Computer Physics Communications 127 (2000) 198–206

Computer Physics Communications

www.elsevier.nl/locate/cpc

Information discovery from semi-structured sources – Application to astronomical literature

Taoufiq Dkaki [a,b], Bernard Dousset [b], Daniel Egret [c], Josiane Mothe [b,d]

[a] *IUT, Université Robert Schuman, Strasbourg Sud, France*
[b] *IRIT, Université Paul Sabatier, Toulouse, France*
[c] *CDS, Observatoire Astronomique de Strasbourg, Strasbourg, France*
[d] *IUFM, Institut Universitaire de Formation des Maîtres, Toulouse, France*

Abstract

Textual information systems provide different kinds of information seeking that answer different user needs. Among them, knowledge discovery systems aim at providing global views and useful patterns from raw information. This paper presents a framework to discover knowledge from semi-structured documents and visualize it through graphical views. An application to astronomical literature is given. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Astronomical literature; Information mining; Information evolution; Trend analysis

1. Introduction

Textual information systems provide different kinds of information seeking that answer different user needs. The goals of these systems are different and so are the techniques used even if some techniques can be shared by different kinds of systems. Information Retrieval Systems (IRS) [11–13] retrieve documents or document parts based on keyword searching. In fact, most of the time the system answers are given in the form of a document reference list and the user has to navigate through that list to access the documents. The user's goal is generally to read the documents s/he considers as relevant. These systems are based on document indexing and on document and query representation matching. The indexing is generally based on statistical and on Natural Language Processing (NLP) theories. The indexing process includes the deletion of stop words, stemming and phrase weighting. Information Extraction Systems (IES) [8,10] have different goals. IES attempt to extract salient facts from unstructured text documents into templates or pre-defined types of information (such as the names of products or of the company headers). The extracted elements can then be directly accessed. These systems use techniques grounded in computational linguistic theory and are based on speech tagging analysis, name entity recognition, and co-reference resolution techniques. Even if these systems generally do not attempt to understand the document contents, the analysis of the text has to be much more complete than the analysis done while indexing text for IR purposes. Knowledge Discovery Systems (KDS) provide global views or patterns of a data set. When applied to documents, they use a structured document representation and attempt to discover some trends and correlations between the structure elements. These systems combine techniques from information retrieval and information extraction (in order to derive a structured representation of the docu-

0010-4655/00/$ – see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0010-4655(99)00509-3

ments) with techniques from data mining (in order to mine the information).

This paper presents some solutions to discover unknown global information from semi-structured documents and gives an example using documents from the astronomical literature. In this paper we first summarize the general goals and phases of knowledge discovery and present the key points of our proposition to adapt this general framework to information discovery from semi-structured documents. We then give some examples of the application of this technology to astronomical literature.

2. Information discovery from documents

2.1. General goals and phases

The goal of information discovery is to find useful and unknown patterns from raw information. The main mining model functions can be grouped together into three groups:

- *Classification*: mapping the information into predefined classes or into clusters constructed according to the information features similarities,
- *Dependencies*: discovering of (weighted) dependencies and relations between fields, temporal dependencies, sequences or regression,
- *Transformation*: summarization.

The raw information can be either data from (relational) database systems (and this is the case for most of the literature in the area) or documents. To achieve the information discovery from documents, we suggest to use the general framework given by the KDD (Knowledge Discovery from Database) technology [6] and to turn it into a general framework for information discovery [2].

Generally speaking, a KDD process can be divided into three stages:

- *Data selection and pre-processing*: This stage consists of collecting data, homogenizing, cleaning and reducing it.
- *Data analysis*: The objective is to mine the cleaned information in order to discover hidden relationships among the data.
- *Interpretation*: The goal of this step is to fulfill the user's needs in terms of knowledge and to allow him/her to take the relevant decisions. This can be done through relevant visualizations.

This general framework that has been defined for factual data from databases can be adapted to be applied to documents and semi-structured documents. Semi-structured documents are documents where some information is semantically pre-defined. That means that some tags in the document itself give clues on what is the information (i.e. HTML documents are semi-structured as some tags such as META, TITLE, ADDRESS, ... are used to mark-up the semantics of some content elements). The framework we define can be applied to any semi-structured documents and indeed has been applied to INSPEC, HTML, ADS, ... documents. It is decomposed into different stages that are presented in the following paragraphs.

2.2. Information selection and information extraction

Information selection. The selected information corresponds to the raw information that will be mined. Indeed, its relevance and exhaustively is a keypoint for the information discovery accuracy. Information retrieval systems can be used to achieve this. The information harvesting can be done through existing servers or databases, either domain oriented such as ADS for the astronomy literature, WPI for patents, etc. It can also be done through intelligent agents on the Web.

Information extraction. The harvested raw information is generally in a specific format. Generally, each document source has its own format and homogenization is needed as a pre-processing task. In addition these formats generally are not appropriate for mining purposes. Before being able to proceed with mining techniques, it is necessary to deeply structure the information and to decide on what elements the mining will be done. The framework we define provides a generalized format that logically reformats the raw harvested information. This format has the advantage of fitting heterogeneous collection requirements. In addition, this format has been defined to allow easy information extraction. The information extraction itself takes advantage of advances in information retrieval indexing and on information extraction methods.

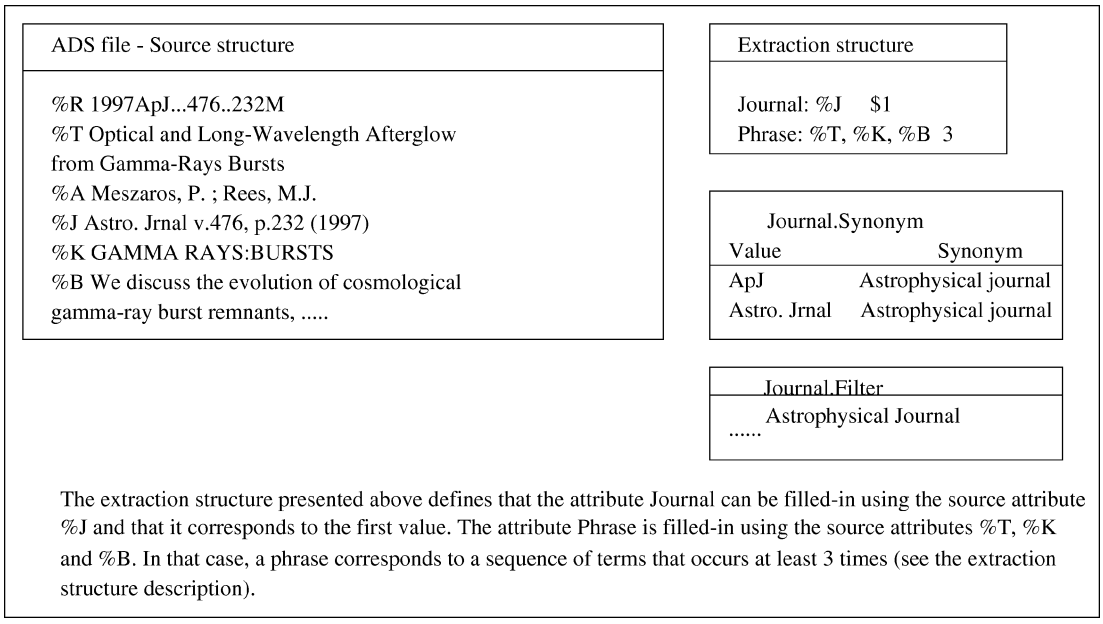


Fig. 1. Extraction schema example.

The *generalized format* is described through the extraction structure, synonym dictionaries and filters (see Fig. 1).

- The extraction structure (or template) provides a direct link between the information source structure and the structure that is relevant for the mining. The lexical and syntactic tags used in each source describe each structure element that can be relevant for the mining phase.
- Synonym dictionaries can be used so that different values can be considered as equivalent while extracting the values from a set of documents according to the template.
- Filters are optional. A filter is a set of values that have either to be omitted (negative filter) or to be the only values retained (positive filter). Filters are used during the extraction phase.

This generalized format provides a simple way to homogenize the extracted information that avoids the physical reformatting of the initial information.

2.3. Information mining

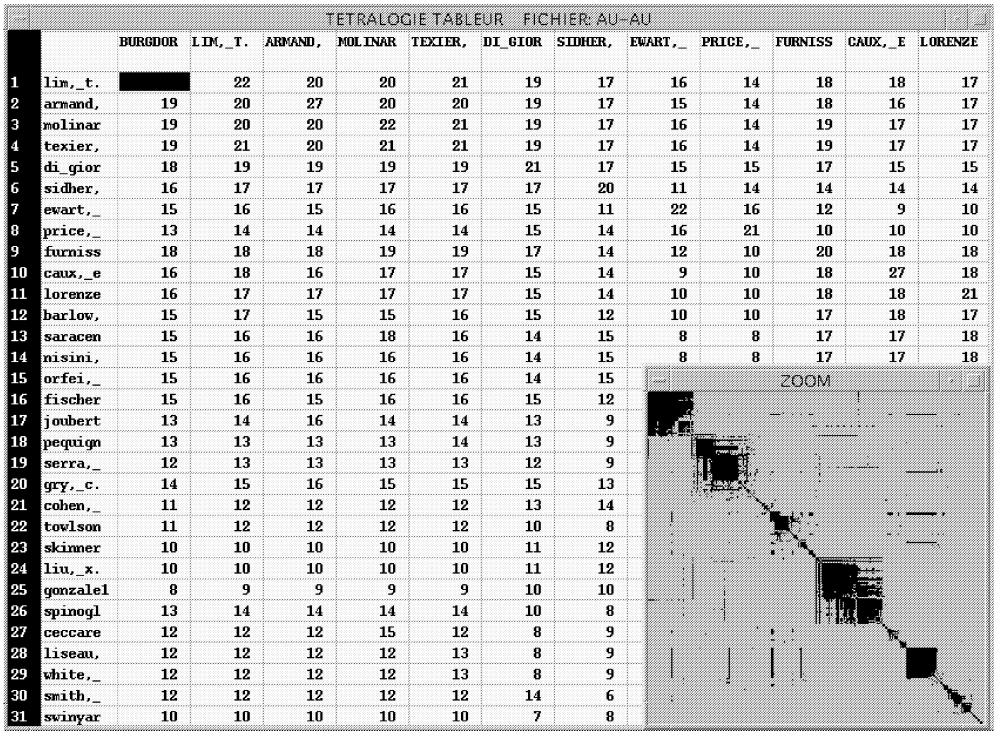
According to the different objectives of classification, correlation detection or summarization, different mining methods can be used. In our approach most of them are based on contingency table processing.

Contingency tables. Contingency tables (see Fig. 1) are the starting point for studying relationships between two kinds of information. In statistics, a contingency table is the representation resulting from an experiment in which the observation performed on the sample studied is categorized according to two criteria. Each cell of the table represents the number of occurrences of a given combination of categories. In our case the sample is the set of harvested documents. The criteria are the different attributes from the extraction structure. The contingency tables which usually cross two kinds of information can be generalized to represent the relationships between more than two kinds of information [2].

Relevant crossings. Depending on what are the crossed attributes, the crossing tables can be used to detect various information correlations. Table 1 gives some examples of relevant crossings with regard to bibliographic documents.

Table 1
Example of relevant crossings

Crossed attributes	Extracted knowledge
Authors–Authors Affiliations–Affiliations	Work team or Collaborative work
Keywords–Date	Evolution of the terminology, of the domain interest
Keywords–Keywords	Sub-domain detection or terminology
Keywords–Authors Keywords–Affiliations	Specific domain of interest of the authors or affiliation
(Keywords–Keywords)–Dates	Evolution of the terminology associated with a sub-domain
(Authors–Affiliations)–Dates	Evolution of the author affiliations



		BURGDOR	LIM,_T.	ARMAND,	MOLINAR	TEXIER,	DI_GIOR	SIDHER,	EWART,_	PRICE,_	FURNISS	CAUX,_E	LORENZE
1	lim,_t.		22	20	20	21	19	17	16	14	18	18	17
2	armand,	19	20	27	20	20	19	17	15	14	18	16	17
3	molinar	19	20	20	22	21	19	17	16	14	19	17	17
4	texier,	19	21	20	21	21	19	17	16	14	19	17	17
5	di_gior	18	19	19	19	19	21	17	15	15	17	15	15
6	sidher,	16	17	17	17	17	17	20	11	14	14	14	14
7	ewart,_	15	16	15	16	16	15	11	22	16	12	9	10
8	price,_	13	14	14	14	14	15	14	16	21	10	10	10
9	furniss	18	18	18	19	19	17	14	12	10	20	18	18
10	caux,_e	16	18	16	17	17	15	14	9	10	18	27	18
11	lorenze	16	17	17	17	17	15	14	10	10	18	18	21
12	barlow,	15	17	15	15	16	15	12	10	10	17	18	17
13	saracen	15	16	16	18	16	14	15	8	8	17	17	18
14	nisini,	15	16	16	16	16	14	15	8	8	17	17	18
15	orfei,_	15	16	16	16	16	14	15					
16	fischer	15	16	15	16	16	15	12					
17	joubert	13	14	16	14	14	13	9					
18	pequign	13	13	13	13	14	13	9					
19	serra,_	12	13	13	13	13	12	9					
20	gry,_c.	14	15	16	15	15	15	13					
21	cohen,_	11	12	12	12	12	13	14					
22	towlson	11	12	12	12	12	10	8					
23	skinner	10	10	10	10	10	11	12					
24	liu,_x.	10	10	10	10	10	11	12					
25	gonzalel	8	9	9	9	9	10	10					
26	spinogl	13	14	14	14	14	10	8					
27	ceccare	12	12	12	15	12	8	9					
28	liseau,	12	12	12	12	13	8	9					
29	white,_	12	12	12	12	13	8	9					
30	smith,_	12	12	12	12	12	14	6					
31	swinyar	10	10	10	10	10	7	8					

Fig. 2. Example of a crossing table and the associated zooming.

Direct correlation detection from contingency tables. By associating some functions with a crossing table, it is possible to directly associate a graphical representation of the detected correlations.

- *Reordering*: the elements of the crossing table are re-ordered so that highly correlated elements are close. A zooming of the table content can then give an overview of the detected groups (see Fig. 2).

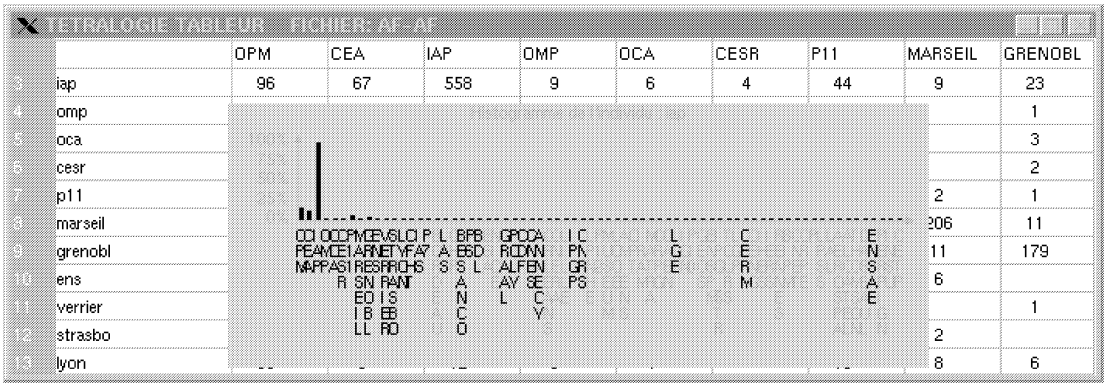


Fig. 3. IAP work group.

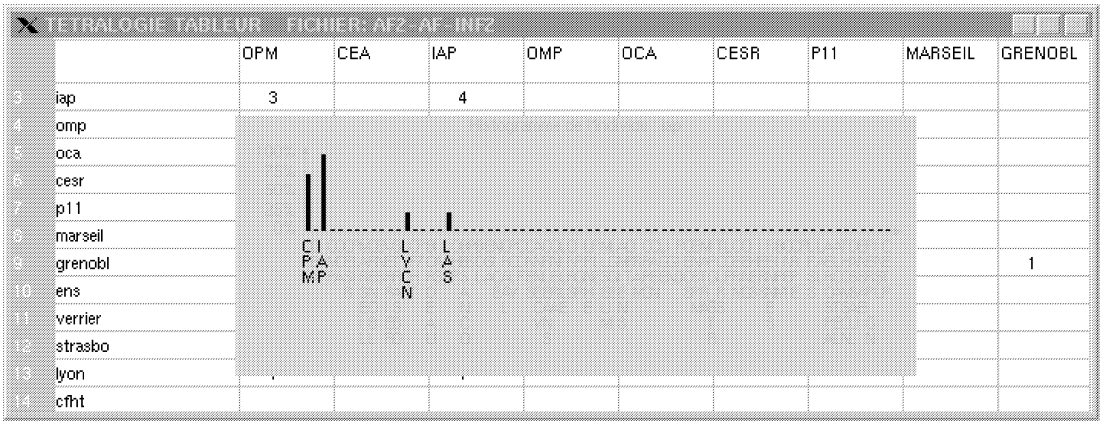


Fig. 4. IAP work group – infrared sub-domain.

- *Histogram visualization*: a table row (respectively a column) is visualized so that the strength of the correlations with the table column values (respectively raws) is shown (see Figs. 3 and 4 as examples).

Typology detection using factor analysis. As indicated above, contingency tables contain quantitative information about one-to-one or one-to-n relationships. Some methods (such as data cube) propose the visualization of these tables as they are in an n-dimension space. One can hardly use or visualize a space whose dimension is bigger than three. Therefore, to reduce the space dimension (which is given by the number of the columns) and still lose the minimum of information (carried by the tables) we use factorial methods [1]. The information is then displayed in spaces which are induced by the eigen vectors associated to the most important eigen values of the variance/covariance matrix of the contingency tables. The spaces maximize the amount (in terms of inertia) of the visualized information. The distance used to calculate the inertia can be either the Euclidean or χ^2 measures. The Euclidean distance permits the visualization of quantitative relationships, and it is closely related with Principal Component Analysis (PCA). The χ^2 distance permits the visualization of qualitative relationships – typologies – related to Correspondence Factor Analysis (CFA). The latter method also permits simultaneous visualizations of both contingency table columns and rows. This allows one to understand the associations that may exits

Table 2
Example of relevant minings

From the following crossings	Can be detected
Keywords–BibCodes	Specific documents Documents covering specific terms Term-document relationships visualization
Keywords–Keywords	Common terms, specific terms Term associations visualization
Authors–Keywords	Domain specificities of the authors

between the two kinds of information (columns and rows).

Table 2 gives same examples of the kinds of information that can be detected from documents.

3. Application to astronomical literature

3.1. Description of the document sample

We have selected through ADS [4] all the papers published from the years 1987 to 1996, for which at least one of the authors was affiliated with a French institute. Note that this process excluded all the papers for which the affiliations were not available from ADS. A systematic effort has been made to manually complete the data set for articles published in volumes of *Astronomy & Astrophysics* – the main refereed journal for French astronomers – for which affiliations are missing in the ADS data base. We obtained 6190 documents. In fact, because of missing affiliations in ADS for a significant fraction of references, this data set is not exhaustive, but can be reliably used as a representative sample. From this data set we extracted 5229 terms from the keyword field, 6455 author names, and 71 different affiliations (after a careful editing to avoid any duplication). The list of authors includes authors affiliated to French institutions, but also all their co-authors, whatever their affiliations are. We did not attempt to link authors to their exact affiliation (all the co-authors of a publication are directly linked to all the affiliations given in that publication.

3.2. Detection of collaborative work

As said before, such knowledge can be detected by crossing the affiliation values by themselves. It is then possible to visualize the correlation a given affiliation has with the other ones and the strength of those links (see Fig. 3). These correlations take into account the whole document set. Using a different filter while building the crossing table, it is possible to directly visualize the same correlations which take into account a sub-set of the documents. As an example, Fig. 4 displays the results obtained by filtering the information on the INFRARED sub-domain.

In Fig. 3 the IAP collaborations are visualized. All the affiliations written in black (OPM, CEA, IAP, OMP, etc.) collaborate with IAP whereas grey ones do not collaborate (according to the initial document set).

One can note (see Fig. 4) that the collaboration between OPM and IAP in the INFRARED sub-domain is relatively much more important than the same collaboration in general (taking all the sub-domains). Most of the labs that collaborate with IAP do not collaborate on INFRARED sub-domain.

3.3. Analysis of the domain evolution

As described in Table 1, the evolution of the domain can be detected by crossing keywords or phrases with the different date values. In fact, most of the time, it is difficult for a non domain-expert to decide whether the evolution detected is a real domain evolution (e.g., more or less interest in a domain) or just a change in the keyword use. Two examples are shown Figs. 5 and 6 (keywords HIPPARCOS and BINARY STARS). The peaks in the use of the HIPPARCOS

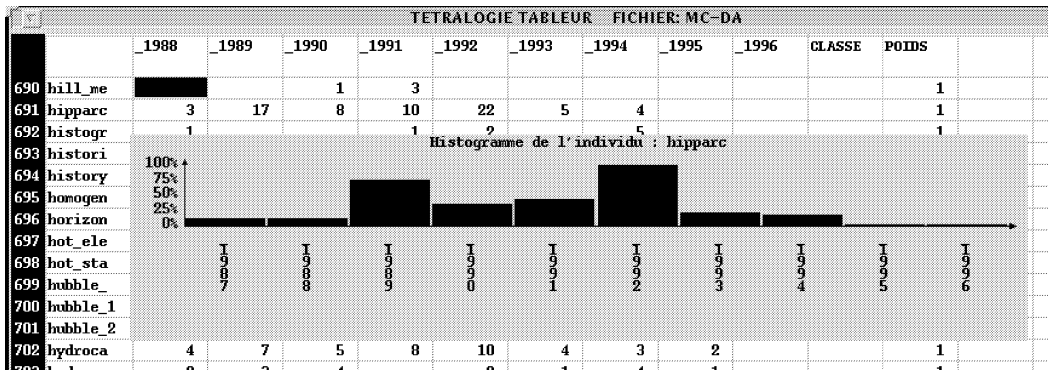


Fig. 5. Evolution of HIPPARCOS interest along time.

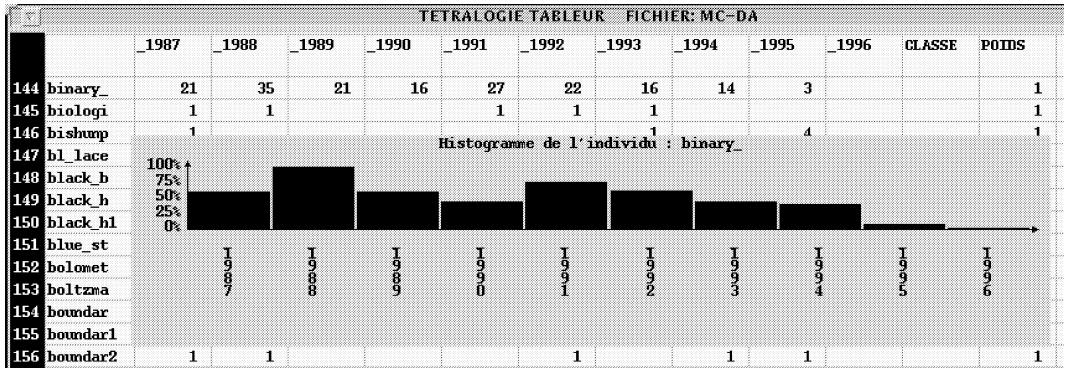


Fig. 6. Evolution of BINARY STARS interest along time.

keyword obviously correspond to key moments of the Hipparcos mission: launch in 1989 and release of intermediary data in 1992. The larger expected peaks of 1997–98 after the release of the final catalogues are not yet included in our sample.

3.4. Specific bias discovery

The document specificities can be discovered by crossing their bibcode with other attributes and by applying a factorial analysis on the resulting crossing table. As an example, crossing the bibcodes and the phrases extracted from the documents can be used to discover the domain specificities of the documents. Fig. 7 shows the results obtained when analyzing the INFRARED sub-domain. For example, one can see that INFRARED SPECTROMETERS is a specific keyword for the documents 1988N89–12527...F, 1989Icar...94...32K and other documents displayed on the top right. In the same way, INFRARED SPACE OBSERVATORY (ISO) make some documents (1990oeob.book..205C, . . .) specific compared to the others in the domain.

A complementary analysis can be done in order to discover the different sub-domains and their specificities according to the keywords usage. Fig. 8 shows the results of a CFA applied to the keywords–keywords crossing. Clear sub-groups appear. They are constituted of papers dealing with:

(1) Interstellar matter and stars,
(2) Galactic structure and external galaxies.

There are a few keywords bridging the gap between these two subgroups: SURVEYS, INTERSTELLAR DUST, INFRARED CIRRUS. These keywords do not discriminate between papers dealing with stellar and galactic studies. On the right hand side, one can

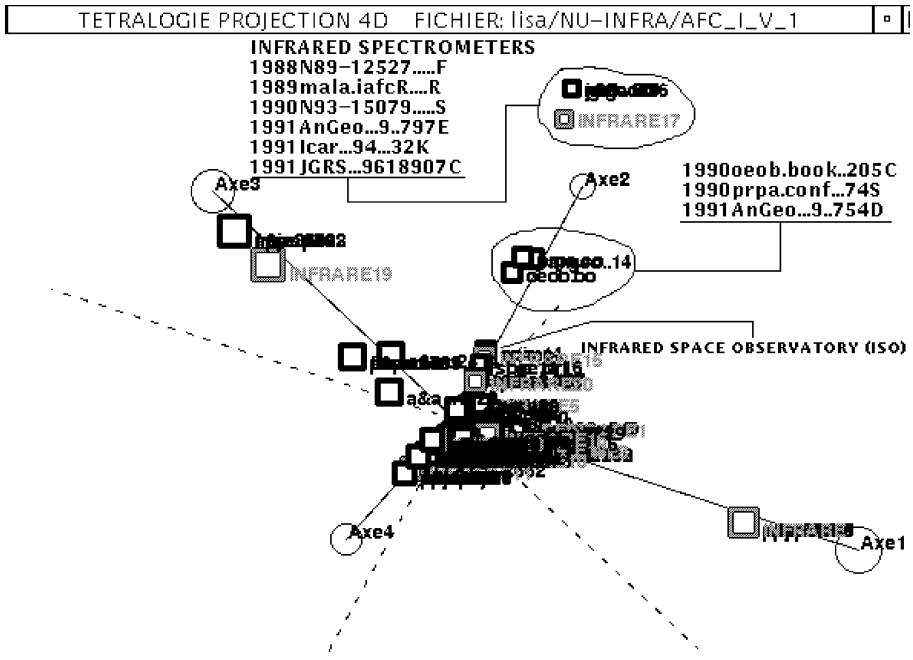


Fig. 7. Document specificities – (INFRARED sub-domain).

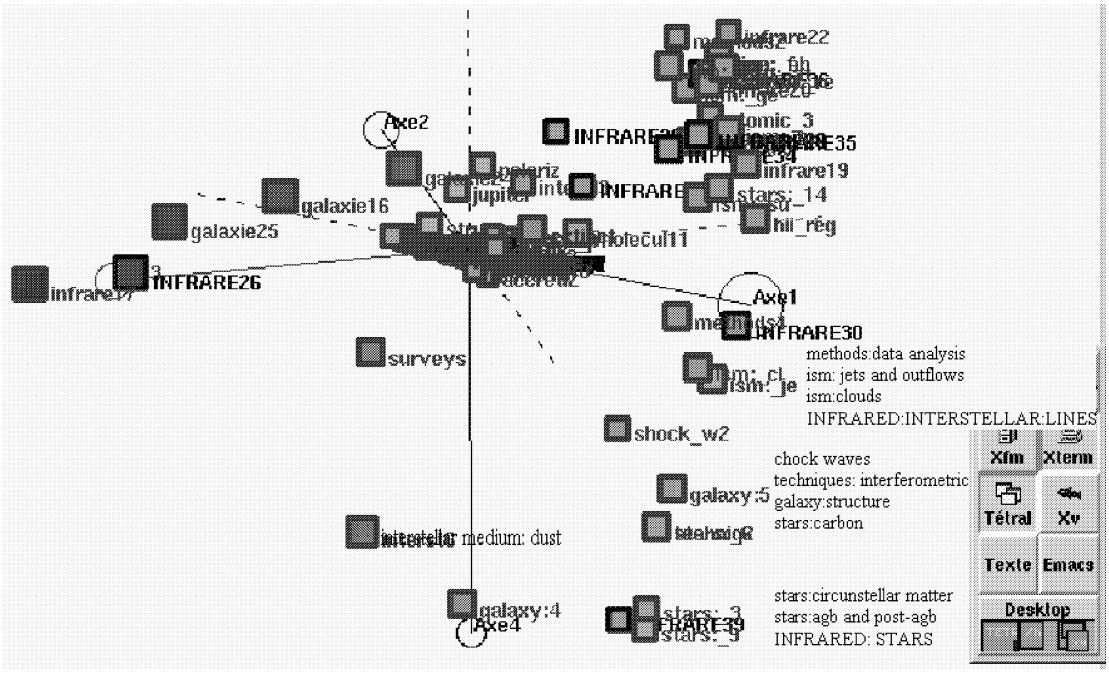


Fig. 8. Term correlations (INFRARED sub-domain).

see lists of keywords for two subgroups within the stellar domain, one connected to the interpretation of the physical phenomena (jets, outflows), the second one linked to the analysis of stellar atmospheres. According to this first analysis, it can then be interesting to visualize the corresponding author names or document contents. The software provides these facilities.

4. Conclusion

Information systems provide efficient tools to create electronic documents and it is necessary to provide efficient tools to retrieve and take advantage of this information. Information retrieval systems make it possible to retrieve documents or document pieces according to a keyword based query. Nevertheless other users' needs have to be answered. More and more often, long lists of documents do not satisfy users; they need global views of the retrieved pieces of information. One of the goals of information discovery systems is to answer this kind of user need. In this paper, we have presented our view of what can be a discovering process from semi-structured documents. Technologies from different fields are used to achieve this. We present a framework that aims to extract the information to mine from different heterogeneous document sources. In addition we present a few methods to extract targeted knowledge from a document set. We give some examples using documents from the astronomical literature.

References

[1] J.P. Benzécri, L'Analyse des Données, Tomes 1 et 2 (Dunod, Paris, 1973).
[2] C. Chrisment, T. Dkaki, B. Dousset, J. Mothe, ISI 5 (3) (1997) 367–400 (ISSN 1247-0317).
[3] D. Egret, J. Mothe, T. Dkaki, B. Dousset, in: Astronomical Data Analysis Software and Systems VII, R. Albrecht, R.N. Hook, H.A. Bushouse (Eds.), 1998, pp. 461–465.
[4] G. Eichhorn, An overview of the astrophysics data system, Experimental Astronomy 5 (1994) 205–220.
[5] G. Eichhorn et al., in: ASP Conf. Series, Vol. 125, Astronomical Data Analysis Software and Systems VI, G. Hunt, H.E. Payne (Eds.), 1997, p. 569.
[6] Fayyad et al., Advances in Knowledge Discovery and Data Mining (AAAI Press, 1996) (ISBN 0-262-56097-6).
[7] J. Mothe, D. Egret, T. Dkaki, B. Dousset, in: Library and Information Services in Astronomy III, ASP Conf. Series, Vol. 153, U. Grothkopf, H. Andernach, S. Stevens-Rayburn, M. Gomez (Eds.), 1998, pp. 69–76.
[8] MUC7, 1998, Message Understanding Conference, DARPA/ITO.
[9] F. Murtagh, A. Heck, Knowledge-Based Systems in Astronomy, Lecture Notes in Physics 329 (Springer, Heidelberg, 1989) (ISBN 3-540-51044-3).
[10] M.-T. Pazienza, Information extraction, A multidisciplinary approach to an emerging information technology, 1997 (ISBN 3-540-63438).
[11] C.J. Van Rijsbergen, Information Retrieval, 2nd edn. (Butterworths, London, 1979).
[12] G. Salton et al., Introduction to Modern Retrieval (McGraw-Hill, 1983) (ISBN 0-07-66526-5).
[13] Trec7, Text Retrieval Conference, D.K. Harman (Ed.), 1998.