Astronomical Data Analysis Software and Systems VII
ASP Conference Series, Vol. 145, 1998
R. Albrecht, R. N. Hook and H. A. Bushouse, eds.

Information Mining in Astronomical Literature with Tetralogie

D. Egret

CDS, Strasbourg, France, E-mail: egret@astro.u-strasbg.fr

J. Mothe[1], T. Dkaki[2] and B. Dousset

Research Institute in Computer Sciences, IRIT, SIG, 31062 Toulouse Cedex, France, E-mail: {mothe/dkaki/dousset}@irit.fr

Abstract. A tool derived for the technological watch (Tetralogie[3]) is applied to a dataset of three consecutive years of article abstracts published in the European Journal *Astronomy and Astrophysics*. This tool is based on a two step approach: first, a pretreatment step that extracts elementary items from the raw information (keywords, authors, year of publication, etc.); second, a mining step that analyses the extracted information using statistical methods. It is shown that this approach allows one to qualify and visualize some major trends characterizing the current astronomical literature: multi-author collaborative work, the impact of observational projects and thematic maps of publishing authors.

1. Introduction

Electronic publication has become an essential aspect of the distribution of astronomical results (see Heck 1997). The users who want to exploit this information need efficient information retrieval systems in order to retrieve relevant raw information or to extract hidden information and synthesize thematic trends. The tools providing the former functionalities are based on query and document matching (Salton et al. 1983; Frakes et al. 1992; Eichhorn et al. 1997). The latter functionalities (called data mining functionalities) result from data analysis, data evolution analysis and data correlation principles and allow one to discover a priori unknown knowledge or information (Shapiro et al. 1996). We focus on these latter functionalities.

A knowledge discovery process can be broken down into two steps: first, the data or information selection and pre-treatment; second, the mining of these pieces of information in order to extract hidden information. The main objectives of the mining are to achieve classification (i.e., finding a partition of the

[1]Institut Universitaire de Formation des Maîtres de Toulouse

[2]IUT Strasbourg-Sud, Université Robert Schuman, France

[3]http://atlas.irit.fr

data, using a rule deduced from the data characteristics), association (one tries to find data correlations) and sequences (the objective is to identify and to find the temporal relationships between the data). The information resulting from an analysis have then to be presented to the user in the most synthetic and expressive way, including graphical representation.

Tetralogie[4] is an information mining tool that has been developed at the Institut de Recherche en Informatique de Toulouse (IRIT). It is used for science and technology monitoring (Dousset et al. 1995; Chrisment et al. 1997) from document collections. In this paper it is not possible to present in detail all the system functionalities. We will rather focus on some key features and on an example of the results that can be obtained from astronomical records.

2. Information Mining using "Tétralogie"

The information mining process is widely based on statistical methods and more precisely on data analysis methods. First, the raw information have to be selected and pre-treated in order to extract the relevant elements of information and to store them in an appropriate form.

2.1. Information Harvesting and Pre-treatment

This step includes the selection of relevant raw information according to the user's needs. It is generally achieved by querying a specific database or a set of databases. Once the raw information has been selected, the next phase is to extract the relevant pieces of information : e.g., authors, year of publication, affiliations, keywords or main topics of the paper, etc. This is achieved using the "rewriting rule principle". In addition, the feature values can be filtered (e.g., publications written by authors from selected countries) and semantically treated (dictionaries are used to solve synonymy problems). These relevant feature values are stored in contingency and disjunctive tables.

Different kinds of crossing tables can be performed according to the kind of information one wants to discover, for example:

Kind of crossing	Expected discovering
(authors name, authors name)	Multi-author collaborative work
(authors name, document topics)	Thematic map of publishing authors
(document topics, authors affiliation)	Geographic map of the topics

2.2. Data Analysis and Graphical Representation

The preprocessed data from the previous step are directly usable by the implemented mining methods. These methods are founded on statistical fundamentals (see e.g., Benzecri 1973; Murtagh & Heck 1989) and their aim is either to represent the pre-treated information in a reduced space, or to classify them.

The different mining functions used are described in Chrisment et al. (1997). They include: Principal component analysis (PCA), Correspondence Factorial

[4]This project is supported by the French defense ministry and the Conseil Régional de la Haute Garonne.

Analysis (CFA), Hierarchical Ascendant Classification, Classification by Partition and Procustean Analysis.

The Result Visualization The information mining result is a set of points in a reduced space. Tools are proposed for displaying this information in a four dimensional space, and for changing the visualized space, or the point of view (zoom, scanning of the set of points).

The User Role In addition to be a real actor in the information harvesting phase, the user has to intervene in the mining process itself: elimination of some irrelevant data or already studied data, selection of a data subset to analyze it deeper, choice of a mining function, and so on.

3. Application to a Dataset from the Astronomical Literature

3.1. The Information Collection

The information collection used for the analysis was composed of about 3600 abstracts published in the European Journal *Astronomy and Astrophysics* (years 1994 to 96). Note that this dataset is therefore mainly representative of European contributions to astronomy, in the few recent years.

In that abstract sample, it is possible to extract about 1600 different authors, and 200 can be selected as the most prolific. Topics of the documents can be extracted either from the title, from the keywords or from the abstract field. Titles have been considered as too short to be really interesting for the study. In addition, the use of keywords was considered too restrictive, as they belong to a controlled set. Indeed, we preferred to automatically extract the different topics from the words or series of words contained in the abstracts.

3.2. Study of the Collaborative Work

The details concerning the collaborative works can be discovered using (author name / author name) crossing and analyzing it.

The first crossing was done using all the authors. A first view is obtained by sorting the author correlations in order to find the strong connexities. The resulting connexity table shows strongly related groups (about 15 groups appear on the diagonal). They are almost all weakly linked via at least one common author (see the several points above and below the diagonal line, linking the blocks). The isolated groups appear on the bottom right corner. This strong connexity is typical of a scientific domain including large international projects and strong cooperative links.

One can go further and study in depth one of these collaborative groups: a CFA of the (main author / author) crossing shows, for instance, some features of the collaborative work around the Hipparcos project in the years 1994-96 (i.e., before the publication of the final catalogues) as can be viewed by grouping together authors having papers co-authored with M. Perryman (Hipparcos project scientist) and L. Lindegren (leader of one of the scientific consortia). The system allows the extraction of 25 main authors (with more than two publications, and at least one with one of the selected central authors) and the cross-referencing of them with all possible co-authors.

3.3. Thematic Maps of Publishing Authors

The details concerning the thematic map of publishing authors can be discovered using (author name / topic) crossing and analyzing it. The significant words in the abstracts have been automatically extracted during the first stage (see 2.1) and they are crossed with the main authors. That kind of crossing allows one to discover the main topics related to one or several authors ; it can also show what are the keywords that link several authors or that are shared by several authors.

For instance, crossing main authors of the 'Hipparcos' collaboration with topical key words, allowed us to discover the main keywords of the 'peripheral' authors (those who bring specific outside collaborations).

4. Conclusion

In this paper, we have tried to show the usefulness of the TETRALOGIE system for discovering trends in the astronomical literature. We focused on several functionalities of this tool that allow one to find some hidden information such as the teams (through the multi-author collaborative work) or the topical maps of publishing authors. This tool graphically displays the discovered relationships that may exist among the extracted information.

Schulman et al. 1997, using classical statistical approaches, have extracted significant features from an analysis of subsequent years of astronomy literature. In a forthcoming study, we will show how the Tetralogie system can also be used to discover thematic evolutions in the literature over several years.

The Web version[5] contains additional figures for illustration.

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 (ISSN 1247-0317)

Dousset, B., Rommens, M., & Sibue, D. 1995, Symposium International, Omega-3, Lipoprotéines et atherosclerose

Eichhorn, G., et al. 1997, in ASP Conf. Ser., Vol. 125, Astronomical Data Analysis Software and Systems VI, ed. Gareth Hunt & H. E. Payne (San Francisco: ASP), 569

Frakes et al. 1992, Information retrieval, Algorithms and structure (ISBN 0-13-463837-9)

Heck, A. 1997, "Electronic Publishing for Physics and Astronomy", Astrophys. Space Science 247, Kluwer, Dordrecht (ISBN 0-7923-4820-6)

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

[5] http://cdsweb.u-strasbg.fr/publi/tetra-1.htx

Salton, G., et al. 1983, Introduction to modern retrieval, McGraw Hill International (ISBN 0-07-66526-5)

Shapiro et al. 1996, Advances in Knowledge discovery and Data Mining, AAAI Press (ISBN 0-262-56097-6)

Schulman, E., et al. 1997, PASP 109, 741