Chapter 18

Information search and classification to foster innovation in SMEs The AREA Science Park experience

F. Neri
Lexical Systems Laboratory,
SYNTHEMA S.r.l., Italy.

Abstract

The AREA Science Park is the first and largest science park in Italy, with more than 70 public institutions, industrial laboratories, and research institutes, as well as 1,700 working people. Its activities are managed and promoted by the *Consorzio per l'AREA di Ricerca*, the public institution specifically designed to foster its development. This consortium has recently contributed to the social and economic development of Friuli Venezia Giulia, in the North-East of Italy. In fact, it has paid an increasing attention to the creation of a network of firms and local universities, supporting them in their research activities, and facilitating the propagation of technological innovations. To achieve its goals, the *Consorzio per l'AREA di Ricerca* has created a Technology Transfer Division (TTD), aiming to promote information and innovation awareness among the local small and medium enterprises (SMEs). Nowadays, these companies are able to monitor technological trends and discover their competitors' strategies by the TTD Search Engine for patents, based on TEMIS Online Miner Light. The SMEs can access this service by a simple Internet browser, search for documents, and then classify the results according to the terminologies they share. Classification of documents is performed according to the Unsupervised and the Hierarchical classification methods, allowing to explore the detected thematic groups and subdivide them into more specific themes. Thus, users can access documents by their topics and have an overview of their contents. A successful way of cutting through the information labyrinth.

WIT Transactions on State of the Art in Science and Engineering, Vol 17, © 2005 WIT Press
www.witpress.com, ISSN 1755-8336 (on-line)
doi:10.2495/978-1-85312-995-7/18

1 The AREA Science Park and its technology transfer division

The AREA Science Park is the first and largest science park in Italy, with more than 70 public institutions, industrial laboratories and research institutes. Today, this science park can count on a wealth of human resources, which add up to more than 1,700 people, mainly devoted to research and development activities.

The AREA Science Park initially focused its attention on hosting highly-qualified research centres and promoting the creation of two important international projects: UNIDO's Centre for Genetic Engineering and Biotechnologies and ELETTRA Synchrotron Light Laboratory. Its activities are managed and promoted by the *Consorzio per l'AREA di Ricerca*, the public institution specifically designed to foster its development. This consortium has recently contributed to the social and economic development of Friuli Venezia Giulia, a North-East Italian region. In fact, it has paid increasing attention to the removal of obstacles between research and industry, encouraging the creation of a network of firms and local universities, supporting them in their research activities, and facilitating the propagation of technological innovations. To achieve its goals, the *Consorzio per l'AREA di Ricerca* has created a Technology Transfer Division (TTD) and a PatLib centre, aiming to promote information and innovation awareness among the SMEs in the North-East of Italy.

Even though the Friuli Venezia Giulia region can rely upon an extraordinary scientific wealth, there is still a wide gap between research and industry: the unstructured support offered by research often falls short of industrial expectations in terms of innovation. In order to fill this gap, which jeopardizes industrial development in this territory, the TTD has adopted a working plan with three main objectives:

- advocating and encouraging integration between research and industry;
- promoting the industrial use of research results;
- increasing the technological development of enterprises to support their competitiveness.

The working plan of TTD aims at very practical actions, based upon three main strategic choices:

- Dynamic actions are taken on the territory to have a direct knowledge of enterprises and their needs.
- A network of skills and services has been set up, relying upon a network of contacts with institutions, research institutes and consulting organisations, to ensure the immediate identification of skills and know-how according to the specific needs of each enterprise.
- A package of on-line services has been activated to encourage the use of Internet to search for information, keep updated and discover new solutions to increase business competitiveness.

The last key action has pushed TTD to activate a new advanced Web service. Nowadays, the SMEs are able to monitor technological trends and discover their competitors' strategies through the TTD Search Engine for patents, based on TEMIS Online Miner Light.

2 TEMIS online miner light, the TTD search engine for patents (TTDSE)

TEMIS Online Miner Light is a complete Internet/intranet application based on SYNTHEMA and TEMIS technologies. It collects and analyzes large sets of data according to morphological and statistical criteria, extracting all patent-relevant terms from documents and indexing them by their most significant terms and phrases. In fact, each patent is identified by specific codes that describe its application areas, its inventor and similar data, as well as by other free textual fields, which are rarely used for classification purposes. The alphanumerical codes are always partially overlapping and redundant, the free textual fields contain instead the true valuable information. Thus, it is not easy, even for an experienced researcher, to recognize the importance of a patent and its relationship with other patents, especially when the corpus consists of hundreds of documents. Text Mining techniques allow to overcome these difficulties, as they classify all documents by *understanding* automatically, quickly and easily the free textual information included in the most relevant fields, that are the abstract, the description, the claims or any other free textual field.

2.1 Data selection

All patents and technical publications, analyzed by the AREA staff, are extracted from Internet or intranet databanks, file repositories or databases, by running simple queries. The selected bibliographic references normally belong to a corpus of documents dealing with a clearly defined subject.

2.2 TTDSE back-end: the knowledge extractor

To extract the most relevant information from corpora, the AREA staff use the Knowledge Extractor, a multi-lingual lexical parser which supports English, French, German, Italian, Spanish, Portuguese and Brazilian. The lexical engine extracts only specific words or phrases from free textual fields, according to Parsing rules, Grammar rules [1–6] and Statistics [7]. It recognizes as relevant terminology only those terms or phrases that comply with a pre-defined set of morphological patterns and whose frequency exceeds a threshold of significance [7]. In fact, a specific algorythm associates an Information Quotient to each detected term and ranks it on its importance. The Information Quotient is calculated taking into account the term, its *Part Of Speech* tag, its relative and absolute frequency, its distribution on documents. The resulting terms are then reduced to their *Part Of Speech* tagged base form and then used as descriptors for documents. Indexation based on terminology detection is extremely reliable

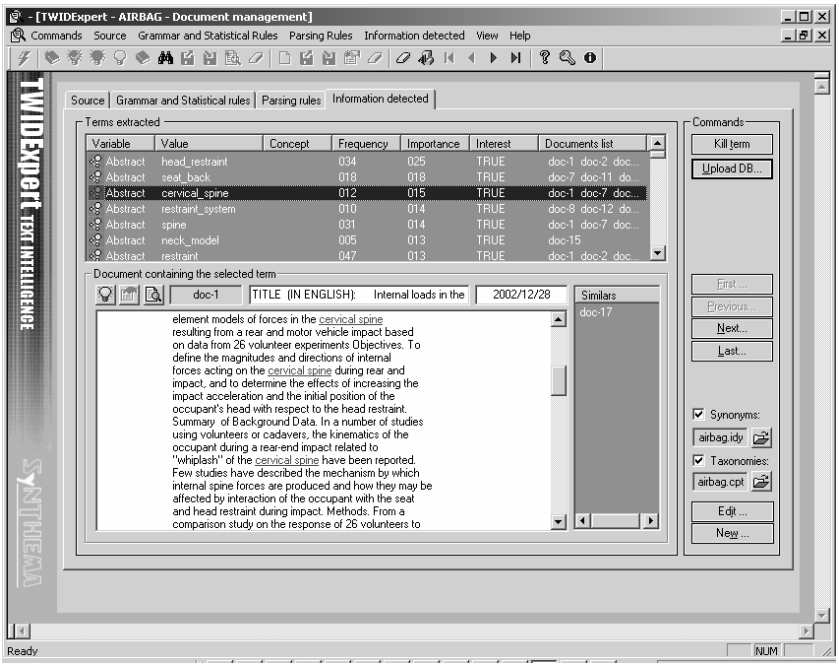

Figure 1: The Knowledge Extractor (information extracted).

for managing any type of documentation, expecially if it is technical and scientific. In fact, unfortunately, few of us have complete knowledge about the world. And, in the consequence of this, the meanings we ascribe to words may differ from those ascribed by others. The same happens with lexical tools capable of syntactic parsing, which have always a limited capability of semantic interpretation and disambiguation, if applied to generic corpora. In such situations, these tools cannot pick out the exact interpretation for all expressions in the language. Besides, main terminology, mostly compound nouns, helps to "understand" the topic, being intrinsically linked to semantics.

2.3 TTDSE front-end: the advanced search engine

Once documents are indexed according to the terminology [6–8] included in them, they can be stored into a database, searched for, accessed and classified into thematic groups.

The Advanced Search Engine allows to perform document searches, based on patents-like/semi-structured document fields stored in the application database. Once a search is completed, the application allows to:

- navigate query results by concepts,
- classify results according to the concepts they share.

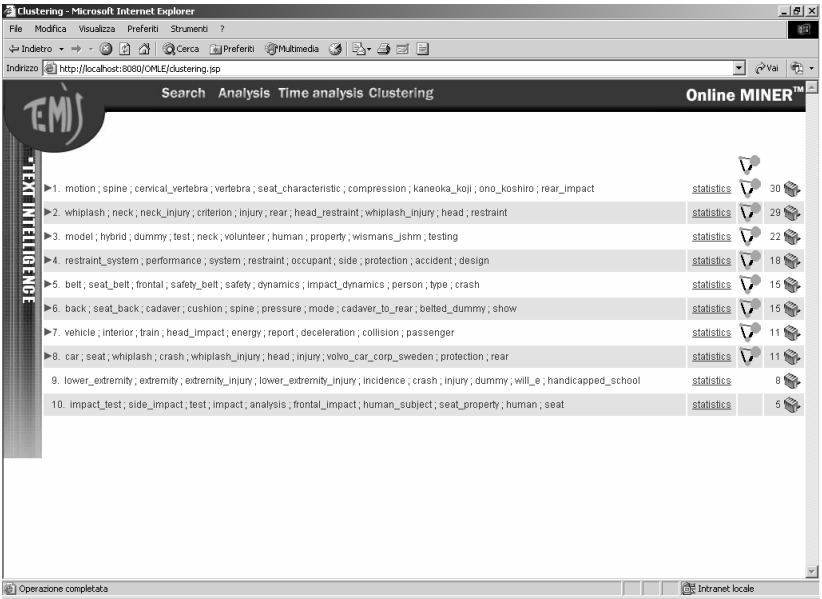

Figure 2: The front-end user interface, classification and navigation by themes.

The classification of documents fulfils the following requirements:

- Unsupervised Classification. The application dynamically discovers the thematic groups that best describe the detected documents, according to the concepts they share. This new approach allows users to access documents by topics, not by keywords.
- Hierarchical Classification. The application makes it possible to explore in depth the thematic groups, subdividing them into more specific themes.

The application provides a visual summary of the analysis. A map shows the different groups as differently sized bubbles (the size depends on the number of documents the bubble contains) and the meaningful correlation among them as lines drawn with different thickness (that is level of correlation). Users can search inside topics and have a look of the documents populating the clusters. The output results can be viewed by a simple Web browser.

3 TTD results

Innovation on the Internet is the Web site conceived by TTD to promote innovation deployment among companies and support their growth and competitiveness. All on-line services are meant to meet any enterprise need for up-to-date information and material by taking full advantage of the Internet potential. The new advanced services are expected to have positive effects on the community in terms of both turnover and employment, as shown in fig. 4.

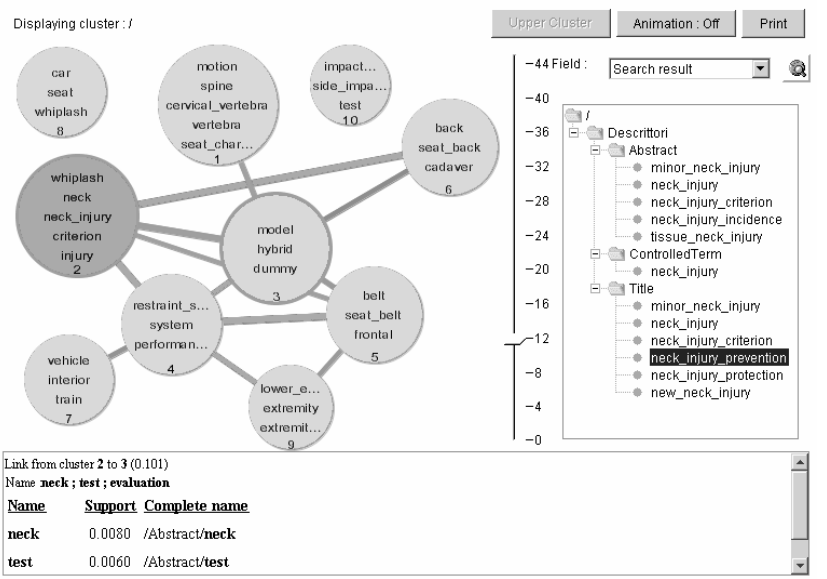

Figure 3: The front-end user interface, the thematic network.

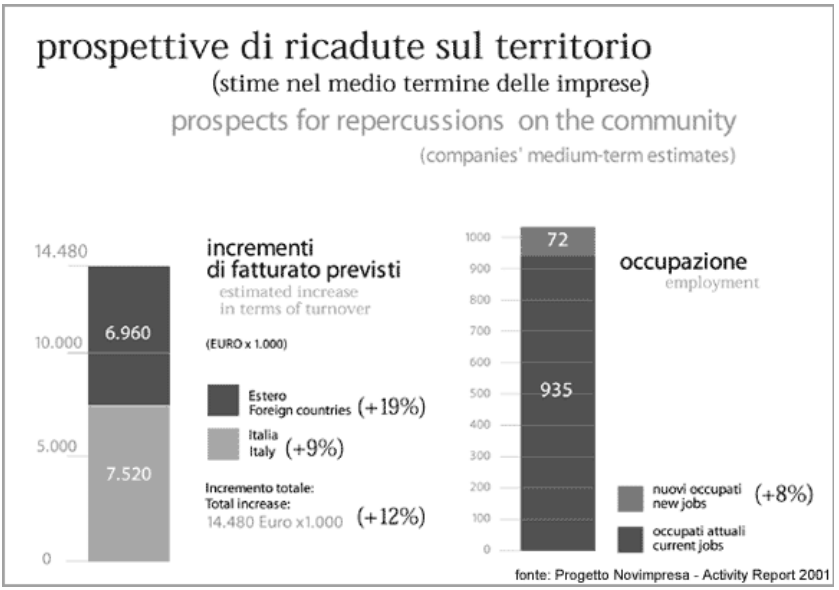

Figure 4: TTD activity and its repercussions on the community.

The TTD Search Engine for patents, based on TEMIS Online Miner Light, provides an overview of textual corpora content, giving an intuitive grid to users and helping them discover hidden and meaningful similarities among documents. It enables the research, analysis, classification and – as a result – the discovery of any information contained and encoded in great volumes of textual data.

References

[1] Raffaelli, R., An inverse parallel parser using multi-layerd grammars, IBM Technical Disclosure Bullettin, 2Q, 1992.

[2] Raffaelli, R., Un ambiente per lo sviluppo di grammatiche basato su un parser inverso, parallelo e seriale, IBM Italy Scientific Centers Technical Report, 1992.

[3] Marinai, E. & Raffaelli, R., The design and architecture of a lexical data base system, COLING'90, Workshop on advanced tools for Natural Language Processing, Helsinki, Sweden, Aug 1990, 24.

[4] Raffaelli, R., ABCD – A Basic Computer Dictionary, Proceedings of ELS Conference on Computational Linguistics, Kolbotn, Norway, Aug 1988, pp. 30–31.

[5] Galli, G., Raffaelli R. & Saviozzi, G., Il trattamento delle espressioni composte nel trattamento del linguaggio naturale. IBM Research Center, internal report, Pisa, Italy, 1992.

[6] Cascini, G. & Neri, F., Natural Language Processing for Patents Analysis and Classification, ETRIA World Conference, TRIZ Future 2004, Proceedings, Florence, Italy, November 2004, pp. 3–5.

[7] Neri F. & Raffaelli R., Text Mining applied to Multilingual Corpora, NEMIS Final Conference, Network of Excellence in Text Mining and its Applications in Statistics, Proceedings, Athens, Greece, Springer Verlag Ed., October 2004, 25.

[8] Elia, A. & Vietri, S., Electronic dictionaries and linguistic analysis of Italian large corpora. *JADT 2000, 5th International Conference on the Statistical Analysis of Textual Data*, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2000.