Information mining : extracting, exploring and visualising geo-referenced information

Claude Chrisment Bernard Dousset Saïd Karouach Josiane Mothe

IRIT
118 route de Narbonne,
31062 Toulouse France
33 5 61 55 63 22

{chrisme/dousset/mothe}@irit.fr

Categories and Subject Descriptors
H.3 [**Information storage and retrieval**]

Keywords
Information mining, Geo-referenced data, Interface, Knowledge Discovery

1. INTRODUCTION
Large-scale analysis to understand a phenomena or to discover the structure of a domain becomes possible thanks to the availability of large sources of information. Many domains such as economy, history or science are involved. *Competitive intelligence* is a core issue in the new world of business. Companies need methods and tools to monitor the activities of its competitors, get information on the market or technologies. In the context of the *web-Graph*, hypertext references are mined in order to determine the authority of the pages or to extract information on host links or other networks. In *information retrieval* (IR), co-citation and co-authoring analysis is studied as a way to monitor scientific activities (White, 2003). Co-authoring analysis can also be used to identify groups of scientists and their inner structure

Graphical interfaces play a important role to display the results of such analysis to users. Graphs are among the most used as linking concepts or elements is the most common mining technique. When geographic referenced data is handled, that is to say data which has location reference within it's structure, the most intuitive way to describe and explain the spatial organisation of a phenomena is to use geographic maps. Geo-referenced data includes statistical data (demography, economy, etc.) and a wide range of information that can be transformed into quantified information. State of the art and evolution of scientific communities and research topics falls in this category. In this paper, we present GeoECD. This platform includes a set of tools that are used to analyse geo-referenced data and documents. Section 2 describes how the information is represented, based on both the document content and other criteria. This representation is used to retrieve specific items and to analyse a collection of documents. In Section 3, we present some of the mining and visualising tools that compose the platform. A case study is presented.

2. CASE STUDY
It is not possible to describe an entire case study here. However, the following figures are part of a real study we made. The objective of the study was to analyse international collaboration of INRA (Institut National de la Recherche Agronomique) based on publications and co-authoring. 28 631 documents were harvested from different databases according to the fact that one of the authors at least was affiliated to INRA.

3. INFORMATION REPRESENTATION
3.1 Multi-dimensional representation
In IR, documents are viewed as bags of weighted words: document indexing results in a one-dimensional information space representation -the free-text space. However, documents and texts express a vast and rich range of information that traditional information indexing does not take into account. We promote an approach in which documents are represented in a multi-dimensional space. Each dimension corresponds to a point of view that may be of interest for the user. For example, considering scientific literature, document dimensions can be “producer of the publication”, “temporal references”, “content”. Some of the dimensions are predefined, others are built from text analysis ; each dimension is organised along hierarchies (specificity/genericity relationships).

3.2 Information extraction
Information extraction provides a wide range of techniques to extract predefined elements from texts. These advanced techniques are useful when information has to be extracted from non structured documents. When considering semi-structured documents, information extraction is eased by markers. Depending of the nature of the extracted information, we applied either indexing-based or information extraction techniques. How we extract specific information from documents is out of the scope of this paper. This has been detailed in (Dkaki, 1996). We rather consider here the results of information extraction. Each document is represented by a set of couples (attribute: list of terms). Since attributes correspond to hierarchical dimensions and terms correspond to elements from these hierarchies, a document representation can also be considered as associations of a document to the corresponding dimension elements.

There is no real specificity for geo-referenced data extraction. Location corresponds to one dimension, like the other dimensions, it has a hierarchical structure. In our case study, the extraction module extract the author's affiliations from each publication.

Table 1 reports dimensions and the number of different extracted values of our case study.

Dimension	Number of different values
Author's country	110
Publication source	3354
Keywords	46330
Date of publication	4 (1998 to 2001)

Table 1. Document dimensions

3.3 Information summarisation

Information (a document collection) is summarised under the form of contingency table for which 2 dimensions are considered. For example Figure 1 shows the results of crossing publication producers (lines) and years (columns). The contingency table depicts the number of publications of each 'category'. A publication for which an author is from Algeria is counted as a publication from Africa.

		1998	1999	2000	2001
Africa	*Algeria*	18	24	21	25
	Benin	2	3	0	3
	Ivory Coast	11	2	7	2
Europe	*Austria*	18	10	26	42
	Belgium	197	163	158	195

⇕ The information can be aggregated, e.g. the detailed data (at the country name level) is aggregated to the continent level

	1998	1999	2000	2001
Africa	31	29	28	30
Europe	215	173	184	237

Figure 1: Example of 2-D contingency table

4. COMBINING METHODS TO ANALYSE AND VISUALISE GEO-REFERENCED DATA

4.1 Maps

Maps are the more intuitive way to display geo-referenced data to users. In our approach, these maps can be used in order to display quantitative data or they can display the results of other mining methods (see section 4.3). They always results from a 2-D table in which one dimension is the *location*.

Figure 2 shows a geographic map. It results from a contingency table that summarises the number of publications per country in the collection. Thus Figure 2 represents the strength of the collaboration INRA has with each country over the 4 years. Each country is automatically coloured according to the value obtained (the brightest, the strongest).

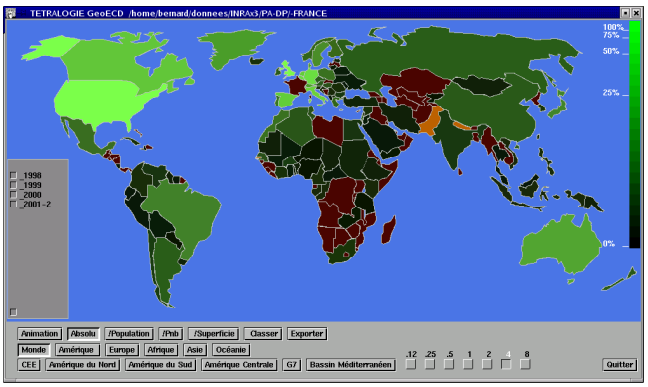

Figure 2: Strength of INRA collaborations

According to the type of information displayed, values can be normalised by parameters such as the Gross Domestic Product.

4.2 Classification modules

The platform consists of several classification modules based on well-known algorithms (Dousset, 2004). This includes the agglomerative clustering method. Figure 3 presents the dendrogram resulting from clustering countries according to the publication date. In that case, the collection has been summarised through a contingency table ; the two dimensions being the country the authors belongs to and the date of publication. The classification module is applied in order to group together countries that have the same behaviour with regard to the collaboration with INRA. The dendrogram can interactively be cut at any level according to the number/size of classes the user prefers.

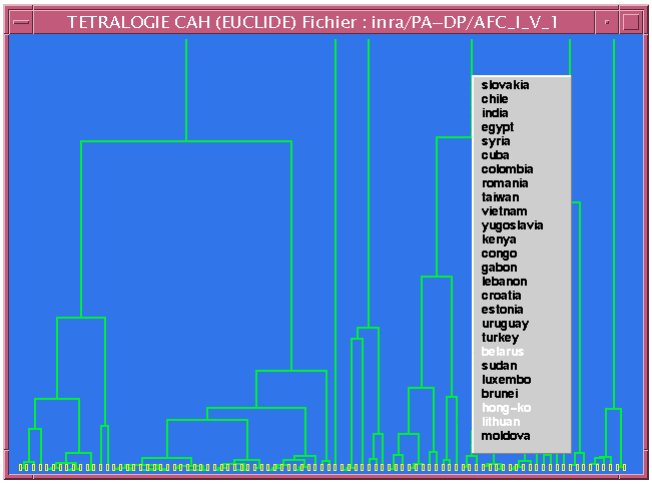

Figure 3. Dendrogram resulting from countries clustering

4.3 Combining mining modules

The result of a classification can be sent to the map module as shown Figure 4. In that case, to each class resulting from the clustering module is associated a colour (see right side of the map).

SIGIR'04, July 2004, Sheffield, UK.

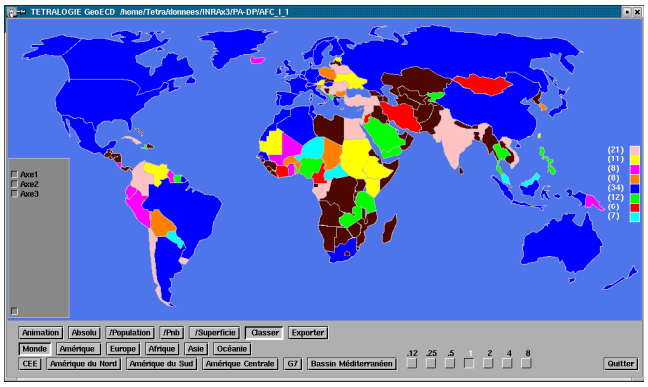

Figure 4: Map visualising the result of country classification

5. CONCLUSION

In this paper, we presented some elements of a platform consisting of different modules used to analyse and visualise geo-referenced data. Documents are first analysed in order to extract pre-defined elements such as location referenced. Doing so a multi-dimensional representation is extracted from each document. The document collection is then summarised under the form of 2-Dimensional tables from which mining modules extract new information. This new global information is graphically displayed. Mining and visualising modules are combined defining a new and useful way to view the documents and the structure of the domain they correspond to.

6. REFERENCES

[1] R.K. Buter, E.C.M. Noyons, Using Bibilometric Maps to Visualise term Distribution in Scientific Papers, Sixth International Conference on Information Visualisation, (2003), 697.

[2] C. Chen, Visualization Of Knowledge Structures, Handbook of Software Engineering and Knowledge Engineering, 2002.

[3] T. Dkaki, B. Dousset, J. Mothe, Mining Information in Order to Extract Hidden and Strategical Information, RIAO, (1997), 32-51.

[4] T. Dkaki, J. Mothe, Extraction et synthèse de connaissances à partir de bases de données hétérogènes, INFORSID, (1996), 287-308.

[5] B. Dousset, Tétralogie, atlas.irit.fr.

[6] V. Geroimenko, C. Chen, Visualizing the Semantic Web XML-based Internet and Information Visualisation, Springer, ISBN 1-85233-576-9, (2002).

[7] J. Mothe et al., DocCube: Multi-Dimensional Visualisation and Exploration of Large Document Sets, JASIST, 54(7), (2003), 650-659.

[8] J.-L. Multon et al., Analyse bibliométrique des collaborations internationales de l'INRA, Journées d'études sur les systèmes d'information élaborée, (2002), CD-ROM.

[9] H.D. White, K.W. McCain, Visualizing a discipline: an author co-citation analysis of information science, 1972-1995, JASIS, 49(4), (1998), 327-355.

[10] H.D. White, Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists, JASIST, 54(5), (2003), 423-434.

[11] M. Zitt, E. Bassecoulard, Development of a method for detection and trend analysis of research fronts built by lexical or co-citation analysis, Scientometrics, 30 (1994), 333-351.