Using of multiple data source for information filtering: first approaches in the MedExplore project.

Emmanuel Nauer, Jacques Ducloy, Jean-Charles Lamirel

CRIN-CNRSet INRIA-Lorraine
Bâtiment LORIA

615, rue du Jardin Botanique

B.P. 239

F-54506 Vandoeuvre-les-Nancy Cedex

E-mail : {Jacques.Ducloy, Emmanuel.Nauer, Jean-Charles.Lamirel}@loria.fr

Keywords

Internet, Information Retrieval System, navigation graph, information statistical analysis, single link clustering, structured data use, multiple source.

Introduction

One of the major challenge of modern information retrieval and of technological survey is the mastering of the use of multiple and various data sources.

In this paper, we first describe the experimental workbench we have set up in the framework of the MedExplore project. This workbench allows both to merge and to cross data issued from multiple funds, including the Internet.

We will detail, an original navigational graph building method based on structured data. This method provides the user with a hypertextual access through thematics ordered by different generality levels. These thematics play the role of guidelines to help the user to formulate a query on the net, whenever his competence level or his type of need in the investigation field.

We will finally give various other samples from the MedExplore project that illustrate the usefullness of the crossing of strutured data. In that part, we will also describe our first heuristics to achieve contextual search on different survey areas.

1. The MedExplore Project

1.1. General overview

The aim of MedExplore is to give a group of experts confronted with an unforeseen field the mastery of its terminological resources together with a synthetic and deeper knowledge of the state of the art of that latter. This task will be achieved by creationing of a system of investigation.

Such a system (see below, figure 1) allows a user to navigate through concept graphs and to manipulate conjointly various pieces of information (large international databases, local source documents, raw information from the INTERNET), written in different languages.

We have chosen to begin our experimentations on biomedical fields because of the large amount of what we call "structuring" funds. Indeed, such funds like MEDLINE, EMBASE or PASCAL possess a homogeneous indexing and also a "quasi" knowledge based representation when they are associated with projects like UMLS.

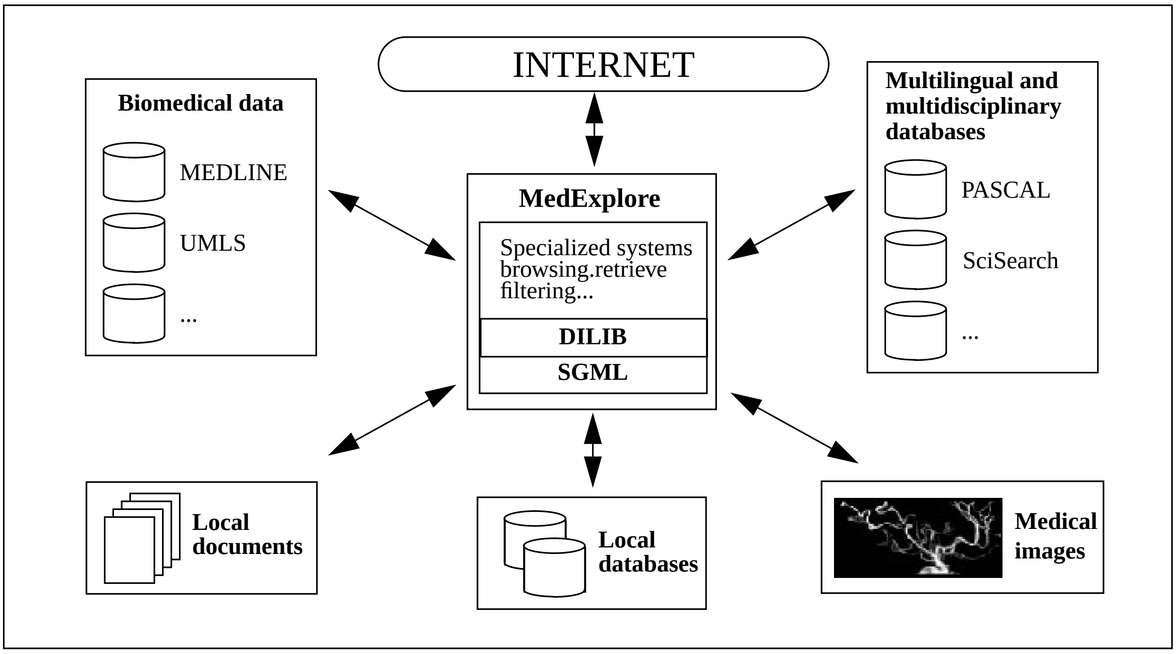

Figure 1 : Overview of the MedExplore project

1.2. Using MedExplore : basic principles

As we have mentioned before, we mainly aim at providing access for different types of users (from the traditional "end-user" to the "expert in data analysis") to a server which deals with different data sources in a coherent and homogeneous way (see figure 2).

For that, we have to solve the different levels of inter-operability between heterogeneous data. SGML/XML brings us a good answer for the "codification - structuration" level. Now we have to deal with more semantic levels. As we work in specialised area, we have chosen to simplify this problem by defining a core vocabulary which contains a limited number of terms (between 100 and 300) and which represents a kind of semantic gateway between the different databases.

Such a lexicon can be generated by automatic tools (for instance clusterisation) and improved by a specialist if necessary.

1.3. First approaches for the use of structured data for information filtering

The use and the crossing of structured data from multiple sources tend to facilitate IR in several ways. We will describe hereafter more precisely some of the main advantages of this approach which was dem-

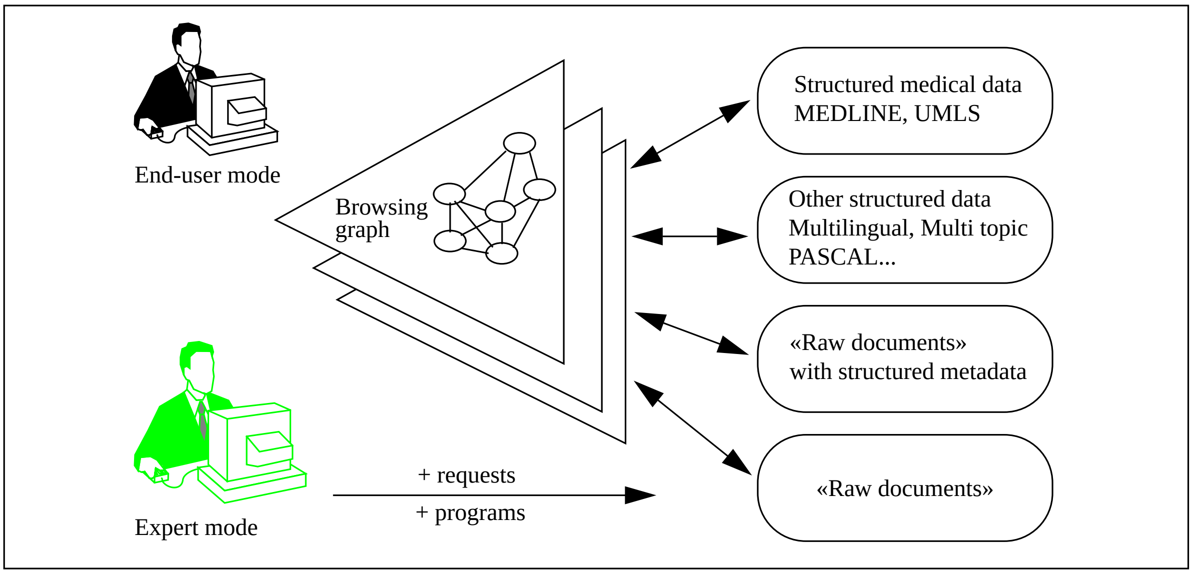
End-user mode
Expert mode
Browsing graph
+ requests
+ programs
Structured medical data MEDLINE, UMLS
Other structured data Multilingual, Multi topic PASCAL...
«Raw documents» with structured metadata
«Raw documents»

Figure 2 : Homogeneous access to the data through a navigation graph

onstrated in our MedExplore experimentations.

Thematic access to WWW

In [NAU97], we have already described how simple bibliographic references issued from Medline allow the automatic building of specialised investigation system. The proposed system was composed of a thematic navigation graph which represents the interface between the user and WWW. The main task of the system was to choose automatically the field vocabulary to lead the user to overcome his limits in terms of vocabulary, knowledge and memory. The user could therefore simply formulate the queries to submit to the search engines through terms selection. This conceptual view of query formulation has been also adopted by AltaVista [ALT97] with the "Refine" options (formerly "Live Topics"), or by Excite [EXC97], which proposed a word set to extend the initial query. Nevertheless, our approach seems more accurate, because it works on field knowledge instead of statistical links. Indeed, it has turned out that these last links are often inappropriate (verbs, idioms,...)

Multilingual access

The complementary use of UMLS allows us to provide the system with multilingual capabilities, allowing then thematic access in different languages. Therefore, the previously described information gateway, trough keywords, also plays the role of finding linguistic equivalences for the selected terms considering the interrogation language. For example, a French user could access to a French thematic graph whilst interrogating transparently the search engine in English. This is a convenient solution for an unskilled user with the English translation of his native query terms.

Complementary information on documents

Remote documents found during query sessions could be dynamically enriched through hypertext inverted links to the graph themes whose role represents complementary information for the user. As a matter of fact, they allow him both to operate accurate thematic linking between the Internet documents and the themes constituting the navigation graph and to retrieve directly bibliographical information which could be useful for his explanation(s).

Generating Dublin Core Metadata.

From the computer point-of-view, the data-crossing process produces a set of resources, such as tables, which can be also used in a generation process.

We are working in that way for the generation of a server allowing browsing through a collection of Medical Images. Each image is described in French with a short set of information :

```
<doc>
    <id>lethor/007_001</id>
    <auteur><e>Lethor JP</e><auteur>
    <specialite>Cardiologie infantile</specialite>
    <tech><e>Radiographie</e><tech>
    <organe><e>Coeur</e><e>Poumon</e></organe>
    <patho><e>Tétralogie de Fallot</e></patho>
    <motif>rupture de patch infundibulaire</motif>
    <age>12 ans</age>
</doc>
```

For this image, we can generate an HTML page whose metadata are the following:

```
<meta  name="DC.title" lang="fr" content="Image : tétralogie de Fallot">
<meta  name="DC.creator" content="Lethor JP">
<meta  name="DC.subject" lang="fr"
       content="Coeur, COEUR (RADIOGRAPHIE ), Poumon, POUMON (RADIOGRAPHIE ),
       Radiographie, Tétralogie de Fallot">
<meta  name="DC.subject"
       content="Heart, Heart_Radiography, Lung, Lung_Radiography, Radiography, Tetralogy of Fallot">
<meta  name="DC.subject" scheme="MESH"
       content=" Heart, Lung, Radiography, Tetralogy of Fallot">
```

Help for technological and scientific survey

The mastering of the analysis of great information resources available on Internet has been, for the last recent years, one the main challenge of technological and scientific survey. The information being on the net, thanks to constant evoluation, allows recent data analysis, in opposition with studies based on classical documentary databases [AND97][ROS93].

Besides, search engines on the net (AltaVista, Excite, Lycos, etc.) suffer from the same defects as classical documentary systems [DUB94] : the absence of a reliant indexing of their documents cause them difficulties to give back relevant documents to the user. In fact, these systems propose most of the times as a query response : - too many documents ;
- documents with low relevance considering the user’s real need.

Therefore, achieving a successful Internet search obliges introducing search mechanisms using knowledge which is specific to the interrogation field. This mechanism, we have call "Contextual Information Filtering", has also been implemented in the framework of the project MedExplore. It is described hereafter (see ???).

2. 2. A technical base for MedExplore : the DILIB workbench [DIL97]

The engineering techniques used for MedExplore is based on the generalisation of SGML codification allowing the use of SGML toolboxes and linguistic modules libraries.

Therefore, we have developed DILIB, an SGML workbench [DUC94] which contains a set of basic components to build Information Retrieval Systems

2.1. SGML, homogenisation of information

We use to convert all information in an SGML markup [ISO86] whose structure is very close to the original one. For instance a downloaded record issued from MEDLINE such as:

```
AN : 96081277
TI : Orthotopic pulmonary valve replacement with a homograft.
AU : Saha K,Iyer KS, Sharma R, Bhan A, Airan B, Venugopal P
CS : Department of Cardiothoracic and Vascular Surgery, All India Institute of Medical Science
JN : J Heart Valve Dis CP : (ENGLAND)
PY : Mar 1995
VO : 4 (2) p187-91
...
```

becomes:

```
<MEDLINE>
    <AN>96081277</AN>
    <TI>Orthotopic pulmonary valve replacement with a homograft.</TI>
    <AU><e>Saha K</e><e>Iyer KS</e><e>Sharma R</e><e>Bhan A</e><e>Airan B</e><e>Venugopal P</e>
    </AU>
    <CS>Department of Cardiothoracic and Vascular Surgery, All India Institute of Medical Sciences</CS>
    <JN>J Heart Valve Dis</JN>
    <CP>(ENGLAND)</CP>
    <PY>Mar 1995</PY>
    <VO>4 (2) p187-91</VO>
    ...
</MEDLINE>
```

In some case, it may be interesting to carry out little transformation on some data sets. For instance, we need to handle multilingual information coming from UMLS whose records look like that :

```
C0017379|ENG|P|L0017379|PF|S0022690|Carriers, Genetic|
C0017379|ENG|P|L0017379|VW|S0044411|Genetic Carriers|
C0017379|ENG|P|L0017379|VWS|S0022684|Carrier, Genetic|
C0017379|ENG|P|L0017379|VWS|S0044407|Genetic Carrier|
C0017379|POR|P|L0436728|PF|S0561010|TRANSPORTADORES GENETICOS|
C0017379|SPA|P|L0447330|PF|S0571612|PORTADORES GENETICOS
```

where C0017379 identifies a unique concept with an English prefered form (*Carriers, Genetic*), various usual forms and some translations.

For an easier data management, it is more convenient to group all information related to a particular concept, which is originally distributed in a table format, into one SGML record like that :

```
<CONCEPT>
    <CUI>C0017379</CUI>
    <TP><PF>Carriers, Genetic</PF>
            <VW>Genetic Carriers</VW>
            <VWS>Carrier, Genetic</VWS>
            <VWS>Genetic Carrier</VWS>
    </TP>
    <VL l="POR">
    <TP><PF>TRANSPORTADORES GENETICOS</PF></TP>
    </VL>
    <VL l="SPA"><TP><PF>PORTADORES GENETICOS</PF>
    </TP></VL>
</CONCEPT>
```

In the same way, we apply such transformation on all the managed data, obtaining an SGML mark-up on which we can apply the associated engineering possibilities.

2.2. Handling SGML information with DILIB

DILIB provides a set of tools to handle SGML or XML elements. They are available at different programming levels.

For instance, if you want to add the key-word "*AIDS*" as an element tagged with <e> to an SGML element which is pointed by "kw" variable in a C program, you have to write:

```
SgmlAddChild (kw, SgmlCreateLeaf("e", AIDS));
```

In the same way, you can use shell commands to handle sets of records. We have introduced a "path pattern mechanism" to specify a set of elements into a given document. For instance, if you want to select records which contain "*AIDS*" as a subpart of keywords and print the corresponding titles, you just have to write:

```
SgmlSelect -g MEDLINE/KW/e#AIDS -g MEDLINE/TI -p @g2
```

(where "-g" is used by analogy with *grep* and *@g2* identifies the 2nd "g" sub-command)

As HTML and Dublin Core deals with SGML, it becomes very easy to generate metadata in a programming environment. For instance, the following program :

```
SgmlNode *meta;
meta= SgmlCreateEmptyMark("META");
SgmlSetAtt(meta, "name", "DC.subject");
SgmlSetAtt(meta, "content", "AIDS");
```

will produce :

```
<META name="DC.subject" content="AIDS">
```

2.3. Information analysis with MedExplore/DILIB

Now, if we want to know the vocabulary which will be appropriate to retrieve relevant documents or to produce significant metadata contents, we have to analyse a set of information.

For that purpose, DILIB also contains a set of basic components to build customised information retrieval systems. These tools allow a global analysis of large sets of information. Up to now, our tools have been mainly based on basic statistical approaches. An illustration of such an approach is the navigation graph building mechanism (clusterisation) described hereafter.

Clusterisation

The navigation graph building is based on a single link clustering method [JAR71]. This method works by iterating on keywords associations issued from the documents, ordered by decreasing relevance. Similarly to Michelet [MIC88], we choose the equivalence coefficient as the statistical indicator to order the keywords associations because it weights the associations importance relatively to their 2 component terms.

This coefficient is given by the following expression :

$$a_{ij} = \frac{f_{ij}^2}{f_i f_j} \qquad (1)$$

f_{ij} being the cooccurence count of keywords i and j in the documents and, f_i and f_j being respectively the i and j keywords occurence counts in the same documents.

However while making experiments, we found that the direct use of the equivalence coefficient on the whole association set, ordered by pertinence [MIC88], almost always inhibits the characterisation of the most general themes of a domain. That phenomenon mainly occurs when the field is strongly structured and composed by numerous very specific and very coherent subfields. Moreover, in some intermediate situations where the subdomains structuration is less strong, the direct use of said equivalence coefficient may lead to an uncontrolled mixture of general and specific themes.

To cope with the above described problems we propose an original two-step clustering method (see figure 3) :

The first step consists in establishing the generality level of the clusterisation by selecting the core associations set of the clusterisation thanks to the keywords cooccurrence count. An associations set with high cooccurrence count will always lead to general themes. Conversely, an associations set with low cooccurrence count will always lead to specific themes.

The second step consists in applying the classical single link clustering algorithm on the selected associations, reordering them by equivalence coefficient.

Finally, this method gives us the opportunity to adapt the navigation graph building to different types of users and to their different needs. The general graph will then be dedicated to the novice users with, most often, no specific knowledge about the explored field, limited vocabulary and general information needs. Conversely, specific graphs will be devoted to field specialists with more technical vocabulary, more elaborated knowledge and more focused needs.

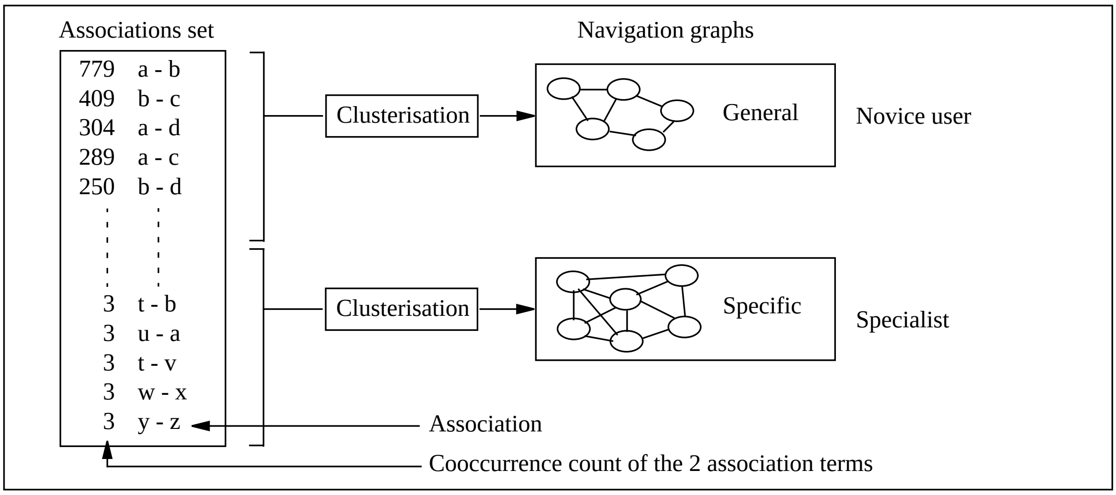

Figure 3 : Original two-step single link clustering

To order the resulting themes of each graph thanks to their generality level, we use a generality indicator which could be described by the below basic formula :

$$g_C = \frac{1}{\mathrm{Card}(C)} \sum_{(i,\, j) \in C} f_{ij} \qquad (2)$$

C being a given themes of the graph and *Card(C)* being the number of associations of the theme *C*.

3. Improvement with MedExplore : Contextual Information Filtering

3.1. General overview of the system

Making an Internet search successful implies solving a typical problem of the IRS : maximizing the number of relevant retrieved documents (recall) whilst minimizing the number of non relevant retrieved documents (noise).

Therefore, we have choosen to use structured knowledge to established a context around the central subject of the research. This technic we have called "Contextual Information Filtering" will limit the response scope and favour the emergence of relevant data.

Our first experimentations in that domain led us to set up a three component system (see figure 4):

The goal of the first component is to detemine the context around the focus of interest : this step is strongly guided by the expected results. Nevertheless, the general principle both uses structured data (sometimes linked with a core vocabulary), and information analysis methods which are for now statistical ones. In a short future, we aim at integrating linguistic methods such as : extraction of nominal groups, thesaurus use, ...

The second component implements the interrogation of the search engines, through multiple queries. At this step, our very first attempts originally use a single query. Unfortunately, our numerous experimentations show that it was very difficult, if not impossible, to find the optimal query. In most cases, adding terms will precise the query and so discriminate relevant documents from the non relevant ones. Nevertheless, the introduction of a general word (which indexes a very large scale of documents on the net) will sometimes alter the results given by the search engines. Indeed, these general words will distort the discriminant context, causing the emergence of non relevant documents. Therefore we choose to submit several well-formed queries and to add a synthesis step in order to suppress the side effect of the awkward "bad" queries.

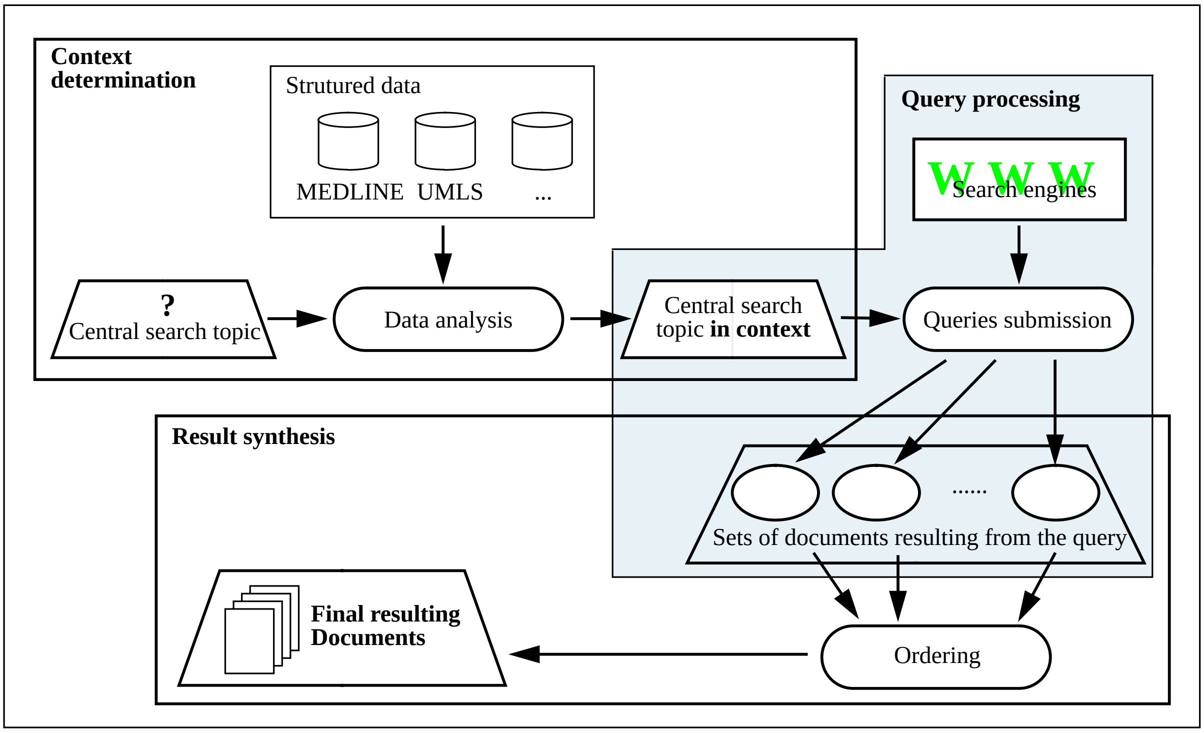

Figure 4 : Contextual Information Filtering System architecture

The last component deals with the results set to provide the user with a list of relevant documents. For that purpose, the weighted mean rank formula proposed in [LAM95] has prooven to be accurate. This ranking formula gives the weighted mean rank $r(\Re, d)$ of a document *d* thanks to a set of queries $\Re$ as :

$$r(\Re, d) = \frac{\sum_{i \in \Re} \alpha_i r(i, d)}{\sum_{i \in \Re} \alpha_i} \qquad (3)$$

where α_i is the occurrence weight of the document *d* in the query i and $r(i, d)$ is a function depending on the rank of the document *d* for the query *i*. When you want to eliminate from the final result the documents which have not been found relevant for one of the queries, the rank function could be set up to give an infinite rank to these documents. Conversely, to keep these documents, the rank function could be set up to attribute them the rank which is the nearest one of the one of the last documents found for the said queries.

We could mention that a single query might be considered sufficient in the very first step of a search session. In this case, the second component of our system will submit only one query to a search engine, and the ranking of the document will be directly given by the search engine query result (the third component will then be useless).

3.2. Samples of contextual information filtering

Large search on a single general keyword

For such an extraction of documents from the INTERNET, the principle consists in associating to the searched term a histogram of the most frequent terms cooccurring with this latter in the MEDLINE records. In databases, which are not indexed with keywords, we can also use full-text words, coming from the abstracts.

For instance, for the keyword "Newborn, Infant", in a local base dealing with "cardiology", the associated histogram (resulting from the abstract) will look like that :

[59] patient - [42] pulmonary - [37] tetralogy - [33] fallot - [29] infant - [27] heart - [25] artery - [24] defect

A single query on a search engine, using this set of words, will then give us the expected results.

Research work of an author

The same kind of technics can be used to search for the work of an author. For instance, the vocabulary associated to Pr. JP Lethor from his bibliography on MEDLINE :

[35] ventricular - [31] volume - [27] left - [26] dimensional - [20] coronary - [19] three - [16] method - [15] patient
[13] defect - [13] image - [12] excised - [11] doppler - [10] artery - [10] echocardiography - [10] tau - [9] pressure

A single query using the author name complemented by this set of words will then give us, again, the required results.

Research area of an author

To generalize the search around a research thematic, a histogram could be obtained thanks to authors working on that thematic or/and by using documents relative to this latter. Unlike the previous cases where the histogram was used to overdefine the initial query, we only made use of the generated histogram without the elements (keyword or author) from which it was built.

This approach seems to be very appropriate to achieve technological survey where one may not always be faced with precise and non-evolutive area.

Conclusion

In the very first steps of the MedExplore project, we have hightlit the usefullness of crossing structured data, coming from multiple sources, by testing and implementing several functionalities of an integrated IRS. This functionnalities include the thematic and multilingual access to WWW, the complementary information about retrieved documents and the generation of metadata for standard web documents.

We now explore in a deeper way and try to generalise one of these functionnalities, which we have called "Contextual Information Filtering". We will attempt to improve our retrieval performances, through different ways :

- the study of different strategies for the "queries submission" part of our Contextual Information Filtering System will help us to determine how to submit a minimal number of queries. We will particularly focused our work on the context definition. For that, we will introduce technics coming from linguistic or even from statistics. We will also use other types of structured data like the ones coming from a thesaurus.

- the experimental determination of the best ordering function will allow us to maximize the recall and to minimize the noise.

- the analysis of the Internet documents with very close technics will lead us to determinate the accurate metadata (from the ones being present in the documents), which could then be used to set up a better context. This context being more suitable to the search engines will then improve the search performance on the net.

Bibliography

[ALT97] AltaVista. Information and search engine access : http://altavista.digital.com/

[AND97] Pascal Andrei. *Elaboration et traitement d'information complexe pour l'aide à la décision stratégique.* Phd Thesis in Information Scientifique et Technique, University of Marne-la-Vallée, 1997.

[DIL97] Information about the DILIB workbench : http://www.loria.fr/DILIB

[DUB94] Jacques Emile Dubois et Belhadri Messabih. Internet-web and ST data management : harmonization and new horizons. In *The Information Revolution : Impact on Science and Technology*, pp 43-56, 14th International ConferenceCODATA, 18-22 septembre 1994, Chambéry, France.

[DUC94] Jacques Ducloy, Jean-Charles Lamirel et Emmanuel Nauer. A workbench for bibliographical or factual data handling. In *The Information Revolution : Impact on Science and Technology*, pp 63-70, 14ème conférence internationale CODATA, 18-22 septembre 1994, Chambéry, France.

[EXC97] Excite. Information and search engine access : http://www.excite.com/

[ISO86] ISO 8879. Standard Generalized Markup Language (SGML), 1986.

[JAR71] N. Jardine et R. Sibson. *Mathematical Taxonomy*. Wiley, Londres et New York, 1971.

[LAM95] Jean-Charles Lamirel. *Applications d'une approche symbolico-connectionniste pour la conception d'un système documentaire hautement interactif : le prototype NOMAD*. PhD Thesis in Computer Science, University of Nancy I, 1995.

[MIC88] Bertrand Michelet. *L'analyse des associations*. PhD Thesis in Information Sciences, University of Paris 7, 1988.

[NAU97] Emmanuel Nauer, Jean-Charles Lamirel. *Environnement d'investigation sur WWW assistance à l'utilisateur par des connaissances fédérées*. In H2PTM'97, pp 101-113, 4th International Conference : Hypertexts and Hypermedia - Products, Tools and Methods, Hermes, September 1997.

[ROS93] Hervé Rostaing. *Veille Technologique et Bibliométrie : Concepts, Outils, Applications*.PhD Thesis in Sciences de l'Information et de la Communication, 1997.