Text Mining for Business Intelligence

Konstantinos Markellos
University of Patras, Greece

Penelope Markellou
University of Patras, Greece

Giorgos Mayritsakis
University of Patras, Greece

Spiros Sirmakessis
Technological Educational Institution of Messolongi and Research Academic Computer Technology Institute, Greece

Athanasios Tsakalidis
University of Patras, Greece

INTRODUCTION

Nowadays, business executives understand that timely and accurate knowledge has become crucial factor for making better and faster business decisions and providing in this way companies a competitive advantage. Especially, with the vast majority of corporate information stored as text in various databases, the need to efficiently extract actionable knowledge from these assets is growing rapidly. Existing approaches are incapable of handling the constantly increasing volumes of textual data and only a small percentage can be effectively analyzed.

Business Intelligence (BI) provides a broad set of techniques, tools and technologies that facilitate management of business knowledge, performance, and strategy through automated analytics or human-computer interaction. It unlocks the "hidden" knowledge of the data and enables companies to gain insight into better customers, markets, and business information by combing through vast quantities of data quickly, thoroughly and with sharp analytical precision.

A critical component that impacts business performance relates to the evaluation of competition. Measurement and assessment of technological and scientific innovation and the production of relative indicators can provide a clear view about progress. Information related to those activities is usually stored to large databases and can be distinguished in: research information stored in publications or scientific magazines and development-production information stored in patents.

Patents are closely related to *Technology Watch*, the activity of surveying the development of new technologies, of new products, of tendencies of technology as well as measuring their impact on actual technologies, organizations or people. Statistical exploitation of patent data may lead to useful conclusions about technological development, trends or innovation (Chappelier et al., 2002).

Traditional methods of extracting knowledge from patent databases are based on manual analysis carried out by experts. Nowadays, these methods are impractical as patent databases grow exponentially. *Text Mining (TM)* therefore corresponds to the extension of the more traditional Data Mining approach to unstructured textual data and is primarily concerned with the extraction of information implicitly contained in collections of documents. The use of automatic analysis techniques allows us to valorize in a more efficient way the potential wealth of information that the textual databases represent (Hotho et al., 2005).

This article describes a methodological approach and an implemented system that combines efficient TM techniques and tools. The BI platform enables users to access, query, analyze, and report the patents. Moreover, future trends and challenges are illustrated and some new research that we are pursuing to enhance the approach are discussed.

BACKGROUND

Patents are closely related to technological and scientific activities (Narin, 1995). They give an indication of the structure and evolution of innovative activities in countries, regions or industries. In this framework, patents are linked to Research and Development (R&D) and can be considered as indicators of R&D activities (Schmoch et al. 1998).

A patent is a legal title granting its holder the exclusive right to make use of an invention for a limited area and time by stopping others from, amongst other things, making, using or selling it without authorization (EPO, 2006). The patent applicant has to provide a detailed technical description of its invention but also mention the points that render it an original application with innovative elements.

A patent can be decomposed and described by several fields (table 1). Each field contains specific information while each patent is described by a code (or in many cases more than one codes) depicting its technical characteristics. These codes are given to patents based on the International Patents Classification system (IPC) or other classification systems. We should also mention that patent documents can be either retrieved from on-line patent databases, or patent databases available on CD-ROMs.

Tools from various vendors provide the user with a query and analysis front-end to the patent data. Some of these tools perform only simple analysis and produce tables, charts or reports e.g. PatentLab II (http://www.wisdomain.com/download.htm), BizInt Smart Charts for Patents 3.0 (http://www.bizcharts.com/patents/index.html), MapOut Pro (http://www.mapout.se/ MapOut.html), etc. Other tools demonstrate enhanced capabilities by using advanced TM techniques e.g. Management and Analysis of Patent Information Text or MAPIT (http://www.mnis.com/mpt.html), VantagePoint (http://www.thevantagepoint.com), Aureka (http://www.micropat.com/static/aureka.htm), Technology Opportunities Analysis or TOA (http://www.tpac.gatech.edu/toa.php), etc.

Table 1. Patent fields in the ESPACE ACCESS database

PN	Priority Number (number of the patent).
AN	Application Number.
PR	Priority Year.
DS	Designated States.
MC	Main Classification Codes.
IC	All Classification.
ET	English Title.
FT	French Title.
IN	Inventor.
PA	Applicant (name of the company depositor).
AB	English Abstract.
AF	French Abstract.

A BUSINESS INTELLIGENCE PLATFORM FOR PATENT MINING

Research and development investment in knowledge discovery and management technologies has made significant progress. However, there still exists a need for an approach that combines efficient and innovative tools for the analysis of patent data, which will guide users (e.g. R&D planners, business analysts, patents analysts, national and international patent offices, economic organizations, national statistical offices, venture capitalists, industrial bodies, etc.) to extract only necessary information and exploit it in an informative way in order to draw useful conclusions.

Our platform was designed to fill this need by quickly analyzing large collections of patents, utilizing multiple algorithms and visualizations, and producing indicators concerning the scientific and technological progress (Markellos et al., 2003). These indicators provide a global understanding of the patent collection and help users to make conclusions about on-going changes and their effects. The methodological approach is depicted in figure 1.

Patents Preparation

The system enables data importing through an easy-to-use dialog box. After downloading a data file in .txt format from MIMOSA search engine a new project to work with can be created. In this step, to reduce the patents representation for efficiency of computation and scalability purposes, while maintaining the maximum of information, several techniques are used. We browse the patent records, read their contents, modify