Ontologies as Background Knowledge to Explore Document Collections

Nathalie Aussenac-Gilles & Josiane Mothe
Institut de Recherche en Informatique de Toulouse, 118 route de Narbonne,
F-31062 Toulouse Cedex 04, France
{aussenac/mothe}@irit.fr

Abstract

This paper introduces a new approach to provide users with solutions to explore a domain via an information space. A key point in our approach is that information searching and exploring takes place in a domain-dependent semantic context. A given context is described through its vocabulary organised along hierarchies that structure the information space. These hierarchies are simplified views on a more complex domain specific ontology, that form a shared and coherent background knowledge representation. So the system benefits from the combination of two innovations that make the exploration of the document space more effective. First, hierarchies (extracted from the ontology) provide with a language and synthetic representation to be explored by the users to express their information need. Additionally, a visual interface presents answers to their queries using multi-dimensional analysis and a global visualisation of document collections. At both stages, ontology is the key structure that guides a meaningful browsing for query formulation and for the document set exploration.

Key-Words

Information exploration, information retrieval, ontology, document visualisation, concept hierarchy, multi-dimensional analysis

1 Introduction

Information access is generally seen as helping users to find the right documents or chunks of documents. This is the common interpretation in Information Retrieval (IR) for which the goal is to retrieve an ordered list of document references according to a natural-like query. A search module in IR matches query and document representations. Different approaches have been proposed in the literature to enhance system effectiveness, specifically methods to improve the document representation or the document–query matching, for instance by query reformulation. Query reformulation can be done before a first search is actually performed. In that case the system adds terms that are correlated with the initial query. These terms can be chosen from external resources such as thesaurus-like information (either generic such as WordNet[1] or domain related such as MeSH[2] for the medical domain) or extracted from the documents themselves (e.g. using term co-occurrence [Smadja,1993]). Then one of the difficulties is to decide whether the reformulation should be done or not. Intuitively, it is more appropriate when none or a few documents would be retrieved using the initial query and the resulting set is likely to be small. Alternatively, query reformulation can start after a first search and in such a case, the terms the system adds to the query are extracted from relevant/non relevant documents [Rocchio,1971]. Relevance feedback implies that at least a few relevant documents are retrieved according to the user's initial query.

With regard to document representation, which is a key point in IR, a common solution is to choose significant sets of -weighted- terms. Several works have investigated a richer representation in order to get better query matching. Natural Language Processing (NLP) is one of the means that have been tested. An alternative way to go beyond bags of words could be to organise indexing terms into a more complex structure than "bags", such as a hierarchy or an ontology. Texts would be indexed by concepts that reflect their meaning rather than words considered as chart lists with all the ambiguity that they convey.

[1] www.cogsci.princeton.edu/~wn/

[2] http://www.nlm.nih.gov/mesh/2003/MeSHtree.html

Intuitively, if indexing terms are organised into a hierarchy, reformulating a query can rely on this hierarchy. More specific or more generic terms can guide the user to find a better formulation of his/her information needs. In this context, ontologies appear as a promising knowledge structure the use of which is worth being validated. One of the key points in IR, which has not been investigated much, is that both the user's knowledge and the knowledge that is implicitly attached to the document collection or to the domain should be more fully considered. We argue that ontologies can be used to achieve this goal.

We promote an approach where information search and exploration take place in a domain-dependant semantic context. A given context (e.g. astronomy) is described through its controlled vocabulary organised along hierarchies which are all extracted from a single and unifying domain ontology. Each hierarchy reveals a given point of view on the domain, that is to say a dimension. Examples of dimensions in the Astronomy field are: astronomical objects, measurement instruments, observatories, journals in astronomy, scientists. Whereas a dimension in IR refers to a list of words extracted from the document contents, a hierarchically organised dimension provides users with a structured organisation of this vocabulary. Its structure gives more meaning to words and presents some domain knowledge which the user is generally missing. Going back to the example of the astronomy field, the astronomical object hierarchy consists in a four level hierarchy with types and sub-types of objects and leaves are the specifics objects (e.g. stars / binary starts / one specific star).

In our approach, the ontology and derived hierarchies provide the query language for users. Not only can the concept hierarchies be browsed by the user, who can select the terms he wants to add to his query, but they also allow them to explore the information space according to different points of view, through the domain vocabulary and its structure. For instance, the exploration focus or query can be expressed through more or less specific terms, according to their position in the hierarchy. Moreover, dimensions and their visualisation define a novel way to provide the users with global views and knowledge of the document collection. A key component of our approach is that the domain ontology allows to define a visual presentation of the entire collection or of a sub-collection based on multi-dimensional analysis, as it is done in OLAP systems [Chaudhuri,1997]. The quantity of information (document) associated to different dimension values is presented to the user so that he has an idea of what he can find in the collection. This information can be displayed at different levels of granularity corresponding to the different levels of the concept hierarchies (using either specific terms or generic terms).

This paper is organised as follows. In section 2, we first describe related work regarding the exploration of document collections based on browsing facilities. In section 3, we explain how ontologies and hierarchies can be used for IR and more precisely for browsing large document collections. In section 4, we present how a domain is represented in our approach through an ontology and associated hierarchies that correspond to document dimensions. The next section shows the benefit brought by the combination of hierarchies and visual interfaces from a user’s point of view. The implemented system, DocCube, has been validated in two domains: economy and astronomy.

2 Related work

The approach we present involves many research areas including document categorisation, information visualisation, text mining and ontology engineering. In this section, we focus on interfaces that aim at helping a user to access collections of documents.

One main goal of the information visualisation is to present abstract data. In the IR field, most of the tools provide the users with search and selection functionalities. The tools provide facilities to organise and view documents according to their semantic content. [Chalmer, 1996] proposes a tool (Bead) to display bibliographic documents as cubes where the documents are placed on the surface in clusters according to the document similarity. In the same way, [Fowler, 1996] presents a tool in the context of web documents. The documents are linked according to their similarity and the user can navigate through this structure. These two interfaces are more related to document clustering and document set browsing. On the contrary, in [Benford, 1995] the query is the staring point. In this interface, the documents are displayed in a 3D space where the axes correspond to key-words given by the user. Similarly, [Dubin, 1995] presents VIBE (Visual Information Browsing Environment) a visual IR system where the retrieved documents are displayed according to the query space. The Isidor interface is based on the same principle, but in this interface the information space is represented under the form of a cone where the axes correspond to query terms [Chevalier, 2000]. The same kind of approach is used in TOFIR [Zhang, 2001] where special view points are defined. These points correspond either to query components or to other interesting points. In these interfaces, the query terms are considered independently and the possible semantic links between them are not visualised. The interface is only used to display the documents that have been retrieved by the system but neither to browse the collection or to query it. [Mukherjea, 1995] presents a system where the documents are display under hierarchies. The user can specify attributes which induce a change in the organisation and the display of the document set. Cat-a-Cone [Hearst, 1997] is an interactive interface allowing the document collection consultation through the search and browsing of category hierarchies to which the documents are associated. The book metaphor is used in this interface in which the left-hand page displays meta-data (title and categories) and the right-hand page displays the document content. Hierarchically organised categories are shown under the form of a ConeTree which can be browsed by the user, when selecting a node, the associated documents are retrieved. The underlying document representation and querying/browsing facilities are similar to the one we use. However, in our interface, several points of view on the documents are considered (several document dimensions), a single one is used in Cat-a-Cone. In addition, in our interface, browsing is not only used to access documents and query the system but to get an overview either of the collection content or of the set of retrieved documents.

3 Knowledge structures and Information Retrieval

3.1 Ontologies: some definitions

The term « ontology » started to be used in artificial intelligence in the early 90’s. The first definitions enhanced their differences with the philosophical definition where ontologies are used to report on the nature and essence of things and beings. Definitions in computer science insist that ontologies are generic and conceptual formal descriptions of the domain entities required to design sharable and inter-operable knowledge-based application [Uschold, 1998]. They provide a sound basis for communications between human and machine agents, or among machine agents. The objective of “defining meaning” urges us to thoroughly distinguish between objects and beings, the symbols that stand for them (words or phrases) and the representations required by an agent that uses this symbol to evoke this being. Concepts correspond to these representations, whether mental or formal. The notion of ontology has evolved quickly: in spite of on-going interest for general ontologies with a universal scope such as Sensus[3], Cyc[4] or universal lexical data-bases such as WordNet, ontologies often now refer to domains, and sometimes even to some tasks, that restrict their scope. This restriction is the right means to manage a good usability: as long as ontologies are models, they are partial specifications of the world, and imply a normalisation according to a given focus. A more realistic trend is to build ontologies depending on the treated context and on the focus of interest for users.

A wide range of knowledge structures hide behind the word “ontology”, which, in practice, may refer to word hierarchies (thesaurus), structured terminologies, terminological knowledge bases or actual formal ontologies. In fact, these structures differ according to their scope (domain specific or generic), their content type (terminological and/or conceptual), the type of semantic relationships that can be represented and their degree of formalisation. The way ontologies are designed, the criteria applied for concept selection and organisation also influence their nature [Guarino, 2000]. In the context of IR, ontologies are not necessarily represented with logic. Formalisation guarantees a semantic interpretation by software agents; it is a means to facilitate the management of concepts as objects, their classification, the comparison of their properties or even just ontology browsing.

[3] www.isi.edu/natural-language/resources/sensus.html

[4] www.cyc.com

3.2 Use of Ontologies to help the IR process

More and more work in IR is trying to improve text indexing or query formulation with the help of ontologies. IR (mainly on the web) can even be considered as one of the favourite application field for ontologies. For instance, ontologies are often presented as silver bullets for the semantic web [Fensel, 2001]. But what is the actual contribution of ontologies to the IR process? Ontologies are expected to bring different targeted gains [Masolo, 2001] that we have categorized and contrasted with currently achieved results as follows:

- Ontologies should improve recall and precision: The relations and axioms in an ontology provide the means to look for some concepts that are not explicitly written in the query. An ontology is precise enough to provide a unique definition for terms, or to deal with synonymy and ambiguity. Experiments carried out using Wordnet with a smart strategy showed that a significant gain is possible on the TREC data-set [Baziz., 2004], although this test had been previously negative [Voorhees, 1994].
- More importantly, with the help of the ontology, users should express their needs more easily: Browsing the ontology leads to the selection of relevant concepts and the definition of a query composed of selected concepts and their description.
- Ontologies should facilitate the IR from various heterogeneous knowledge sources.

Among others, the PICSEL project illustrates most of these gains [Reynaud, 2000]. A formal domain ontology mediates the information exchange between heterogeneous knowledge sources on the web and a user that looks for some information in these sources. The user feels like searching in a single information base. The system interface guides query formulation and adaptation in order to answer to the exact request or to similar ones. Recall is much higher then because if a request has no answer, the system calculates the closer formula that has an answer. Query adaptation and reformulation as well as translation towards each specific wrapper is made possible because the ontology is represented in a formal language. The cost for such a successful application is quite high : the core ontology dialogs with as many specific ontologies as there are knowledge sources. Each of them is built up by hand from database schemas or files structure analysis. The core ontology is a kind of rich gathering of all the specific ones.

To sum up, let us focus on the specific IR tasks that we are concerned with: document indexing and retrieval and exploration of document sets. These tasks are not much carried out with the help of ontologies yet. It is all the more surprising as these kinds of applications suit quite well with the use of ontologies when:

- The domain is closed, and covers the set of documents in the information stage.
- These applications just require a concept hierarchy because query extension or specialisation with the help of specific relation types is hazardous and difficult to use for a naive user.
- The classes of users are known well enough to anticipate their needs and their views in terms of domain knowledge.

We propose an approach where hierarchies, built up from ontologies, are used in a more unusual and promising way in combination with visualisation tools for a guided exploration of the information space, as shown in section 5. Our first hypothesis bears on ontology design: domain ontologies are all the more relevant for document exploration as they are built up from text analysis. However, how the ontologies are build up from the texts is out of the scope of this paper. Some solutions exists and are mentioned in section 4. Our second hypothesis is that ontologies provide a rich background from which various hierarchies may be extracted according to specific points of view. The use of hierarchies for document exploration is twofold. First, each hierarchy may correspond to a dimension to explore the information space. Then, hierarchies can help to compare several user's queries selected at a more or less abstract level along several dimensions. This enables the user to go through the information space according to several new and original insights that would have been too complex to express otherwise.

4 Structuring a domain through dimensions

Documents can be seen under many dimensions (or points of view) that could be used in order to extract some knowledge from their content. IR often relies on a single dimension, corresponding to the key-word list whereas dimensions can be of a great help. Providing a variety of dimensions and the possibility to combine them is a means to suggest relevant ways to explore a document collection or a set of retrieved documents that meet a given information need. First, some information can be derived to help the user to decide whether to access a given document or not. In addition, the various dimensions can be used to guide the information need refinement by providing information about the content of the document collection, and overviews based on different points of views. A system based on such dimensions for collection browsing would require two kinds of technologies: advanced visualisation tools and efficient facilities to build up adequate dimensions and use them for visualisation. Our research works aim at this kind of tools.

4.1 Multiple document dimensions

In IR, document indexing generally results in document representations that are based on a one-dimensional information space, the free-text space, which represents the document content. In doing so, the documents are viewed as bags of (weighted) words. However, texts express a vast and rich range of information that traditional information indexing does not take into account. Indeed, unlike information extraction, information indexing does neither consider the semantics carried by terms nor the role or the type of terms. Until now, this type of information is under used in IR. Some systems use meta-information, but mostly in order to filter documents that could be potentially retrieved based on their content (e.g. Web search engines and bibliographic servers such as Science Citation Index).

Meta-information and content are not the only document dimensions that users can be interested in. For example, in a science monitoring activity, other focus topic or document dimension could be “used techniques”, “discovered products” or “temporal references”.

4.2 Building domain-based ontologies from text corpora

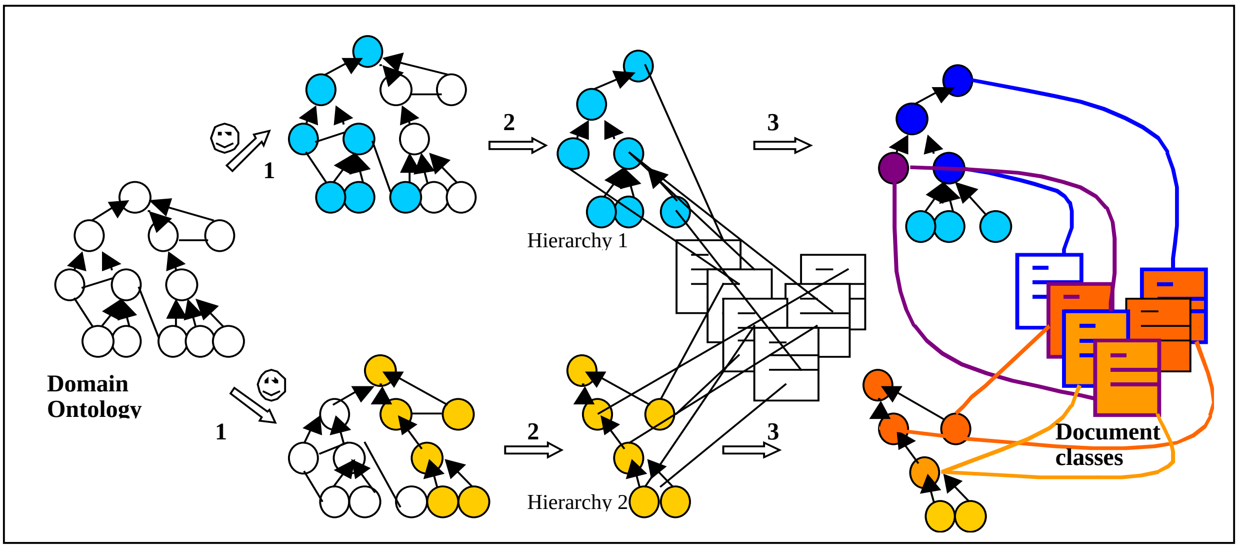
We refer to domain ontologies as defined in section 3: they provide a shared and formal description of the concepts and relationships of a domain grounded on the knowledge and terminology in use in this domain. In our application, domain coverage depends on the document collection. We suggest to build up or reuse an ontology very close to the vocabulary and knowledge in these documents. Indeed, in the past five years, several research works led to the definition of methods and tools for ontology engineering from texts. This trend raises a major hope: an easy and partly automated process to structure and maintain knowledge representations in keeping with the knowledge as expressed in representative texts of the domain. In this scope, NLP tools play a major role, for the identification of domain terms and concepts as well as semantic relationships.

Term extractors may be based either on syntactic principles, like Syntex [Bourigault, 2002] and Nomino [David, 1990], or on statistic principles as Ana [Enguehard, 1995] or Tétralogie [Chrisment, 1997]. Conceptual clustering tools like Zellig [Habert, 1996] or learning based tools like Asium [Faure, 1999] put together noun phrases that share syntactic dependency relations. The resulting clusters must be manually analysed to define semantic classes. *Relation extractors* usually are based on linguistic patterns such as Prométhée [Morin, 1999] or Caméléon developed in our team [Séguéla, 1999]. Patterns are applied on tagged text files to find out phrases where the lexical relation appears. Most of these systems are based on M. Hearst principles: patterns and relations may be either general or domain dependant [Hearst, 1992]. The use of these tools may require some linguistic skills but it gives significant information for structuring domain knowledge.

More complex platforms include several of these tools and support a modelling process for ontology engineering. KAON is a standard workbench for the analysis of texts in German and English [Maeche, 2003]. Terminae is a general workbench to build up ontologies from large text corpora in French or English [Aussenac, 2000]. The Terminae methodology suggest steps and heuristic principles to apply linguistic tools (Syntex and Caméléon), to explore their results, to identify and structure knowledge in an ontology. Validation is made through a formal representation in a description logic.

4.3 From domain ontologies to hierarchies

Ontologies can be used in an intuitive way to index documents. For the document categorisation task, a concept from an ontology is viewed as a category [Terrier, 2001]. Semantic relationships and inferences can be used in order to make more meaningful indexing. However, using ontologies as a querying language in a browsing way is more tricky. Indeed, the complexity of ontologies can make an interface too difficult to use. We promote an approach in which ontologies are transformed into hierarchies that are more easily browsed and used as a query language (see Figure 1). Each hierarchy corresponds to a dimension on which documents can be mapped.



According to his own point of view, each user *selects* a sub part of the ontology (steps1) and *defines a hierarchy* with it (step2). This can be done once for all by the system manager who knows the different dimensions users may be interested in. The hierarchy nodes are concepts with their labels and synonyms. After projection of the terms onto the documents where they occur, hierarchies are automatically connected to documents. A document can be associated to several classes/hierarchy categories, depending on the terms that occur in the texts. When associated to a given class, a document is implicitly associated to the parents of this class.

Figure 1 : Ontology as a shared representation for Hierarchy definition and document classification

In our approach a dimension corresponds to a domain hierarchy, which can be viewed as a sub-part of the domain ontology. Then, in a formal hierarchy, concepts are organised into classes and sub-classes along a genericity dimension. This genericity relation can be defined from the IS-A relation or any other kind of relation in the ontology. For instance, our reference ontology contains relations between the concepts *country* and *laboratory: laboratory BELONGS-TO country* and relations between *author* and *laboratory* : *author WORKS-IN laboratory* (see Figure 2). From the user's point of view, all these concepts are associated to the *publication* concept because the author's affiliation (country or laboratory) can be viewed as the producer of a publication. So, in the hierarchy, the two above relations are used to define a genericity dimension where *country* is the generic for *laboratory* which is in turn the generic for *author*. All the publications of authors from the same country are considered as the publications from this country. In a similar process, some concepts may be considered as equivalent from the user's point of view. They may be gathered and form a single class in the hierarchy.

Additionally, the various labels of each concept in the ontology are gathered as synonym terms. So a hierarchy defines a controlled vocabulary organised according a specific point of view on the domain. These hierarchies are defined once for all for each domain by the application manager, according to the main possible users' focuses.

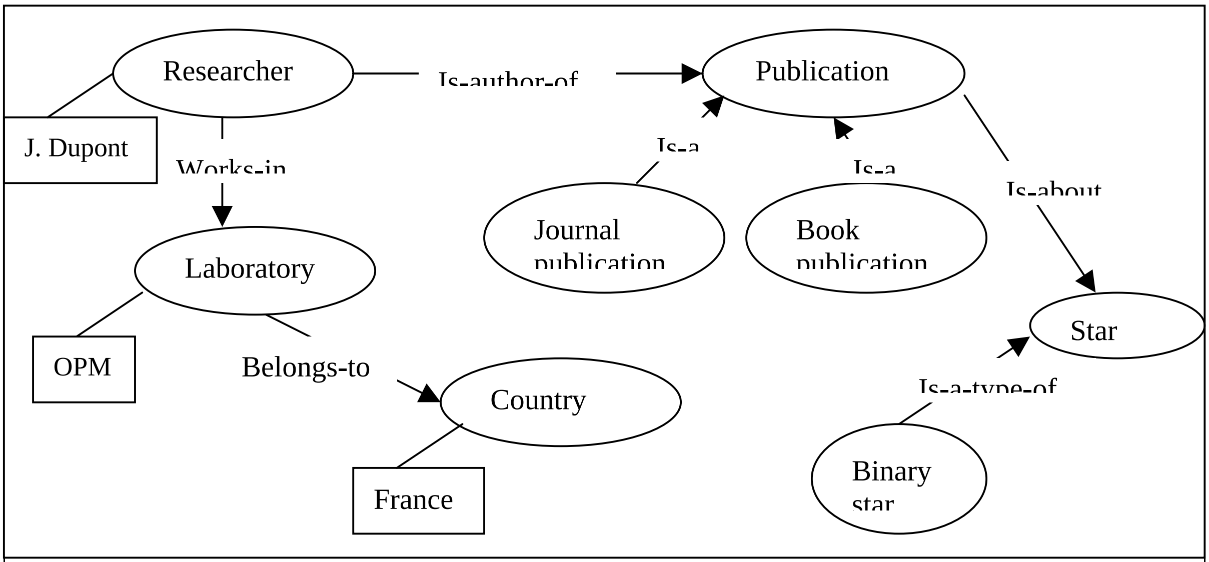


Figure 2 : Sample of an ontology used for bibliographical investigations

In the interface, hierarchies are shown in two different ways at two stages of the collection exploration: for query formulation and for the exploration of a document set (either the collection or a subset of it, e.g. retrieved documents) (see section 5).

4.4 Association of documents to concept hierarchies

Concept hierarchies and documents can be associated in a intuitive way by considering each concept as a category. Some approaches have been proposed in order to consider the hierarchical structure of the categories as opposed to a flat categorisation [Weigend, 1999]. We developed our own method to associate a document to the concepts from the different concept hierarchies in which the hierarchical structure is taken into account [Mothe, 2003]. In our approach a document can be associated to several hierarchies and to several concepts in a given concept hierarchy as soon as the content matches the concept representative. This method included learning strategies. In this paper we will not describe this method in details as it can be found in [Mothe, 2003] and evaluations in [Mothe, 2003b] ; we will rather consider that the association is made, whatever the method used.

5 Information access browsing an information space

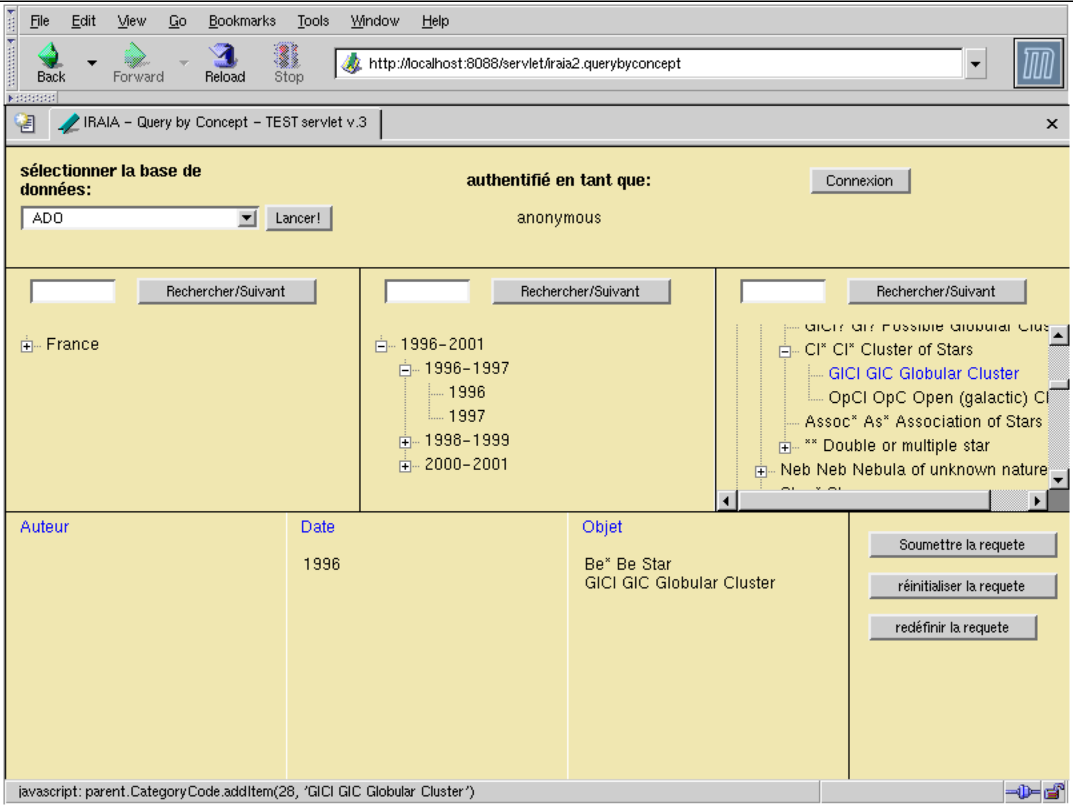
5.1 User information space

A domain is materialised by a specific ontology, from which a set of independent hierarchies (dimensions) can be defined according to well-known users' points of view. Given a domain, a user defines its own information space. It is composed of a selection of hierarchies or dimensions among the set of possible ones. This selection depicts his focus of interest, and lead to identify the associated documents. The fact that a document set is associated to given hierarchies could be seen as restrictive. On the contrary, because the user defines the information space himself, information exploration is more flexible: the dimensions are defined through guidelines and according to one's interests.

In practice, to start with, the user has to decide the information space he wants to move within by selecting a set of dimensions. Doing so, the user has potential access to the documents associated with these dimensions. This document set corresponds to the current collection, which is necessarily domain oriented.

In a validation project of our approach, we have defined two domains. One is related to Economics, the other one to Astronomy, but additional domains can be added. The Economic domain has been structured according to usages in economic institutes (EuroStat, Ifo): the "indicators", "industry", "country" and "date" dimensions have been kept and described with corresponding concept hierarchies. The Astronomy domain covers science monitoring activities. The domain has been structured according to the type of queries sent to Astrophysics Data System[5] server : the “Astronomical Object”, “Author”, “topic”, "journal of publication" and “date” dimensions have been kept. Different information spaces can be of interest when exploring a document set composed of bibliographic references: for example, the topic/author/date can provide information on the distribution of the publications, but it is also possible to query the system combining query terms from any dimensions.

5.2 Browsing concept hierarchies for information access



http://localhost:8088/servlet/iraia2.querybyconcept

IRAIA – Query by Concept – TEST servlet v.3

sélectionner la base de données: ADO

authentifié en tant que: anonymous

France

1996-2001
1996-1997
1996
1997
1998-1999
2000-2001

GlCl? Gl? Possible Globular Clus
Cl* Cl* Cluster of Stars
GlCl GlC Globular Cluster
OpCl OpC Open (galactic) Cl
Assoc* As* Association of Stars
** Double or multiple star
Neb Neb Nebula of unknown nature

Auteur	Date	Objet
	1996	Be* Be Star GlCl GlC Globular Cluster

javascript: parent.CategoryCode.addItem(28, 'GlCl GlC Globular Cluster')

Example of concept hierarchies (astronomical objects, authors, and dates) that may be browsed. A resulting user’s query consists of a set of selected terms in these hierarchies.

Figure 3: Formulating a query by browsing dimensions

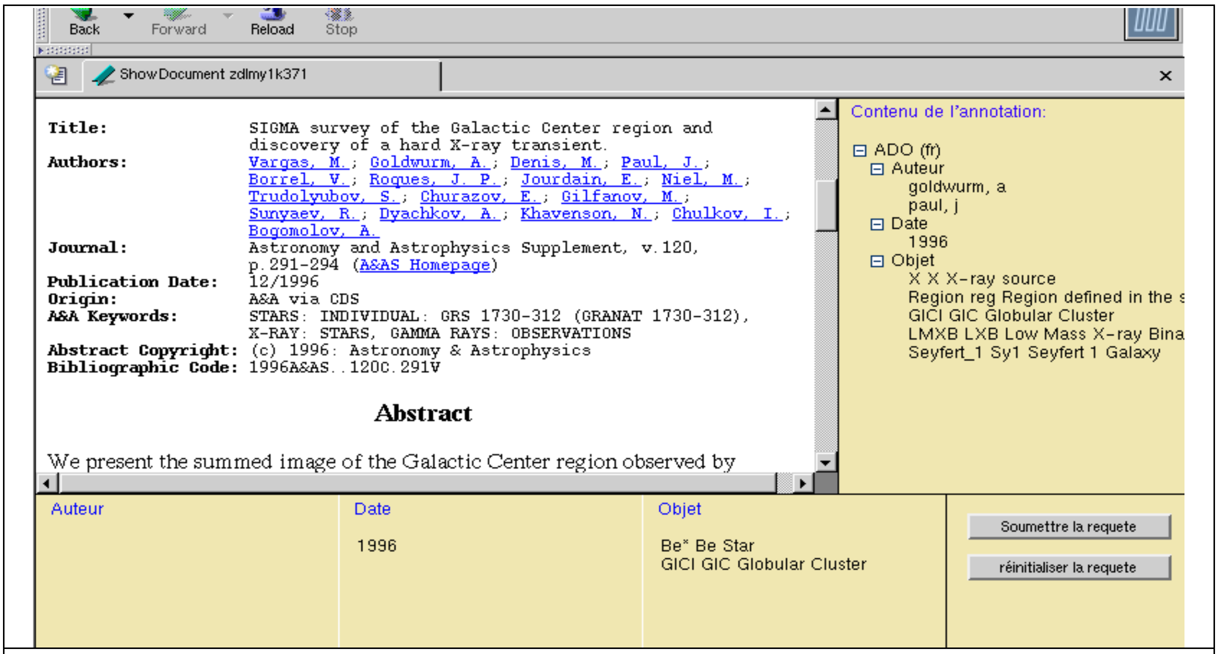
It often happens that users have some difficulty in formulating the initial query that corresponds to their information need. Several reasons can explain these difficulties:

- They have no idea whether their vocabulary will match the collection content.
- They have difficulties in deciding the level of detail they should use, general versus specific terms.
- Their proper vocabulary in the domain is limited.

Dimensions give direct access to domain knowledge. A user can easily formulate an initial query that corresponds to his needs by picking up concepts from the hierarchies (Figure 3). Because concepts are organised in a hierarchical way, the user can browse the dimension parts that correspond to his need to access to specific terms. In addition, hierarchies provide orientation to users, as they know at each step the information context. A domain is not limited to three dimensions, however, according to users, too many dimensions make browsing more difficult.

Once the query is processed, the system retrieves the corresponding documents and displays their URL organised according to the query terms. After the user has selected a document, it is displayed in a new window (Figure 4). One of the important points is that the link between the current document and the dimensions is kept. On the right side of the screen, the concepts from the different dimensions which represent the document are displayed.

[5] http://adsabs.harvard.edu/ads_abstracts.html



Title: SIGMA survey of the Galactic Center region and discovery of a hard X-ray transient.
Authors: Vargas, M.; Goldwurm, A.; Denis, M.; Paul, J.; Borrel, V.; Roques, J. P.; Jourdain, E.; Niel, M.; Trudolyubov, S.; Churazov, E.; Gilfanov, M.; Sunyaev, R.; Dyachkov, A.; Khavenson, N.; Chulkov, I.; Bogomolov, A.
Journal: Astronomy and Astrophysics Supplement, v.120, p.291-294 (A&AS Homepage)
Publication Date: 12/1996
Origin: A&A via CDS
A&A Keywords: STARS: INDIVIDUAL: GRS 1730-312 (GRANAT 1730-312), X-RAY: STARS, GAMMA RAYS: OBSERVATIONS
Abstract Copyright: (c) 1996: Astronomy & Astrophysics
Bibliographic Code: 1996A&AS..120C.291V

Abstract

We present the summed image of the Galactic Center region observed by

Contenu de l'annotation:

- ⊟ ADO (fr)
  - ⊟ Auteur
    - goldwurm, a
    - paul, j
  - ⊟ Date
    - 1996
  - ⊟ Objet
    - X X X-ray source
    - Region reg Region defined in the s
    - GlCl GlC Globular Cluster
    - LMXB LXB Low Mass X-ray Bina
    - Seyfert_1 Sy1 Seyfert 1 Galaxy

Auteur	Date	Objet
	1996	Be* Be Star GlCl GlC Globular Cluster

Figure 4: Visualising a document

Relevance feedback is commonly used in IRS. In our approach, relevance feedback is user directed. He can directly select the terms from the current document that correspond to his information need in the right side of the window. Another option for the user to reformulate his query is to come back to the full dimension descriptions.

Because dimensions are built up from the domain ontology, a concept is considered with all its synonym terms. A query concept can then be shown together with its generic concept or any other related one so that the user can get more documents presented with a more generic or any related view.

5.3 Multi-dimensional analysis of a document set

Generally, users have no precise idea of what they can find in a document collection. In that context, they have difficulties to decide which terms to use to describe their information need so that they can find out the information they are looking for. Many query modifications may be necessary to achieve their goal. Unfortunately, apart from the librarians, users have not necessarily the skills to reformulate in a efficient way their queries. Query reformulation requires some know-how and guidance. For instance, testing whether more generic terms recall some missing documents without introducing too much noise is a classical strategy. But the choice of the right terms may have a significant influence. If these terms are already shown in a hierarchy, or if some equivalence between terms is automatically calculated after the request is formulated, this work is made easier and more effective.

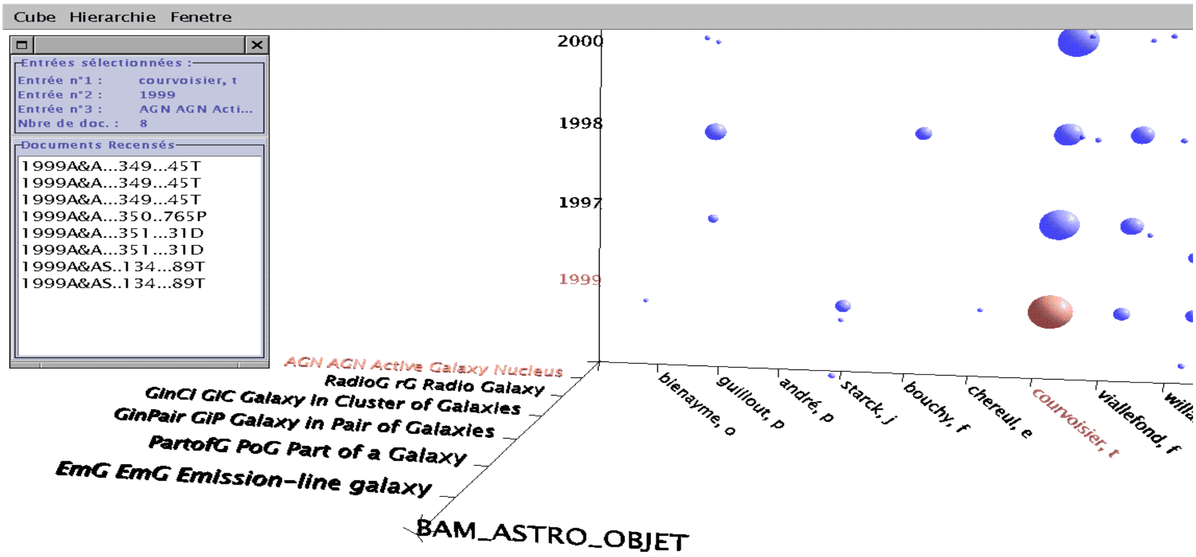
We propose an advanced visualisation that aims at helping the user when browsing or querying a collection. This visualisation allows him :

1) To display global views of either a document collection or documents retrieved after a query.
2) To take into account various aspects of the documents.
3) To explore graphically the collection through these various dimensions (multi-dimensional analysis).

A major feature of our approach is that it benefits from domain dependent hierarchies to reach these purposes. Indeed, the domain ontology and thus associated dimensions provide information on the terms that depict the context of the collection and then give orientation to the users. However, finding the concepts that describe the information need is not enough, as the user does not know if any document or too many documents are exactly described by the chosen concepts. Users are helped in deciding the level of generality/speciality they should use according to the document distribution along the domain hierarchies.

Multi-dimensional analysis is the way to provide users with such global information [Chaudhuri, 1997]. A graphical interface is generally associated with multidimensional analysis allowing the visualisation of the figures in 2 or 3 dimensions simultaneously.

The DocCube interface that we developed implements these functionalities and the screenshots presented in this paper are taken from this interface.



Entrées sélectionnées :

Entrée n°1 : courvoisier, t

Entrée n°2 : 1999

Entrée n°3 : AGN AGN Acti...

Nbre de doc. : 8

Documents Recensés

```
1999A&A...349...45T
1999A&A...349...45T
1999A&A...349...45T
1999A&A...350..765P
1999A&A...351...31D
1999A&A...351...31D
1999A&AS..134...89T
1999A&AS..134...89T
```

2000
1998
1997
1999
AGN AGN Active Galaxy Nucleus
RadioG rG Radio Galaxy
GInCl GIC Galaxy in Cluster of Galaxies
GinPair GiP Galaxy in Pair of Galaxies
PartofG PoG Part of a Galaxy
EmG EmG Emission-line galaxy
BAM_ASTRO_OBJET
bienaymé, o
guillout, p
andré, p
starck, j
bouchy, f
chereul, e
courvoisier, t
viallefond, f
willi

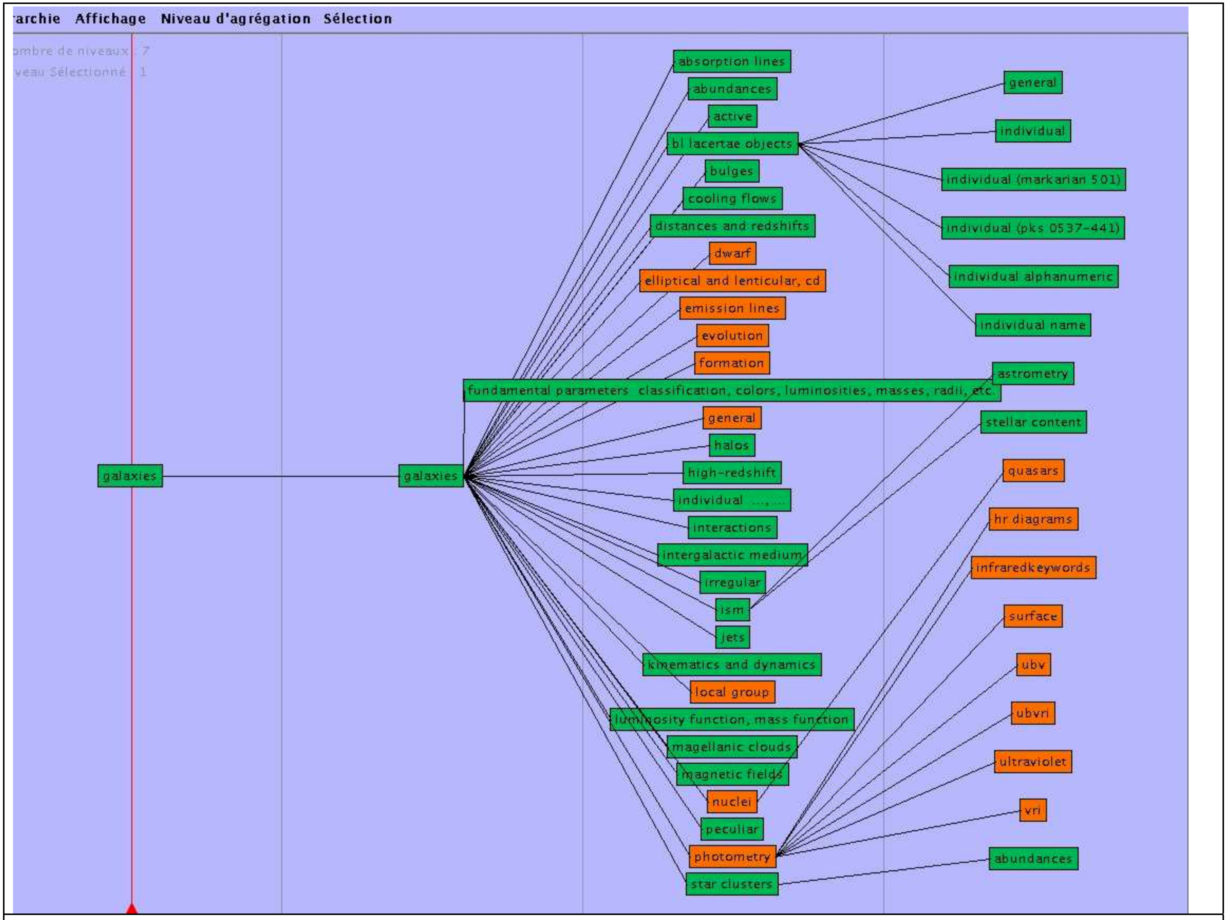
In DocCube, the data cube dimensions depict the context whereas the facts correspond to the number of documents associated with each node. The size of a sphere is proportional to the number of associated documents. The scale used to display the cube can be interactively changed in case of large/small number of documents. In the above screenshot, dimensions are the years of publication, the authors and the type of objects (Galaxy part).

Figure 5: Data Cube – Example of DocCube used in Astronomy.

The key component of the multidimensional analysis is the data cube concept associated with the OLAP operators [Chaudhuri, 1997]. A data cube allows multidimensional views of the data and can provide users with overviews of large data sets. These views can interactively be changed in particular using different levels of aggregation that corresponds to the levels of detail the data is viewed with. This mechanisms has been developed for databases for which a dimension corresponds to an attribute that can be depicted in a hierarchical way. We generalised this method to handle large document collections [Mothe, 2003]. In that case, dimensions correspond to hierarchies as depicted in section 4. A node depicts the number of associated documents. Doing so, the data cube gives an overview of the dispersion of the document set (Figure 5). By using this overview, a user can check if the level of detail he chooses is too deep (when a too small set of documents is associated with the concepts he is interested in). On the contrary, the level may be too general if large sets of documents are associated with the selected concepts. In such cases, the user knows he should modify his query before accessing the actual document contents. Modifying the level of detail of the data cube is done through the drill-down and roll-up operators. The 3D view provides additional type of knowledge: correlations between dimension values are directly displayed to the user.

- ***Roll-up:*** Subsets of units can be rolled-up (i.e. aggregated) into a single object (from children nodes to parent nodes). In our approach, the roll-up operator allows the user to consider more general terms in the concept hierarchy. In that case, the documents attached to children of a given concept are counted as attached to the concept itself.
- ***Drill-down:*** Drill-down is the opposite of roll-up. Drilling in one dimension down a deeper level of the hierarchy allows looking at the detailed data. In our approach, the drill-down operator corresponds to a refinement of the visualisation, considering the specific terms.
- ***Item selection:*** Additionally, the user can select some sub-trees only from a given hierarchy. This does not correspond to a query in a traditional IR system. Instead, this process is useful when large hierarchies are involved for the user to give focus to some parts (Figure 6).

- The ***slice function*** is used to set one of the dimensions and to obtain a two dimensional view of the results (Figure 7). In DocCube, this operation results in a graph that sets documents along two dimensions and represents them in the form of circles proportional to their number.



Sub-trees in red have been deselected by the user whereas green ones correspond to his selection. The second level of aggregation has been chosen.

Figure 6: Selecting a granularity level for a hierarchy and subparts of a hierarchy

A two-dimensional view is easy to analyse in a short amount of time. The user knows directly from it what query terms are useless because no document is attached to them (empty columns on figure 7). He then knows which terms will retrieve large sets of documents. From this information, he can decide to change the level of detail he has chosen (using roll-up and drill-down operators). These two-dimensional views are easily interpretable: with a single glance, the user evaluates the distribution of the document collection among the concepts. In Figure 7 for example, the user can see that some authors (axis X) have no connected available information regarding some of the astronomical objects (axis Y) e.g. cepheid variable star. Additionally, he can see that some of the objects provide a lot of documents (e.g. object of unknown nature).

Moreover, a single view corresponds to the result of many trials if a traditional IR system would be used. Users can decide whether they want to access the documents from a node or not (the result of a traditional IR query). Additionally, they are provided with overviews of a grid of query results.

A click on the sphere or circle makes it possible to get to the corresponding documents.

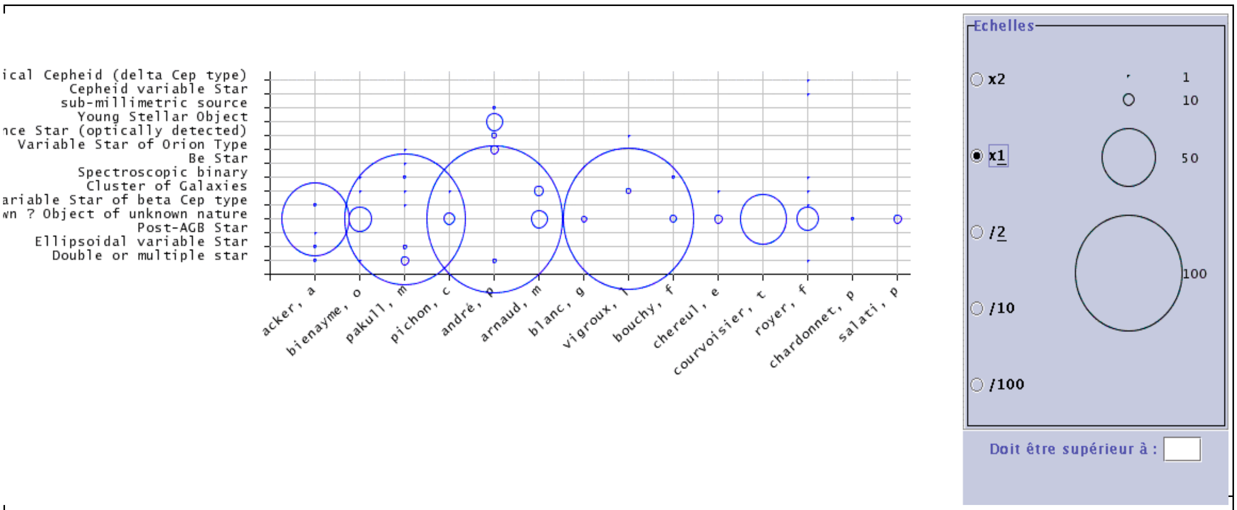


Figure 7: Slice operator – example of DocCube

6 Conclusion

The approach and tool we have developed aim at providing orientation to users who look for information in a set of documents. A search is done within a context, which is described by an ontology from which hierarchies of concepts are extracted. Each hierarchy describes a specific aspect a user can be interested in. Because the hierarchies are domain oriented, they help users in formulating their information need and they make them increase their knowledge of the domain vocabulary. The ontology holds for a controlled vocabulary during the indexing process. In addition, query reformulation based on thesaurus-like knowledge (dimensions) is easily implemented.

In DocCube, raw information access is not the only facility offered to users. In addition, they are provided with global views of the document collection. In that case, ontologies are used to structure the collection. Moreover, multi-dimensional analysis, as it is done in OLAP systems, allows users to analyse the collection before they decide to access some of the documents. The level of detail is chosen by the user in the ontologies and helps him to decide the level of detail that his query should have. This use of ontologies to define axes along which a query can be specified is completely original.

The first developments we have done are based on existing document collections in the domains of economics and astronomy. The approach proves to be appropriate to visualise either a document collection or a set of documents resulting from a query. In the case of the Web, such an approach could help to structure results according to a domain or to a user's point of view (using either content or usage-oriented ontologies). In fact, collecting document collections can be an intermediary step to give some meaning to sets of web-extracted documents. This proposal contributes to the progress towards a “semantic web”, mainly thanks to information space exploration and visualisation, where a new advantage of ontologies is taken into account. The meaning does not come only from reading concept definitions or from following relations between concepts. It is brought by considering globally each hierarchical dimension as a point of view that helps to focus in more or less detail on properties of the information space. The ontology provides a “volumic” reading grid, and each hierarchical dimension is a semantic scale to explore up and down a part of this grid.

The approach we developed will be all the more pertinent as it is based on an intelligent indexing that accurately matches texts and concepts from the hierarchies. In fact, the traditional query-text matching is transferred to the indexing level as the querying and the search languages are the same. Ontologies offer the advantage to organise terms and concepts in a coherent, meaningful and consensual representation. Another important statement underlying our approach is that domain knowledge in the ontology should be stable enough. These points correspond to two main issues for which different partial solutions already exist. The first range of difficulties is that building an ontology is a complex and time-consuming task: experts (domain and ontology experts) often manually do it. Exploring texts with some statistical or semantics based tool is an efficient way to design terminological structures, and particularly ontologies [Maeche, 2003]. This trend has recently benefited from the synergy between research in various disciplines such as text mining, knowledge acquisition from texts, NLP, corpus linguistics or even terminology. Although the idea is not new, recent advances in NLP have led to a better expression of the underlying theoretical issues in corpus collections. Text mining approaches combined with information extraction techniques can help the ontology designer by automatically determining what are the important terms and what are the relationships between terms. The second type of problem comes from the evolution of domain knowledge, for example new terms appear, other terms are no longer used. It is possible to help the designer to update the ontology by text analysis as well (e.g. analysis of term frequency along time). In fact, these directions of investigation, that are shared by IR and ontology engineering communities, will form one of the main points of our future works. Finally our future works will be devoted to conduce users' studies in order to evaluate further this interface. We will measure the user acceptability, the time required, the number of attempt and the answer quality gained through DocCube.

7 References

Aussenac-Gilles N., Biébow B., Szulman N., (2000) Revisiting Ontology Design: a method based on corpus analysis, Knowledge engineering and knowledge management: methods, models and tools, *Int. Conf. on Knowledge Engineering and Knowledge Management*. LNAI Vol 1937, Springer Verlag, 172-188.

Baziz M., Aussenac-Gilles N. et Boughanem M.. (2003) Désambiguïsation et expansion de requêtes dans un SRI : Etude de l'apport des liens sémantiques, *Revue des Sciences et Technologies de l'Information (RSTI) série ISI*, Ed. Hermes, 8(4/2003), 113-136.

Benford S.,Snowdon D., Greenhalgh C., Knox I., Brown C., (1995) VR-VIBE: A Virtual Environment for Co-operative Information Retrieval, *EUROGRAPHICS*, 349-360.

Bourigault D., (2002) Upery: un outil d’analyse distributionnelle étendue pour la construction d’ontologies à partir de corpus, *Traitement Automatique des Langues (TALN)*, 75-84.

Chalmers M., Ingram R., Pfranger C., (1996) Adding Imageability Features to Information Displays, *ACM Conference on User Interface Software and Technology*, 33-39.

Chaudhuri S., Dayal U., (1997) An overview of data warehousing and OLAP technology, *ACM SIGMOD Record*, 26(1), 65-74.

Chevalier M., (2000) ISIDOR: a visualisation interface for advanced information retrieval, *Inter. Conference on Entreprise Information Systems*, 414-418.

Chrisment C., Dkaki T., Dousset B., Mothe J., (1997) Extraction et Synthèse de Connaissances à partir de Données Hétérogènes, *Ingéniérie des Systèmes d'Information*, 5(3), 367-400.

David S., Plante P., (1990) Termino version 1.0, Report, Centre d’Analyse de Textes par Ordinateur, Université du Québec à Montréal.

Dubin D., Document analysis for visualization, *Inter. Conference on Research and Development in Information Retrieval*, 199-204, 1995.

Dumais S., Chen H., (2000) Hierarchical classification of web content, *Inter. Conference on Research and Development in Information Retrieval*, 256-263.

Enguehard C., Pantéra L., (1995) Automatic natural acquisition of terminology, *Journal of Quantitative Linguistics*, 2(1), 27-32.

Faure D., Nedellec C., (1999) Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System ASIUM, *European Workshop, Knowledge Acquisition, Modelling and Management (EKAW’99)*, 329-334.

Fensel D., (2001) Ontologies: a silver bullet for Knowledge Management and Electronic Commerce, Berlin, Springer Verlag.

Fowler R. H., Fowler W. A. L., Williams J. L., (1996) 3D Visualization of WWW Semantic Content for Browsing and Query Formulation, WebNet.

Guarino N., Welty C., (2000) A Formal Ontology of Properties, *Inter. Conference on Knowledge Engineering and management*. Springer Verlag. LNAI 1937, 97-112.

Habert B., Naulleau E., Nazarenko A. (1996) Symbolic word clustering for medium-size corpora, *Int. Conference on Computational Linguistics*, 490-495.

Hearst M.A., (1992) Automatic Acquisition of Hyponyms from large Text Corpora, *Int. Conf on Computational Linguistic (COLING*).

Hearst M.A., Karadi C., (1997) Cat-a-Cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy, *Inter. Conference on Research and Development in Information Retrieval*, 246-255.

Maedche, A., Staab, S., (2003) Ontology Learning, In S. Staab & R. Studer (eds.) Handbook on Ontologies in Information Systems. Springer.

Masolo C., (2001) Ontology driven Information retrieval: Stato dell'arte. Report of the IKF (Information and Knowledge Fusion) Eureka Project E!2235. LADSEB-Cnr, Padova (I).

Morin E., (1999) Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique, *TAL (Traitement Automatique des Langues*),40(1), 143-166.

Mothe J., Chrisment C., Dkaki T., Dousset B., Egret D., (2001) Information mining: use of the document dimensions to analyse interactively a document set, *European Colloquium on IR Research*: ECIR, 66-77.

Mothe J., Chrisment C., Dousset B., Alaux J., (2003) DocCube: Multi-Dimensional Visualisation and Exploration of Large Document Sets, *Journal of the American Society for Information Science and Technology*, JASIST, Special topic section: web retrieval and mining, 54 (7), 650-659.

Mothe J., Hubert G., Augé J., Englmeier K., (2003) Catégorisation automatique de textes basée sur des hiérarchies de concepts, *Journées Bases de Données Avancées*, 69-87.

Mukherjea S., Foley J.D., Hudson S., (1995) Visualizing complex hypermedia networks through multiple hierachical views, *CHI*, 331-337.

Reynaud C., Aussenac-Gilles N., Tort F. (1998) A support to domain knowledge modelling: a case study, Information Modelling and Knowledge Bases IX, H Kangassalo, P.J Charrel (Eds.), IOS Press, Amsterdam, V. 45, Frontiers in AI and Applications, 35-50.

Rocchio, J., Relevance feedback information retrieval. In Gerard Salton, editor, The Smart retrieval system| experiments in automatic document processing, pages 313-323. Prentice-Hall, Englewood Cliffs, NJ, 1971.

Séguéla P., (1999) Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés, *Terminologies Nouvelles* n°19, 52-60.

Smadja F., (1993) Retrieving Collocations from Text: Xtract, Computational Linguistics, 19(1), 143-178, Oxford University Press.

Swan R.C., Allan J., (1998) Aspect Windows. 3-D visualizations, and indirect comparisons of IRS. Inter. Conference on Research and Development in Information Retrieval, 173-181.

Ternier A., Rousset M.-C., Sebag M., (2001) Combining Statistics and Semantics for Word and Document Clustering, IJCAI2001 *Workshop on Ontology Learning*.

Uschold M., (1998) Knowledge Level Modelling: Concepts and Terminology, *The knowledge engineering review*, 13(1), 5-29.

Voorhees, E. (1994) Query expansion using lexical-semantic relations, *Proceedings of the 17th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, 61-69, Dublin, Ireland.

Weigend Andreas S., Wiener Erik D., Pedersen Jan O., (1999) Exploiting Hierarchy in Text Categorization, *Information Retrieval Journal*, ISSN:1386-4564, Volume 1 , Issue 3, 193 - 216.

Zhang J., (2001) TOFIR: A tool of facilitating information retrieval – introduce a visual retrieval model, *Information Processing and Management*, N.37, 639-657.