars with functionality for: the dissemination of relevant information to community members in the form of recommendations, which are based on collective profiles and behaviour; very quick on-line annotations on research results published by members of the community; carrying out community services such as peer review, which requires the annotation of on-line manuscripts by reviewers, and the sharing of these annotations among editors, authors, and others; as well as enabling community members to learn from, contribute to, and collectively build upon the community's knowledge.

The digital archives to which CYCLADES users have access to are those adhering the *Open Archives Initiative*³ (OAI). 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 possess a *metadata* record describing the documents properties and content. In particular, the format of the metadata records should be Dublin Core⁴. The metadata record consists of several attributes describing author, title, abstract, etc. of a document. 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. In CYCLADES we gather these records periodically from the archives.

In order to use the CYCLADES environment and its functionality users have to register with the system. As the user interface is Web-based, users simply enter a valid email address in the registration Web page. After the user has carried out these steps, the user is registered and has a login name, password, and (empty) home folder. CYCLADES provides a folder-based environment (Figure 1 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, external documents, ratings and annotations. In particular, users may orga-

³<http://www.openarchives.org>.

⁴<http://dublincore.org>

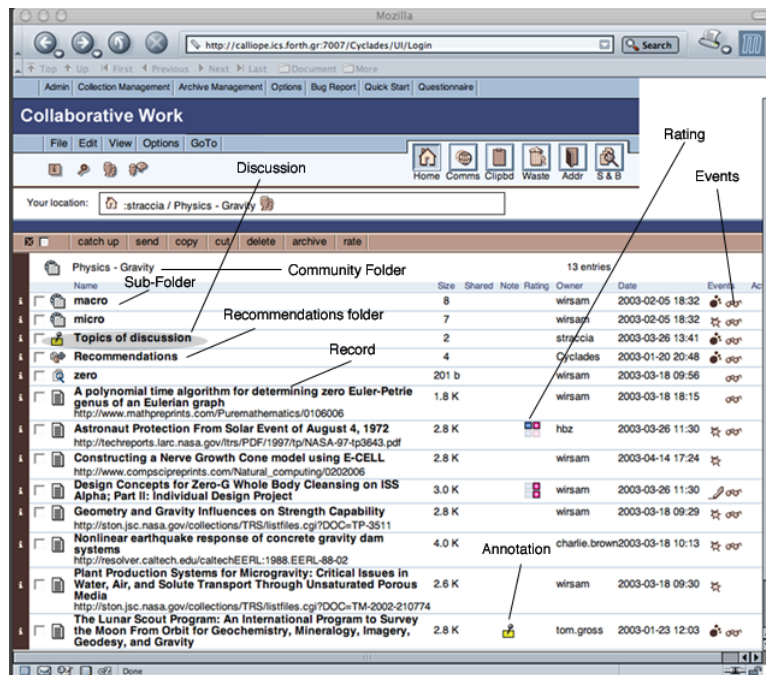


Figure 1: User interface: folder content

nize 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. This environment is an extension of the Basic Support for Collaborative Work environment (BSCW-<http://www.bscw.de>) [BAB⁺97]⁵. Users can then login in order to access their home folder or add some additional user information such as full name, affiliation, postal address, phone and fax numbers, and so forth. Furthermore, users can start working with folders and folder contents. They can create *private folders*, *community folders*, as well as *project folders* and add subfolders and documents to them. Private folders and community folders can contain metadata records and queries; project folders additionally can contain any type of document (e.g., PDF-files, slides, text files). The contents of private folders can be seen, accessed, and changed only by its owner. The contents of community and project folders can be seen, accessed, and changed by the community and project members respectively. Users can add documents in various ways: they can store records they found in a search and browse activity, they can upload documents from the local hard disks, they can add links, and so forth. Community folders may also contain discussion forums where notes may be exchanged in threaded discussions (similar to news groups). All documents can be rated on a simple scale. The median of the ratings is then attached to the document and visible for any user who has access to the respective document. Additionally, annotations containing free text can be added to documents. Annotations can also be seen by any user who has access to the respective document.

In order to become member of an existing community users have two options: either they browse a list of names and descriptions of open communities and subscribe to them according

⁵It is not accidental that the BSCW development team was member of the project.

to their interests or they get invited by community managers of closed communities. After a user has become a member, the community folder is visible and accessible from the user's home folder. If users leave a community, the community folder is removed from their home folder, but is still available for the other community members. Only when the last community member leaves the community, the community folder is actually deleted.

In order not to lose shared activity in the collaborative DL environment, mutual awareness is supported. Indeed, the CYCLADES environment provides two features for informing users about activities in the system: event icons and activity reports. Event icons are attached to the individual documents and show recent actions that were performed on the documents (e.g., creation, change, read). Activity reports are sent out daily via email; they contain information about changes to documents since the last report was sent out.

Besides the possibility to search into the folders a user has access to, advanced functionality of searching records in the various collections accessible from within CYCLADES are provided. Users can issue a query and are allowed to store selected records within the 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 specified collections; (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 one of his folders, without specifying a query, and the goal of the system consists in finding all records, relevant to the 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 topic a folder is about is build automatically from the folder content (we call the topic of a folder, *folder profile*), and that records are delivered to the user on-demand.

An important service in CYCLADES is the automatic delivering of recommendations to a user. The key point is that 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. For instance, Figure 2 shows the recommendations related to the 'Physics-Gravity' folder, deemed by the system as relevant to this folder. In particular, the user Dian is considered by the system to have overlapping interests with 'Physics' and 'Gravity'. CYCLADES provides several types of recommendations. A user folder may get recommendations of (i) collections, suggesting to the user that the recommended collections contain relevant data with respect to the folder and, thus, it may worth to search within them; (ii) metadata records, in this case the system is suggesting that the recommended records are relevant to the folder content; (iii) communities, suggesting to the user to join communities, as they may deal with user related topics of interests; and (iv) 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, as he may have overlapping topics of interests.

In the following we give an idea to the reader on how our recommendation algorithms work and will devote a special attention to the recommendation of users, which is indeed an important way in CYCLADES to relate users with similar interests to each other and may help users to get aware of other users sharing similar interests.

The recommendation of collections to a folder consists in automatically determine the archives in which to search for relevant records to the folder. It is accomplished by means of a technique called *automated source selection* (see, e.g. [Fuh99]). Roughly, it consists in (i) the computation of an approximation of the content of each OAI compliant archive (statistical information about the content of an information source, see [CC01]) registered in

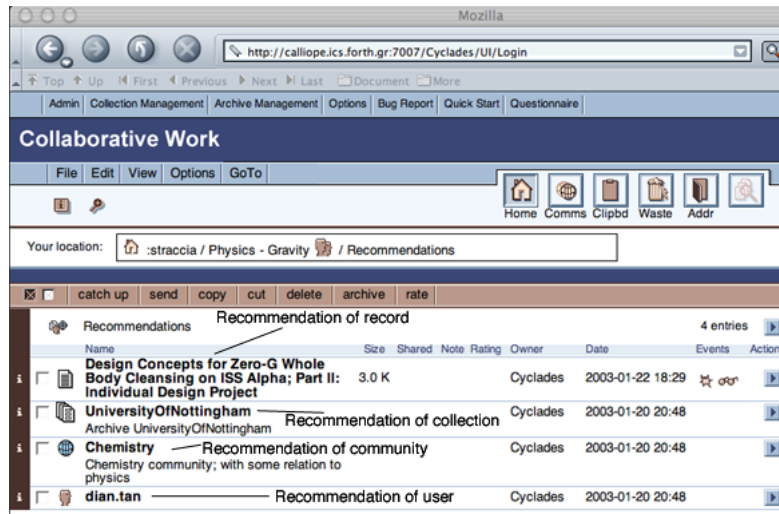


Figure 2: User interface: recommendations to a folder

CYCLADES; and (ii) the selection of those archives deemed as most relevant to a folder, by relying on the approximations of the archives' content and on the folder profile. Both steps are done periodically (please refer to [CS03], for a detailed description).

The recommendation of records to a user folder consists in automatically determine the records deemed as relevant to a folder and is accomplished roughly in four steps as follows (please refer to [AS03], for a detailed description): (i) select a set of most similar folders to the user folder, according to a similarity measure between folder profiles (see next section); (ii) from this set, determine a pool of possible recommendable records, which consists of the records belonging to the similar folders ; (iii) for each of the records in the pool compute a recommendation score; (iv) select and recommend a subset of records with highest score, and not yet recommended to the user folder.

The recommendation of communities to a user folder consists in automatically determine the communities, dealing about topics relevant to the topic of the user folder. It is accomplished roughly similarly to the recommendation of records (the reader may refer to [RS02], for a more detailed description). While for the recommendation of records and collections, we have conducted experiments in order to evaluate the effectiveness of the adopted algorithms, this is not yet the case for the recommendation of communities, to which we will devote major effort in future work.

In the following section, we will detail the algorithms used for user recommendation. We also report some experimental results of their effectiveness.

3 User recommendation and experimental evaluation

In the following, consider a set of users u_k , a set of folders F_i and a set of available metadata records d_j distributed in the folders. 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 also rate a document within a folder he has access to. 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 folder–document pair (i, j) and indicate it as $r_{ij} = \text{mean}_{k \geq 1} \{r_{ijk}\}$.

All records are indexed according to the well-known vector space model [SM83]. With $d_j = \langle w_{j1}, \dots, w_{jm} \rangle$ we indicate its indexed representation, where $0 \leq w_{jk} \leq 1$ is the ‘weight’ of term t_k in the record d_j . The *folder profile*⁶ (denoted f_i) for folder F_i is computed as the *centroid*, or average, of the records belonging to F_i , i.e. $f_i = (1/|F_i|) \sum_{d_j \in F_i} d_j$, thus, it is represented as a vector of weighted terms as well, i.e. $f_i = \langle w_{i1}, \dots, w_{im} \rangle$. The *user profile* of a user u (denoted p_u) is built as the *centroid* of the user’s folder profiles, i.e. if \mathcal{F}_u is the set of folders belonging to the user u , then $p_u = (1/|\mathcal{F}_u|) \sum_{F_i \in \mathcal{F}_u} f_i$. Therefore, the user profile is represented as a vector of weighted terms as well, i.e. $p_u = \langle w_{u1}, \dots, w_{um} \rangle$. By relying on the vector representation of records, folders profiles and user profiles, we can easily determine a similarity measure between them. Indeed, the similarity among two vectors (whether records, folder profiles or user profiles) is computed as *content* correlation (denoted $CSim(\cdot, \cdot)$) and is the well-know *cosine* angle among the two normalized (norm-2) vectors, i.e. is the scalar product between two vectors, e.g. $CSim(f_1, f_2) = \sum_{k=1}^m w_{1k} w_{2k}$. Another correlation among folders can be determined by taking the ratings issued by users into account only. This similarity is called *rating similarity* of two folders F_1 and F_2 (denoted $RSim(F_1, F_2)$) and is determined using the *Pearson correlation coefficient* [BHK98], i.e. $RSim(F_1, F_2) = \frac{\sum_{d_j \in \mathcal{D}} (r_{1j} - \bar{r}_1) \cdot (r_{2j} - \bar{r}_2)}{\sigma_1 \cdot \sigma_2}$, where \bar{r}_i is the mean of the ratings $r_{i1} \dots r_{in}$, and σ_i is their standard deviation. The combined similarity among two folders is then obtained by taking into account the content similarity and the rating similarity. In what follows, the *combined similarity* or simply *similarity* (denoted $Sim(F_1, F_2)$) between two folders F_1 and F_2 will be determined as a linear combination between their content similarity and their rating similarity, i.e. $Sim(F_1, F_2) = \alpha \cdot CSim(f_1, f_2) + (1 - \alpha) \cdot RSim(F_1, F_2)$, where $0 < \alpha \leq 1$.

3.1 User recommendation algorithm

The goal of the user recommendation algorithm is, given a folder F (called *target folder*) of user u , to recommend to F (and, thus, to user u) those users, which by the system are thought to have overlapping interests with the topic addressed by the folder F (and, thus, may be related to user u). We have analyzed three different algorithms, with increasing level of effectiveness. We start with the first one implemented in the CYCLADES system. It is described in [RS02], but no experimentation has been worked out before. The recommendation algorithm follows a four-step schema, like for the record recommendation case:

1. select the set $MS(F)$ of s -most similar folders to F , according to a similarity measures. We can use either $CSim$, $RSim$ or the combination of both (Sim), but in [AS03] we have already observed that Sim works better, so we use it here as well;
2. from this set of similar folders, determine a pool $P_U(F)$ of candidate users to be recommended, i.e. let $P_U(F)$ be the set of users being owners of the folders in $MS(F)$;
3. compute the *recommendation score* for each possible recommendable user, i.e. for each user $u_k \in P_U(F)$ determine the *user hits factor* $h(u_k) = |\{F_i : F_i \in MS(F), F_i \in u_k\}|$ (the number of folders F_i judged as similar to the target folder F belonging to the

⁶The folder profile is a machine representation of what a folder is *about*.

same user u_k). For each user $u_k \in P_U(F)$ the recommendation score p_{u_k} is computed as follows: $p_{u_k} = h(u_k) \cdot \sum_{F_i \in MS(F), F_i \in u_k} Sim(F, F_i)$;

4. recommend to folder F , the top- n ranked users, ranked according to the recommendation score.

The intuition behind Step 3 is that the more a user appears among the owners of the top- s similar folders, the more he is considered as relevant to the target folder.

The second algorithm is a variation of the first one in which Step 3 is replaced with:

3. for each user $u_k \in P_U(F)$, consider the profile of u_k , p_{u_k} , and compute the recommendation score as the similarity between the user profile p_{u_k} and the target folder profile f , i.e. $p_{u_k} = CSim(f, p_{u_k})$;

The intuition here is to use the user profile of recommendable users $u_k \in P_U(F)$ directly in place of the folder profiles of similar folders.

Finally, the third algorithm does not consider the set of similar folders, but just compares the target folder profile against all user profiles. Note that in this way, no ratings are taken into account. Therefore, we remove Step 1, and in Step 2, the pool of candidate users, $P_U(F)$, is given by all users known to the system, and Step 3 is as in algorithm 2.

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 pairs (F_i, u_k) , where F_i is the target folder and u_k is a user having folder F_i in his folder hierarchy. Second, *user recommendations* are given for each of the folders of the test set. Finally, we *analyse the results*.

As to date, neither there is yet a significant corpus within the CYCLADES system built by real users ⁷, 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.

Corpus: The corpus was selected from the *Open Directory Project* hierarchy of (ODP or DMOZ) ⁸. 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 ⁹. 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 [BP98] algorithm). 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} ¹⁰. 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. Totally, we have 500

⁷To date, CYCLADES has 59 users sharing 118 folders, which contain 284 records with 35 (non default) ratings.

⁸<http://dmoz.org>

⁹<http://www.google.com>

¹⁰Note 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.

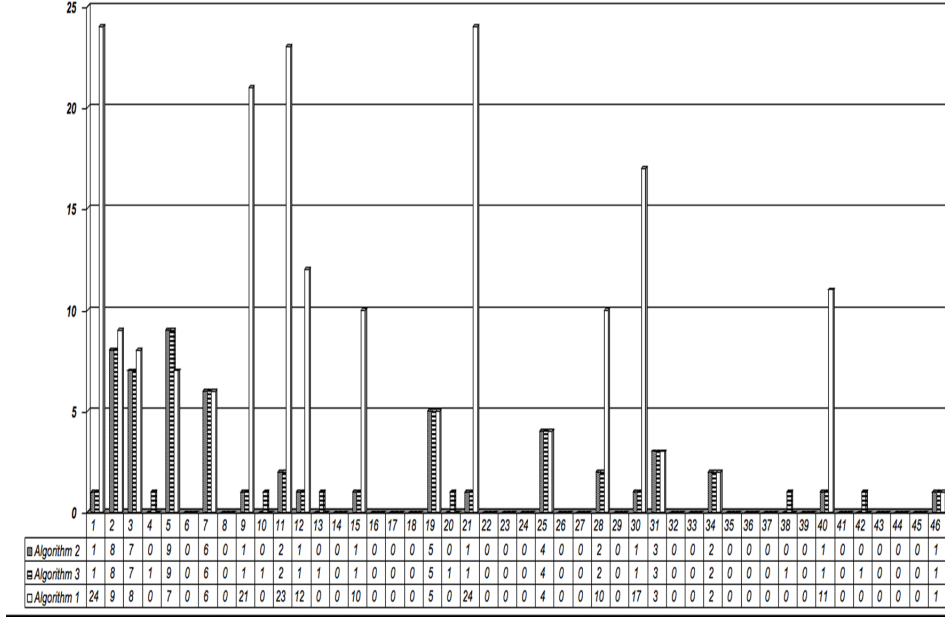


Figure 3: Rank position of correctly recommended users

folders distributed among 46 users, sharing 8083 documents, each having at least one rating associated. As the initial set of folders a user has access is small, we constructed a random algorithm, which taken a folder F and a user u having access to F , randomly selects a subtree of F to which u has access as well. For *Sim*, we used $\alpha = 0.5$ and $s = 100$.

Test set: To create the test set, we considered the set of all users \bar{U} of the corpus, which have access to at least two folders. For each of these users $u_k \in \bar{U}$ (46 users), we randomly choose a folder F_i of u_k . The set of chosen folders $\bar{F} = \{F_i\}$ (35 folders) and the users u_k , i.e. (F_i, u_k) forms the test set. Totally, we have 46 users to be recommended distributed over 35 folders.

Evaluation. For each target folder $F_i \in \bar{F}$, where (F_i, u_k) belongs to the test set, we compute the set of recommended users $\{u_k\}$ and ranked them according to their recommendation score p_{u_k} . In this rank, we highlight the rank position of user u_k . If the recommendation score is 0, user u_k is ranked 0. Additionally, we determine their average rank, precision and recall. Totally, we performed 35 (number of folders) tests.

3.2.1 Results analysis

The results of our experiments are summarized in Figure 3, and Table 1, where for each test pair (F_i, u_k) we report the rank position of user u_k , according to our recommendation algorithms 1, 2 and 3, respectively. It turns out that algorithm 3 performs best, algorithm 2 is second, while the less effective one is algorithm 1. As we can see, for algorithms 2 and 3, in many cases the correctly recommended user is highly ranked. For all three algorithm, we also computed *recall* and *precision*, by recommending the top- n users for each target

Table 1: Effectiveness measures

	algorithm	top-1	top-2	top-5	top-10
Precision	1	0.03	0.03	0.03	0.03
	2	0.23	0.16	0.08	0.05
	3	0.40	0.24	0.11	0.07
Recall	1	0.02	0.04	0.11	0.24
	2	0.17	0.24	0.30	0.39
	3	0.30	0.37	0.43	0.52
F1	1	0.02	0.03	0.05	0.06
	2	0.20	0.19	0.13	0.09
	3	0.35	0.29	0.18	0.12

algorithm	average rank
1	10.94
2	3.11
3	2.58

folder, where $n \in \{1, 2, 5, 10\}$. Precision is the fraction between correctly recommended users and the total amount of recommended users, *i.e.* $Precision_n = \frac{|CorRec_n|}{35 \cdot n}$, where $CorRec_n$ is the set of correctly recommended users among all 35 tests and n is the number of recommended users for each test. Recall, is the fraction between correctly recommended users and the number of test users, *i.e.* $Recall_n = \frac{|CorRec_n|}{46}$. Of course, the more users we recommend (*i.e.* n increases), the more correctly recommended users we have (recall improves), but the less precise we are (precision decreases). We also report the standard measure $F1_n = \frac{2 \cdot Precision_n \cdot Recall_n}{Precision_n + Recall_n}$, which gives us an estimate of the combination between precision and recall. It turns out that algorithm 3 is the most effective one and, interestingly, that recommending just the top ranked user is the most satisfactory compromise between precision and recall.

4 Conclusions

In this paper, we describe the Digital Library environment CYCLADES that is not only an information source 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 also recommend communities, collections and users as well. Particular attention has been paid to the user recommendation part and some experiments showing its

effectiveness.

Acknowledgments

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

- [AS03] Henri Avancini and Umberto Straccia. Personalization, collaboration, and recommendation in the digital library environment CYCLADES, 2003. Submitted.
- [BAB⁺97] R. Bentley, W. Appelt, U. Busbach, E. Hinrichs, D. Kerr, K. Sikkell, J. Trevor, and G. Woetzel. Basic support for cooperative work on the world wide web. *International Journal of Human Computer Studies*, (46):827–846, 1997.
- [BHK98] John S. Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)*, pages 43–52, Madison, Wisconsin, USA, 1998.
- [BLG99] Kurt Bollacker, Steve Lawrence, and C. Lee Giles. A system for automatic personalized tracking of scientific literature on the web. In *Digital Libraries 99 - The Fourth ACM Conference on Digital Libraries*, pages 105–113, New York, 1999. ACM Press.
- [BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.
- [CC01] Jamie Callan and Margaret Connell. Query-based sampling of text databases. *ACM Transactions on Information Systems*, 19(2):97–130, 2001.
- [CS03] Leonardo Candela and Umberto Straccia. The personalized, collaborative digital library environment CYCLADES and its collections management. In J. Callan, F. Crestani, and M. Sanderson, editors, *Multimedia Distributed Information Retrieval*, pages –. Springer Verlag, 2003. To appear.
- [FFS⁺01] Daniel Faensen, Lukas Faulstich, Heinz Schweppe, Annika Hinze, and Alexander Steidinger. Hermes: a notification service for digital libraries. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 373–380, 2001.
- [FM01] Edward A. Fox and Gary Marchionini. Digital libraries: Introduction. *Communications of the ACM*, 44(5):30–32, 2001.
- [FSG00] Lourdes Fernandez, J. Alfredo Sanchez, and Alberto Garcia. Mibiblio: personal spaces in a digital library universe. In *ACM DL*, pages 232–233, 2000.
- [Fuh99] Norbert Fuhr. A decision-theoretic approach to database selection in networked IR. *ACM Transactions on Information Systems*, 3(17):229–249, 1999.
- [Mou96] Alexandros Moukas. *Amalthea*: Information discovery and filtering using a multiagent evolving ecosystem. In *Proceedings Practical Applications of Agents and Multiagent Technology*, London, GB, 1996.
- [Roc99] Luis Mateus Rocha. Talkmine and the adaptive recommendation project. In *ACM DL*, pages 242–243, 1999.
- [RS02] M. Elena Renda and Umberto Straccia. A personalized collaborative digital library environment. In *5th International Conference on Asian Digital Libraries (ICADL-02)*, number 2555 in Lecture Notes in Computer Science, pages 262–274, Singapore, Republic of Singapore, 2002. Springer-Verlag.
- [SM83] Gerard Salton and J. Michael McGill. *Introduction to Modern Information Retrieval*. Addison Wesley Publ. Co., Reading, Massachusetts, 1983.