Project Overview: EUROgatherer - A Personalized Information Gathering System

Umberto Straccia
I.E.I. - C.N.R.
Pisa, Italy
http://faure.iei.pi.cnr.it/~straccia

1 The EUROgatherer project

The aim of this short note is tho describe the EUROgatherer project. The project is a 20 month project of the european Telematics Programme and will start in January, 1998, involving the following partners:

- I.E.I. - C.N.R., Pisa - ITALY, **Coordinator**
- Italia Online SpA, Milan - ITALY
- Rank Xerox Research Center, Grenoble - FRANCE
- Eurospider Information Technology AG, Zurich - SWITZERLAND
- Xarxa CINET SL, Barcelona - SPAIN
- University of Dortmund, Dortmund - GERMANY
- Dublin City University, Dublin - IRELAND

1.1 Rational of the project

A tremendous amount of news and information is created and delivered over electronic media. This has made it increasingly difficult for individuals to control and effectively manage the potentially infinite flow of information. Ironically, just as more and more users are getting on-line, it is getting increasingly difficult to find information unless one knows exactly where to get it from and how to get it. Tools to regulate the flow are urgently needed to prevent computer users from being drowned by the flood of incoming information. Traditional information retrieval systems concentrate on retrieval of unstructured texts of static documents. Information filtering systems have instead been applied to document streams, such as newswire, news groups, and electronic mail. Information gathering is a new field which combines features from information retrieval, information filtering, natural language and knowledge representation, and applies it to the new domain of documents structured in various forms (hypertext, MIME, etc.) and different formats (text, PostScript, GIF, MPEG, etc.). This field has recently seen a significant growth and an enormous popularity with the appearance of several search engines, such as Altavista, Lycos, Yahoo, Excite, Harvest, which help in finding material on the Web. These systems regularly scan the Web to produce indexes to be used in answering queries from users. They provide a generalized service of indexing digital collections accessible through the Internet. In essence, they index textual documents, structured in HTML pages, and provide a search and retrieval service

to the Web community at large, but do not provide any personalized support to individual users. Indeed, they are targeted towards a general and generic user, and therefore they are oriented to answering queries crudely rather than to learning the long-term requirements idiosyncratic to a specific user and selecting and organizing material for him/her accordingly. The technology of information gathering can be applied to a huge number of on-line services, assisting for instance in the selection of books or other archived documents from libraries, news items from press agencies, television station and journals, or documents from administrative bodies. The niche for personalized, prioritized information as an alternative to the uniform newspaper or television broadcast media available today is likely to be the first application domain in which personalized information gathering systems become widespread.

The EUROgatherer project aims at designing and implementing a system which provides a personalized information gathering service and is based on software agent technology. In particular, the goals of the project are:

1. to filter and control the potentially unlimited flux of information from sources to end-users;
2. making information available to people in the appropriate *form*, *amount*, and level of *detail* at the *right time*;
3. to reduce the time spent by the users in knowing regarding: info availability (what, when, where), info structure, info organization, info retrieval services, info access languages and modalities.

The EUROgatherer system will be able to provide the following functionalities:

1. to acquire and retain an interest profile of the user and act upon one or more goals based on that profile;
2. to act, autonomously, pursuing the goals posed by the user irrespective of whether the user is connected to the system where the agent is based;
3. to access a variety of information sources;
4. to create meaningful abstractions of the retrieved documents and classify them appropriately on the basis of their structure and content according to an internal classification scheme, based on user profiles; and
5. to support a relevance feedback mechanism which permits the user to provide the system with feedback on how relevant the retrieved documents are.

The system will collect documents in the following domains in parallel:

- monitoring of frequently changing information sources. The system will monitor at regular intervals URLs that are updated in fixed (or random) intervals for changes. If such changes do exist and are significant then the user will be notified.
- continuous flow of information environment. The system will periodically monitor URLs which generate a continuous flow of information. It will analyze the retrieved documents and select only the proper ones.
- web documents. The system will not search the Web itself, but will utilize existing indexing engines and perform a meta-search in order to discover documents that are, broadly, of interest to the user. Then, the system will further analyze the retrieved documents in order to select those closer to the users preferences.

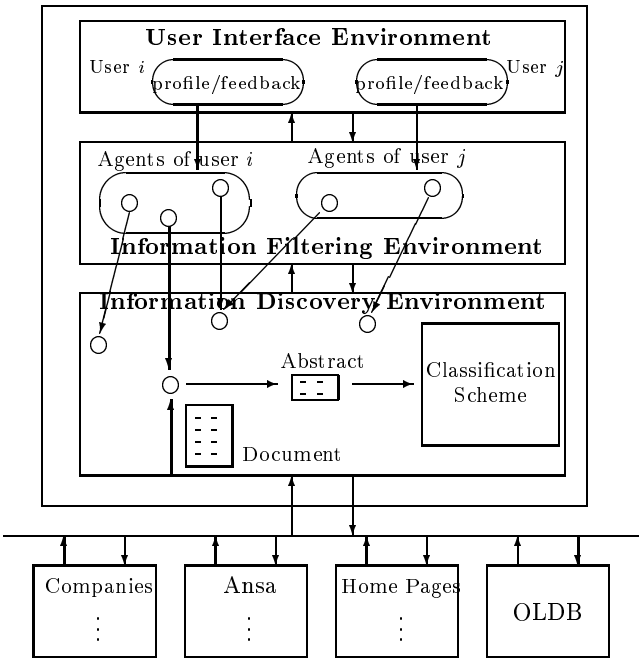

Figure 1: The EUROgatherer system architecture

- on-line Data discovery. The system will access on-line data bases in order to discover data/documents that are of interest to the user.

From the architectural point of view, the project aims at developing an agent-based multilayer system architecture. The system architecture is composed of three layers: the *User Interface Environment*, the *Information Filtering Environment* and the *Information Discovery Environment* (see Figure 1).

Two different species of software agents will be developed: *information filtering agents* and *information discovery agents*. The information filtering agents will be responsible for the personalization of the system and for keeping track of (and adapting to) the interests of the user. The information discovery agents will be responsible for finding, fetching, abstracting and classifying the actual information that the user is interested in. They are utilizing existing Web search engines to find documents (a type of meta-search).

The interactions between the user, the information filtering agents and the information discovery agents are described in terms of a penalty/reward strategy, according to whether the retrieved documents are relevant to the user's needs.

One important aspect of the system architecture is the separation of information filtering and information discovery environments. In the proposed system architecture the personalization of the information, i.e., the information filtering, should be decentralized at the user level, while the information discovery should run on an on-line server. This design choice has a number of advantages:

1. in a multiple user environment, each user will have his/her own set of filtering agents, but they will be able to share their discovery agents;

2. it provides the ability to support real off-line operations; and

3. the introduction of several processing levels between the actual information and the user achieves a greater flexibility in utilizing other novel forms of filtering or other forms of discovery.

Finally, the user interface environment will support the following functionalities:

1. the user profile acquisition by the system;
2. an interactive presentation of the documents retrieved by the system to the user; and
3. the communication of user feedback to the system on how relevant the retrieved documents are.

2 Related literature

1. The Yahoo index. http:// www.yahoo.com
2. The world-wide-web worm. http:// www.cs.colorado.edu/mcbryaan/wwww.html.
3. The Lycos, the catalog of the internet. http://www.lycos.com
4. The metacrawler multi-threaded web search service. http://www.metacrawler.com.
5. K. Decker, V. Lesser, et al., Macron: An Architecture for multi-agent cooperative information gathering. In CIKM Conference, Workshop on Intelligent Information Agents, 1995.
6. Y. Labrou and T. Finn, A semantics approach for kqml - a general purpose communication language for software agents. In Proc. of Conference on Information and Knowledge Management 1994. MIT press, 1994.
7. B. Grosof and al., Reusable architecture for embedding rule-based intelligence. In CIKM Conference, Workshop on Intelligent Information Agents, 1995.
8. A. ORiordan and C. Buckley, An intelligent agent for high-precision information filtering. In CIKM Conference, Workshop on Intelligent Information Agents, 1995.
9. R. Armstrong et al., Webwatcher: A learning apprentice for the world-wide web. In Proc. of the Symposium on Information Gathering from Heterogeneous, Distributed Environments. AAAI Press, 1995.
10. H. Lieberman Letizia, an agent that assists web browsing. In Proc. of the IJCAI-95. AAAI Press, 1995.
11. M. Balabanovic and Y. Shoham, Learning information retrieval agents: Experiments with automated web browsing. In AAAI Technical Report SS-95-08, Proc. of the 1995 AAAI Spring Symposium Series, 1995.
12. B. Sheth and P. Maes, Evolving agents for personalized information filtering. In Proc. of the ninth Conference on Artificial Intelligence for Applications, 1993. IEEE Computer Society Press, 1993
13. N. Belkin and B. Croft, Information filtering and information retrieval. Communications of the ACM, 35, No. 12, 1992.
14. G.S. Jung and V.N. Gudivada, Autonomous tools for information discovery in the world-wide web. Technical Report CS-95-01, School of Electrical Engineering and Computer Science, Ohio University, Athens, OH, 1995.