COLLABORATIVE INFORMATION GATHERING

M. Antonietta Grasso, Uwe M. Borghoff, Natalie Glance, Jutta Willamowski

Rank Xerox Research Centre, Grenoble Laboratory
6, chemin de Maupertuis. F-38240 Meylan, France
E-mail: <lastname>@grenoble.rxrc.xerox.com

Key Words: collaboration, constraint-based knowledge brokers, knowledge pump, information gathering.

Abstract: The combination of push and pull information retrieval technology can benefit from collaborative information gathering schemes. In this paper we outline an approach that combines a pull technology (*Knowledge Brokers*) with a push technology (*Knowledge Pump*) in a community-based environment. Appropriate toolkits supporting synchronous collaboration on the World Wide Web are briefly introduced. A *PlaceWare*-based implementation illustrates the synchronous construction of queries and the publication of search activities and results supported by the push component.

1 Introduction

Information retrieval technology is commonly positioned as push or pull, where push is the passive receiving of information and pull is the active search of it. We propose in this paper that the two are complementary and fit well into a community-centered scenario. This already happens in our everyday working life: either we need some particular information and actively look for it - most of the time on the basis of suggestion from people we know to be an expert of the field - or we receive spontaneously information from the environment - mostly from social and working interactions.

The constraint-based *Knowledge Brokers* model [ABP+95, ABP96] is typical of pull technologies: it provides a uniform interface to search engines [BoSc96, CBC97], asynchronously searched by issuing queries [BCW96]. In the current implementation the system does not provide any facility either to involve other people in the definition of the query, to evaluate the results, or to publish them (beside basic functionality to export them in HTML format).

The *Knowledge Pump* [GAD98] is typical of push technologies: it provides a community centered environment supporting the (asynchronous as well as synchronous) flow and use of knowledge among the community members on the basis of their interactions and recommendations.

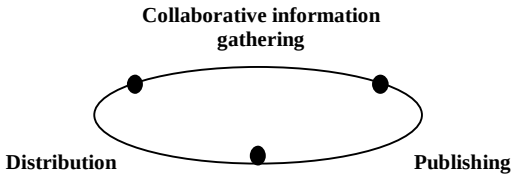

Fig. 1: Life cycle of information searches

As depicted in Figure 1, the activity of searching is one step of a more complex social process:

- each information search is not a stand-alone entity, but is part of a permanent chain where advice about *where*, *what* and *how* to search comes from other people [COT+97];
- information searches are shared among people by updating team membership, consulting, communicating and putting the results in shared archives.

In particular, the outcomes of the case study reported in [DaJe93a, DaJe93b] provide some insight into how information search processes are shared. Four basic models were identified:

1. sharing results with other members of a team;
2. self-initiated broadcast of interesting information encountered in search results: this pattern is used when unexpected but interesting tidbits

are encountered and circulated to people across the organization expected to find them useful. Very often the relevant information is extracted before passing it along. This pattern implies understanding the profiles of the people as well elaborating a proper format for circulating it (e.g. a spreadsheet with financial information). In effect a member of several communities publishes and distributes information where appropriate;

3. acting as a consultant, handling repeated or ad hoc information search requests made by others. Communities ask a member with the proper profile to act as a consultant;

4. archiving potentially useful information into a group repository.

An interesting point raised in the case study is that only rarely the results were not shared in any way.

On the base of these social observations, in this paper, we propose that Knowledge Brokers, extended by the Knowledge Pump, can be made useful as a cooperative tool provided that it is integrated in a community-oriented environment in such a way that the queries and results become permanent shared resources (see Figure 2).

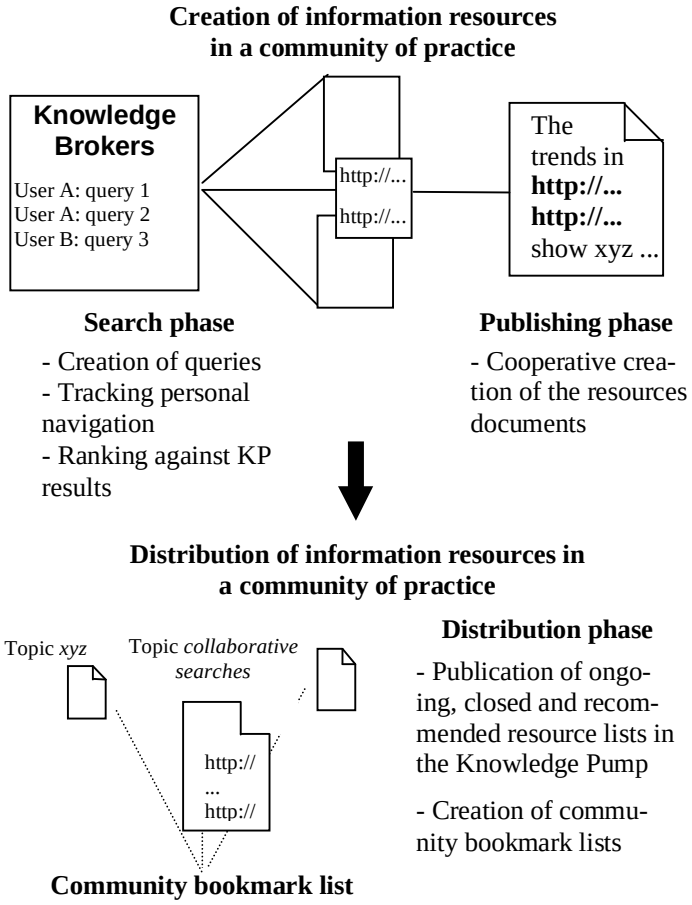

Fig. 2: Different types of collaboration

The paper is organized as follows. Section 2 describes the push and pull technologies used in the system and Section 3 presents how the two technologies can be integrated in a community-based environment. Section 4 describes some of the toolkits that could be used for combining a push and a pull information retrieval technology through the support of synchronous collaboration on the World Wide Web. Section 5 outlines the prototype and Section 6 concludes the paper presenting future direction of research.

2 Information *push* and *pull* technologies

2.1 Knowledge Pump

The goal of the Knowledge Pump is to help communities, defined by their common interests, more effectively share knowledge, be it in the form of documents or new ways to get work done. On one hand, users can submit ratings of items they come across in their everyday work via the Pump. On the other hand, users receive personalized recommendations of items contributed by others. Users are encouraged to classify items into domains of interest, which provides a first level of filtering and helps move the focus onto the community.

The Pump personalizes recommendations in two ways. First, it uses automated collaborative filtering [Res94] to get a baseline prediction of a person's likely interest in the item: effectively, automated collaborative filtering uses statistical methods to match items to people by first matching people to each other. Secondly, it refines its prediction by weighing more heavily the ratings submitted by those people whose opinion the user most values (as identified explicitly by the user).

The collaborative search environment will provide additional information regarding a user's interests and preferences, and, in turn, the user profile data gathered by the Pump can be used to inform collaborative search sessions.

2.2 Knowledge Brokers

The Knowledge Brokers are a hierarchically structured set of agents that allow—using an attributed-based graphical user interface—to access and query a variety of heterogeneous data repositories. A layer of so-called *wrappers* hides the idiosyncrasies of the different data repositories together with their differences in the search syntax, the ranking schemes, and the formats of the corresponding result sets. In some sense, a user conducting queries using Knowledge Brokers gets the look and feel of a "*federated*", homogenized database.

Local constraint solvers and adaptive filtering techniques are used to complement—and/or extend—the query functionality of the search subsystems of the data repositories involved.

Therefore, the Knowledge Brokers system is a kind of meta-search engine realizing the pull metaphor. It is beyond the scope of this paper to describe the system in more detail. Elsewhere (http://www.xerox.fr/research/ct/research/cbkb.html) the interested reader will find a publication list and on-line papers on Knowledge Brokers.

3 Towards an integrated environment

The user of any system can be seen as a single user sitting in front of a machine, or, when enlarging more the view, as a member of an organizational structure. Indeed, when looking more carefully, each user can be seen as member of a *community of practice* as defined in [LaWe91].

When the context of the user activities is enlarged and seen in terms of belonging communities, this can lead to the design of more powerful systems, because:

- information about the structure of communities can be used as meta-information, e. g. to suggest groups of users sharing the same interests;
- active participation in the activities of the communities can be explicitly supported, e.g. providing awareness mechanisms about activities of other community members;
- the notions of reciprocity, trust and reputation [GAD98] can be leveraged, e.g. as in the Knowledge Pump.

Among the possible metaphors that can be used to host a community in the virtual space, the metaphor of a physical space has been studied as an effective way of supporting cooperation of virtual communities. The MUDs environments [Cur93, Cur97] support this metaphor and despite there having been first conceived as a recreational space, they have been studied for their cooperation possibilities also in work settings.

What is particular to MUDs is the familiar space based metaphor in which users can manipulate and share information in a synchronous and asynchronous manner. This allows all the collaborative aspects of physical libraries, while adding three main advantages:

- geographically distributed cooperation;
- awareness about others' activities;
- persistent state of shared related information (both meta-information about search format and search content).

So far no complete MUD environment is available on the Web, because of all the limitations the Web infrastructure has in terms of supporting synchronous applications. However some support in developing cooperative applications using the physical metaphor is starting to become available. As we will see this has influenced our choice of the PlaceWare toolkit as development environment. In fact even if PlaceWare is not a real MUD, it offer support for building applications where the physical metaphor can be fully exploited.

3.1 Information generated by the Knowledge Brokers

Usage of the search engine generates two kinds of information: *about* the user and *for* the user. Moreover, the process of searching itself generates a third set meta-information which reflects how a person uses the system: what she looks for, which information sources she finds relevant, which topics are unfamiliar, which topics are of on-going interest.

The information generated *for* the user are the search results enhanced by personalized information in terms of relevance, accuracy and appropriateness.

The meta-information about the search process describes the steps taken to get to the information. This meta-information includes not only the choice of the sources to search over, but also a record of set of queries, keywords and any other manipulations supported by the system.

3.2 Information generated by the Knowledge Pump

The Pump also generates several types of information both for and about the user that can help both guide searches and cull search results. For instance, automated collaborative filtering uses person-person correlation to predict ratings. This person-person correlation could be used to help rank the advice of one person with respect to another in a collaborative session.

Secondly, some of the results of a search may already be in the Pump database, in which this information can be used to help rank the results to a query.

Finally, the Pump maintains user profiles of its members, information that is explicitly entered by the user and consists principally of:

1. the user's domains of interest; and

2. those people whose opinion the user most values. This explicit information can be combined with the implicit calculation of user preferences and interests to improve the collaborative search process.

3.3 The community based integrated environment

The third element of the integrated system is a community-based environment featured to represent a *community of practice* [LaWe91] in the virtual world. In our integrated system it is the glue that enables the two systems to enrich each other by linking and elaborating the information generated in a "community-oriented" way (see Figure 3).

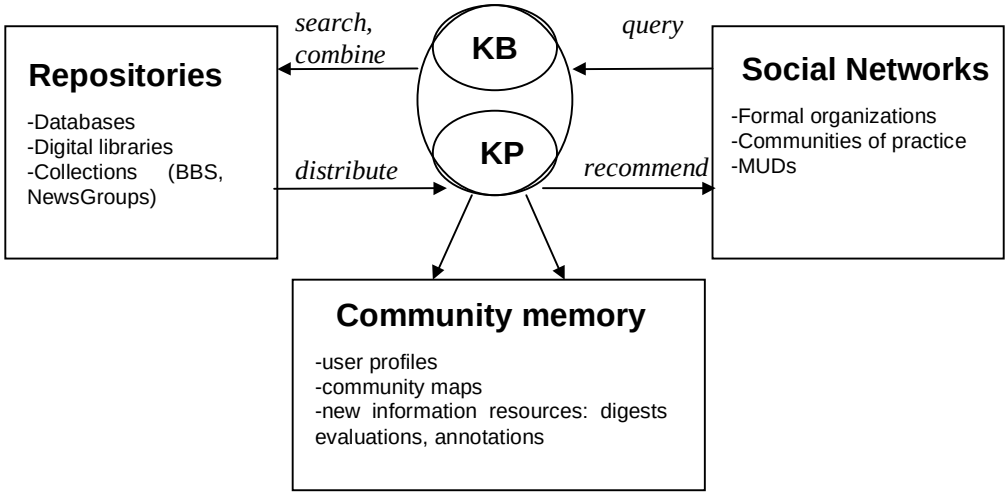

Fig: 3 Community-based environment

Shared sessions of the Knowledge Brokers

Instead of being a single user search engine Knowledge Brokers becomes a shared tool of the environment that provides shared manipulation of queries and results. To share the two across the network makes possible various kinds of interaction. For instance, experts can be brought in and asked about how to better use the system. Sessions can be stored in the environment, allowing cooperation with other users to be both synchronous and asynchronous. These two ways to share a session can be particularly useful in the case of the expert-peer cooperation because requests of help (shaped as ongoing sessions) can be responded to when the expert is available.

Awareness about activities of the community

The Knowledge Pump can be used as a channel to let the other members of the community be aware of each other's ongoing activities. This allows

1. a better communication among the members of the community, because duplicate efforts can be avoided
2. peripheral participation, because activities of the others can be "seen" without direct involvement
3. and forms of spontaneous cooperation because the electronic environment brings in a mode for synchronous activities.

Submission of search objects in the Knowledge Pump

The outcomes of search sessions can be made part of the community memory by publishing them as complete documents and letting them flow to members of possibly more communities than those involved in its construction. At this stage, comments and other reviews of the search results can then be used to refine future iteration of the search.

Ranking of results based on previous evaluations of the documents

Each user of the Knowledge Broker can indirectly benefit from the work previously done by other members of the community in evaluating documents and information, by matching the results of the queries with results already evaluated by others. This evaluation can give feedback about how documents are judged in certain communities of competence; moreover the judgments can take into account the relationship among the evaluator and the user issuing the search.

4 Supporting cooperation on the World Wide Web

The World Wide Web is more and more seen as a platform for deploying and testing groupware applications despite the inherent technical problems of the underlying infrastructure. There is a key motivation in this: the possibility to embed collaborative applications in a multi-platform client-server architecture already in use in everyday life overcomes technical barriers to the usage of groupware applications, such as the heterogeneity of platforms and the dynamic boundaries of user groups [Borenstein92]. For this reason we have chosen to develop our environment on top of the World Wide Web and have investigated the possible technical solutions in this context.

We have identified five technical dimensions we need to support in our environment for collaborative information gathering (see Fig. 4).

Communication	Awareness	Sharing	Persistence	Security
Support of text, audio and video channels	Event representation	Event notification	Permanent storage of intermediate states	Access control

Fig. 4: Technical dimensions for groupware on the Web.

One of the major obstacle to the development of collaborative applications supporting both synchronous and asynchronous cooperation on the Web is that the HTTP protocol is stateless and there is not a permanent connection allowing notification of changes from the server to the client and from the client to the server.

To solve this problem three approaches can be undertaken:

- modifying the HTTP protocol;
- developing a specialized client;
- relying on an additional infrastructure.

Of the three approaches only the third one does not impact on the critical mass problem [Gru88], because it doesn't affect the Web advantages of openness and multi-platform availability.

Currently there are several ongoing projects in this direction.

Habanero™

Habanero is a framework from NCSA for constructing distributed collaborative software tools. Using Habanero, programs that were designed for a single-user can be relatively easily recast as multi-user collaborative tools by means of a rich API for the sharing of objects and floor control. Habanero is written in Java and is therefore inherently multi-platform, however it targets only applications and not applets.

Java

As of the time of writing, Java has not yet published a set of API for supporting cooperation, but they have been planned its first release by the end of 1997. As described at the Web site, Java Collaboration will allow "for interactive two-way, multi-party communications over a variety of networks. There will be two releases, Part 1: Sharing of collaboration-aware applications, and Part 2: Sharing of collaboration-unaware applications".

MetaWeb

MetaWeb [CSCW97 paper] is a Java class library that supports the development of collaborative applications and includes an API that provides functionality of session management and application-specific message exchange.

MetaWeb provides three abstractions for supporting the cooperation of Web-based applications: user, location and session. The *user object* is a representation of an application instance connected to the MetaWeb system; *session objects* maintain the current members of a session, a description and a list of potential users; *location objects* have an associated list of sessions.

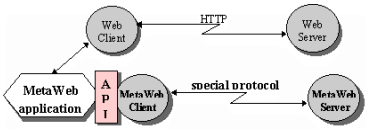

Fig. 5: MetaWeb architecture (source: [TKW97])

PlaceWare

PlaceWare is a company, which is working on a development kit supporting the construction and deployment of shared and persistent Java-based applications on the Web. PlaceWare is a client/server architecture where the collaborative applications are developed splitting them in an applet part and a server part. The infrastructure comprises a PlaceWare server in charge of supporting the communication among the client side and the server side of each collaborative application. The PlaceWare infrastructure supports the concept of attaching applications to a *place.* The concept of place can be used to provide awareness about who are the other users accessing the cooperative application and to enforce access control by checking that only a specific set of users can access an application belonging to a certain place. Moreover the toolkit supports the development of distributed objects, i.e. objects cooperatively manipulated across multiple sites. Each application can maintain a persistent state, which allows the users to enter at

any moment and to obtain the history of the cooperative usage of the application.

5 Prototype description

A first prototype of the system is being developed, in order to test its usability in the context of the European project Campiello (http://www.i3net.org). The aim of this project is to integrate the communities of local inhabitants of artistic towns, the tourists and the cultural managers. To this purpose a methodology of early user involvement in the definition and evaluation of the systems has been defined.

Both directions of exchange of information between KB and KP are implemented in the prototype, though not addressing the extended set of features identified in the Section 2.

The community environment is inherited from the KP environment: it is a defined set of meaningful communities of practice where users are registered as belonging to one or more of them and have "affinities" with other users of the system.

To each community a virtual *search area* is associated, shaped with rooms where people can meet and maintain searching sessions.

5.1 Collaborative usage of the Knowledge Brokers

KB is being extended in order to allow multiple users to cooperatively create the queries and rank the results (see Figure 6). Each room of the search area can host a session issued around a topic. The users navigating the area have feedback about the ongoing activities and the users currently involved. When joining a session each user obtain the current state of the session in terms of query list and associated results. To each query is associated the name of the user who created it.

Synchronous channels of communication (chat and audio link) allow the users to discuss around the search process in terms of problem specification and evaluation of the results. Given the permanent nature of sessions, it is possible to work cooperatively, but in an asynchronous manner. Annotation objects can be attached to result lists and items in order to provide unstructured feedback to the other users.

To each session there is also an associated area available for adding relevant information sources, provided by humans instead of being results of session query. A session is therefore conceived to be a logical folder to find and maintain all the relevant information sources to address a specific research.

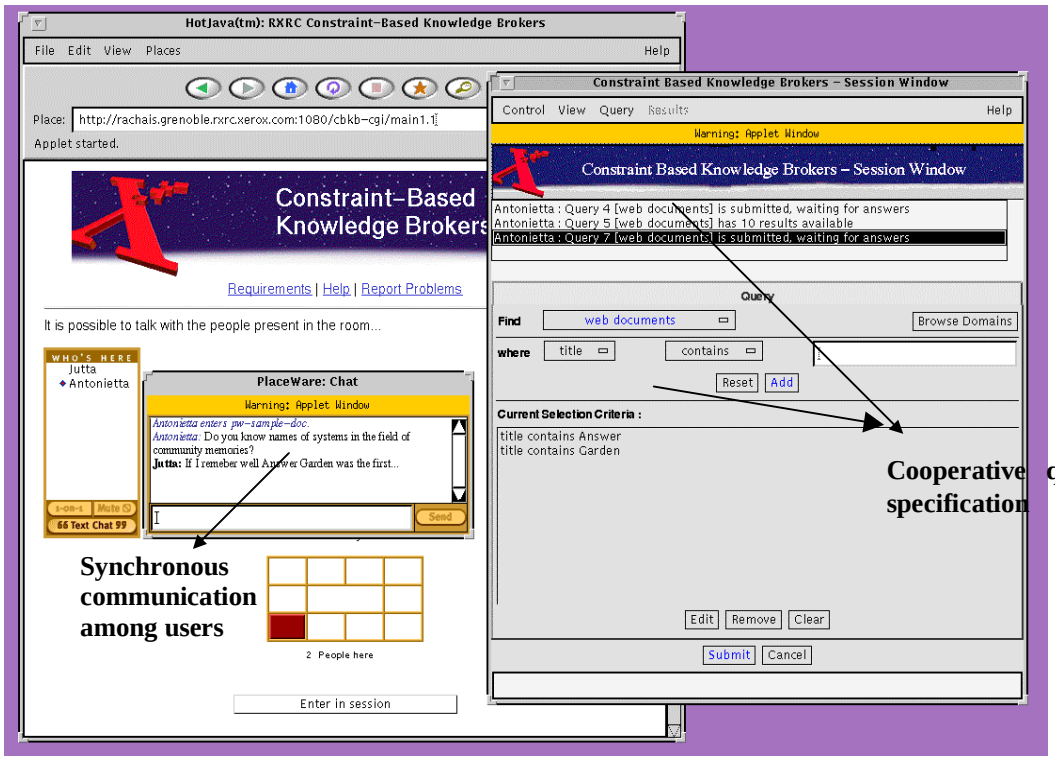

Fig 6: Session sharing

5.2 Publication of ongoing searches and recommendations of search results

Ongoing searches can be "published" in the appropriate community page managed by the KP (see Figure 7). This page is a unique entry point to the activity of a certain community and its aim is to support peripheral participation even in remotely distributed groups. Each concluded session is archived and available for future reuse or updates. Moreover concluded sessions can be "exported" the collection of results as an HTML document. These documents can also be recommended to appropriate communities. The recommendation to a broader audience with respect to the group who created the document closes a first loop in the life of the search, while opening a new one, where the enlarged audience to which the collection is submitted can participate in refining, evaluating and updating the collection of results.

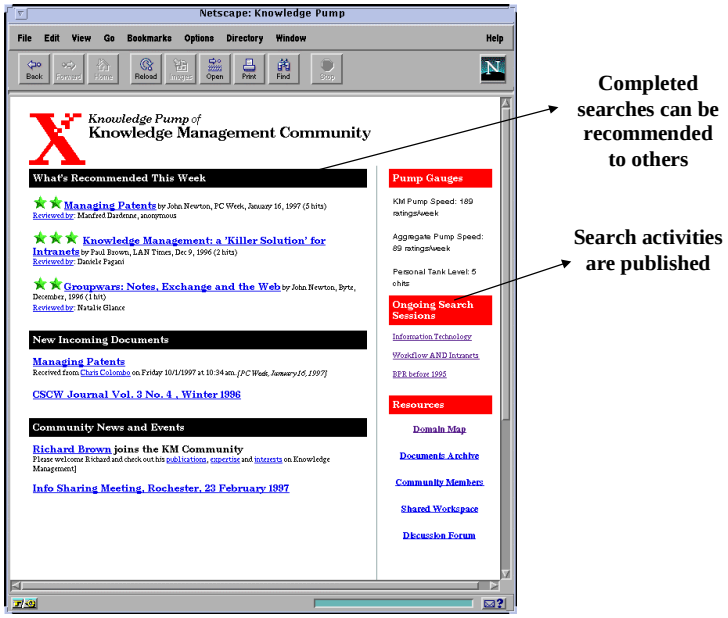

Fig. 7: Collaborative searches in the integrated environment

6 Conclusion

This paper described the first prototype of a combination of Knowledge brokers and the Knowledge Pump systems. Both systems represent a member of the so-called *pull* and *push* information retrieval models. A *PlaceWare*-based implementation showed the feasibility of such a combination.

6.1 Future work

The next steps of this work will explore:

1. the exploitation of the meta-information generated by the search engine to inform the user profiles used in the KP environment;
2. the usability of the approach in real settings, like the tourist settings of the Campiello project;
3. the interface aspects that can facilitate/hinder the cooperation.

In the next section some issues related to the interface are outlined.

6.2 User interface issues

An important part of the system is related to the sharing and cooperative manipulation of the query and result objects. This raises the issue of how best represent objects of this kind.

In the framework of the *Digital Libraries project* at PARC some research has been conducted about information access and visualization. This area has investigated the search process, browsing, and visualization of large text collections, summarization, and automatic detection of thematic structure.

Relevant work in this area are discussed in [RPH+95] includes:

- *search and browsing techniques (Scatter/ Gather*, *TileBars* and *Murax*);
- an architecture for 3D information visualization (*Perspective Wall*).

Moreover some research has been conducted about the implicit structure in spatial layouts of information and the automatic recognition of emergent structure. This study has observed a set of conventions in user constructed spatial arrangements of information objects and developed an architecture for recognizing these common spatial structures [MSC94].

A current ongoing research direction in Knowledge Brokers is the usage of graphical representations of the information objects (both queries and results). In this framework the collaboration could be obtained by sharing the same graphical billboard of objects and remotely manipulate them by giving structure to the set of results.

References

[ABP+95] J.-M. Andreoli, U. M. Borghoff, R. Pareschi, J. H. Schlichter: Constraint Agents for the Information Age. *J. Universal Computer Science* **1**:12, 1995, 762-789. Electronic version available at http://www.iicm.edu/jucs/

[ABP96] J.-M. Andreoli, U. M. Borghoff, R. Pareschi: The Constraint-Based Knowledge Broker Model: Semantics, Implementation and Analysis. *J. Symbolic Computation* **21**:4, 1996, 635-667

[Bor92] Computational Mail as Network Infrastructurefor Computer-Supported Cooperative Work. In J. Turner and R. Kraut (eds.), CSCW92. Proceedings of the Conference on Computer Supported Cooperative Work, Toronto, Canada, October 31-November 4, 1992, New York: ACM Press,pp. 67-74.

[BoSc96] U. M. Borghoff, J. H. Schlichter: On Combining the Knowledge of Heterogeneous Information Repositories. *J. Universal Computer Science* **2**:7, 1996, 512-532. Electronic version available at http://www.iicm.edu/jucs/.

[BCW96] U. M. Borghoff, P.-Y. Chevalier, J. Willamowski: Adaptive Refinement of Search Patterns for Distributed Information Gathering. In: Verbraeck, A.*: Proc. Int'l. Conf. EuroMedia/WEBTEC*, 1996, London, U.K., December, San Diego: The Society for Computer Simulation, pp. 5-12.

[CBC97] B. Chidlovskii, U. M. Borghoff, P.-Y. Chevalier: Towards Sophisticated Wrapping of Web-based Information Repositories. In: *Proc. 5th Int'l. RIAO Conf.*, Montreal, Canada, 1997, pp. 123-135.

[COT+97] A. Crabtree, J. O'Brien, M. B. Twidale, D. M. Nichols: Talking in the Library: Implications for the Design of Digital Libraries. In: Proc. of the *2nd Int'l Conference on Digital Libraries*, Philadelphia, PA, USA, 23-26 July.

[Cur97] P. Curtis: Mudding: Social Phenomena in Text-Based Virtual Realities In: P. Agre, D. Schuler (eds.), *Reinventing Technologies, Rediscovering Community*, Ablex Publishing Corporation, London, 1997.

[DaJe93a] V. O'Day, R. Jeffries: Information Artisans: Patterns of Result Sharing By Information Searchers. In: *Proc. COOCS'93*, Milpitas, CA, USA, 1993, pp. 98-107.

[DaJe93b] V. O'Day, R. Jeffries: Orienteering in an Information Landscape: How Information Seekers Get From Here to There. In: *Proc. INTERCHI '93*, 1993, pp. 438-445.

[GAD98] N. Glance, D. Arregui, M. Dardenne: Knowledge Pump: Supporting the flow and use of knowledge in networked organizations. In: Borghoff U, Pareschi, R. (eds.), *Information technology for knowledge management*, Springer-Verlag, to appear..

[Gru88] J. Grudin: Why cscw applications fail: problems in the design and evaluation of organizational interfaces. In *Proceedings of the International Conference on Computer-Supported Cooperative* Work, Portland, OR, September 1988. New York: SIGCHI/SIGOIS ACM, pp. 85-93.

[LaWe91] J. Lave, E. Wenger: *Situated Learning Legitimate Peripheral Participation.* Cambridge University Press, 1991.

[MSC94] C. Marshall, F. Shipman, J. Coombs: VIKI: Spatial Hypertext Supporting emergent Structure. In *Proceedings of the 1994 European Conference on Hypermedia Technologies*, Edinburgh, Scotland, Sept. 18-23.

[RIS+94] P. Resnick, N. Iacovou, M. Suchak, Peter Bergstrom and John Riedl: GroupLens: An open architecture for collaborative filtering of netnews. In: Proceedings of the Conference on Computer Supported CooperativeWork, Chapel Hill, ACM-Press, pp. 175-186.

[RPH+95] R. Rao, J. Pedersen, M.Hearst et al.: Rich Interaction in the Digital Library. Communications of the ACM, 38(4), 1997, pp. 29-39.

[TKW97] J. Trevor, Thomas Koch and G. Woetzel: MetaWeb: Bringing synchronous groupware to theWorld Wide Web. In J. Hughes et al. (eds.) *ECSCW'97. Proceedings of the Fifth European Conference on Computer Supported Cooperative Work*, 11th-15th September, Lancaster, UK. Dordrecht: Kluwer AcademicPublishers, pp. 65-80.