Intelligent access to document archives

Luciana Bordoni
ENEA-Studi

INTELLIGENT ACCESS TO DOCUMENT ARCHIVES

Luciana Bordoni
ENEA-Studi, Via Anguillarese 301, I00060 S. Maria di Galeria (RM), Italy

E-mail: bordoni@casaccia.enea.it

INTRODUCTION

Researching information has always been a crucial problem both in the area of paper and computer documents. Thanks to the Internet explosion, an enormous amount of information has been made readily, easily and economically accessible to users. Progress is being made in the existence and the use of information technology in libraries. Up to the beginning of the '90s information technology was used in libraries only for automating traditional library functions: purchases, cataloguing, catalogue consulting, lending, statistics, etc. Information technology tools have evolved, from mainframes to servers. Client-server architecture has asserted itself and so have graphic interfaces. All this has shown it was necessary for traditional library tools to evolve. Even the consulting techniques within libraries, which were traditionally limited to the material owned and kept in the library, are changing. It is now possible for users to refer to and possibly process information found not only in the material owned by the library, but also found by accessing the net. The multimedia or virtual libraries make it possible to manage documents in a differently than before. This new organisation gives the end user a new angle on the resources available in libraries. The end user is no longer a passive consumer of documents, but an active researcher. The contents of the virtual library come from documents, which, in turn must be available for access from any part of the organisation. The new systems for accessing information make use of Artificial Intelligence models and methods. They are capable of creating *user models* that allow the system to adapt itself to the characteristics and needs of the specific users of the system. Information filtering systems are interfaces located between the information source and the user. They can determine the relevance of a document by comparing it to an explicit model of the user's information interests.

These systems base their modelling activities on a dynamic classification model of the user by means of a series of stereotypes. Besides gaining some knowledge, the system is also able to infer data to obtain new knowledge from the one contained in the model. An information filtering system based on user modeling means that the user need not read all the documents provided by the source, but only the potentially interesting ones. While reading a document, the user can express an opinion on its relevance, which makes it possible to adapt the model to the subjects that in time, the user seems to appreciate or not appreciate. In this paper the methods and models capable of creating user models and of filtering information, in order to contribute to an intelligent access to document archives, are presented.

SEARCHING IN A DIGITAL LIBRARY

Libraries exist in many forms and are of many types. In computing, code libraries have been a part of the world of software engineering. With multimedia technologies we now have image libraries, audio libraries, and even digital video libraries. We might also think of libraries when we refer to collections that now reside in databases, knowledge bases, text bases, or the World Wide Web (WWW).

The phrase "digital libraries" evokes a different impression in each reader. To some it simply suggests computerisation of traditional libraries. To others, who have studied library science, it calls for carrying out of the functions of libraries in a new way , encompassing new types of information resources ; new methods of storage and preservation; new approaches to classification and cataloging. To many computer professionals, a digital library is simply a distributed text-based information system, a collection of distributed information services, a distributed space of interlinked information, or a networked multimedia information system. The digital libraries are seen as the space in which people communicate, share, and produce new knowledge, and also as support for learning.

In order to make effective use of any kind of electronic information, as found in a digital library, the organisation and manipulation of information by content is a crucial component.

Several projects aim to improve user information seeking by closely coupling text search with visualisation methods to display the result of the searches [1], other projects allow users to directly manipulate data. Today's digital libraries that depend on WWW are increasingly able to take advantage of the fruits of human-computer interaction research.

In the following will be shown as the application of user modeling technology to information filtering improves the usability and usefulness of the WWW.

USER MODELING IN INFORMATION FILTERING

The WWW has emerged as a new type of information space. Its lack of central control mechanisms leads to many new interesting features but at the same time has the potential danger that users can drown in irrelevant information. Being lost in space and overloaded with information [2] are two problems users confront: there is more information out than a single user can manage. The potential benefits of the WWW will not be realised if users cannot retrieve information easily and efficiently. The principal techniques that have long served users to retrieve information are browsing and searching. Selective search becomes critical because formulating queries to retrieve the desired information is difficult and requires considerable skill; for example, the use of search engines in the WWW is not standardised. New strategies are needed to deal with information space such as WWW: users need active support to determine if potentially information exists, where the information is located, how to retrieve the information when it is located. To satisfy these needs the users need selection systems that can be easily manipulated to match their personal interests. Belkin and Croft [3] compared two major methods for information selection: filtering and retrieval. For effective filtering to occur, accurate information about both the document space and the user must be maintained. An important part of an information filter is a user model to predict what a target user would you like to filter. Information

filtering is based on profiles that describe either individual or group preferences. Such profiles often represent long-term interests of the user.

Several applications are focused on developing products that are based on user modeling to improve the filtering and interactive search experience in Web-based environment.

A user model is a knowledge source which contains explicit assumptions on all aspects of the user that are relevant to the interactive behaviour of the software application. These assumptions concern aspects such as the user's knowledge and interests. A user modeling component, on the other hand, is the part of the software application whose function is to incrementally construct a user model; to store, update, and delete entries in this model; to maintain the consistency of the model and to supply other components with assumptions about the user. User modeling shells are general user modeling components that may provide a representation scheme for assumptions about a user's beliefs and interests, an inference mechanism for this representation, a mechanism for detecting inconsistencies in the assumptions, a belief revision mechanism, etc. Such a shell can then be filled with domain specific inference rules as well as with domain- specific default assumptions about any user, and then be integrated into a software application. The most well known user modeling shells are BGP-MS[4], GUMS[5], TAGUS [6], UMT[7].

User modeling technology (UMT) has also been integrated into information filtering system. The IF-UMT (Information Filtering based on UMT) [8] presents a prototype system developed in an project on the application of the UMT shell to information filtering. IF-UMT aims at providing user-oriented information filtering to users accessing Internet services. Others information filtering systems content-based are the following: Amalthea [9], Letizia [10], PROFILE [11], SIFTER [12], WIFS [13].

In the next section an experiment of the system WIFS (Web-oriented Information Filtering System), carried out to evaluate user satisfaction, will be illustrated.

AN EXPERIMENT AT ENEA

The system WIFS uses an approach for user modeling based on stereotypes [14]. A stereotype is a description of a prototypical user of a given class. A case-based approach [15] for the task of automatically inferring user stereotypes is used. Moreover an artificial neural network is implemented. This hybrid architecture has been tested and compared to some traditional approaches [7, 5] and it has proved to be more precise [13]. The filtering algorithm assigns a score to each document calculated as the similarity between its representation, the user model and the query.

An experiment at ENEA (Italian Agency for Energy, New Technology and Environment) was performed in order to evaluate and compare the performance of WIFS with standard information retrieval techniques. We have tried out several tests to determine the efficacy of using an information filtering system based on user modeling methodology.

Seven experts working at ENEA in different topics of the environmental domain, have scored a set of documents obtained by each query session of WIFS. Each user started with the preliminary interview managed by the system. As next step, each user

input 15 queries, then personally analysed all filtered documents and gave a relevance feedback (value between –10 and +10). After each filtering process the relevance ordering of the collection is obtained. A statistical results demonstrate that the WIFS improves the performance of Alta Vista by more than 30% both in precision and recall and in the ordering capabilities.

CONCLUSIONS

The use of systems based on Artificial Intelligence methods and models, capable of creating user models and of filtering information, greatly contributes to an intelligent access to document archives and offers a valuable support when building intelligent digital libraries. Introducing these advanced systems will open important prospects in libraries, as they will facilitate the promotion and production of the cultural heritage they possess.

REFERENCES

1. HEARST, M. Tilebars: Visualization of term distribution information in full text information access. Proceedings of CHI'95, ACM New York, pp 59-66.
2. SCHICK, A.G.; GORDON, L.A. and HAKA, S. Information Overload: A Temporal Approach. Accounting Organizations and Society, 15 (3) 1990: pp. 199-220.
3. BELKIN, N.J. and CROFT, W.B. Information Filtering and Information Retrieval: Two Sides of the Same Coin. Communications of the ACM, 35 (12) 1992: pp.29-38.
4. KOBSA, A. and POHL, W. The user Modeling Shell System BGP-MS. User Modeling and User-Adapted Interaction, 4 1995: pp.49-106.
5. FININ, T.W. GUMS-A General User Modeling Shell. In A. Kobsa & W. Wahlster (Eds): User Models in Dialogue Systems, Springer, Berlin 1989.
6. PAIVA, A. and SELF, J. TAGUS- A User and Learner Modeling System. Proceedings of the fourth International Conference on User Modeling, 1994 pp.43-50.
7. BRAJNIK, G. and TASSO, C. A shell for developing non-monotonic user modeling systems. Int. I. Human Computer Studies, 40, 1994: pp. 31-62.
8. MINIO M. and TASSO, C. User Modeling for Information Filtering on Internet Services: Exploiting an Extended Version of the UMT Shell. Workshop on User Modeling for Information Filtering on the World Wide Web, Fifth International Conference on User Modeling, Hawaii, 1996, pp.9-14.
9. MOUKAS, A. Amalthaea: Information Discovery and Filtering Using a Multiagent Evolving Ecosystem. Applied Artificial Intelligence, 11, 1997: pp. 437-457.
10. LIEBERMAN, H. Letizia: An Agent that Assists Web Browsing. Proceedings of IJCAI-95, Montreal, Canada, 1995: pp. 924-929.
11. AMATI, G., D'ALOISI, D., GIANNINI, V. and UBALDINI, F. An Integrated System for Filtering News and Managing Distributed Data. FUB Tech. Report., 1996.
12. MOSTAFA, J., MUKHOPADHYAY, S., LAM, W. and PALAKAL, M. A multilevel approach to intelligent information filtering: Model, system and evaluation. ACM Transactions on Information Systems, 15(4) 1997: pp. 368-399.

13. AMBROSINI, L., CIRILLO, E. and MICARELLI, A. A Hybrid Architecture for User-Adapted Information Filtering on the World Wide Web. Proceedings of the 6th International Conference on User Modeling UM-97, Chia Laguna, Sardegna, 1997.
14. RICH, E. Users are individuals: individualizing user models. International Journal of Man-Machine Studies 18 1983: pp. 199-214.
15. KOLODNER, J. Case-Based Reasoning. San Mateo, CA: Morgan Kaufmann, 1993. 668p.