Knowledge discovery in bibliographic collections using concept hierarchies and visualization tools. Application to the astronomy domain.

Josiane Mothe (IRIT, IUFM, Toulouse), Daniel Egret (CDS, Strasbourg),
Claude Chrisment (IRIT,UPS, Toulouse), Kurt Englmeier (DIW, Berlin),
Taoufiq Dkaki (IUT URS, Strasbourg), Soizick Lesteven (CDS, Strasbourg)

ABSTRACT

This paper presents new methods for knowledge extraction and visualization, applied to datasets selected from the astronomical literature. One of the objectives is to detect correlations between concepts extracted from the documents. Concepts are generally meta-information which may be defined a priori, or may be extracted from the document contents and are organised along domain ontologies or concept hierarchies.
The study illustrated in the paper uses 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, for the years 1996 to 2000. The study presents new approaches for visualizing relationships between institutes, co-authorships, scientific domains, astronomical object types, etc.

Keywords: text mining, bibliographical database, concept hierarchies, visualization

1. INTRODUCTION

We introduce a new approach to provide users with solutions to explore a document collection. 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 different concept hierarchies that correspond to different points of view (Englmeier et al., 2001). 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. Additionally, the hierarchies of the domain provide a new way to explore the information space, using multi-dimensional analysis and global visualisation of the document collections for knowledge discovery purposes.

Generally speaking an information retrieval system (IRS) aims at retrieving all the relevant documents (and only the relevant ones) according to a user ‘s need. Many IRS are used for access to public documents:
- web search engines,
- library information systems,
- or, in the astronomy domain, the NASA Astrophysics Data System (ADS; Eichhorn 2002) which provides a general access to the published astronomy literature through a comprehensive abstract service.

Generally when a user knows what is in the collection, how it is structured, what he is searching for and how it can be described, he has no real problem to find it. But if one of this element is missing, exploration or discovery strategies may become necessary.

In our view, it is important that a system provides orientation to the user by structuring the information space and informing him on the collection content. In the literature, we can distinguish three ways that are used to organise a document collection:

1. *Indexation* : the principle is to associate some descriptors to the document content. These descriptors are used as keys during the retrieval process. The descriptors can be “free text” (that kind of descriptors can be obtained by an automatic indexing process) or they can be chosen among the terms of a controlled vocabulary, i.e. list of pre-defined terms (as it is the case for the key-word system used jointly by the major astronomical journals).

2. *Meta information*: information on the information. They correspond to "factual" data (which can generally be associated to the document without ambiguity): e.g. author names, journal, date of publication, etc. This meta information can be used as query components (e.g. all the publication written by Mr Dupont in 2002). A specific item, in the astronomy domain, is the bibcode (Schmitz et al. 1995) which uniquely describes a bibliographic reference and is used by SIMBAD, NED, ADS, and many other databases. This item encodes in it-self different meta information including the journal and year of publication.
3. *Categorisation*: the documents are associated to predefined classes that organise the knowledge of the domain. These classes can be organised along hierarchies. Categorisation can be viewed as the indexation according to a controlled vocabulary: if the controlled vocabulary is under the form of a taxonomy, which is generally the case, each concept can be seen as a class for the categorisation (see Yahoo!, Dewey). In the astronomy domain, examples are: lists of documents produced by the members of an institute (often posted as preprint lists); or lists of articles produced by users of an observatory or a space mission.

In this communication, we propose a new approach that combines this three methods of organisation.
We propose to categorise each document according to different concept hierarchies (CH). Each CH corresponds to a facet of a document: a meta information (authors, date of publication, etc.) or its content. That means that all the facets of a document are considered in an homogeneous way i.e. a hierarchical description of the facet. The facets correspond to query component and to mining components in order to help the user understanding better a targeted document set.

2. EXPLORING 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). 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 from these author names, for the period 1996 to 2000. (5 years) : 9838 articles have been extracted. See Egret et al. (2002) for more details about the sample.
In a subsequent step, we extracted the abstracts and key-words, from ADS. Affiliations come from the French directory for the authors of the basic list, and from 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 select domain ontologies, organized as 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 user in specifying his 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:

1. One of the facets of the documents corresponds to the *keywords*. The key-word hierarchy corresponds to the thesaurus which is used by the major astronomy journals (*The Astrophysical Journal*, and the other AAS journals; *Astronomy & Astrophysics*, etc.). There are three levels, from general topics (Solar system, Stars,

Galaxies) to more specific (e.g.: galaxies – star clusters – abundances). The advantage is that the collection (extracted from ADS) already uses, in general, the elements from this thesaurus in the "Keyword" field. 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).

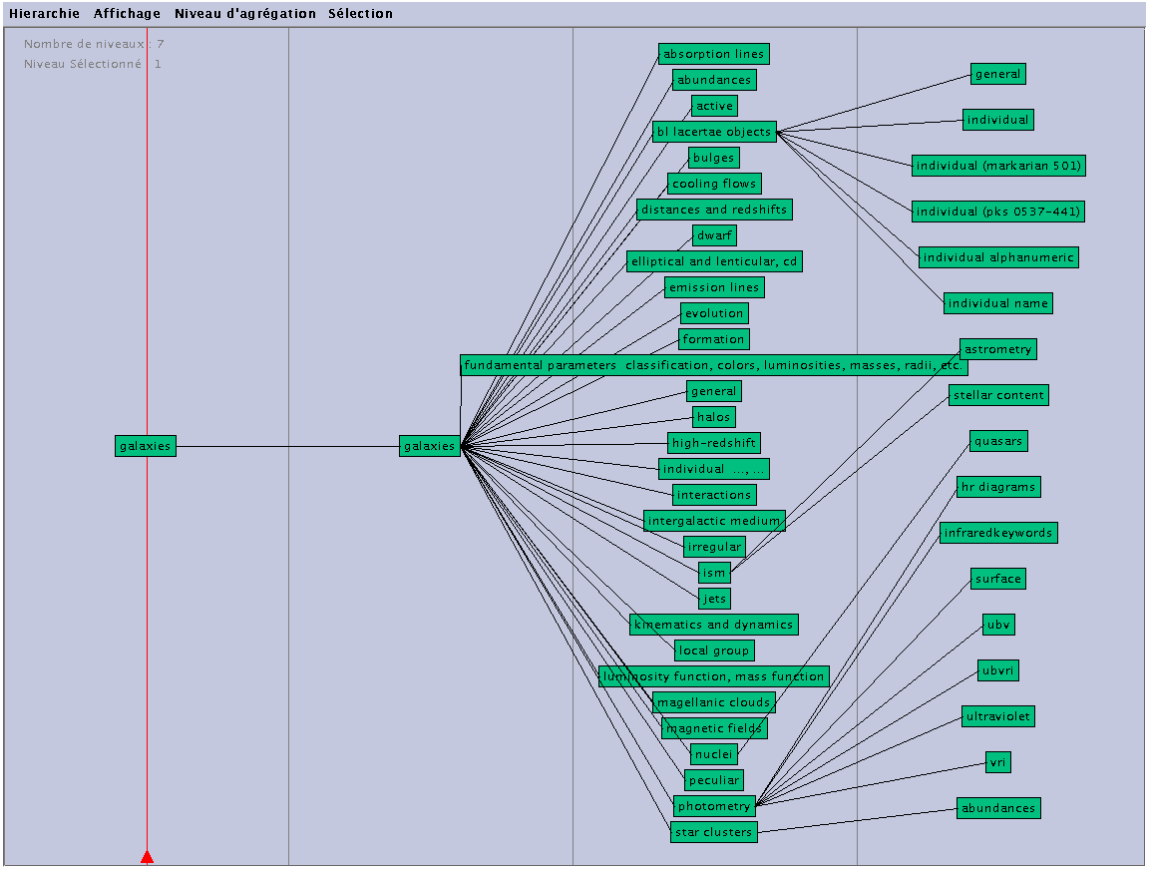


Figure 1. Sample extracted from the keyword concept hierarchy (three levels).

2. The second facet is *authorship and affiliation*. The hierarchy has three levels: the leaves correspond to the author names, then the organisation they belong to (affiliation), and finally the country in which the organisation is set. A document may be assigned to several leaves (all co-authors).

3. Interoperability between the ADS abstract service and the SIMBAD data base (Genova et al. 2001) makes possible to categorise a document on the basis of the *astronomical objects* studied or mentioned in the document. With regard to objects, we have used the hierarchy which is used for classifying object types in Simbad. The upper level corresponds to the general types of objects (stars, galaxies, etc.); the next level corresponds to more specific object types (pulsating variable, quasar, etc.). Finally some types have the individual object names as children (e.g. individual stars).

4. Another hierarchy corresponds to the *journals*. The individual journal is easily extracted from the *bibcode*. Additionally journals are organised 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).

5. Finally, the last hierarchy correspond to the *date of publication*. Publication year is also directly extracted from the *bibcode*. The "father" level is built by grouping together consecutive years. For a period of six years, we create three groups of 2 years. The root level corresponds to the entire period of time (6 years).

3. INFORMATION PROCESSING

3.1 Mapping document collection towards the concept hierarchies

The first goal 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 key 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 text of the abstracts. Additionally, keywords have been automatically added to the corresponding concept hierarchy on the basis of their occurrence in the key-word 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.

- Authorships and affiliations: affiliation from the basic list of French astronomers was derived rather directly from the directory. Affiliations of co-authors were extracted from the ADS, when available, and implied a lot of efforts on semantic and filtering.

3.2 Contingency tables : multi-dimensional representation of the document collection

Contingency tables are a way to transform non-numerical information into numerical information. They have been shown to be efficient to represent summarised information in the case of databases (Fayyad, 1996) and they are the starting point of many mining function (Lebart, 1998). Classification, clustering, factorial analysis (principal component analysis, correspondence analysis) are easily performed on contingency tables. Additionally, multidimensional analysis can be performed on these structures. Moreover, a contingency table is a basic representation from which many other representations can be derived.

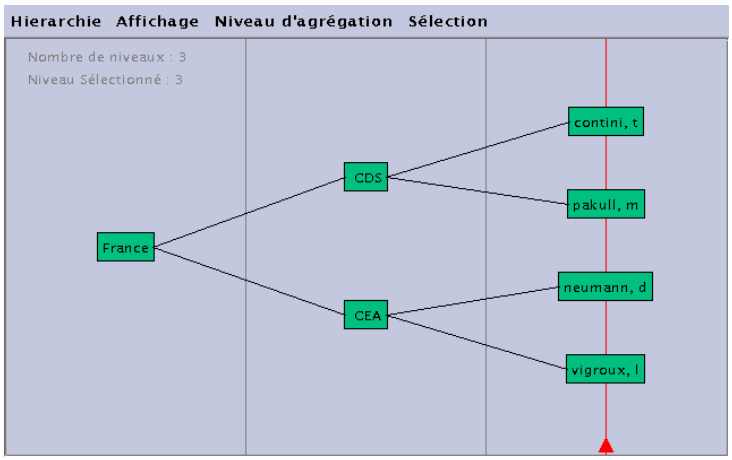
A contingency table is obtained by dividing up a population according to two variables, I and J. The columns of the table correspond to the modalities (or values) of the variable J, whereas the lines of the table correspond to the modalities of I. The table could be viewed as the characterisation of the lines (objects) according to the columns (the characteristics). In fact, the two variables play symmetric roles and can be treated the same way. In statistical applications, the intersection Tij of a row i and a column j corresponds to the number of objects in the population for which the variable I has the value i and the variable J has the value j simultaneously. This principle can be extended to different document representations taking into account the document dimensions (Mothe et al. 2001).

A traditional document representation can be obtained based on a contingency table where the two variables are document references and keywords. The contingency table expresses then the term frequency for each document. When the two variables correspond to the author names, we get a traditional co-authoring matrix. In fact any facet of the documents can be taken into account to generate contingency tables in order to summarise a document set according to these facets.

In the case of hierarchical facets as defined in our approach, the contingency tables are constructed according to the leaves of the hierarchies, then, the contingency tables at upper levels can be deduced automatically (see figure 2).

		CDS		CEA	
		contini, t	*pakull, m*	*neumann, d*	*vigroux, l*
Star	*Variable Star*	2	6	1	6
	Peculiar Star	3	9	1	4
Galaxy	*Cluster of Galaxy*	6	1	4	3
	Galaxy Nucleus	12	7	6	14

a) Contingency table crossing authors and objects.



b) Authoring facet: changing the aggregation level from author names to affiliations

		CDS	CEA
Star	*Variable Star*	8	7
	Peculiar Star	12	5
Galaxy	*Cluster of Galaxy*	7	7
	Galaxy Nucleus	19	20

c) Contingency table after rolling-up the authoring facet (affiliation level instead of author name level).

Figure 2. Example of contingency tables –extract- in the case of hierarchical variables

The contingency tables are a starting point for different methods to visualise a document collection in a global way and to discover knowledge from a collection.

4. GLOBAL VISUALIZATION OF DOCUMENT COLLECTIONS

The 2D representation corresponds to a key component of the interface (see Figure 3). It displays a global view of the documents related to the concepts associated with two selected concept hierarchies and is directly derived from contingency tables.

The two axes of the graph correspond to the dimensions, whereas the size of the circles represents the number of documents that have been categorised in the corresponding category. Thus, the size of the circles directly provides the users with the information on how important are the different elements (dimension values) for the analysed collection. For example, considering the two dimensions “organisations authors belong to” and “object type”, if circles of equal size are shown for a given organisation whatever the topic is, that means that all the topics correspond to a domain of equal interest for that organisation. On the opposite if a single institute is present for a given topic, this can mean that this topic corresponds to a specificity of this institute. This type of information is particularly interesting when trying to detect what are the other institutions, or companies, working in the same area or possible contributors to a given field.

In the case of a large hierarchy (composed of a lot of concepts), the user can select / delete some concepts. Additionally, at any time it is possible to change the level of aggregation of a hierarchy. For example the visualisation can use the author name level rather than the organisation level (read vertical line). This operation corresponds to what is called roll up and drill down: these operators refer to the possibility for the user to change interactively the level of aggregation of the analysed data. Rolling-up, the user goes up the levels and obtains more general information; the opposite operation corresponds to a drill-down. In the interface, that allows the user to refine the level of details he needs for the global document visualisation, before eventually querying the system in order to access the document content.

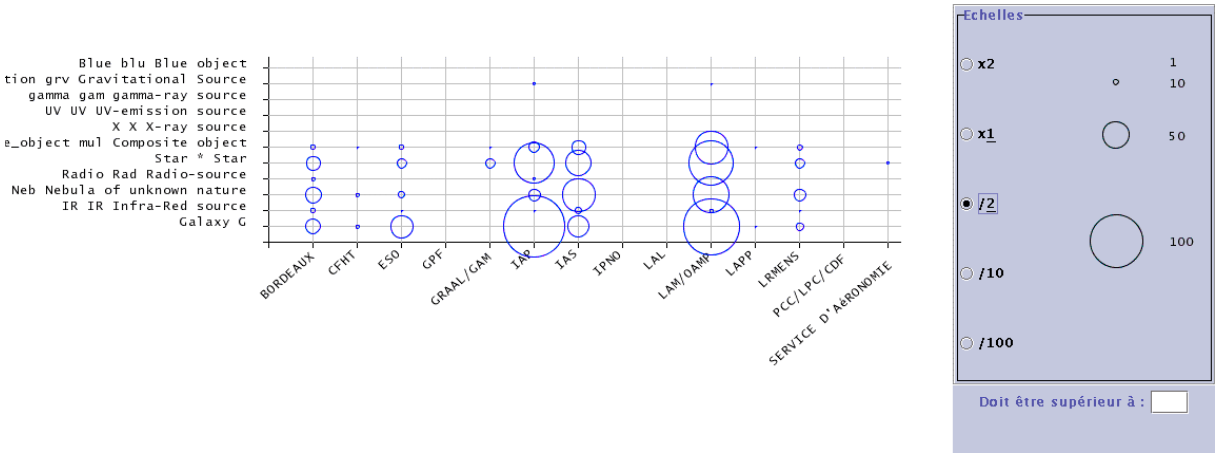


Figure 3. Example of detailed view extracted from the graphical overview of the document collection. Two dimensions are crossed here : 'organisations authors belong to' and 'object type' (see text).

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 or lines on figure 3). He then knows what are the terms that will retrieve large set 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: in a single glance, the user evaluates the distribution of the document collection among the concepts. Additionally, he can see that some of the indicators provide a lot of documents.
Moreover, a single view corresponds to the result of many trials using a traditional IR system. Users can decide whether he wants 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.

Another tool is provided in the system in order to get graphical overviews of the document contents.

5. NETWORKING RELATIONSHIPS BETWEEN CONCEPTS AND INFORMATION ITEMS

The networking tool is used to graphically display the relationships that are automatically extracted from the analysis of the document contents. The same starting point is used (contingency tables) in order to discover the strong and relevant relationships that exist between document facets or components. Given one element (e.g. an author name) and one related document facet (e.g. organisations authors belong to'), a graph is automatically computed that indicates the strength of the links that exists between these elements. The basis of the process is that the more two elements are correlated, the thicker is the link between these elements.
Figure 4 displays the results of the analysis of the contingency table that crosses the author names (one only was kept, *Combes, F.*) and the organisations. The resulting network has been mapped on a French map. The nodes of the network correspond to the different organisations and to the author name. The affiliations for which a link with the given author name exists have been highlighted.

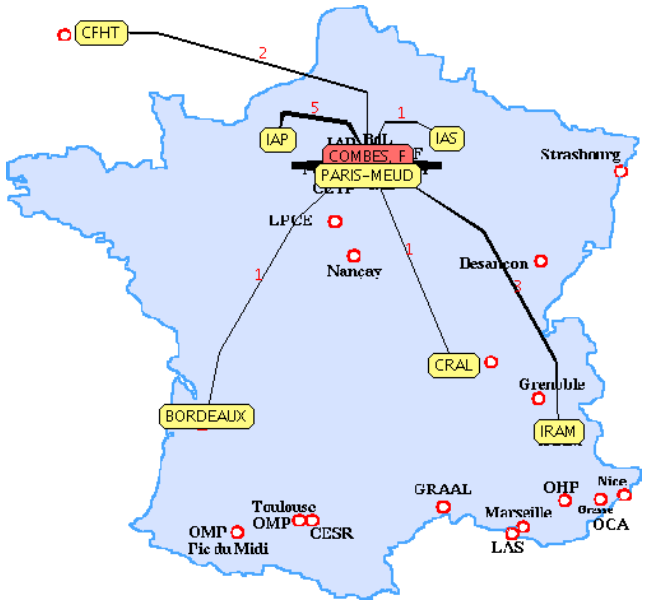


Figure 4. Network of co-authorships : from one author (F. Combes, affiliated to LERMA, Observatoire de Paris), links are shown to the French institutes where co-authors are affiliated. A similar network can be drawn for the international collaborations.

Another example of a network is shown Figure 5. The contingency table taken into account in that case crosses author names (lines and columns). Each node of the network corresponds to an author name. The number of publications corresponding to each author follows her/his name in the resulting graph. Co-authorship is represented through more or less thick links. In the graph figure 5, the members of one institute only have been considered (using a filtering process). This tool allows one to detect if one element is not correlated to the other (e.g. an author has no co-author in his organisation, he may work on a different topic or has external co-authorships). At the same time a dense network part corresponds to the core of the facet that is studied.

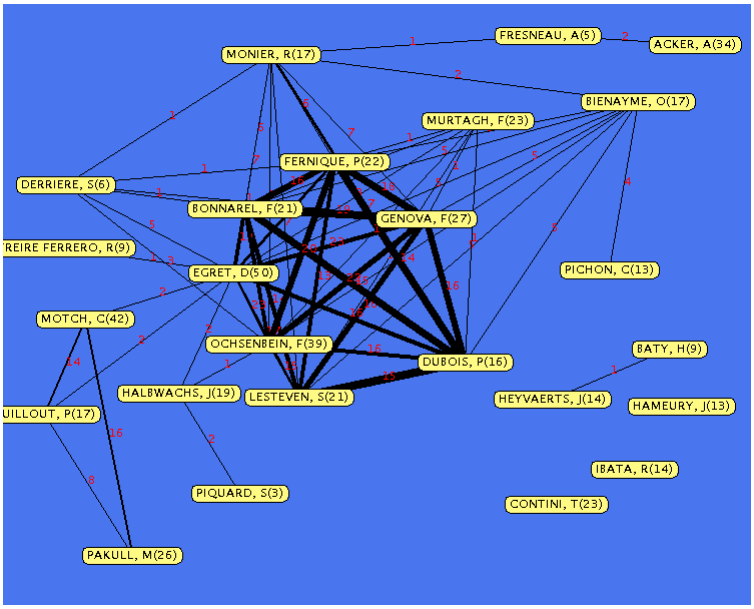


Figure 5. Network of co-authorships within one single institute (Observatoire de Strasbourg) : the width of the lines is correlated with the corresponding number of documents (co-authored by the two linked persons). The densest part of the network is dominated by the 'caravan' of authors from the Strasbourg Data Center (CDS) who published frequently articles with several members of the team.

Co-authorships is quite intuitive to analyse, but networking relationships can be discovered whatever the document facets taken into account. Depending on the type of information one want to discover, he has to select first the facets of the documents and eventually the elements (e.g. the authors from CDS), he wants to focus on.

6. CONCLUSION

The new approach we present in this paper aims at helping users to discover knowledge from targeted document sets, that is to say to discover useful and unknown patterns from raw information. Because of the textual nature of the documents, pre-processing is mandatory before any mining technique can be applied. Pre-processing aims first at extracting meta information or any other document facets from the document content that can be relevant in a user's point of view. Then, documents are mapped towards pre-defined hierarchical structures that depict each facet. Finally, contingency tables are used in order to summarize this information. Whatever the facets that are taken into account (meta information, content, key-words, etc.), there are depicted and used in a homogeneous way (concept hierarchy and contingency tables).
Based on this document set representation, our approach provides two novel ways to represent graphically the results of automatic document facet mining. The first one is a 2-D representation which provides a global view of the documents related to the concepts associated with two selected dimensions or facets. This 2-D representation is easy to analyse in a short amount of time : in a single glance, the user evaluates the distribution of the document collection among the concepts. Users can decide whether he wants to access the documents from a node or to change the level of detail he has chosen. The second tool provides the users with networks of relationships between concepts. Given one element and one related document facet, a graph is automatically computed that indicates the strength of the links that exists between these elements.

ACKNOWLEDGEMENTS

This study has made use of the DocCube system developed at IRIT and which is a module on top of IRAIA system, mothe@irit.fr, http://www.irit.fr/~josiane.mothe
IRAIA : European project IST-1999-10602, http://iraia.diw.de, kurt@diwsysv.diw-berlin.de
We would like to thanks Didier Barret (SF2A) and Alberto Accomazzi (ADS) who provided the raw data.
This study has made use of the NASA Astrophysics Data System (ADS) operated by SAO. This study has made use of the SIMBAD database operated by CDS, Strasbourg.

REFERENCES

1. Egret, D., Mothe, J., Dkaki, T., Barret, D., 2002, in Proceedings of the annual meeting of the Société Française d'Astronomie et d'Astrophysique (SF2A), F. Combes & D. Barret (Eds), in press
2. Eichhorn, G., 2002, The NASA Astrophysics Data System, http://ads.harvard.edu/
3. Englmeier, K., Mothe, J., 2001, Trustworthy personal assistance: a design objective for interactive agents, 7th Americas Conference on Information Systems, Association for Information Systems, (CD-Rom), Boston.
4. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., 1996, Advances in Knowledge Discovery and Data Mining, AAAI Press, ISBN 0-262-56097-6.
5. Genova, F. et al. 2001, the CDS information hub.
6. Lebart, L., Salem, A., Berry, L., Exploring textual data, Kluwer Academic Publishers, ISBN 0-7923-4840, 1998.
7. Mothe, J., Chrisment, C., Dkaki, T., Dousset, B. , Egret., D., 2001, Information mining: use of the document dimensions to analyse interactively a document set, pp 66-77, European Colloquium on IR Research: ECIR.
8. Mothe, J. et al. 2002, Visualisation globale de collections de documents sous forme d'hypercube - Le système DocCube", Mothe, J., Chrisment, C., Alaux, J., Journées francophones d'Extraction et de Gestion des Connaissances, EGC, Hermès, pp 131-142 ,2002.
9. Schmitz, M., Helou, G., Dubois, P., LaGue, C., Madore, B., Corwin, H.G. Jr, and Lesteven, S., 1995, in Information & On–line Data in Astronomy, Egret & Albrecht (Eds.), Kluwer Acad. Publ., p. 259.
10. SF2A, 2001, Annuaire de la Société Française d'Astronomie et d'Astrophysique.