Library and Information Services in Astronomy III
ASP Conference Series, Vol. 153, 1998
U. Grothkopf, H. Andernach, S. Stevens-Rayburn, and M. Gomez (eds.)

Analysis of Evolutionary Trends in Astronomical Literature using a Knowledge-Discovery System: Tétralogie

Josiane Mothe

IRIT, Université Paul Sabatier, Toulouse, France;
IUFM, Institut Universitaire de Formation des Maîtres, Toulouse, France

Daniel Egret

CDS, Observatoire astronomique de Strasbourg, Strasbourg, France

Taoufiq Dkaki

IRIT, Université Paul Sabatier, Toulouse, France;
CDS, Observatoire astronomique de Strasbourg, Strasbourg, France;
IUT Strasbourg Sud, Université Robert Schuman, France

Bernard Dousset

IRIT, Université Paul Sabatier, Toulouse, France

Abstract. Databases of electronic abstracts provide a wealth of information for studies about covered topics, collaborative works, or short-term evolution trends.

In this paper, we report a study of a subset of the ADS database (Astrophysics Data System) of astronomy abstracts, restricted to 6190 articles for which at least one of the authors is affiliated to French institution, in the range 1987 to 1996.

In a next step, Tétralogie is used for analysing evolutionary trends. For that, some data analysis mining functions, such as the Principal Component Analysis or classifications methods, are used to discover the most significant combinations of keywords, clusters of similar papers or outliers presenting extreme characteristics in the analysis.

In addition to allowing us to discover global dependencies between the document contents, the applied methods allow one to discover features on the short-term evolution of topics along the years covered by the dataset: what are the growing fields, the emerging centres of interest, and those which gradually become “out of fashion”.

In a similar way, it allows one to discover the evolution of the collaboration between the several teams involved in projects during the covered period.

1. Introduction

More and more information is available in electronic form so that information seeking is a crucial task. At the same time, the increase in importance, range and volume of the information makes information seeking a more and more complex task. Sophisticated information retrieval systems (IRS) have to be designed in order to answer a wide range of users' needs. When dealing with textual information, information retrieval technology can be used to retrieve documents according to a user's query expressed in natural-like language form (see Salton & McGill 1983). Most of the systems based on that technology present the search results to the users in the form of a ranked list of documents.

In any case, when querying large sets of information, a document list provided as a result can be of limited value for the user: even ranked in a probable relevance order, the retrieved documents are generally too numerous so that it is too time-consuming for the user to fully exploit them. It would be much more useful for him to be first provided with overviews of the retrieved information or with some patterns induced from the information. For that, efficient information summarization techniques have to be applied in addition to efficient interfaces to display the results.

In this paper, we present new ways to display the retrieved documents to users. The main goal is to allow them to discover strategic but hidden information from the raw documents without having to read them. The system mines the retrieved documents and provides the user with graphical views of the correlations that have been discovered between the documents themselves or between some of their components (authors, topics, author affiliation, etc.), possibly as a function of time. The system developed, named Tétralogie (see Chrisment et al. 1997), has been applied to a sample of astronomical literature. Two complementary studies have been realized: one provides a global analysis of the set of documents whereas the second one provides a deep analysis of a sub-domain.

The first task is to select the documents on which the mining has to be done. It is achieved by querying any IRS or any dedicated server. The next step consists in mining this selected information in order to discover unknown but useful patterns.

In the next section we present the way Tétralogie achieves these tasks. Then we present the results obtained when mining a sample of documents from astronomical literature.

2. Knowledge Discovery Steps using the Tétralogie System

Tétralogie is a knowledge discovery system from textual information sets. Its main goal is to allow a user to answer questions such as: what are the specificities of a domain? Who are the main researchers from a domain and what are their relationships, how do those relationships evolve with time, etc. It makes it possible to discover correlations that exist between the document features.

To achieve that, a series of complementary processes are used.

2.1. Information extraction

This task is used to extract some relevant elements from the raw information (authors names, authors affiliations, publication dates, topics, ...) in order to analyse them. Information extraction in Tétralogie is based on the tags used in the documents. For example, INSPEC provides, among others, AUTHORS and JOURNAL tags that can be used; SGML documents have relevant tags as well. In addition, Tétralogie provides an efficient phrase extraction module that allows to extract terms from full text components such as the document title, the abstract or the document itself. To summarize, this step provides several semantic document representations (according to their authors, keywords, etc.).

2.2. Information filtering

Subsets of the extracted values can be used as filters. Doing so, a user can indicate which values he is interested in (positive filters) or which values he is not interested in (negative filters).

2.3. Information reduction

The goal of this task is to reduce the extracted information: keeping a global information instead of all the detailed information.

Tétralogie uses 2 or 3 dimensional contingency and disjunctive crossing tables. In a table, each dimension corresponds to a kind of extracted element (author name, affiliation, journal, publication date, etc.). If T is a contingency table, T_{ij} corresponds to the number of documents where the values i and j co-occur.

2.4. Information mining

The goal of the information mining step is to achieve different mining functions (Fayyad et al. 1996): classification, dependencies or sequences (temporal dependencies) discovering.

Tétralogie mining functions are based on statistical data analysis methods (see e.g., Benzecri 1973; Murtagh & Heck 1989). All these methods use a contingency table as input. The table rows are the elements to be analysed, and they are initially represented in the multi-dimensional column space.

Principal component analysis (PCA) and Correspondence Factorial Analysis (CFA) are used for representing the elements in a more convenient space, with a reduced number of dimensions.

Hierarchical Ascendant Classification and Classification by Partition are used to classify the elements according to various similarity measures.

2.5. Result presentation

Graphical tools are used to display the mining results: for example, the result of information mining obtained using CFA can be represented as a set of points in a reduced multi-dimensional space. Tools are proposed for displaying this information in a four-dimensional space, and for interactively changing the point of view (zoom, rotation) for the visualised space.

3. Global Analysis of a Set of Documents: Articles with French Affiliations (1987–1996)

3.1. The document sample

Document selection. We have selected through ADS (Eichhorn et al. 1995) all papers published in the years 1987 to 1996 for which at least one of the authors was affiliated in a French institute. Note that this process excluded all papers for which the affiliations were not available from ADS. A systematic effort has been made to complete manually the dataset 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. We insist that, because of missing affiliations in ADS for a significant fraction of references, this dataset is not exhaustive, but can be used reliably as a representative sample.

Information extraction. From this data set we extracted 5229 terms out of the keyword field, 6455 author names, and 71 different affiliations (after a careful editing for avoiding any duplication).

Note that the list of authors includes authors affiliated in French institutions, but also all their co-authors, whatever their affiliation is. We did not attempt to link authors to their exact affiliation.

3.2. Analysis of the field evolution

It is possible to analyse the evolution in several ways. One can focus on the fields of interest and their evolution with time. This can be done by analysing the evolution of the frequency of selected keywords along the years, or by finding newly appearing keywords.

Topic evolution. Let us take the examples of the keywords HIPPARCOS or BINARY STARS (see Fig. 1 and 2).

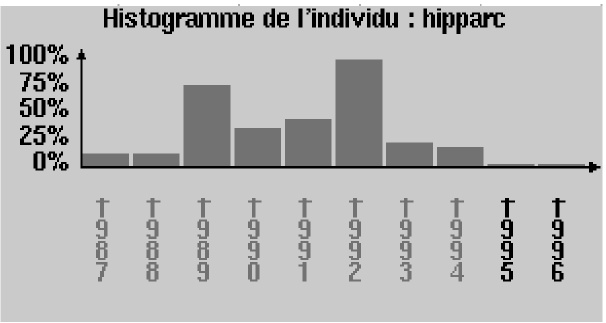


Figure 1. Evolution of HIPPARCOS interest versus time.

The peaks in the use of the HIPPARCOS 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.

According to the interpretation, it may be sometimes difficult to decide if it is a real effect (decrease of activity in a given field) or simply a trend in the way Editors are using the keywords.

A complementary indication is given by the list of terms which appear only in the most recent years: in our data set this includes, e.g., the two following terms: "ISM: MOLECULES" and "STARS: CIRCUMSTELLAR MATTER". This shows the increasing importance of studies on the interstellar medium and its chemical composition.

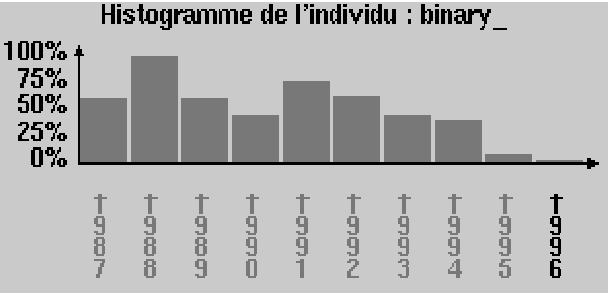


Figure 2. Evolution of BINARY STARS interest versus time.

Collaborative work evolution. One possible way to analyse collaborative work is to cross author names and affiliations to discover the strong correlations. In complement, evolution of these collaborations can be studied by crossing the author names and the affiliations for several consecutive year bins. In the present study we have used the three following bins: 1988–90, 1991–93, 1994–96. Note that only affiliations corresponding to French institutions are used in this study.

Figure 3 provides an example of evolution of the collaborative work for a given author (Ferlet). The histograms display the relative number of papers in which affiliation from given French institutes appear. (IAP appears the largest number of times, and is normalised to 100%, as the author himself is affiliated there).

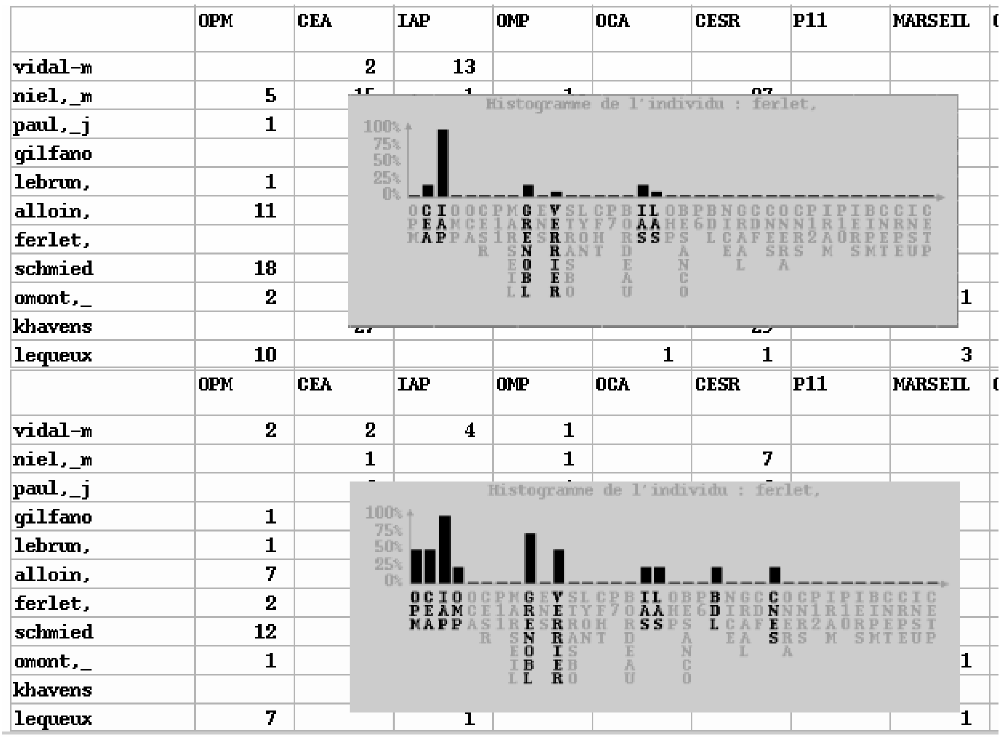
4. Specific Analysis of a Subdomain

Different kinds of information can be discovered according to a given topic. First of all, it can be interesting to know what is the specific vocabulary of the domain or what are the topics strongly related to that domain. Then, one can be interested in knowing who are the authors of the domain or what are the laboratories or observatories which are really implied in that domain.

To proceed, it is first necessary to define precisely the domain of interest, that is to say to determine a set of terms that could filter the documents related to the studied domain.

4.1. Term selection

Depending on the user's knowledge on the domain, it can consist either in selecting manually all the words that are known to be related to the domain, or



	OPM	CEA	IAP	OMP	OCA	CESR	P11	MARSEIL
vidal-m		2	13					
niel,_m	5							
paul,_j	1							
gilfano								
lebrun,	1							
alloin,	11							
ferlet,								
schmied	18							
omont,_	2							1
khavens								
lequeux	10				1	1		3

	OPM	CEA	IAP	OMP	OCA	CESR	P11	MARSEIL
vidal-m	2	2	4	1				
niel,_m		1		1		7		
paul,_j								
gilfano	1							
lebrun,	1							
alloin,	7							
ferlet,	2							
schmied	12							
omont,_	1							1
khavens								
lequeux	7		1					1

Histogramme de l'individu : ferlet,
100%
75%
50%
25%
0%
OPM CEA IAP OMP OCA CESR P11 MARSEIL GRENOBL ENS VERRIER STRASBO LYON CFHT P7 BORDEAU IAS LAS OHP BESANCO P6 BDL NICE GRAL CDF CNES ONERA CNRS P12 IRAM P10 IERS BIPM CNET CRPE INSU CETP
Histogramme de l'individu : ferlet,
100%
75%
50%
25%
0%
OPM CEA IAP OMP OCA CESR P11 MARSEIL GRENOBL ENS VERRIER STRASBO LYON CFHT P7 BORDEAU IAS LAS OHP BESANCO P6 BDL NICE GRAL CDF CNES ONERA CNRS P12 IRAM P10 IERS BIPM CNET CRPE INSU CETP

Figure 3. Evolution of collaborations for author FERLET

in starting from a seed set of significant words and automatically finding all strongly related terms.

Automatic selection. According to one important word of the domain, it is possible to use the links it has with the other words in the documents in order to determine the other important words of the domain. The process is based on a crossing table (terms vs terms). The result of the crossing can be sorted so that the most strongly related terms appear close.

As an example, we have used "INFRARED" as a seed term, for obtaining a list of 110 keywords, strongly connected to the 33 keywords containing the character string INFRARED. Those keywords are expected to be relevant to the astronomical observations in the Infrared wavelength.

4.2. Term correlations

Using a CFA, it is possible to detect the main groups in the field, that is, the specificities of the field content. It is also possible to visualise the correlations that exist between those sub-fields.

Clear sub-groups appear, consisting of papers dealing with:

1. Interstellar matter and stars
2. Galactic structure and external galaxies

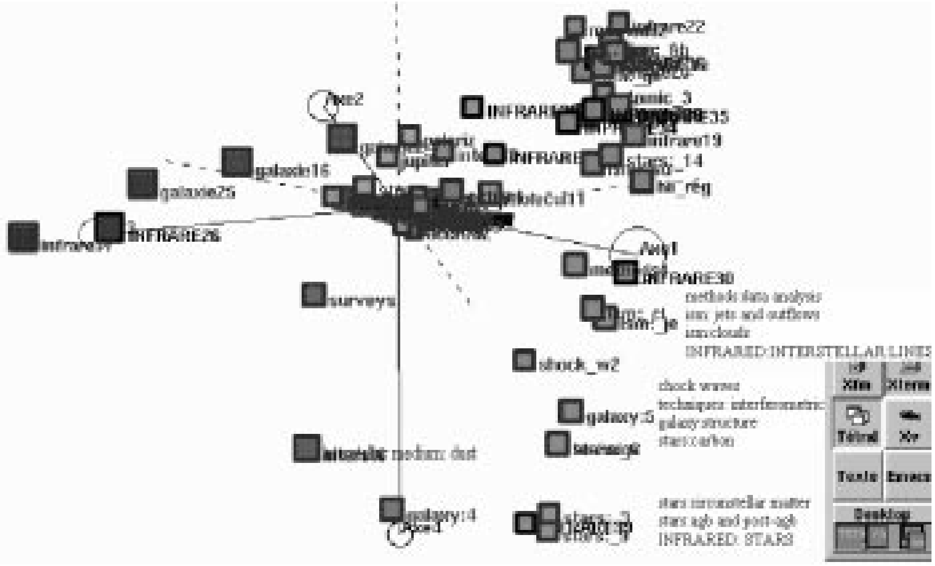


Figure 4. Term correlations (INFRARED subfield).

In Fig. 4, 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 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.

4.3. Specificities of authors with respect to domains

The crossing of the keywords related to INFRARED, with the most prolific authors is used to trace the expertise domain of some of the authors.

A CFA using this crossing allowed to discover a specific group of authors (at the bottom of Fig. 5) which appear all connected to "TITAN" (a keyword related to the INFRARED domain): these authors are found as those most closely connected to the Infrared observation of TITAN (in the French affiliation dataset).

5. Conclusion

In a first paper (Egret et al. 1998) presented at the ADASS VII Conference, we have shown how the Tétralogie system (http://atlas.irit.fr/), based on data analysis methods, can be used to mine astronomical information. We analysed a sample of abstracts from the astronomical literature and showed some results of a study of the collaborative work around some large observational projects.

In this paper we tried to show how a knowledge discovery system can be used for discovering evolutionary trends in a data set extracted from the astronomical literature. The main steps of the analysis we have presented can be summarized as follows:

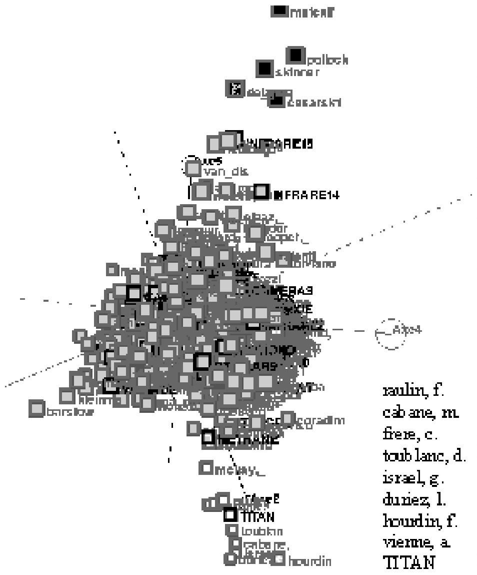


Figure 5. Term-author crossing. The specific group connected with the keyword TITAN is listed.

- Efficient pre-treatment of textual information;
- Use of a panel of mining methods (classification, factorial and procustean analyses);
- graphical representation;
- user interaction (filtering, mining, visualisation);
- validation and interpretation of the discoveries.

References

Benzecri, J. P. 1973, L'analyse de données, Tome 1 et 2, Dunod Edition

Chrisment, C., Dkaki, T., Dousset, B. & Mothe, J. 1997, ISI vol. 5(3), 367-400 (ISSN 1247-0317)

Egret, D., Mothe, J., Dkaki, T. & Dousset, B. 1998, in *Astronomical Data Analysis Software and Systems VII*, R. Albrecht, R. N. Hook & H. A. Bushouse, eds., ASP Conf. Ser. 145, 461-465

Eichhorn, G. et al. 1995, in *Astronomical Data Analysis Software and Systems IV*, R. A. Shaw, H. E. Payne, & J. J. E. Hayes, eds., ASP Conf. Series vol. 77, 28

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R. (eds.) 1996, Advances in knowledge discovery and data mining, Menlo Park, CA: AAAI Press (ISBN 0-262-56097-6)

Murtagh, F. & Heck, A. 1989, Knowledge-based systems in astronomy, Lecture Notes in Physics 329, Heidelberg: Springer-Verlag (ISBN 3-540-51044-3)

Salton, G. & McGill, M. J. 1983, Introduction to modern information retrieval, New York: McGraw Hill (ISBN 0-07-66526-5)