Computational system for the textual processing of industrial patents

G. M. Caputo & N. F. F. Ebecken
COPPE – Federal University of Rio de Janeiro, Brazil

Abstract

This work presents a study related to the application of text mining methods in Brazilian industrial patents (INPI databases) as a tool of competitive profit for companies supplying technology. The main objective is to discover, using patents abstracts analysis, the main developers' companies in all technological areas, helping in the competitors' identification. The patent mining is capable of discovering, through the research aiming perception, new industrial trends helping to create strategies that anticipate the demand and to get a competitive profit through technological innovations.
Keywords: text mining, patent analyses, competitive intelligence.

1 Introduction

Innovation is perhaps one of the main subjects interest for the majority companies that desire to keep updated in the current market. To innovate demands to know the competitions strategies, main producing and the consumers' real necessities, the technological trends knowledge that anticipates the demand and the market [1]. To reach such level, Competitive Intelligence (CI) techniques are used. CI provides organizations evolution and competitive advantage utilizing computational techniques and qualified professional. These techniques can understand involved sectors and assist in decision making.

From this dispute for innovations and organization copyright protection, constantly, new products are invented, new ideas appear and so that the industrial property law on the product or idea is preserved, new patents are deposited.

These industrial patents have the products details, so an analysis detailed on these patents, make possible visualize the technological trends and understand

the competing organization intellectual environment and from the results, get a competitive profit through innovations.

Patents documents are a large technician and commercial knowledge resource in technician progress term, market trends and copyright. These documents analyze is a strategically advantage study subject. It provides technological prospection, planning, management, programs formularization and evaluation and an important R&D vehicle for institutions.

Considering that patents information great are in literal form, text mining application consequently has great utility for knowledge better understanding included in these documents and provide resources to Competitive Intelligence application.

However, this technique implementation for patents database is special standards dependent, differently of other documents, because it demands not to lose the referring information to particularities field in patents documents, as for example, the fields TITLE and ABSTRACT. Therefore, it was studied patents manipulation techniques utilizing text mining that promoted better database exploitation, and whose results provide great quality.

2 Brazilian industrial patens

In Brazil, the National Institute of Industrial Property (INPI) [2] is responsible for prescribed and protecting the marks and patents copyrights right. These documents have an international standard format, and classified in accordance with Patents International Classification (CIP), that allows the fast recovery for technological area, through indexation.

Patents can be classified in eight main sectors:

Table 1:   Patents classifications sections.

Section A	Human Necessities
Section B	Process Operations; Transport
Section C	Chemistry and Metallurgy
Section D	Textile and Paper
Section E	Fixed constructions
Section F	Mechanics Engineering /Illumination/Heating
Section G	Physics
Section H	Electricity

The complete classification symbol for a specific technique is constituted by symbols represented by: Section (as previously), Class (composed number for two numbers), Subclass (capital letter), group and Sub-group, as for example:

A Section
01 Class
B Subclass
1/00 Main Group or 1/24 Sub-group

With this classification, patents easily are located such as the access to one determined technology type.

It is extremely important for any country the availability of an agency for patent filed management, guaranteeing the inventor right, therefore this protection encloses particular business, organizations and others. The increase patents number can directly influence the economy, and through innovation, technology evolves and increases commercialized products amount and innovations for exportation, adding value to the products.

2.1 Patents and competitive intelligence

Many subject trends can be identified through patent frequency analysis. For example, necessities, concerns, fashions and customs. This can be taken analyzing assigned patents in a period [3, 4].

In this point, one can observe that patents documents volume increment a classification indicates a technological trend, or either, research aiming, the world-wide level, and/or posterior consequences in market. This valuable source of technological information became one thus important strategically essential for companies that desire to keep updated on its sector technological development, and making possible the level technician rise, because in patents is contained as scientific concepts as process details.

Moreover, utilizing patents analysis it is possible to identify which effectively are its main competing and understand in which technological areas and which products these intend to invest. It can provide information on strategic plans, financial risk projection, competitors' plans and products, products and innovations values real conscience and market strategy aspects identification. This all helps a company when entering in a highly competitive market

In general, this analysis aims extract knowledge processing data under different criteria, depending on the information type that one wants to have access.

This process can be executed from some perspectives [5], as for example, analysis statistics, citation analysis [6, 7], data mining [8] and text mining.

Some studies are dedicated to reach this intention, as can be verified in [9, 10], where are applied bibliometric analysis in carbon patents nanotubs to measure the activity growth of nanotubs carbon industry and its links with science.

Patents analysis has been greatly used in new technological trends identification and new necessities forecast, in some industrial sectors, as druggist, prolific and services rendering, therefore anticipating new products and taking in consideration industrial development along the years, it is possible to make a technology forecast and discover how it gradual evolves.

3   Text mining

Text mining is a process that involves documents clustering based on their concept and the used data are not structure [11]. In Brazilian patents case,

available in INPI, fields that have such format are the TITLE and the ABSTRACT. For these fields, the KDT phases (Knowledge Discovery from Text) had been applied, that includes patents pre-processing and pos-processing [12].

Patent data pre-processing demands a text detailed preparation to be used [13], where meanings similarities and terms relevance are detached, and the terms dictionary [14] can drastically be reduced. Therefore, a Stopwords list was used contends term without semantics relevance, as conjunctions and articles.

Moreover, a list containing less frequent terms and most frequent, (that did not have a discriminatory degree in contents), had been removed of the clustering process. Patents' pre-processing also includes Stemming application [15], adapted for Portuguese language, to reduce morphologic variations to a same radical.

A patents text mining tool was developed to support its diverse fields, as the TITLE, ABSTRACT, INTERNATIONAL CLASSIFICATION and ASSIGNEE.

The tool is capable to cluster only selected patents ABSTRACT, list grouped documents and show the INTERNATIONAL CLASSIFICATION and the ASSIGNEE that more appear in grouped patents.

4 Case study

For the current study, an industrial patent set had been collected from INPI's site. The database has patents that have the term "petroleum" in ABSTRACT patent field. Those patents that had been deposited between the years 2001 and 2005 had been chosen to compose the database because it presents the most current innovations. In total, 319 patents documents had been found with such specifications.

Analyzing patents INTERNATIONAL CLASSIFICATIONS that appear more frequently, indicating bigger development in areas:

- E21B: Fixed Constructions. Earth or rock drilling; Mining**.** Obtaining oil, gas, water, soluble or meltable materials or a slurry of minerals from wells.
- C10G: Chemistry; metallurgy**.** Petroleum, gas or coke industries; technical gases containing carbon monoxide; fuels; lubricants; peat. Cracking hydrocarbon oils; production of liquid hydrocarbon mixtures, e.g. by destructive hydrogenation, oligomerisation, polymerisation. Recovery of hydrocarbon oils from oil-shale, oil-sand, or gases; refining mixtures mainly consisting of hydrocarbons; reforming of naphtha; mineral waxes.
- C10L: Chemistry; metallurgy. Petroleum, gas or coke industries; technical gases containing carbon monoxide; fuels; lubricants; peat. Fuels not otherwise provided for; natural gas; synthetic natural gas obtained by processes not covered by subclasses c10g, c10k; liquefied petroleum gas; adding materials to fuels or fires to reduce smoke or undesirable deposits or to facilitate soot removal; fire-lighters

The main sub-group appeared between INTERNATIONAL CLASSIFICATION was E21B 37/06:
- E21B 37/06: Methods or apparatus for cleaning boreholes or wells using chemical means for preventing or limiting the deposition of paraffins or like substances

Amongst selected patents, the company that have more assigned industrial patents is "Petróleo Brasileiro S.A. - Petrobras (BR/RJ)" with 26 assigned patents and a partnership with "Tecnologia LTDA (BR/RJ)".

Although patents have various fields, only "ABSTRACT" had been used in text mining. Seven clusters were selected from the database. Table 2 presents the obtained results.

Table 2:   Main keywords and number of documents.

Cluster	Patents Numb.	Keyword
1	49	Composition; gas; derivatives; formation; mixture; content; presence; Fischer-Tropsch; natural; water.
2	43	Feeding; fuel; catalytic production; inclusion; mixture; relationship; change; lubricant; total.
3	40	Treatment; content; liquid; chain; load; substance; pressure; gas; temperature; fiber.
4	56	Gas; system; valve; production; injection; fluid; pressure; formation; equipment; exit.
5	13	Liquid; marking; dyes; dieletro; alquila; independent; maximum; absorption; substitution; antraquinona
6	77	Water; production; system; sea; layer; floating; platform; upper; pressure; lower.
7	41	Device; system; tubing; energy; chain; electricity; structure; communication; gas; fluid.

Verifying the main keyword of each cluster superficially and analyzing the meaning and content of each one, the contents difference of the database document grouped by patents text mining system can be noticed.

Observing clusters, it can be noticed that detected contents are respectively:
- Fischer-Tropsch products and derivatives use for fuel, oil and natural gas formation and mixture.
- Fuel, chemistries substantiate rude oil catalytic product and lubricant production methods and products.
- Hydrocarbon, substances and oil liquids derived load treatment processes, through temperature and pressure.
- Storage mechanisms and GLP and fluids injection pump and retain control.
- Liquid petroleum products marking through dyes.

- Petroleum in liquid surfaces for manipulation systems and equipment.
- Devices and methods for operation and communication in oil wells, as electricity conduction and fluid injection.

As a previous clustering documents classification doesn't exists, it is not possible to get a system rightness degree relation. However, some conclusions can be taken and used to support competitive intelligence. It takes in consideration some factors as: clusters grouped documents set, keywords set, and clusters results INTERNATIONAL CLASSIFICATION and ASSIGNE.

5   Conclusions

Text mining application results in industrial patents documents obtain extreme efficiency and can provide precious information in respect the technological trends.

Such fact can disclose a change in studies focus and industrial and technological developments in petroleum derived components, techniques and use. This text mining application can be used in any intellectual content and be converted in knowledge, using any keyword or different query to take industrial patent databases.

Moreover, the study results prove the viability to execute technological trends search using available resources in industrial patents and text mining tools.

Brazilian patents, assigned in INPI, store great information source and so its database correct manipulation can supply great competitive advantage in companies and knowledge source for research.

To better discuss the presented study case an expert in the petroleum area was necessary. Perhaps much more hidden knowledge could be extracted and the interestingness and actionability of this knowledge used for competitive advantage.

Acknowledgements

References

[1]   Mogee, M. 1991. Using patent data for technology analysis and planning. *Research-Technology Management*, 34, 43-49.
[2]   INPI, http://www.inpi.gov.br
[3]   NARN, F., NOMA, E., 1987. Patents as indicators of corporate technological strength. Research Policy, 16, 143-155
[4]   Yoon, B., Park, Y, 2003, "A text-mining-based patent network: Analytical tool for high-technology trend", *The Journal of High Technology Management Research,* 15, Seoul, South Korea, 37-50, September.

[5]   Deboys, J., 2004, Decision pathways in patent searching and analysis, *World Patent Information*, 26, 83-90.
[6]   Holl, B., Jaffe, A., Trajtenberg, M., 2000, *Market value and patent citation: A first look.* NBER Working Paper Series, Cambridge, MA.
[7]   Michel, J., Bettels, B., 2001, "Patent citation analysis: A closer look at the basic input data from patent search report", *Scientometrics*, vol. 51, no. 1, pp. 185-201.
[8]   Han, J. & Kamber, M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, 2001.
[9]   Gupta, V., Pangannaya, N., 2000. Carbon Nanotubes: Bibliometric analysis of patent. *World Patent Information*, v. 22, 185-189.
[10]   Karki MMS. Bibliometric analysis of patents: implications for R&D and industry, emerging trends in scientometrics. In: Nagpul PS et al., editor. New Delhi: Allied Publishers, 1999.
[11]   ZANASI, A., Text Mining: the new competitive intelligence frontier. *In VST2001 Barcelona Conference Proceedings – IRIT*, Spain, 2001.
[12]   Fattori M., Pedrazzi G., Turra R., 2003, Text Mining applied to patent mapping: a practical business case.
[13]   Zanasi, A., 2005, Text Mining and its Applications to Intelligence, CRM and Knowledge Management. 1 ed. Great Britain, WIT Press.
[14]   Salton, G., Automatic Text Processing, Addison-Wesley, 1989.
[15]   Porter, M.F., *An algorithm for suffix stripping*, Program, vol. 14, pp.130-137, 1980.