KNOWLEDGE EXTRACTION ON INTERNATIONAL MARKETS FROM PATENT BASES: A STUDY ON GREEN PATENTS

Wagner Vianna Bretas
wbretas@gmail.com
Instituto Federal de Educação,
Ciência e Tecnologia Fluminense -
IF Fluminense

Alline Sardinha Cordeiro Morais
amorais@iff.edu.br
Instituto Federal de Educação,
Ciência e Tecnologia Fluminense -
IF Fluminense

Henrique Rego Monteiro da Hora
dahora@gmail.com
Instituto Federal de Educação,
Ciência e Tecnologia Fluminense -
IF Fluminense

Edson Terra Azevedo Filho
edsonterrafilho@gmail.com
Universidade Estadual do Norte
Fluminense

ABSTRACT

Goal: This article aims to propose a model for stratifying technological information from meta-data contained in international patent bases, capable of supporting the strategic decision making that potentiates actions directed to foreign trade.

Design / Methodology / Approach: This applied research was based on the KDD - Knowledge Discovery in Databases methodology and carried out a study focused on green patents. Patent bibliographic data published in the Patent Cooperation Treaty (PCT) from 2003 to 2012, focusing on alternative energies, more precisely on biofuels, were obtained from the Derwent database, with the search string based on the Green Patents IPC Inventory, published by the World Intellectual Property Organization (WIPO). After treatment and sanitization, more than 36,000 resulting records were performed under C4.5 algorithm, denominated J-48 from the software Weka, resulting in Brazil as the destination country.

Results: A decision tree was established, in which Mexico was highlighted as the main discretionary country. It was also verified the adhesion of the other emerging countries, which, along with Brazil, compose the BRICS.

Limitations of the investigation: The proposed model is limited to areas that show intensive use of technology in products and processes.

Practical implications: It could be inferred that the proposed method can help companies to identify international markets more sensitive to a certain technology, from a free database, reliable and capable of being used by micro and small companies.

Originality / Value: In scientific communication, it is not easy to find Data mining applied to Patent database, and in this study, BRICS cluster were identified in Green patents WIPO deposit.

Keywords: Intellectual Property; Sustainable Development; Decision Support Systems; Foreign Trade; Knowledge Discovery in Databases.

1. INTRODUCTION

The business strategies that use decision support systems grow every day. In this respect, the concept of Business Intelligence – BI – or Competitive Intelligence refers to the process of transforming data into information, from which knowledge is extracted, and applied to decision making. According to Chau and Xu (2012), the growing popularity of Web 2.0 has led to the exponential growth of user-generated content, both in volume and in meaning. The challenge is to search for the right information, whether in volume, accuracy, cost of procurement and time adequacy. According to Chen et al. (2012), business intelligence and analysis have emerged as an important area of study for professionals and researchers, reflecting the magnitude and impact of data problems being solved in contemporary business organizations.

However, given the complexity of conducting business studies at an international level, the option to seek secondary data contained in free and accessible public databases is an interesting alternative. There is evidence of the importance of developing a consistent business intelligence methodology, which contributes to making decisions regarding the expansion of markets to other countries, the expansion of technological R&D frontiers, and the increase of global productive chains, which is accessible and customizable, regardless of the size of the business that will use it.

Among the existing databases, the patent base has been the subject of many studies to analyze the evolution dynamics of technology. Frietsch and Schmoch (2010) point out that the proliferation of patent-based studies can be observed in recent years, but that increasing internationalization and globalization also require an adaptation of patent analyzes to this new world order. By the refined organization and international standardization of information contained in patents, patent bases constitute more than a document repository for mere priority verification on the examination of the merits of a new technology that one wishes to protect. The set of information contained therein, if well explored by scientometric tools, constitutes a relevant tool for the strategic management of companies. In this regard, Goldschmidt and Passos (2015) point out that "The value of stored data is typically linked to the ability to extract higher-level knowledge from them."

The interest in protecting an invention beyond the territorial boundaries of the country in which the R&D was given can be interpreted as indicating the interest in the international exploitation of that technology, as well as its higher economic value, according to Leydesdorff (2008). It is by the Patent Cooperation Treaty (PCT), of which 152 countries are currently signatories, that the original deposit (unionist priority) can enter the national phase for its potential protection in each of the countries that have been nominated as a destination. The set of protections in the target countries (Designated Offices), along with the unionist priority, after the period required for merit procedure and evaluation, compose the Patent Family.

The database used in this work is composed of a subset of the patent applications published in the PCT with their respective families. To carry out a case study, it was decided to cut green patents in the area of alternative energies, focusing on biofuels. This decision was based on the results found by Bretas et al. (2018), which, based on the Inventory IPC Patent Green, WIPO (2017) - World Intellectual Property Organization, found that this was then the field of environmentally friendly technologies (Environmentally Sound Technologies - EST) with more applications published under the PCT.

Breitzman and Mogee (2002) discuss different business situations in which the use of patent analysis is appropriate. The authors discuss techniques for strategically managing the portfolio of a company's patents, evaluating the technologies it develops, identifying companies interested in acquiring licenses for these technologies, as well as opportunities for cross-licensing or even patent donations for universities for tax deduction.

Shih et al. (2010) use patent citation analysis and patent families for their R&D management, identifying core technological competencies and assessing more influential international players in a specific technology area. They identify key inventors in competing companies with a view to attracting them. They perceive opportunities for mergers and acquisitions with companies with technological competence complementary to theirs. They promote a valuation of companies based on the impact of their patents. Finally, they conclude that the combination of different patent analyzes based on co-citations and patent families would lead to strategic, tactical and point-of-competitive actions of competitive intelligence, capable of answering questions related to competitors and future technological scenarios.

Similarly, Liu and Shyu (1997) analyzed the patent bases and developed a unique technology enhancement scheme, which worked not only as a roadmap but also as a guide to strategic planning and forecasting technological trends, supporting future decisions.

In the same lines, Lee et al. (2009) proposed the use of patent data to evaluate business opportunities, based on technological capabilities of companies, categorizing such opportunities in monitoring, collaboration, diversification and benchmarking. Thus, it became possible to perceive the

trends that the development of innovations is taking, aiding the direction of investments in technologies denominated future carriers.

The work of Liu and Shyu (1997), Lee, and Yoon and Park (2009) advance the exploration of the technological information present in the patent bases in a structured way, proposing models of exploitation of their data. Despite these advances, it is still possible to note the lack of tools and methodologies that support decisions for managers that do not focus on technology but the business itself. In addition, all work focuses on strategies that are especially applicable to large companies that conduct R&D and already have a patent portfolio. The central focus of the work has been the development of technological Roadmaps, identifying protagonists, technological trends, key inventors and a detailed view of competitors' performance.

The model proposed here is intended to enable the manager to extract market information from a highly technological base. It is precisely in this gap that the efforts of this work are based. Because it is sought to associate a set of different techniques to solve a problem, one can consider that its core is the development of its own methodology, with a demonstration of its application by means of a case study. Therefore, special attention was given to the details of the now developed technical procedures.

Thus, it is possible to resort to a structured, reliable, world-wide, free access data source to see promising markets for its products, promising products for their market, and business partnerships for export and import.

Silveira et al. (2018) understand that patenting tasks can induce or stimulate industries in the some sectors, to exploit the technical knowledge contained in patents, obtained by third parties as a valuable source of technological information and low cost, and which is capable of feeding a company on its own new products and processes research and development.

All of this, regardless of the business focus, is technology development that is aimed to market without the need for market research with primary data collection, without resorting to technologists and, especially, being applicable to any business size.

Therefore, the objective of this work is to propose a model to stratify technological information from meta-data contained in international patent bases, capable of supporting the strategic decision making that potentiates actions directed at foreign trade, whether they are export or import, as well as the internationalization of Research, Development and Innovation (RD&I) activities.

2. RELATED WORK

Yan and Luo (2017) propose to compare network maps of technological fields, created from patent analyzes, observing the differences and similarities in the structural properties of these maps. In order to identify the best techniques to explain the distance measures between different classes of patents, they concluded that the best maps are based on standardized likelihood measures and inventor diversification.

Using specific tools, Uhm et al. (2017) propose a method for forecasting technology from text mining techniques on patent bases, using the R data language and the interval estimation method, which they call IEM.

Ajay et al. (2015) present the Intelligent Patent Analysis Tool (IPAT) free software tool. This tool, based on user-defined parameters, retrieves public patent available data by Google Patent Search, and presents the top fifteen results in an Excel spreadsheet. Its proposal is to contribute to the process of technology evaluation, players monitoring, and change trends understanding.

Tekić et al. (2015) describe the Patent Search and Analysis for Landscaping and Management (PSALM), which is a software tool developed for competitive intelligence based on patent data. This tool collects and analyzes patents bibliographic parameters and performs text mining and clustering from patents deposited on USPTO.

Milanez et al. (2017) propose a method for the development of patent indicators based on text mining applied to patent claims, stating that such a method can contribute significantly to the technological prediction analytical process, monitoring processes, and competitive intelligence studies, by using more accurate and reliable key terms than those used in titles and abstracts.

Despite the evolution in approaches to discover trends, Seo et al. (2016) criticize the identification of opportunities for innovation that rely on the analysis of generic technology trends, without considering whether such opportunities are feasible for a target company. Thus, they proposed a systematic approach to identify viable opportunities, depending on the internal capacity of a particular company.

Jun et al. (2018) sought to dissect a particular technology in interdependent technological clusters. For this, they performed a multivariate multiple regression modeling.

Observing companies from the potential investors' viewpoint, Motta et al. (2015) present a patented-scientometric approach to support project selection processes by seed capital funds, favoring the judgment of non-financial criteria, especial-

ly those related to technology, market, divestment, and team. Using the scientific data published in the Web of Science (WoS) database and patent data in the Derwent Innovation Index (DII), they evaluated these non-financial criteria in a case study, applied to the most important project in the CRIATEC fund of the Brazilian National Economic and Social Development Bank (BNDES, acronym in Portuguese). They concluded that such a method can be extrapolated to support business incubator programs, eligibility to locate technology parks or to receive funding from government support programs.

3. METHODOLOGY

This research can be classified as an applied nature for decision support purposes, using a qualitative approach regarding the analyzes derived from its results, and quantitative in relation to the parameters analyzed. It is presented as a bibliographical documentary research. Finally, a study from the cut-out is presented for punctual application of the methodology developed here.

Technical Procedures

The technical procedures used refer to the knowledge discovery methodology in databases, called KDD - Knowledge Discovery in Databases, whose steps are presented in Figure 1.

Based on the KDD steps, the development of this research pervaded the steps outlined in Figure 1, whose methodological details are described below.

Data Acquisition and Data Selection

Data were obtained from the Derwent Innovation Index, maintained by WoS – Web of Science. Such a database is composed by more than 40 patent-issuing authorities and has proved to be sufficiently complete to account for the records and their necessary attributes to what is intended, including patent families.

There was a time cut, limited to a period of 10 years, including requests published under the PCT between 2003 and 2012. The year 2012 was used as a more recent cut due to the need to wait for patent families published at that time from the national publications of the destination countries, from the local depository (unionist priority). At the moment of the time cut of this study, a local deposit had a period of 12 months to enter the international phase (PCT). After 30 months (international phase), the national phases of each destination country were entered, and wait for an average time of more than 18 months for publication, according to Table 1. Currently, the international phase was reduced to 18 months, having incorporated the 12 months between the local storage and the entry into the PCT.

In order to carry out a study, it was decided to perform a technological cut contemplating a segmentation of environmentally friendly technologies (Environmentally Sound Technologies – ESTs) set by the United Nations – UN. To this end, it was employed the IPC Patent Green, created by WIPO, which were listed and categorized IPC codes (International Patent Classifications) of ESTs, and it was available at https://www.wipo.int/classifications/ipc/en/green_inventory/index.html.

By analyzing the amount of PCT green published patent applications, a new cut was carried out, listing topic and its subtopics, considering the increase of deposits on the temporal interval. It was noticed that the area of alternative energies and Subarea biofuels was the most representative, and its corresponding IPCs are compared in Table 2.
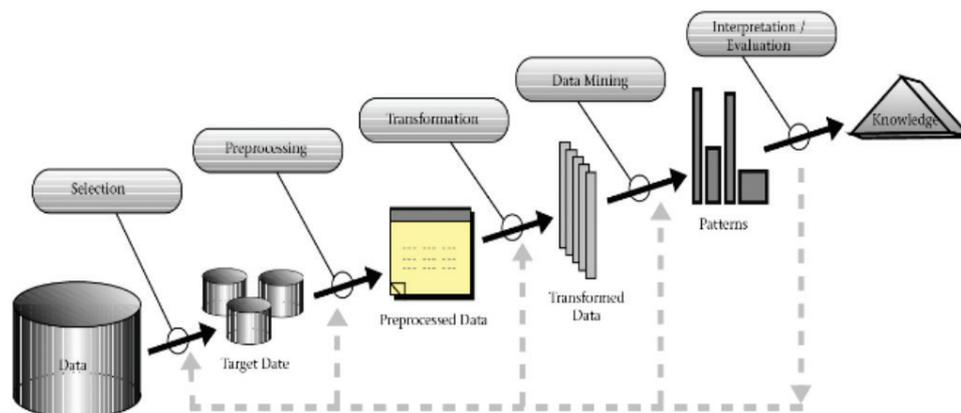

Figure 1. An Overview of the Steps That Compose the KDD Process.
Source: Adapted from Fayyad et al. (1996).

Table 1. Summary of PCT Steps

Table 2. Green Patents IPC Inventory Extract.

ALTERNATIVE ENERGIES	IPC
. Biofuels	C10L 5/00, 5/40-5/48
. . Solid Fuels	C10B 53/02, C10L 5/40, 9/00
. . . Biomass roasting	C10L 1/00, 1/02, 1/14
. . Chemicals	C10L 1/02, 1/19
. . . Vegetable Oils	C07C 67/00, 69/00, C10G, C10L 1/02, 1/19, C11C 3/10, C12P 7/64
. . . Biodiesel	C10L 1/02, 1/182, C12N 9/24, C12P 7/06-7/14
. . . Bioethanol	C02F 3/28, 11/04, C10L 3/00, C12M 1/107, C12P 5/02
. . Biogas	C12N 1/13, 1/15, 1/21, 5/10, 15/00, A01H
. . Genetically Modified Organisms	C12N 1/13, 1/15, 1/21, 5/10, 15/00, A01H

Source: WIPO (2017).

For the acquisition of raw data, the search string was applied to the Derwent base, comprising all the aforementioned indentations in order to obtain a set of 36,316 requests of PCT, as shown in Figure 2, with their patent families.
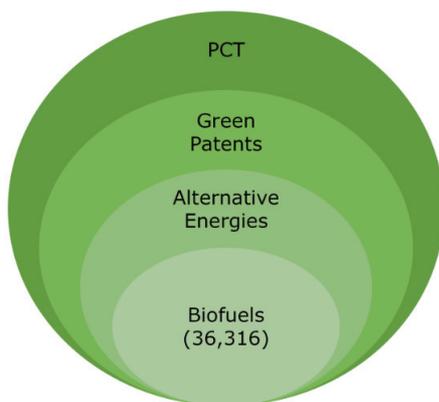
Figure 2. Data Cuts.

The complete results download records (full records) totaled an ordered list consisting of 24 attributes. Then come the data cleaning process, purging unnecessary records to the study in question.

Data processing and data cleaning

The attributes derived from the records obtained in the Derwent database contain, but not explicitly, the information necessary for conducting mining. Therefore, computational procedures were applied in order to extract from each record the parameterized information for the composition of the database for data mining - next step of the KDD. The detail of the parameterized information and its respective attribute that served as source to obtain it is associated in Table 3.

Table 3. Parameterized information and its source attributes.

Parameterized information	Source atribute
Country of origin (Unionist Priority)	PI - Prioriity Application Information and Date
Unionist priority of the PCT application	PI - Prioriity Application Information and Date
IPC complete	PD - Patent Details
ICP section	PD - Patent Details
IPC Class	PD - Patent Details
IPC subclass	PD - Patent Details
151 columns, indicating each country member of the PCT*	DS - Designated States

For the composition of the database, incomplete records that did not show any of the information listed in Table 4 were eliminated. With this, a base with 36,316 records and 47 attributes each was obtained. This is because, of the 151 PCT countries presently associated with the study temporal-cut, only 41 have been designated as destination countries of one of the patent families.

Algorithm parameters

Weka software, which implements C4.5 algorithm, needs to be parameterized to run an instance and generate a result. Table 4 shows the most relevant of them. Parameters not shown were defined as the default recommended values in software.

Table 4. Parameters applied to WEKA software.

```
=== Run information ===
Scheme:        weka.classifiers.trees.J48 -C 0.1 -M 50
Relation:      Base36316Final
Instances:     36316
Attributes:    47
Test mode:     10-fold cross-validation
=== Classifier model (full training set) ===
```

-C 0.1 - Set confidence threshold for pruning in 0.1;

-M 50 - Set minimum number of 50 instances per leaf;

Relation – Just a name for the base;

Instances –Quantity of registers, organized in lines;

Attributes - Number of features of each register;

Test mode – Validation mode. The database is split in n parts, and when one part is used to generate the decision tree, the other n-1 parts are used to test the results. It is performed n times, in a way that each n is used as to generate the model and test it.

4. RESULTS AND DISCUSSION

By applying the parameters described previously in the methodology, the following decision tree is presented in Figure 3. The aim was to associate the attributes that would adhere to the fact that one of the analyzed requests has or not Brazil as a destination for its protection.

The quality of the result of this mining is shown by confusion matrix analysis, which presented 85.16% of the correctly classified instances (precision), as presented in Table 5.

Table 5. Confusion Matrix

		True Value	
		No	Yes
Expected value	No	25,934	1,962
	Yes	3,4291	4,991

Observing the decision tree in Figure 3, it was noticed that only the attributes related to destination countries were discretionary. There was no discrimination regarding the technological areas (translated by the IPCs), nor related to the year of publication of the requests, nor was the country of origin evidenced as a point of relevance. It can be inferred that, given the narrow technological cut (biofuels), the differentiation between technologies was very subtle. Moreover, the adopted time scale (10-year cut) may have been less than or equal to the time needed for the maturation of such technologies.

Finally, because of the development in foreign trade logistics, the origin of the technologies may not have been an impact factor for their limitation to geographically close markets.

Among the destination countries, Mexico was the most discretionary country, when the outcome as a destination of the biofuel-related technologies was Brazil. Such adherence
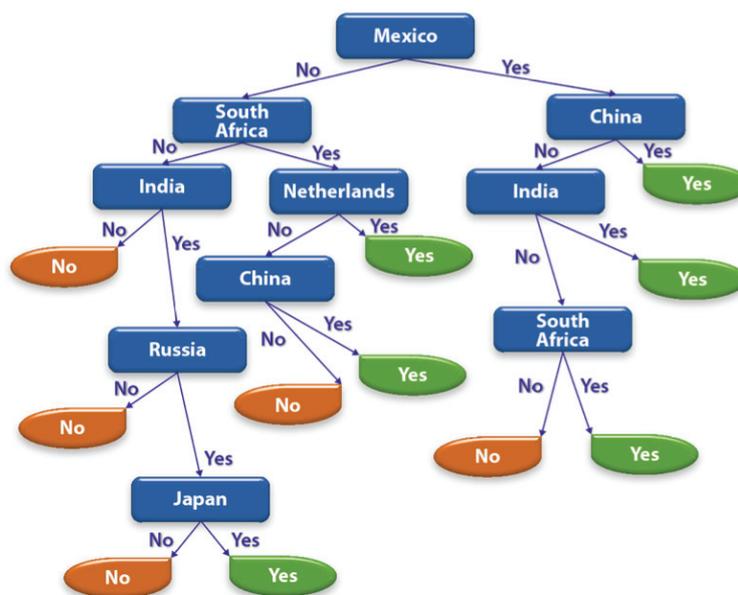

Figure 3. Decision tree of PCT EST patent deposits which have designated Brazil as destination.

between these two markets can be corroborated when one explores the quantitative data of the same decision tree. In the branch of the tree where Mexico is a destination, there is also the outcome of Brazil as target in 76.7% of the 36,316 records analyzed. Similarly, where Mexico did not compose the family of patent applications analyzed, in 97.8% of cases, Brazil was not a listed target either.

Another point that should be highlighted is the emergence of the other emerging countries that make up the BRICS in this decision tree, which has given rise to a new bibliographic and documentary research to elucidate it. Fulquet and Pelfini (2015) show that emerging powers, notably the BRICS, have been redefining the architecture of international cooperation in a global context of growing demand for energy. This can also be seen when analyzing the UN initiative to hold, in 2007, the International Biofuels Forum. This forum brought together the main emerging economies (Brazil, China, India, and South Africa), the European Union and the United States, with the aim of promoting the sustained use and production of biofuels around the world, including seeking to create standards and codes for bioenergy products, in order to consolidate and facilitate world trade.

The absence of Russia in this forum, as well as its little quantitative expression in the decision tree of Figure 3.3, is corroborated by MacFarlene (2006), when he throws light on the question: "Is Russia an emerging power?" And he concludes in his study that the maintenance of Russia's sovereignty and the recovery of its economic position is more evident than its effective growth.

5. CONCLUSION

By identifying which countries have the highest number of patent applications in a given technological area, it can be inferred that those markets are heated for such technology. However, this limited analysis could be extrapolated to purely descriptive statistics. What has been shown is that the relationship between certain technology markets in a specific period can lead to the identification of opportunities for business expansion.

In this work, it could be exposed that the BRICS economic cluster works with green patents, except for Russia. It means that when a company wants to protect an Environmentally Sound Technology in Brazil, it must look for other countries, because there is an 85% precision pattern identified for them.

The applicability of the method to companies of any size and the use of a free, accurate and cohesive world demonstrate its potential for integration. In spite of this, the limita-

tion of its application can be denoted when looking at areas where there are weak indications of the use of scientific and technological knowledge to base business, or in areas exclusively focused on providing services, to the detriment of product development and productive processes.

As a suggestion for future work, a real case of application of the prospecting method proposed to a group of companies that have real interest in foreign trade is being operated, along with the association of this study with national foreign trade strategies for the country.

REFERENCES

Ajay, D.; Gangwal, R. P.; Sangamwar, A. T. (2015), "Ipat: A Freely Accessible Software Tool for Analyzing Multiple Patent Documents with Inbuilt Landscape Visualizer", Pharmaceutical Patent Analyst, Vol. 4, No. 5, pp. 377–386.

Breitzman, A. F.; Mogee, M. E. (2002), "The Many Applications of Patent Analysis", Journal of Information Science, Vol. 28, No. 3, pp. 187–205.

Bretas, W; Morais, A.; Hora, H. et al. (2018), "Roadmap Tecnológico De Patentes Verdes Como Subsídio Estratégico Ao Empreendedorismo Sustentável", Sustentabilidade e Responsabilidade Social em Foco, Vol. 4., Poisson.

Chau, M.; Xu, J. (2012), "Business Intelligence in Blogs: Understanding Consumer Interactions and Communities", Mis Quarterly, Vol. 36, No. 4, pp. 1189–1216.

Chen, H.; Chiang, R. H. L.; Storey, V. C. (2012), "Business Intelligence and Analytics: From Big Data to Big Impact", Mis Quarterly: Management Information Systems, Vol. 36, No. 4, pp. 1165–1188.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996), "From Data Mining to Knowledge Discovery in Databases", Ai Magazine, Vol. 17, No. 3, pp. 37–37.

Frietsch, R.; Schmoch, U. (2010), "Transnational Patents and International Markets", Scientometrics, Vol. 82, No. 1, pp. 185–200.

Fulquet, G.; Pelfini, A. (2015), "Brazil as A New International Cooperation Actor in Sub-Saharan Africa: Biofuels at The Crossroads Between Sustainable Development And Natural Resource Exploitation", Energy Research & Social Science, Special Issue On Renewable Energy In Sub-Saharan Africa, Vol. 5, pp. 120–129.

Goldschmidt, R.; Passos, E. (2015), Data Mining, São Paulo, Elsevier Brasil.

Wipo (2017), Ipc Green Inventory, disponível em: <https://Www.Wipo.Int/Classifications/Ipc/En/Green_Inventory/Index.Html>. Acesso Em: 14 Dez. 2018.

Jun, S.; Wood, J.; Park, S. (2018), "Multivariate Multiple Regression Modelling for Technology Analysis", Technology Analysis & Strategic Management, Vol. 30, No. 3, pp. 311–323.

Lee, S.; Yoon, B.; Park, Y. (2009), "An Approach to Discovering New Technology Opportunities: Keyword-Based Patent Map Approach", Technovation, Vol. 29, No. 6–7, pp. 481–497.

Leydesdorff, L. (2008), "Patent classifications as indicators of intellectual organization", Journal of the American Society for Information Science and Technology, Vol. 59, No. 10, pp. 1582–1597.

Liu, S. J.; Shyu, J. (1997), "Strategic Planning for Technology Development with Patent Analysis", International Journal of Technology Management, Vol. 13, No. 5/6, p. 661.

Macfarlane, S. N. (2006), "The 'R' In Brics: Is Russia An Emerging Power?", International Affairs, Vol. 82, No. 1, pp. 41–57.

Milanez, D. H.; Faria, L. I. L.; Amaral, R. M. et al. (2017), "Claim-Based Patent Indicators: A Novel Approach To Analyze Patent Content And Monitor Technological Advances", World Patent Information, Vol. 50, pp. 64–72.

Motta, G. S.; Quintella, R. H.; Garcia, P. A. A. (2015), "Patento-Scientometric Indicators for The Selection of Projects By Investment Funds", Vine, Vol. 45, No. 3, pp. 446–467.

Seo, W.; Yoon, J.; Park, H. et al. (2016), "Product Opportunity Identification Based on Internal Capabilities Using Text Mining and Association Rule Mining", Technological Forecasting and Social Change, Vol. 105, pp. 94–104.

Shih, M.-J.; Liu, D.-R.; Hsu, M.-L. (2010), "Discovering Competitive Intelligence by Mining Changes in Patent Trends", Expert Systems with Applications, Vol. 37, No. 4, pp. 2882–2890.

Silveira, F.; Machado, F. M.; Romano, L. N. et al. (2018), "Strategy for The Development Of Agricultural Machinery: Importance Of Patent Analysis", Brazilian Journal of Operations & Production Management, Vol. 15, No. 4, pp. 535–544.

Tekić, Z.; Drazic, M.; Kukolj, D. et al. (2015), "Psalm – Patent Mining Tool for Competitive Intelligence", Tehnicki Vjesnik - Technical Gazette, Vol. 22, No. 6.

Uhm, D.; Ryu, J.-B.; Jun, S. (2017), "An Interval Estimation Method of Patent Keyword Data for Sustainable Technology Forecasting", Sustainability, Vol. 9, No. 11, pp. 2025.

Yan, B.; Luo, J. (2017), "Measuring Technological Distance for Patent Mapping", Journal of The Association for Information Science and Technology, Vol. 68, No. 2, pp. 423–437.