# User Modeling and Instance Reuse for Information Retrieval Study Case : Visually Disabled Users Access to Scientific Documents

**JERIBI Lobna, RUMPLER Beatrice, PINON Jean Marie**

LISI – INSA of LYON
Bâtiment Blaise Pascal
7, avenue Jean Capelle
F69621 Villeurbanne Cedex France
Lobna.Jeribi@lisi.insa-lyon.fr,
Beatrice.Rumpler@insa-lyon.fr

**Abstract**

In this paper we present an Information Retrieval System (COSYDOR) designed for sightless students or researchers, to improve their access to scientific information. Our approach consists in reusing search instances in order to help these users when querying COSYDOR.
Thus, In the first part we define a search instance model, including our proposed user model. Then we present our approach of Case Based Reasoning applied on this model and we describe the instances reuse process. In the last part we briefly describe our prototype and show the experimental evaluations results.
Keywords : Knowledge retrieval, User modeling, CBR, Educational system, Information retrieval.

## Introduction

The main goal of this research is to improve Information Retrieval Systems by enabling them to generate search outcomes that are relevant and customized to each specific user. Our proposal advocates the use of Case Based Reasoning (CBR) during the information retrieval process.
When conducting a search, the system retrieves a previous similar search experience and traces back previous human reasoning and behavior and then replicates it in the current situation.
Thus, users information retrieval experiences or instances are saved to be reused in future similar cases. The resulting cooperative memory is utilized for user query expansion.
In order to improve the information retrieval experience, we propose in this paper to conceptualize and model both the user profile, and the information retrieval process. This leads us to define some similarity functions between user profiles and information retrieval situations.
The reuse and exploitation of past experiences serves to enrich the initial user query by words from documents found in similar cases. Unlike the classical *Rocchio* method, these documents are those already judged as valid by users with similar profile and in similar search situation. The value this method brings to the user is an increasing relevance of the search outcomes while reducing user interaction with the system.



## Related Works

Some related works of context definition and experience reuse were proposed in the early literature. RADIX project (Corvaisier F, Mille A and Pinon J.M 1999) proposes the modeling of internet navigation sessions carried out by the user. These models are reused in order to suggest similar sessions to the user. CABRI-N (Smaïl M 1999) is a personalized image retrieval system. Smaïn proposes a modeling of user strategy during an information retrieval process. Retrieval sessions are memorized and reused to improve user strategy search.
In this paper, we present briefly a modeling study of the user during a search session, and a representation of a search instance or situation. Then, we present our approach of instance reuse for query expansion. Afterwards, we expose our project context and our application case. Lastly, we present the results of our first tests and the prospects for evaluations.

## User Modeling

Intelligent information systems aim to automatically adapt to individual users. Hence, the development of appropriate user modeling techniques is of central importance. Algorithms for intelligent information agents typically draw on work from the information retrieval (IR) and machine learning communities. Both communities have previously explored the potential of established algorithms for user modeling purposes (Belkin N, Kay J, and Tasso, 1997), (Webb G 1998). However, "work in this field is still in its infancy" (Billsus D, Pazzani M 1999).

### User Knowledge

Intelligent systems for information access are typically aimed at assisting a user in his search for interesting or useful information. A large variety of agents that make use of machine learning techniques have been developed and presented in literature (e.g. (Pazzani, M., Billsus D 1997)).

Most of this work focuses on the acquisition of a precise model of the user information need. However, in order to build truly useful information retrieval systems we also need to be aware of the user's knowledge.
To define the specific user knowledge, that is exploited during a search session, we have exploited cognitive approach results (Allen N 1991). We have classified the user knowledge into four knowledge ***categories***, ranked according to their *evolution* degree.

- *Cultural knowledge* (features having little or no evolution)
- *Professional knowledge* (features having long term evolution)
- *System knowledge* (features having mean term evolution)
- *Search knowledge* (features having short term evolution), related to the current search session.

Some features varies from an application case to another. These variations depend on the document, the search system and the user types. For example, in image retrieval systems, the search preferences include specific features as: plan, luminosity, contrast, colors, etc. Our specific model, concerning our use context, is presented in the next part.
User knowledge features presented above, constitute a generic model of user profile, which specific models are related to the application cases used.

## Representation Formalism

The chosen representation formalism is the vector model. It is the formalism commonly used in both communities: information retrieval and instance based reasoning. The vector model presents several advantages in processing similarity between vectors:
Let $U=<U_1, U_2, U_3, U_4>$, be the vector representation of U, where $U_i$ represents the $i^{th}$ category of user knowledge U.
$U_i = \{a_{i.j}\}$ ; $\forall j \in [1, n]$; $a_{i.j}$ represents the $j^{th}$ attribute of the category $U_i$;
$a_{i.j} \in \{v_k\}$ ; $\forall k \in [1, n]$; $v_k$ represents the $k^{th}$ possible instance of $a_{i.j}$

## Similarity Function

We propose to memorize user retrieval experiences, in order to reuse them, when users have "similar" profiles. Thus, our first goal of formalizing the user model, is to define the "distance" between user profiles.
We define the similarity function $S_U$, between two user profiles $U_i$ and $U_j$

$$S_U(U^i, U^j) = \frac{\mu\, s_1(U_1^i, U_1^j) + \nu\, s_2(U_2^i, U_2^j) + \kappa\, s_3(U_3^i, U_3^j) + \lambda\, s_4(U_4^i, U_4^j)}{\mu + \nu + \kappa + \lambda}$$

Where:
$U^p_i$ is the vector representing the $p^{th}$ category of $U_j$ , which is the profile of the user j.
$s_p$: similarity function between vectors of the $p^{th}$ category of U
$s_p \in [0, 1]$ ; $\mu, \nu, \kappa, \lambda \in [0, 1]$
$\mu$, $\nu$, $\kappa$, $\lambda$, represent the parameters enabling to "contextualize" the similarity.
Details about this similarity function are presented in (Jéribi L, Rumpler B and Pinon J.M. 2001b)

## Search Instance Modeling

The results of various studies on search instance (Jéribi L, Rumpler B and Pinon J.M. 2001b), make highlight of following features of a search instance:

- The user profile represented by U
- The user information need expressed by a query, represented by Q
- The documents solutions represented by D
- The evaluations E of relevancy of the documents D, given by the user U

Referring to the problem resolution field, the initial problem in information retrieval system is represented by the user profile U, and his query Q. Collected documents D represent the solution to the problem, and E the solution evaluation.
We propose a formal description of a search instance, carried out by a user, during an information retrieval session, (vector representation) as follows:
Instance = <U, Q, D, E>

- U: represents the user features during the search session: $U=<U_1, U_2, U_3, U_4>$
- Q: As defined in information retrieval vector model (Salton, G 1986) giving the space E including all the corpus terms: $E= <t_1, t_2, t_3, ..., t_n>$, Q is the weighted term vector representing the user query, in the space E: $Q = <a_1, a_2, a_3, ..., a_n>$ ; $a_i$ corresponds to the weight of the $i^{th}$ term of the query
- D = $\{d_i\}$ ; D is a set of documents $d_i$ ; $i \in [1, p]$ ; p: number of documents evaluated by the user ; p = |D| ; $d_i = <b_{i.1}, b_{i.2}, b_{i.3}, ..., b_{i.n}>$ ; $d_i$ is the vector representation of a document in the defined space E corpus. $b_{i.j}$ represents the weight of the $j^{th}$ term of the document $d_i$. $d_i$ is the weighted term vector representing the document (or a part of the document) that the user have chosen to evaluate
- E represents the evaluation given by the user U of the relevance of D according to Q

# Experience Reuse for Query Expansion

## Knowledge Base for Relevance Feedback

Our proposal is based on the *Rocchio* method of query expansion. The principle approach of the proposed solution, consists in "completing" the documents used for the query expansion issued from the current relevance feedback, by the evaluated documents extracted from the previous search instances. From these documents - coming

from both sources- we apply the *Rocchio* approach of query expansion (Rocchio J 1971).
Thus, the terms added to the user query result from the documents which has just been evaluated (relevance feedback) or of the documents resulting from the instance base, evaluated previously by the user or other users being in similar search contexts, and having similar profiles.
However, these two sources are not independent, since the documents evaluated during the previous search iteration, represent also an instance contained in the memory of instances.
The interest of our proposal is primarily justified when no relevance feedback is made by the user during his search session. In this case, the reuse of the instance base constitutes an interesting alternative for the query expansion. Moreover, this enables to give a certain "freedom" to the user, because the relevance feedback is not any more an obligatory step to help him to reformulate his information need.
Nevertheless, the instances reused cannot contribute in the same way for query expansion. For this, we propose to "contextualize" this contribution, according to the "confidence degree" of the reused instance. This concept will be detailed in the following paragraph.

### Integrating Learning in Rocchio Approach

As presented above, our system allows two types of learning :
Long term learning thanks to the instance memory. The reuse approach and instance based learning allows the user to benefit from the aid of the system without having to interact and evaluate during each session, the collected documents (contrary to traditional methods of query expansion based on relevance feedback).
However, our solution is optimal when the number of instances of the instance base is significantly important. In the contrary case, the system functions as *Rocchio* based on relevance feedback. The effectiveness of the instance base is well exploited when the user population using the system have common interests and carry out exploitable common search then other users. This is a classical constraint in the co-operative systems.
Short term learning thanks to the training by reward / penalty of the search instances. The system evolves according to the failure / success of the proposed solutions.

## COSYDOR Project and Experiments

### Project Context

This research is carried out within the framework of a national French project of the Rhone-Alpes department entitled: "pedagogic information access systems for sightless users: use of speech and sound". From existing virtual libraries specialized in engineering sciences (scientific documents), the goal of this project is to produce an "intelligent" tool of information retrieval, adapted to the visually defective users. At this level, we are interested on the one hand in the design of an information retrieval system allowing to help the user to better formulate his information need on the one hand, and to offer to him a better access to the documents.
These specific users have particular information access difficulties, which is added to the traditional constraints of information retrieval (query formulation difficulties, irrelevant answers, etc). These accessibility difficulties are accentuated in the context of scientific documents (Braille transcribing and/or voice synthesis). This work concerning the accessibility aspects (Design of accessible IHM, transcribing Braille and/or vocal), was carried out in collaboration with the other partners of the project, and is detailed in (Jéribi L, Rumpler B and Pinon J.M. 2000b). A work, in close collaboration with "sight deficient" scientists, enabled us to note that it becomes crucial for these users to have dedicated systems, taking in account their use profile. It becomes tiresome to these handicapped users to exploit the current information retrieval. This is due not only to the accessibility problems, but also to the imposed use logic of these systems (excessive interaction with the user, logic of navigation, visual criteria, etc).

### Sight Deficient User Profile

In this work, we applied the user model (§ 2) to the specific case of the sight deficient users, under a university context using scientific documents. In a study on the behavior of these users (Bergère T, Portalier S 1998), we highlighted the following features:

- The visual user group ($U_1$): Sighted, Sight deficient, blind user
- The user category or function ($U_2$): Student, Teacher, Researcher, Engineer
- The field of interest ($U_2$): Mathematical, Natural science, Biology, Biomedical field, etc.
- The knowledge about the system ($U_3$)
- The documentary preferences ($U_{4.1}$): the date of publication, the author, the document support
- The finality or goal of the search ($U_{4.2}$): bibliographical synthesis, technological survey, project study, etc.

These features are taken into account to evaluate the similarity of user profiles, by our prototype COSYDOR.

### COSYDOR Prototype Presentation

Our prototype COSYDOR (Cooperative SYstem for DOcument Retrieval) is based on Intermedia de Oracle 8i. We enriched Intermedia by an intelligent layer (developed in java) enabling the users query expansion and the management of the instance base . Intermedia is a textual DBMS, using linguistic tools (thesaurus, lexicon, etc.) for documents and queries representation. The choice of this tool results from a comparative survey on several information retrieval systems in (Jéribi L., Rumpler B. and Pinon J.M. 2000a). Intermedia proved to be most relevant in our context. One of its advantages, is to offer paragraph

extraction functionality, enabling to present document "views" during the document restitution to the user. This makes the user evaluation more precise on the one hand, and makes easier the access to the long documents for the sight deficient users on the other hand.

### Experiments and Evaluations using Test Corpus

In order to test and to evaluate the contribution of our system, we have used a TREC corpus of test. TREC[1] (Text Retrieval Conference) is an American organization which provides a corpus of tests and common procedures of analysis of performance. Among this base of tests, we used and indexed a whole of 7000 documents, in format HTML Moreover, the tests quantitative results enabled us to note the significant contribution to the system performance, of the weights assigned to the terms of the initial query.
The second part of our evaluation consists on the comparison of COSYDOR performances versus other information retrieval systems using manual and automatic query expansion. The results of these systems were provided to us by TREC. First experimentation of these comparative tests are shown in next figure.

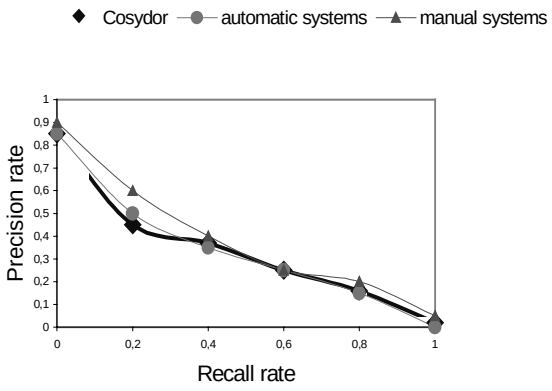


Figure 1: Comparison of COSYDOR to other query expansion systems

The tests show that manual query expansion results remain always better (according to the recall/ precision rate performance) then automatic query expansion. However, the curves show that COSYDOR presents very close performances to those of the automatic systems. Our results are optimal for high recall rates. Our current work consists on handling more experiments in order to make better justifications of these comparison results.
The first evaluation results, although very encouraging, must be moderated by the limited number of tests (ten queries). Moreover, the users who contributed to the tests have very close profiles, and were previously initiated to the system use. It would be then interesting to carry out these tests on a broader sample of users, having different profiles and having a visual handicap.

[1] http://trec.nist.gov

## Conclusion

The goal to build a knowledge base making "permanent" the user evaluations on search results, represents our first work motivation.
The search instances memory offers several improvements, compared to the *Rocchio* method of query expansion, based exclusively on the relevance feedback.
The value this method brings to the user is an increasing relevance of the search outcomes while reducing user interaction with the system.
This method has been implemented in the COSYDOR (Cooperative System for Document Retrieval) prototype based on *Intermedia* (Oracle 8i). Tests and evaluations have been performed using the test corpus of TREC. The results of these analyses show a significant improvement of performance in the first search iterations compared to the *Intermedia* benchmark.
This study has been conducted as part of a Rhone-Alpes regional project (French project) in which visually disabled persons are the main target users. This research represents a tremendous contribution to these users in universities, considering the difficulties they face when using traditional information retrieval systems (heavy system interaction, accessibility problems, etc).

## References

Allen N. 1991 Cognitive Research in information science: implication for design. *Annual review of information science and technology,* 1991, vol. 26: 3-37

Belkin, N. Kay, J.and Tasso, C. 1997 Special issue on User Modelling and Information Filtering . *User modelling and User adapted Interaction*, 1997, vol 7(3): 313-331.

Bergère, T., Portalier, S. 1998. Modélisation du comportement de l'utilisateur déficient visuel. Workshop NTI SPI & santé. Chassey le Camp (France).

Billsus, D., Pazzani, M. 1999 A hybrid User Model for New Story Classification. Proceedings of the Seventh International Conference on User Modelling (UM'99), Banff, Canada : 20-24

Corvaisier, F., Mille, A.and Pinon, J.M.1999 – Recherche assistée de documents indexés sur l'expérience (RADIX): Mesures de similarité des épisodes de recherche sur le WEB. *IC'99 Ingénierie des connaissances*.

Jéribi, L., Rumpler, B. and Pinon J.M. 2000a Personalised information retrieval in specialised virtual libraries. *New Library Worl review,* MCB press, VOL 101 N° 1153 : 21-27.

Jéribi, L., Rumpler, B. and Pinon J.M. 2000b Intelligent System for document retrieval and access to scientific documents for visually deficient users. 12-14 Avril 2000. *Conférence internationale de Recherche d'Information Assistée par Ordinateur, RIAO'2000: Accès à l'information multimédia par le contenu*, Paris (France). ISBN 2-905450-07-X, : 870-884.

Pazzani, M., Billsus, D. 1997. Learning and revising user profiles: The identification of interesting web sites. *Machine learning* 27: 313-331.

Rocchio, J 1971. Relevance feedback in formation retrieval. In Gerard Salton editor*, The SMART retrieval system: Experiments in Automatic Document Proceedings,* pages 313-323. Prentice Hall, 1971.

Salton, G. 1986 Text-retrieval systems, *Communication of the ACM*. July 1986, n° 7, p. 648-655.

Smaïl, M. 1999 Recherche de régularités dans une mémoire de sessions de recherche d'information documentaire, *InforSID 2-4 juin 1999, actes des conférences, XVIIème congrès*

Webb, G. 1998 Special issue on Machine Learning for User Modelling*. User Modelling and user Adapted Interaction*, vol 8 :1-2, Kluwer Academic Publishers.