Processing data stream by relational analysis

Ilhème Ghalamallah, Aziz Grimeh and Bernard Dousset

IRIT-SIG
Paul Sabatier University, 118 route de Narbonne, 31062 Toulouse cedex 9
{ghalamal, grimeh, dousset}@irit.fr

Abstract. In the business intelligence (BI) context, the majority of the strategic information comes from relational sources and the relevance of extracted knowledge usually depends on considering data evolution and their interactions. The multidimensional approach (nD) may bring forth a solution to identify and understand the underlying structures or strategies. But non-expert users get easily lost. Knowing that our team is experienced in knowledge extraction, we have already a platform called Tetralogie that is specialized for strategic scanning and another tool called Xplor which is dedicated to business intelligence. As a consequence, we provide a unified system to generate and manage relational data and extract implicit knowledge, whose content and format are adapted to decision-makers that are not experts in nD or BI.

Keywords: Knowledge extraction, business intelligence, relational analysis, evolution.

1 Introduction

In the context of the strategic scanning, Tétralogie is a tool particularly adapted to the macroscopic analyses. Indeed, it is able to detect the strong signals, the weak signals and tendencies from a corpus of documents collected for a precise subject. The elaborate information results, represents a synthesis obtained by various methods of data analysis and diffused via graphic visualizations. But because of the different strategic analyses that we have already carried with this software, it appeared that the end users of the produced analyses want, in addition to the general and strategic aspect (general knowledge), more precise views on certain points. In order to satisfy their specific needs for more precise information on elements, which they have already identified (competition, new products or processes, potential partners,…) or in order to discover other elements. Many experts and decision makers are demanding for more details while processing the elements that represent their traditional environment. These elements should contain more precise information about key words, the different actors, the prospective partners and markets that they're coveting for.

In addition, we propose for our macroscopic analyses a computerized decision-making system with perspective to automate the on-line processing of relational information and to propose analysis and navigation tools oriented to business intelligence (BI). This system provides strategic analyses on corpora of textual information resulting from the most various sources like: on line databases, Cds, the visible and invisible Web, the news, the press, linking sites, intern databases, etc. The proposed system gives the possibility to decision makers to perform their investigations without the participation of an analyst. The interaction between the System and decision makers can be on various levels such as choosing variables, extracting and selecting useful data, the choice of the analysis to be deployed and the visualization of the results.

Our principal goal is to reduce the cost of elaborate information and to increase the reactivity of the decision makers so we offer them a new interactive approach with their strategic data. We also offer query formulation assistance, and the development of new statistical graphs that aim to interpret evolutionary data.

2. Proposition

According to the report "Martre" the BI is defined as the group of the coordinated actions for research, treatment, distribution and protection of information that are useful for the economic actors, taking in consideration their individual and collective strategies. The system that we propose covers the essential treatments of the various phases of this cycle. We define the cycle of the BI as follows:
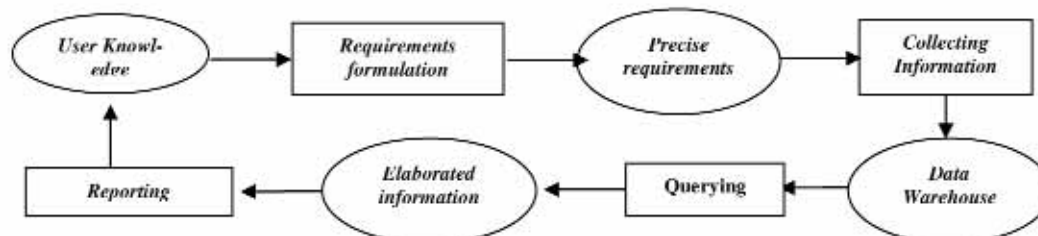

Fig. 1. Business intelligence life cycle

For this cycle we propose four keys steps (processes):

2.1 Requirements Formulation

In this step the user express their needs and requirements. So, he can identify and classify the targets that he's searching for, but his expressed needs are sometimes irrelevant or difficult. The objective of BI is to let them know the space of information that exists and specially the one that would be useful for them.

2.2 Information collect

Collecting information is based on different and various information sources like: on line databases, CDs, visible and invisible Web, copyrights, intern databases, the news, the press, etc.

The process of the collecting and preparing information in order to be analyzed is presented in five steps:

- Sources Identification: this process consists of identifying the sources that might contain the desired information.

- Information identification: this step helps in information retrieval and to identify useful information in the sources selected in the previous step in order to nourish the analysis.

- Structuring and homogenizing information: specific metadata for each database is established in order to describe in a standard form their initial format. Various databases could be simultaneously treated in the same analysis, even if it possesses different formats.

- 3D Matrix / list of evolutionary relations: these equivalent data structures allow to map the relation between different variables of the analysis, by integrating systematically the temporal variable. Our system proposes a generational model and the different relations generated are:

❖ Non oriented links: presence/ absence, contingence, co-occurrences, weighted co-occurrences, etc.

❖ Oriented links: Citation, acquisition, action, inclusion, etc.

- Generation of relational base: we have integrated a tool that permits to generate database information automatically, with which the user can select different granularities and variables and the intersection matrix.

These five steps are covered by our system.

2.3 Requesting

The interrogation can be made in two approaches:

Either by using a predefined query that represents a synthesis that shows how to perform strategic uses of relational databases.

Or, by formulating user queries with the help of a query formulation tool.

2.4 Reporting

The "reporting" functions are essential to accomplish successful production presentation in BI and to convince the decision makers by a readable, relevant and concise document. In addition to the great classical graphics (histograms 2D and 3D, boxplot, pie charts, straight regression line,…), we intend to integrate specific visualization techniques to each type of request like (evolutionary 2D and 3D histograms, comparative or cumulative 2D and 3D histograms, weighting by using external data, geographical charts, relational graphs, classifications, transitivity,…). This group of possibilities should offer for each decision maker better visualization to discover, and then communicate strategic information that aims to be integrated in their personalized report analysis. In figure 2, we present some examples of the graphics that our system offers.
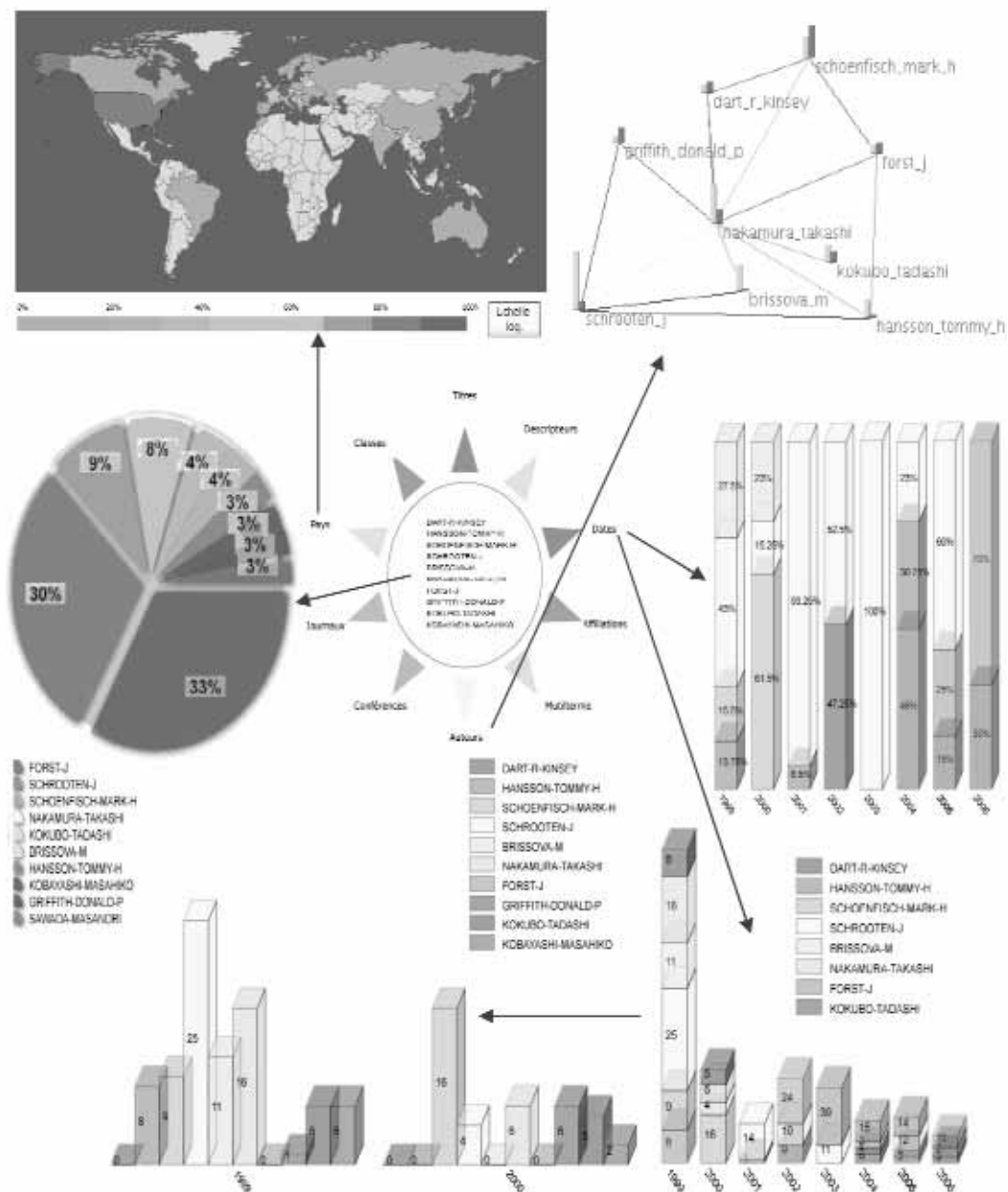

Fig. 2. Various representations of graphical outputs

Looking at the fig 2, our systems would allow the user to have more clear views and representations for the results of his analysis. For example, the star is the first result that the decision maker obtains from his analysis, and starting from this star, the user could obtain several representations according to his needs.

Here the star represents the analysis theme (authors) that is extracted from a relational database, and shows all the elements related to them in function of date (publications, authors, journal, country, affiliation).

Our system will offer the user different possible graphical outputs that would meet each kind of user needs in order to have more adapted results which the user can interpret in the best way.

In our example, we have presented a transitivity graph that shows the relation between an author and another author. The knowledge extracted in this graphical representation is the group of collaboration between authors. For each author, the user can see all these collaborations either in a precise period or in general.

Conclusion

We have presented, in this article a new tool that could cover all available sources of electronic information, such as: documentary databases that exists on-line or in a Cd Rom (like Medline, Inspec, Current content, Biosis, Pascal, Sci, Chemical abstract,…), the Web pages, the news groups, links to web sites, the press, the copyrights (Esp@cenet, Uspto, Derwent, Inpi,…), structured and non structured databases. Knowing that the principal source of strategic information is a relational source, the organization of the data which we propose will make it possible to users to navigate easily in their analyses using some techniques that they expertise (Internet, the descriptive statistics, filtering, and functions of reporting). As the generation of a new database, on a targeted subject, takes only a few minutes, a decision maker, even if he's a distant, could be informed very quickly on the recent evolutions of his environment and on the strategies induced by the ruptures or the updated emergent structures. Lastly, the cost of this approach can be limited: many sources are free, the expert is not necessary and the application server can be used by many users at the same time.

References

Chrisment, C., T.Dkaki, J.Mothe, B.Dousset (1997). "Extraction et synthèse de connaissances à partir de bases de données hétérogènes", Ingénierie des Systèmes d'Information, vol. 5, n°3, pp. 367-400.

Dousset, B. (2005). "Intégration des méthodes interactives de découverte de connaissances pour la veille stratégique", Mémoire d'habilitation à diriger les recherches, Université Paul Sabatier, Toulouse.

Frank S.C. Tseng, Annie Y.H. Chou (2005). "The concept of document warehousing for multi-dimensional modeling of textual-based business intelligence", National Science Council, TAIWAN.

Roux C., D. Sosson, B. Dousset (2004). XPlor : un outil d'investigation en ligne sur des données relationnelles. (VSST).

Sosson, D., M. Vassard, B. Dousset (2001). "Portail pour la navigation en ligne dans les analyses stratégiques. " Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 347-358, (Barcelone, Espagne).