Library and Information Services in Astronomy IV
July 2-5, 2002, Prague, Czech Republic
B. Corbin, E. Bryson, and M. Wolf (eds)

Exploring Bibliographic Collections using Concept Hierarchies.

Josiane Mothe

IRIT, Université Paul Sabatier, and IUFM, Toulouse, France
mothe@irit.fr

Daniel Egret

Centre de Données astronomiques de Strasbourg, UMR 7550, Observatoire Astronomique de Strasbourg, 11 rue de l'Université, 67000 France
egret@simbad.u-strasbg.fr

Claude Chrisment

IRIT, Université Paul Sabatier, Toulouse, France
Claude.Chrisment@irit.fr

Kurt Englmeier

DIW, Berlin, Germany
kurt@diwsysv.diw-berlin.de

Abstract. This paper presents a new approach for browsing and mining document collections. The study is based on a data collection of 10,000 articles extracted from the ADS, corresponding to all papers authored or co-authored by French astronomers during the period 1996 to 2000. A key point in our approach is that information searching and exploring take place in a domain-dependent semantic context. A given context is described through its vocabulary organized along different concept hierarchies that correspond to different points of view. These hierarchies structure the information space. Moreover, they provide the query language for users and allow them to explore the vocabulary of the domain before they express their information need which is helpful when the information need is not well defined.

1. Introduction

Information retrieval systems (IRS) generally aim at retrieving all the relevant documents (and only the relevant ones) according to a user's need. Examples of such systems are: web search engines, library information systems, or, in the astronomy field, the NASA/ADS abstract service (Eichhorn et al. 2002). Whatever the system used, one of the problems users have to face is expressing their information needs with respect to the collection content. Generally when the user knows what is in the collection, how it is structured, what he is search-

ing for, and how it can be described he has no real problem expressing his query. But if one of these elements is missing, it can be more problematic. One solution to this problem is that the system provides orientation by structuring the information space and displaying information on the collection content. *Document indexing* is the most common way to structure or organize a document collection. Document indexing associates descriptors to each document content. The descriptors are then used as keys during the retrieval process. Another way to structure a document collection is to take into account the *meta–information* associated with each document. Meta–information corresponds to information on the information. It generally consists of "factual" data that can be associated to the document without ambiguity: e.g. author names, date of publication, etc. Meta–information is generally combined with the descriptor representation during the retrieving process. More specifically, meta–information can be used either to filter the retrieved documents according to some descriptors (e.g. filter the publications written by Mr. Dupont in 2002) or to categorize the documents (according to their common meta–information value). Additionally, documents can be *categorized* according to their content. In that case two documents are grouped together if their content (i.e. their descriptors) are similar enough. In practice, indexing using a controlled vocabulary —i.e. a list of pre-defined terms— can be seen as an alternative way to categorize documents: each concept can be seen as a class for the categorization. Examples of such an approach are the Dewey classification system, or the *Yahoo!* directory. The categories - generally organized under the form of a taxonomy - organize the knowledge of the domain.

In this communication we propose a new approach that combines the three methods of organization described in the previous section. We propose to categorize each document according to different *concept hierarchies*. That means that each facet of a document (authors, content, etc.) is considered in a homogeneous way and according to a hierarchical description of the facet. With this approach, information searching and exploring take place in a domain-dependent semantic context. A context is described through its vocabulary organized along different concept hierarchies that correspond to different points of view. These hierarchies structure the information space.

When formulating their information needs, users are guided in choosing the relevant terms. They are provided with the controlled vocabulary of the domain under the form of hierarchies. They can browse these concept hierarchies and select different terms that will compose their queries. Additionally, global visualizations help users to get an overview of the document collection content and of the distribution of the information according to different elements of document description (e.g. per author, per date or per topic). From this overview, and according to the number of documents associated with each element, users can decide the level of detail that is the most appropriate to express the query (general terms vs specific terms). The interface provides the users with interactive maps for which they decide the focus. These maps display the links between a descriptor (e.g. author name, object, method) and the other descriptor values.

2. The Astronomy Domain

2.1. Document collection

In the following, we demonstrate our approach through the exploration of a data collection of about 10,000 articles extracted from the NASA/ADS, corresponding to all publications for which at least one author is a French astronomer, during the five-year period 1996 to 2000. French authors are defined here as being listed in the Directory of the French Astronomical Society (Société Française d'Astronomie et d'Astrophysique, SF2A) in the release published in 2001 (SF2A 2001). In all, 1023 names have been extracted from the on-line version of the *Annuaire*. Most of the names correspond to scientists or students working in a French astronomical institute (irrespective of nationality) or to French postdocs working temporarily in a foreign institute.

The document collection has been built by querying the ADS abstract service for these author names, for the period 1996 to 2000, and 9838 articles have been found (see Egret et al. (2002) for more details). In a subsequent step we extracted the abstracts and keywords from the ADS. Affiliations come from the French directory for the authors of the basic list and from the ADS (although this information is frequently absent or incomplete) for the co-authors.

2.2. Concept hierarchies

In our approach, a domain (e.g. astronomy) is composed of several domain *ontologies* that correspond to complementary descriptions or facets of the same set of documents. For exploring the astronomy domain we have organized these domain ontologies under the form of concept hierarchies. For instance, one concept hierarchy is : author name - institute of affiliation - country. Each concept hierarchy corresponds to a point of view that may interest the user, e.g. search of articles signed by one author, publication list of an institute, international collaborations, etc. A hierarchy describes one aspect of the document and defines a controlled vocabulary. The controlled vocabulary helps the users in specifying their information need and limits the risk of ambiguity of the language. A document can be associated to different hierarchies, and to several nodes from a given hierarchy.

For the study developed in this paper, we have adopted five concept hierarchies which we describe below:

– One of the facets of the documents corresponds to the *keywords*. The keyword hierarchy is based on the thesaurus used by the major astronomy journals (*The Astrophysical Journal* and the other AAS journals: *Astronomy & Astrophysics*, *MNRAS*, etc.). There are three levels, from general topics (e.g. Solar system, Stars, Galaxies) to more specific (e.g. Galaxies – Star clusters – Abundances). Our collected documents from the ADS generally carry a "Keyword" field using this thesaurus. When it is not the case (e.g. for documents not belonging to the core journals), we have developed automatic techniques in order to associate the documents to the hierarchy. These methods have been developed in the framework of a European project at IRIT (IRAIA; Mothe et al. 2002).

– The second facet is *authorship* and *affiliation*. The hierarchy has three levels: author names, then the organizations they belong to (affiliation), and

finally the countries in which the organizations are based. A document may be assigned to several levels (co-authors).

– Interoperability between the ADS abstract service and the SIMBAD data base (Genova et al. 2002) makes it possible to categorize a document on the basis of the astronomical objects studied or mentioned in the document. With regard to objects, we adopted the hierarchy which is used for classifying object types in Simbad. The upper level corresponds to general types of objects (stars, galaxies, etc.); the next level corresponds to more specific object types (pulsating variable, quasar, etc.). Finally some types contain the individual object names (e.g. individual stars).

– Another hierarchy corresponds to the journals. The journal in which the article is published is extracted from the ADS metadata. Additionally, journals are organized either regarding the type of source (journal, conference, international conference) or according to the country in charge of the source (e.g. national conference associated to the corresponding country).

– Finally, the last hierarchy corresponds to the date of publication. Publication year is also part of the ADS metadata. The "father" level is built by grouping together consecutive years. The root level corresponds to the entire period of time (five years).

2.3. Mapping document collection towards the concept hierarchies

The first step of document processing is to associate each document to the corresponding nodes in the different concept hierarchies. In practical terms, in the case of our document collection from the astronomy domain, the problems to be solved were the following:

– Keywords: assigning documents for which no keywords were given to the keyword concept hierarchy was made on the basis of word frequencies in the texts of the abstracts. Additionally, keywords have been automatically added to the corresponding concept hierarchy on the basis of their occurrence in the keyword field. Adding a keyword implies advanced processes in order to determine the level to which the new keyword has to be added in the hierarchy.

– Authorship and affiliations: affiliations for the basic list of French astronomers were derived directly from the directory. Affiliations of co-authors were extracted from the ADS, when available, and implied a specific effort on semantics and filtering in order to improve homogeneity.

3. Accessing documents through concepts and concept hierarchies

The first step, when using our system, is to decide which domain the user is interested in. In the following, examples are taken from the astronomical domain, but the system may in principle handle different domains simultaneously. To each domain corresponds its own hierarchies that provide the knowledge of the domain. The browsing interface presented here is part of the IRAIA project, supported by the European Commission under the 5th Framework Program[1].

[1]IST-1999-10602, IRAIA Getting Orientation in Complex Information Spaces as an Emergent Behavior of Autonomous Information Agents.

When selecting a domain, the corresponding Concept Hierarchies (CH) are displayed to the users who can start browsing them in order to define their focus of interest.

3.1. Browsing concept hierarchies

Concept hierarchies correspond to a key point of the interface. They correspond to the language of mediation between the users and the document collection.

When interacting with the system, users first define the document facets they are interested in by selecting the corresponding CH which are displayed as a result of the selection. This corresponds to a type of customization of the interface where only the user's points of interest are displayed. Then, the users can browse these CH to visualize the corresponding vocabulary. Compared to traditional systems for which the user is asked to formulate a query according to his own a-priori knowledge of the domain, this system provides guidance to the user in formulating his information need by browsing the concept hierarchies.

The browsing facility works like a file manager, where 'folder' or 'concept' can be opened in order to have a look at the children-nodes. A branch can be developed in order to see all the specific concepts or closed if it corresponds to a concept the user is not interested in. At any moment, the user can select a term that will be added to the current query. The main advantage of this process is that the user never loses the context of his search because each term is located according to a given concept.

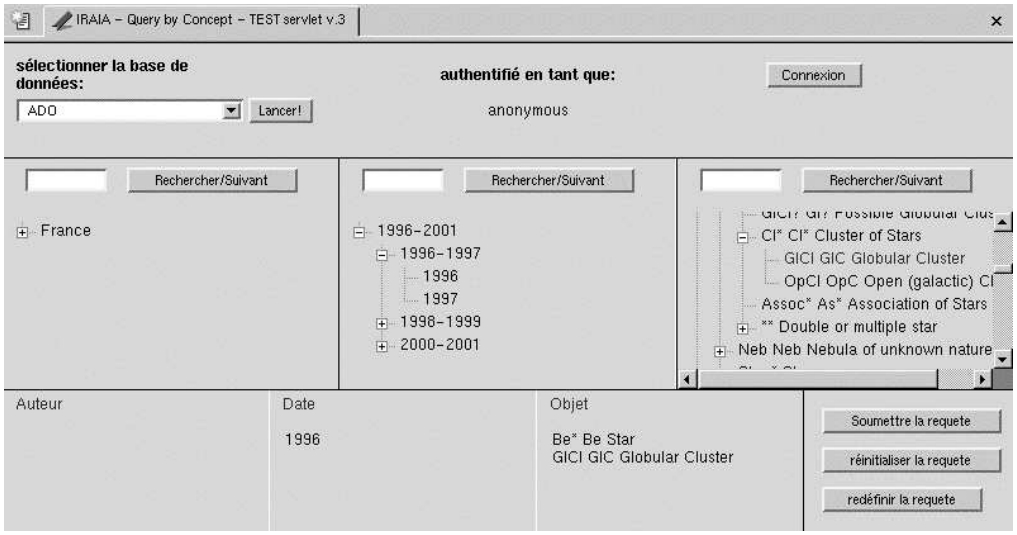

Figure 1. Browsing concept hierarchies in order to express the user's need.

3.2. Browsing retrieved documents

When the query has been submitted, the system selects the relevant documents and the result is displayed to the user. Instead of displaying a list of documents, as traditional information retrieval systems do, the retrieved documents are categorized according to the query terms. The result is displayed under the

form of folders, each of them containing the documents that correspond to a certain combination of query terms (the *AND* operator is used between the query terms) —see Figure 2.

This provides guidance to the user who has an overview of what has been retrieved and why it has been retrieved, i.e. what query terms have led to retrieve a specific document. According to the query terms the folder represents, the user can open it in order to visualize the corresponding documents (Figure 2).

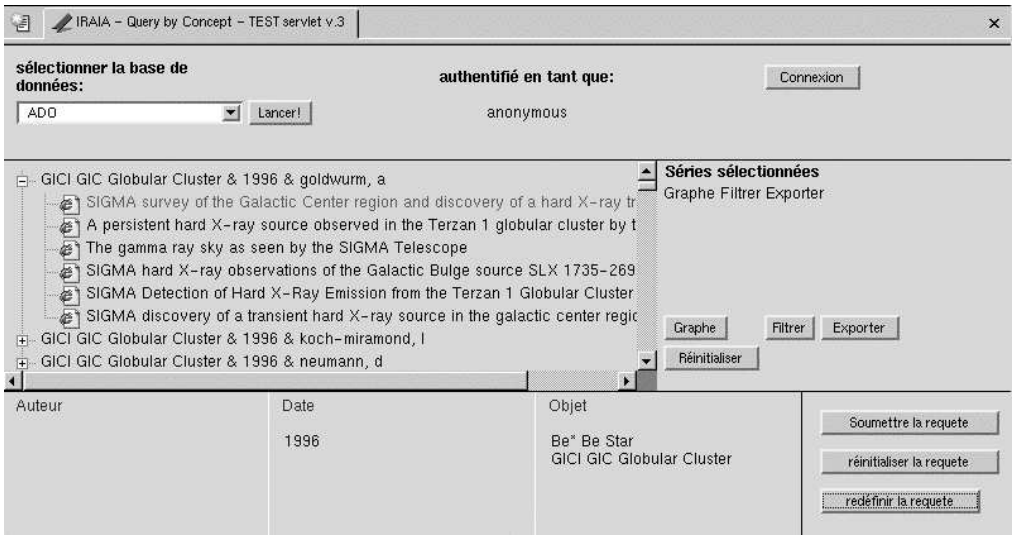

Figure 2. Browsing a cluster of retrieved documents. The selected folder contains six documents from the same author. The query parameters (here: year and two keywords) are visible in the bottom panel.

Finally, the user can decide to select one of the documents in order to visualize its contents. In this case, the application automatically queries the NASA/ADS, and the usual ADS abstract is displayed in the web browser (Figure 3). In addition, for each of the CH, the list of concepts associated with the document is displayed (right part of the screen).

3.3. Query reformulation

While browsing the results, the terms associated with the current document are displayed for each of the CH. Selecting some of these terms automatically adds them to the query. This feature corresponds to *directed relevance feedback* : terms that are associated with relevant documents are suggested to the user who can decide to add them to his initial query. The user will probably find terms that were not in the initial query but which are highly relevant regarding his information need (see Figure 3). The reformulated query can be then used in a new retrieval process.

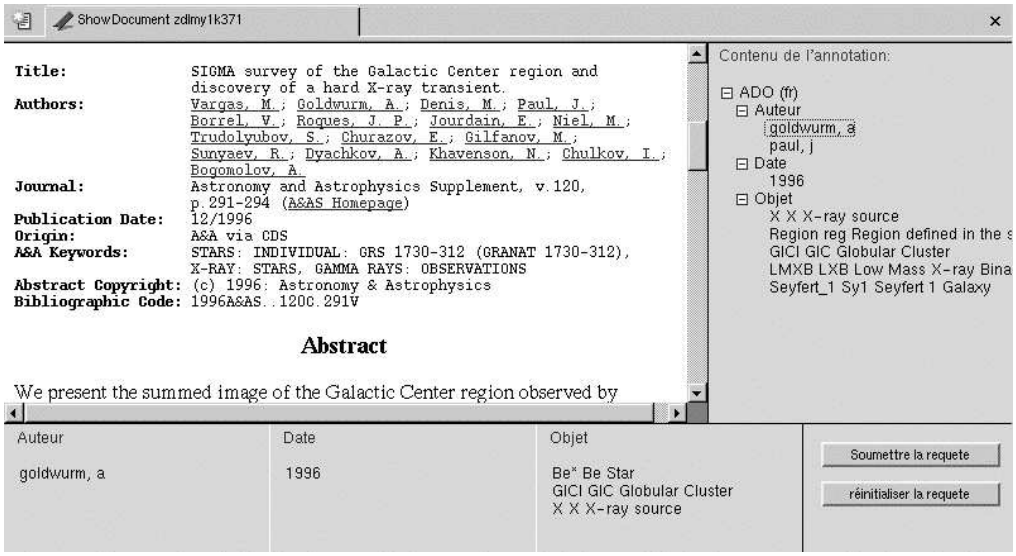

Figure 3. Visualization of one specific document and query reformulation. The document abstract from ADS is presented. The corresponding categories, visible in the right panel, can be used for query reformulation. They are organized according to three facets: 2 authors, 1 year and 5 objects.

4. Exploring the document collection

Another tool is provided by the system in order to get graphical overviews of the document content. More specifically, this tool provides the user with global views of the documents related to the concepts associated with two selected concept hierarchies and of the distribution of the document collection.

This information is displayed under the form of a 2D representation (see Figure 4). The two axes of the graph correspond to the selected CH, whereas the size of each circle represents the number of documents that have been mapped or categorized in the corresponding category. Thus, the sizes of the circles directly provide the users with information as to how important are the different elements (dimension values) are for the analyzed collection. For example, considering the two CH "organizations authors belong to" and "object type", if circles of equal size are shown for a given organization whatever the topic is, that means all topics are of equal interest for that organization. In contrast, if a single institute is present for a given topic, this can mean that this topic corresponds to a speciality of this institute. This type of information is particularly interesting when trying to determine which are the other institutions, or companies, working in the same area, or to search for possible contributors to a given field.

In the case of a large hierarchy (composed of a lot of concepts), the user can select or delete some concepts. Additionally, at any time it is possible to change the level of aggregation of a hierarchy. For example the visualization can use the author name level rather than the affiliation level. This feature allows users to refine the level of details they need for the global document visualization, before eventually querying the system in order to access the document content.

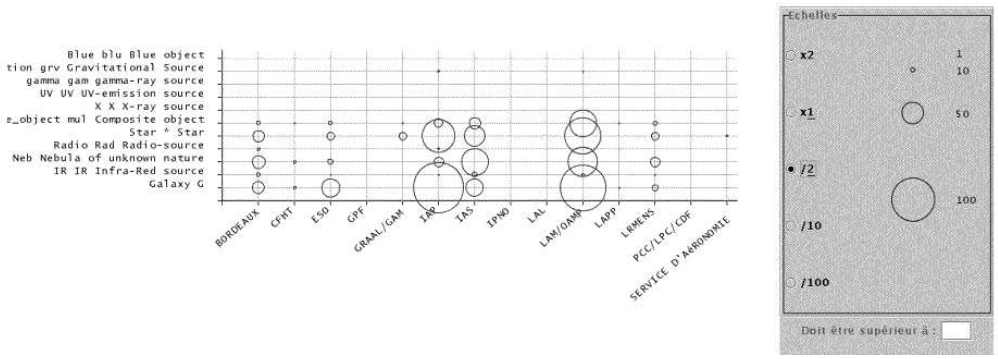

Figure 4. Example of a detailed view extracted from the graphical overview of the document collection. Two dimensions are crossed here: 'organizations authors belong to' and 'object type' (see text). This study makes use of DocCube, a module developed for the IRAIA project

A two-dimensional view is easy to analyze in a short amount of time. The user learns directly from it which query terms are useless because no document is attached to them (empty columns or lines in figure 4), and, conversely, what are the terms that correspond to large sets of documents. From this information, the user can decide to change the level of detail used for visualization.

5. Conclusion

We have presented in this paper a new approach for exploring document collections. After a first processing, which is necessary for extracting homogeneous metadata representing several facets of the documents, the information space specific to the domain (in this case, the astronomical domain) is represented through several concept hierarchies.

Based on this representation, our approach provides a novel way to browse and mine the document collection. It is possible to browse the concept hierarchies, select relevant levels of visualization, display groups of documents, view the query results, and add new terms for query reformulation. A two-dimensional representation provides a global view of the documents related to the concepts associated with two selected concept hierarchies, or facets. This new approach globally helps users express their information needs and become oriented in the information space.

The system presented here is a prototype developed as part of a research program. It is currently not available publicly for online query.

Acknowledgments. This study has made use of the DocCube system developed at IRIT, as a module on top of the IRAIA system (European project IST-1999-1062, see http://iraia.diw.de). We would like to thank Didier Barret (SF2A) and Alberto Accomazzi (ADS) for their kind help in providing the raw data for this study. This study has made use of the NASA Astrophysics Data System operated by SAO, and of the SIMBAD database operated by CDS, Strasbourg.

References

Egret D., Mothe J., Dkaki, T. 2002, to appear in Proceedings "Semaine de l'Astrophysique Française, Paris 2002", Editions de Physique.

Eichhorn G. et al. 2003, in this Conference (ADS)

Genova F. et al. 2003, in this Conference (CDS)

Mothe J., Chrisment C., Alaux J. 2002, DocCube : multi-dimensional visualization and exploration of large document sets, to appear in special issue of Jasist "Web Retrieval and Mining".

SF2A 2001, Annuaire de la Société Française d'Astronomie et d'Astrophysique.