Intelligent retrieval in virtual libraries for education and training

Lobna Jéribi, Béatrice Rumpler and Jean Marie Pinon

LISI-INSA de Lyon
20, Avenue Albert Einstein
F-69621 Villeurbanne Cedex FRANCE
Tel: 04 72 43 83 92
Fax: 04 72 43 85 18
E-mail: ljeribi@ifhamy.insa-lyon.fr, {rum, pinon}@if.insa-lyon.fr

Abstract

We are interested in managing documents in virtual libraries specialised in the engineering science field (ESF). Our goal is to design an intelligent tutoring system for personalised information retrieval depending on the user's interests. The proposed system is specially designed for people having sight deficiency. These specific persons use slow devices in order to access textual information, so they have a critical need for systems able to retrieve relevant information quickly. For these reasons, we need intelligent systems to delegate tasks and to make decisions. Our goal is to define a multiagent system for retrieving and filtering documents in these specific virtual libraries. We aim at improving answer quality by personalising the retrieval and by performing user's profile management. The multiagent solution proposed in this work is based on co-operative and adaptive information filtering and retrieval.

1. Introduction

As a consequence of the development of new communication technologies and powerful computers, we can note an increasing development of virtual libraries in various fields.

We are interested in managing documents in virtual libraries. Our principal goal is to design an intelligent tutoring system for personalised information retrieval depending on the user's interests.

Our work is a part of a project concerning the engineering science field (ESF). This project focuses on the contribution of the virtual library in the fields of education and research.

The proposed system is specially designed for people having sight deficiency. These people use slow devices in order to access textual information, so they have a critical need for systems able to retrieve relevant information quickly. For these reasons, we need intelligent systems to delegate tasks and to make decisions.

Our goal is to define a multiagent system for retrieving and filtering documents in virtual libraries specialised in ESF. We aim at improving answer quality by personalising retrieval and by way of managing the user's profile [JER 97].

In the multiagent approach, agents incorporate innovative and useful concepts such as autonomy, collaboration and co-operation, learning for adaptation and evolution. We particularly use the learning algorithms provided by the artificial life field [MOU 96]. The multiagent solution proposed in this work is based on intelligent information filtering and retrieval support systems.

In the first section of this paper, we present the project context and the different teams involved. Then we describe the general architecture of our system. In the third part, we present our multiagent system architecture, and briefly describe the agent's roles. The functioning and the detailed information-seeking process is described in the fourth part. Finally, we show our prototype architecture and justify our choices. We conclude with the perspectives of our solution.

2. Project Context[1]

This work is a part of a project called Intelligent access systems to engineering science field for blind users . From existing virtual libraries specialised in engineering science field (ESF), we aim to design and develop an intelligent tutoring tool for document retrieval appropriate to sight deficient people. This project includes various features such as document retrieval and management, document transcoding, the blind's cognitive behaviour, etc. This research is performed in collaboration with various teams specialising in complementary fields[2]. The whole project is composed of parts complementing each other, as shown in Figure 1.

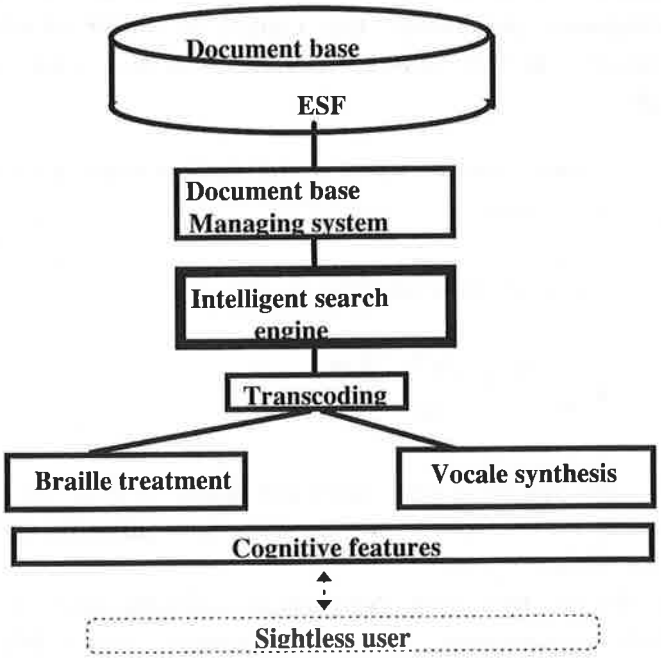

Figure 1: Project presentation

[1] Project financed by « Région Rhône Alpes France : « Nouvelles technologies de l'information et formation de pointe dans le secteur SPI et santé .

[2] Handicap Mission (UCB Lyon I University) - Psychological cognition laboratory (Lyon II University) - LISI (INSA LYON) - GEOD (ENSIMAG- Grenoble)

Our team deals with Document management and intelligent engine design for information filtering and retrieval. We provide the link between the documentary base part and the document transcoding part. At this level, we are interested, on the one hand, in designing a search engine which integrates intelligent aspects in order to bring pertinent and precise results to the user. On the other hand, we aim to carry out document base management mechanisms, which enable fast information access.

Moreover, in order to be exploitable by blind users, collected documents should be transcoded by Braille terminals or a voice synthesiser depending on user's preferences.

Thus, we collaborate with other research centres specialised in vocal synthesis/ recognition and Braille treatment. They provide Human Machine communication by voice and Transcoding of visual and non textual information to accessible formats for blind users.

A research centre specialised in cognitive psychology is also involved in this project. This team deals with Cognitive aspects of Human Machine Interface for blind users. The blind's behaviour modelling enables to integrate user specificity during the document retrieval phase.

3. General Presentation of our System

A library specialised in ESF is a base of scientific documents dealing with various fields such as mathematics, physics, chemistry, and computer sciences... A major access difficulty comes from the different formats of the documents, such as PDF, TEX, PostScript, HTML, XML, etc.

Another difficulty is that these scientific documents contain mathematical and chemical formulae, graphs, diagrams, and tables, that cannot be automatically translated by Braille terminals or vocal editors. In fact, automatic translation brings understanding ambiguity as shown in this example:

Vocale synthesis	" x equals minus b plus square root of b squared minus four ac, over two a "
Mathematical Braille	<x=-b+sqrt(b^2-4ac)/2a>
Image mode	$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$

In this example, the sightless user does not know if over two a refers to the whole expression or just the last sub-expression. Some work is currently being carried out to solve these accessibility problems.

Concerning the Braille treatment, translation software from TEX to Braille is being developed. Nevertheless, automatic translation remains impossible because the symbols defined by users are not represented by Braille terminals. Thus a human contribution during the translation process seems to be necessary.

As regards vocal synthesis, TEX to English translation already exists provided by the AsTer software. This software is the result of work performed in USA by T.V.Raman [RAM 97].

These points will not be detailed further in this paper. They are studied by the other teams previously mentioned. This paper will focus on the intelligent interface (Figure 2).

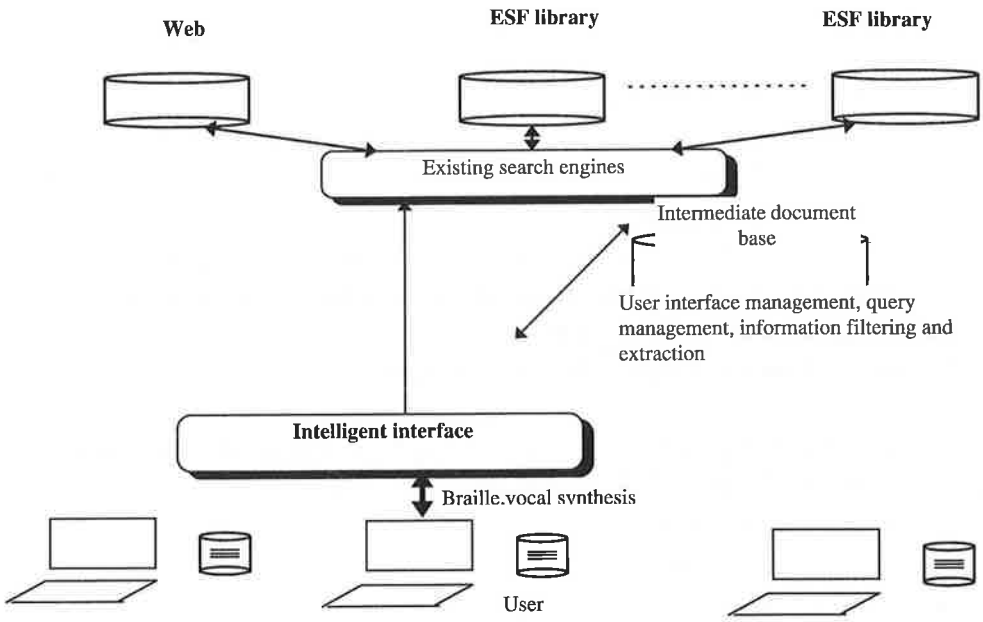

Figure 2: General architecture of our system

As mentioned in the previous paragraph, there are problems concerning the scientific document accessibility to blind users. Thus, we propose to retrieve these documents from existing libraries - using an existing meta search engine - and to file them in an intermediate document base in order to standardise their formats in Tex and then to translate them into a text readable by Braille terminals or vocal synthesis.

The essential part of our work concerns the elaboration of an intelligent interface existing between the document base and the final user. This software layer enables, on one side, to manage the user interface, the user profile, the query. On the other side, it allows the communication with existing search engines and with the intermediate documentary base.

In the next part, we will detail the approach used to design and implement such intelligent interface. Then we will present the detailed system architecture and its functioning.

4. Architecture of the Proposed System for Filtering and Information Retrieving

We are particularly interested in the elaboration of a personalised retrieval system using intensive filtering, depending on particular user interests or user profile. This profile is initially defined by the user itself, and will be then refined and managed. Thus, our system will be based on various software entities having different specialisations:

- document retrieval,
- extraction of relevant information from within documents,
- document comparison and ranking according to the user profile,
- user profile management.

These entities we propose must be autonomous to make decisions (e.g. to keep a document, analyse the information relevancy etc.); however they should communicate with each other. Moreover, the environment that our system manages (the documentary base and the user behaviour) is completely evolving, so these entities require the capacity to evolve and adapt, by the use of appropriate learning algorithms.

To do so, a multiagent approach seems to be perfectly appropriate. A multiagent system is able to manage the dynamic aspects of the environment and the required concurrency. In fact, if we refer to the definition of 'agent' given by specialists [FER 95] [DER 97], agents are software entities having three properties:

- Autonomy: goal definition and reaching capacities without any human control.
- Collaboration: possibility to cooperate and exchange information with other agents or humans in order to reach a specific aim.
- Learning: ability to adapt to the environment.

So, the proposed multiagent system is a set of specialised agents, collaborating with each other. Its architecture is composed of three major agent layers :

1. User communication layer:
   - ◊ query treatment,
   - ◊ information extraction and restitution.
2. Information filtering and user profile management layer.
3. Document retrieval layer.

The multiagent system architecture is shown in Figure 3.

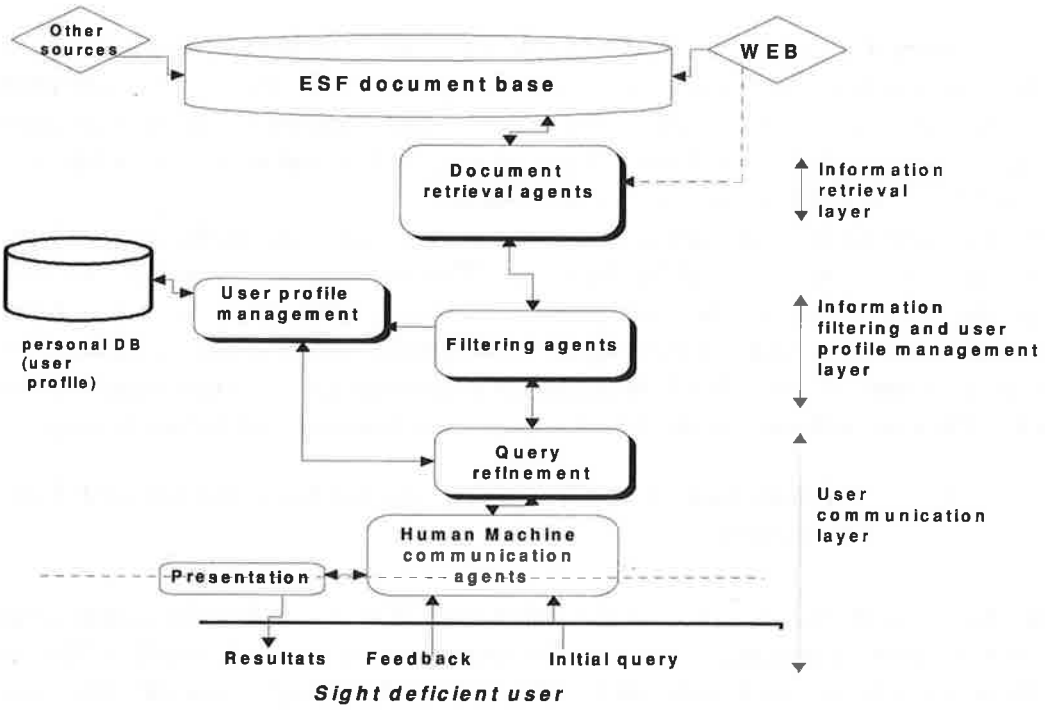

Figure 3: Multiagent architecture of the proposed system

In the next part, the agents and their roles are briefly presented. The detailed information retrieval process is described in section 5.

4.1 User communication layer

This layer is composed of query refinement agents, document presentation agents and human-machine communication agents.

a) Query refinement agents

They perform:

- ♦ query linguistic analysis:
  - ◊ morpho-syntactic analysis,
  - ◊ semantic analysis based on a thesaurus.
- ♦ query expansion and enrichment:
  - ◊ finding the different concepts included in the user query,
  - ◊ generation of different queries enriched from the initial user's query.

b) Document presentation agents

The collected and filtered documents will be presented to the user. The document presentation agents enable us to extract information from documents and then translate the extracted information before presenting it to the final user.

c) Human machine communication agent

These agents collect relevance feedback associated to the documents. The relevance feedback is a user evaluation of the relevance of each presented document. The learning process of the system is based on the document relevance feedback (see section 5).

4.2 Information filtering and user profile management layer

From the collected documents, a personalised document filtering (based on the user profile) will eliminate non relevant documents. The management user profile agents use learning algorithms to evolve and adapt their knowledge about the user profile (§V.6).

4.3 Information retrieval layer

This layer is composed of document retrieval agents. These agents call a meta search engine to perform the document retrieval. The retrieval is based on queries sent by query refinement agents.

5. Description of the Information Seeking Process

The document seeking process include six steps, as shown in Figure 4. The seeking process consists of learning loops (short term learning). The number of loops is reduced when the knowledge base of the user profile is accurate and exhaustive see section 6).

When the sight-deficient user expresses a query, he/she can use a keyboard, a Braille terminal or a vocal recognition system. His/her query will then be translated and treated by query refinement agents .

In order to obtain better documents answers, we perform the query refinement step. During this step, we first analyse the query and then enrich it with appropriate keywords.

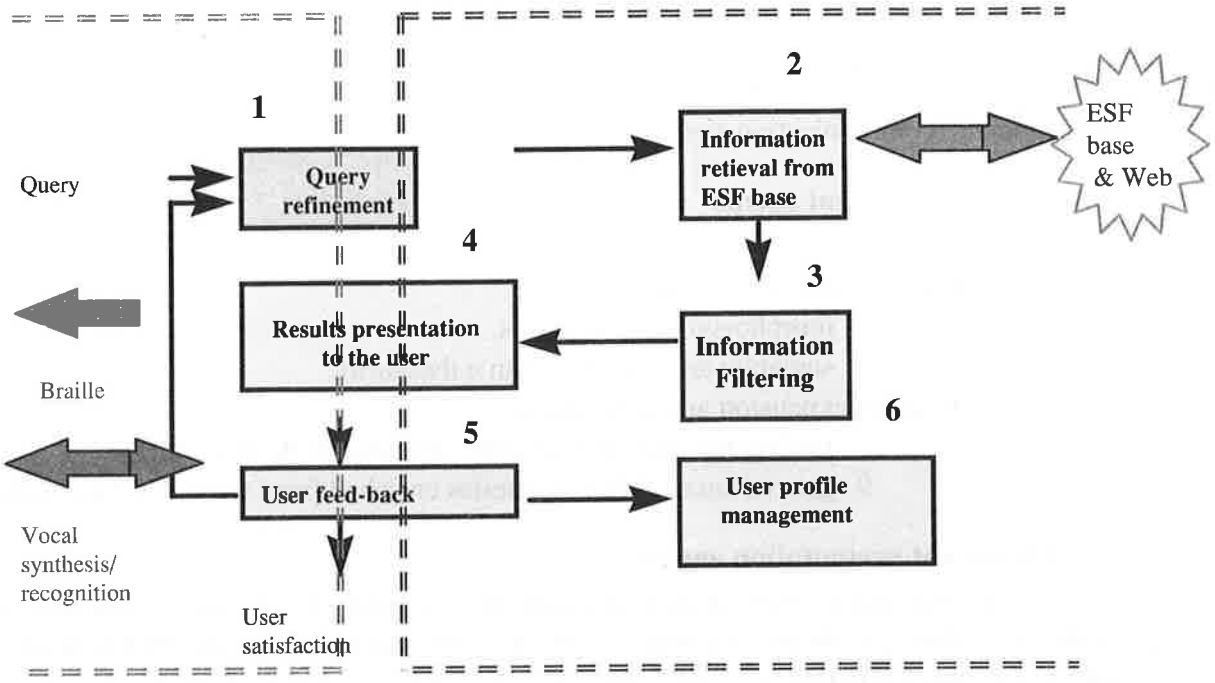

Figure 4: Different steps of the search process

5.1 Query refinement

Given a blind user, searching in a documentary base, express a query.
Example of a query:

Relation between object Petri nets and coloured Petri nets ?

a) Linguistic analysis

A textual analysis based on a linguistic approach such as SPIRIT or SENSE [LUB 97] will perform the query syntactic and morphologic analysis and eventually multilanguage translation, in order to find term associations.

Results of the morpho-syntactic analysis: *Object petri nets , coloured petri nets , petri nets*

After the morpho-syntatic analysis step, a query semantic analysis based on a thesaurus will be performed to find adjacent terms (synonyms, generic terms, specific terms, etc.) of the various terms included in the user query.

Results of the semantic analysis:

Nets: deadlocks, stochastic
Object: oriented, persistent
Petri nets: transitions, tokens, places, tool, marking, formalism
Object Petri Nets: analysis, distributed, concurrency, specification, graphical

Thus, the semantically associated term to the term 'object' are 'oriented' and 'persistent'. These terms will be used in the query enrichment step.

b) Query enrichment and expansion

After the three steps of the textual analysis (morphologic, syntactic and semantic) of the query, the various concepts figured out by the semantic analysis will be used to generate enriched queries [JER 97].

5.2 Document retrieval from ESF library

Enriched queries are sent to document retrieval agents. These agents call a meta search engine to perform the document retrieval from the ESF libraries. We have chosen to reuse an existing search engine, including some intelligent components. This meta search engine will be optimised by our intelligent interface. In the literature, different intelligent robots are proposed such as *Harvest* or *CIG Searchbots* or *Autonomy* , and they all use a multiagent approach based on adaptive form recognition. But these robots can be considered as laboratory prototypes.

Currently, one of the most commonly used meta search engines is *Savvy-Search*. On the basis of a unique query, the *Savvy-Search* engine calls 24 search engines at the same time, such as AltaVista, Excite, Infoseek, Lycos, Opentext, Yahoo etc. *Savvy search* makes a choice among them according to some criteria such as the keywords of the query, the traffic estimation, the index response time, the server workload. *Savvy Search* is the meta search engine that is used in our prototype.

The documents collected by the search engine are then analysed and filtered by agents in order to remove the irrelevant documents and to pick up the relevant ones.

5.3 Document filtering according to user profile

The document filtering phase is based on four steps: document vector representation, filter constitution, document ranking, and information filtering. These steps are detailed in the following paragraphs.

a) Document vector representation

There are three major information representation paradigms: statistical, linguistic and structural.

- *Statistical approach:* it emphasises statistical correlation of word counts in documents and document collections. Salton [SAL 86] [SAL 94] [SAL 96] describes the use of statistical schemes such as vector space models for document representation and retrieval. Another example is Latent Semantic Indexing which captures the term associations in documents [FOL 92].

• *Linguistic approach:* it is usually based on a three-step analysis (morphologic, syntactic and semantic). The semantic approach could involve the use of a thesaurus and encoded relationships among terms.

• *Structural approach:* it takes advantage of the structural information typically available in structured documents (e.g. titles, headlines; etc.).

We propose to use (for the moment) the statistical approach to analyse documents for vector representation. In order to perform the document vector representation, various terms or keywords included in documents are extracted, and their frequency is processed according to the following formula:

weight = TF * IDF

TF: the term frequency in the document
IDF: number of documents including this term.

A document is represented by a vector of terms. Each term (or keyword) of the vector is associated to a weight. A weighted keyword vector has the following structure:

Keyword	Keyword						author1	author2	links
Weight	Weight						Weight	Weight	Weight

b) Filter constitution

The user profile is set of weighted keyword vectors that were considered very relevant in previous search sessions. So, during the current session, if we find a vector very similar to the user's query, this vector would be called a typical vector representing the typical document to be retrieved. These vectors will represent the filters of the current session. So, the collected documents are kept if they are similar or near to filters, and are removed otherwise. The vectors proximity process is presented in the next paragraph.

c) Document ranking

Documents are ranked according to their proximity to the filter extracted from the user profile. The proximity is defined by as follows [SAL 94]:

$$\text{Proximity } (V_i, V_j) = \sum_k w_i^k w_j^k$$

w_i^k is the weight of the term K for the vector V_i
Vector length is normalised.

Example:

V1 :

object	petri	nets
0.13	0.22	0.25

V2 :

Coloured	petri	nets
0.02	0.21	0.24

The terms 'coloured' and 'object' are not considered here, because they exist in only one vector.
Proximity (V1,V2) = 0.22*0.21 + 0.25*0.24 = 0.1062

d) Information filtering

The ranked documents are then filtered according to the next rule [SAL 86]:

if (proximity > threshold) then (keep the vector)

The documents kept are sent to the information presentation agent.

5.4 Result presentation to the user

Filtered documents are sent to information presentation agents in order to be evaluated by the user. These evaluations enable the system to refine its results further and to enrich its knowledge about the user's interests.

a) Textual translation of documents

This step is essential for sightless users. Documents are textually translated and figures are described to avoid ambiguity and to be understandable by sightless users. Currently, some translations can be processed automatically (mathematical or chemical formulae), but others can only be performed manually (schemata, graphs, etc.).

b) Information extraction

We associate an abstract to the translated documents. In order to avoid cluttering up the user, we restore only the document abstract. At this level, we suppose that only structured documents are processed. Thus, creation of the abstract will be performed by a structure based approach, using the title, heading, abstracts, keywords of the document. When the user receives these documents, he/she gives his/her feedback or evaluation about the relevance of each document.

5.5 User evaluation

The learning process of our system is based on evaluation of the results given by the user. The user marks the documents. The mark is scaled from zero to four:

0 out of subject
1 irrelevant
2 fairly relevant
3 relevant
4 strongly relevant

This feedback will be used to update the vectors in the user profile. The queries which have generated relevant results will be the basis for the next loops during the retrieval phase. These queries are sent to the query refinement agents to be enriched and refined again.

5.6 User profile management

The knowledge base concerning the user profile is a base of keyword weighted vectors. Relevance feedback is associated with each vector. Thus, quality of the results will depend on the exhaustivity and accuracy of this base.

During an initialisation phase, the user fills in a form in order to give minimal knowledge about his/her interests. This knowledge will be enriched and refined during the use of the system. These refinements are performed by learning algorithms.

During an information search session, the vector representing the user's interests is refined gradually depending on the user's evaluation of the intermediate results. This is called short term learning. Nevertheless, as time goes by, the user's interests could change. This evolution of interests is managed by learning mechanisms based on genetic algorithms, called the long term learning.

a) Short term learning

According to the user feedback, vectors representing the user profile are refined: their keyword weights are readjusted as shown in the next example:

F *a filter which contributed to a presentation of a document* ***D*** *to the user*
f *the user feedback given by a user for the document* ***D***

The vector representing the filter F is changed proportionally to the feedback received [SHE 94]:

$$w_{i+1}^{k} = w_i^{k} + \alpha * f * w_d^{k}$$

α *represents the agent sensibility to the user feedback.*
w_i^k *is the weight of the term k of the filter vector F*
w_d^k *is the weight of the term k of the document vector D*

The weight of each keyword of the filtering vector is modified according to the document feedback and the learning percentage.

The resulting effect of this readjustment is that terms or keywords already existing in the filter have their weight updated proportionally to the feedback. The new terms will be added to the filter.

b) Long term learning and genetic algorithms

The concepts of genetic algorithms is not developed in this paper. In fact, we only reuse the results in our specific learning algorithm. However, we briefly present here the general principles [MOU 96] [SHE 94].

The most competitive filter, which often contribute to producing relevant results, have high scores. They generate other vectors, very similar to the initial ones, called genetic outsprings. The outsprings represent modified versions of their parents. These outsprings will replace vectors with low scores, in order to keep constant the number of vectors in the system.

Thus, only vectors with high performance are kept up or subsists. The genetic outsprings inherit the major features of their parents, but their differences enable the system to look for new solutions. Thus, by natural selection, only relevant vectors subsist.

6. First implementation of our Multiagent System

We have developed a prototype which incorporates the essential steps of the process described in this paper. For the implementation, we choose a Java environment including the constraints specified by the multiagent systems (object oriented, distributed, concurrent, etc.) and at the same time, an emergent well known environment language.

The Java language enables us to develop applets. These applets constitute our multiagents system for information retrieval and filtering. Java is an object oriented language allowing the management of multitask environments. Agents communicate with users and with the document base representing the virtual library. As a first test, we use the WWW environment.

- The communication with the database is established via the TCP/IP protocol, a server manages the database access through the gateway JDBC (JAVA Database Connectivity) - ODBC (Open Database Connectivity).
- The communication with WEB servers through a meta search engine (*Savvy Search*) called by a navigator such as Netscape or Internet Explorer.
- Finally, the communication with the user is performed directly through Java applets.

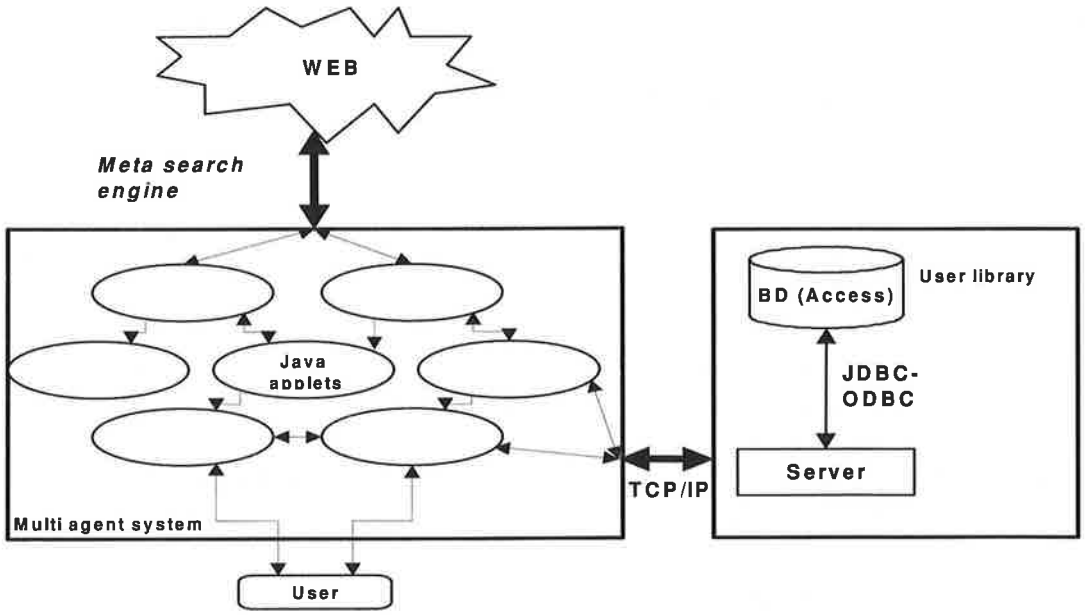

Figure 5: Prototype architecture

7. Conclusion

Information explosion in virtual libraries makes information retrieval very difficult for users. The need for intelligent retrieval tools is stressed in our case, regarding information access difficulties of sight deficient users. Moreover, scientific documents can not be directly

accessed by users, and specific translation and filtering are required (Braille and vocal synthesis devices).

In order to solve these problems, we first have studied the specific needs of sightless users, and examined existing tools for information retrieval. Then, we fixed our interest on the definition of an intelligent system able to analyse and to manage user profiles. Thus, we have chosen the multiagent approach as the basis of our system. This approach includes the ability to manage the evolution, the adaptability, and the concurrency of task processing. In our system, agents are specialised in information retrieval, information filtering, user profile management, query refinement, and human-machine communication.

This research was specially devoted to an intelligent retrieval tool in special fields (engineering science) for special users (sight deficient). Nevertheless, considering our approach, this work can be generalised for many other kinds of users and virtual libraries.

Moreover, in this work we do not implement a semantic approach to extract information and to summarise documents. This point will constitute one of the next steps of our research work in this project.

8. Acknowledgements

We wish to express our gratitude to professors Jean Caelen, Claude Decoret and Serge Portalier, and other collaborators in this project.

9. References

[BAL 96] **Balpe, J.P., Lelu, A., Papy, F., Salah, I.** *Techniques avancées pour l'hypertexte.* Edition Hermès, Paris, 1996 (Techniques de l'information) ISBN 2-86601-522-3

[DER 97] **Derdudet, D.** *La révolution des agents intelligents.* Internet Professionnel, Mai 1997, No 9 pp 74-80

[ETI 95] **Etizioni, O., Weld, S.** *Intelligent Agent on the Internet: Fact, Fiction and Forecast.* 1995. IEEE Expert, p44-49.

[FER 95] **Ferber.J.,** *Les systèmes multiagent vers une intelligence collective.* 1995, InterEdition, Paris.

[FOL 92] **Foltz, Dumais**. *Personalized Information Delivery:An analysis of Information Filtering Methods..* 1992. Communication of the ACM 35(12), pp 51-60.

[JER 97] **Jeribi, L.** *Recherche et filtrage d'information dans les hyperdocuments: approche multiagent.* 1997. Mémoire de DEA, LISI - Insa Lyon, France.

[LUB 97] **Lubkov,M.** *La percée du language naturel.* Archimag, avril 1997, No 103.

[MOU 96] **Moukas, A.** *Amalthaea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem.* Proceedings of the Conference on Practical Application of Intelligent Agents & Multiagent Technology, London, 1996.

[RAM 97] **Raman, T. V.***Net surfing without a monitor.* Scientific American (Internet special). March 1997. Http ://simon.cs.cornell.edu/Info/People/raman/raman.html

[SAL 86] **Salton, G.** *Text-retrieval systems.* Communication of the ACM. July 1986, N°7, p 648-655.

[SAL 94] **Salton, G., Allan, J., Buckley, C.** *Automatic structuring and retrieval of large text files.* Communication of the ACM. February 1994, vol 37.

[SAL 96] **Salton, G., Allan, J., Buckley, C.** *Automatic text decomposition and structuring* . Information proceeding &management, vol.32, no.3, pp. 127-138.

[SHE 94] **Sheth, B.** *Newt: a learning approach to personalised information delivery.* Master thesis 1994.