IBIMA Publishing

mobile

Communications of the IBIMA

Vol. 2010 (2010), Article ID 592641 , 154 mini pages

Contact author: Jose Aldo Diaz-Prado, e-mail: jadiaz@itesm.mx

Title

Corporate Technology Intelligence Research System through Recycling Public Patent Databases

Authors

Jose Aldo Diaz-Prado, Arturo Lopez-Pineda and Marco Polo Cruz-Ramos

Abstract

Investment in technology is a sensitive issue in modern corporate decision making processes. Developing software for analyzing the trends of any

kind of technology requires the extraction of patent information available through web, but also to clean, transform and load it into a standard database technology. In this report an ECTL process

is described in detail for analyzing the overall results and establishing a framework for future technology intelligence research.

Keywords: Patent analysis, Data extraction, Technology Intelligence.

Introduction

One of the main goals of businesses around the world is

to develop successful technologies and applications that can be sold in the market by the same company or by third parties, but research and technological developments are becoming more

competitive in a global and regional scale causing the costs of these activities to increase.

Those companies that expect to become leaders on their fields are looking at the time to the

patent world. Patents have a very important role in the technological development of a company, because they determine the useful life of the technology being used, propose new development areas and

can provide a strategic advantage compared to its competitors. Having software that can analyze the most important trends in the technology market is an extremely important issue,

since many companies are investing huge amounts of money in the research and development of new technologies; hence, the more they know about their

investments, the best they can predict their success.

Nevertheless, the growth of information systems has derived in a significant increase of the data available

for possible analysis in favor of research and economic purposes. Normally, in these data sets many anomalies arise within them, due to the way that they were acquired: human error, or to a deficient

storage model, etc. In order to perform information analysis process, establish relationships, and perform clustering or technology intelligence processes; the information must be reliable,

complete and accurate as possible, as mentioned by Oliveira et al (2006).

The anomalies within the information can be identified and eliminated from the stored

data through the process of data cleaning (DC). DC looks forward to overcome the problems related to the quality and integrity of the data. This process requires of a framework created by a

knowledge expert of the data domain. Since an external intervention is needed, the DC process is also referred as a semi-automated process proposed by Fisher et al (2008); even though

considerable efforts have been carried out in order to try to fully automate it, as the work carried out by Simitsis (2003). The remainder of this document is organized as follows: Technology

Intelligence section explains the steps needed for creating an Information Technology Research System, and why the Extraction, Cleaning, Transformation and Loading (ECTL) process is important.

The ECTL process section explains the extraction, cleaning, transformation and loading process, required for building an information system repository. Patent Information Systems section refers to the

construction process of the patent information system, detailing its general design and database and Patent Visual Analysis explains the graphical tool developed for the correct visualization of the patterns.

Finally, some conclusions are made; including as well a brief description of the future work.

Technology Intelligence

In the evolution of the intelligence related to

Intellectual Property (IP), the following steps are identified:

- *Patent Search*. The process of identifying specific criteria within the overall database of patents; the purpose is to

select specific patents related to a certain technology.

- *Patent Mining*. In this step companies study their own patent portfolio and the ones from their competitors, with the idea of identifying

relationships between research and the technology market.

- *Patent Landscaping*. Involves all the previous steps but it also adds the Patent Mapping activity, which

provides a perspective of the issues that influence certain domain of technology, including the leading actors and its competitors, as well as the speed of the development

of new products in the market.

- *Technology Intelligence*. This is a broader vision than IP, it involves risks and opportunities detection in the technological market. It is a

tool for decision making of the trends for research and investments that a certain company should do.

Extraction, Cleaning, Transformation and Loading process

Based on the statement “Knowledge is power”, it can be inferred that having awareness

and understanding of the process a company undergoes, allows it to take better decision and thus perform actions that have outcomes align to the companies desires and expectations. Yet, information

is available in many facets. We are today awash in data, primarily collected for governments and business purposes. Automation produce an ever-growing flood of data, now feeding such vast ocean

that we can only watch the swelling tide.

All companies have to take decisions about which actions suit best their interests. *Informed decisions* – those

made with knowledge of current circumstances and likely outcomes- are more effective than *uniformed decisions*. Today there are three processes for harvesting high quality-data: data modeling,

which reveals the relevant information within the collected data ocean; data surveying, which looks at the shape of the ocean, allowing inferring where the relevant information might be found;

and data preparation, which cleans the relevant data, by removing the dirty data.

It is valuable to provide accurate, on time, and useful information when addressing

the companies' strategic problems. That is the main reason why information is vital. Nevertheless, the value of that information is proportional to the scale of the

problem it addresses. Relevant information is
expensive to collect: it takes time, money, personnel, effort, skills, and insight knowledge of the data to discover proper information. If the cost of

discovery is greater than the value gained, the effort is not profitable.

The Extraction, Cleaning, Transformation, Load Process (ECTL) process is required to

construct an information system repository. Figure 2 shows the overall process, described by Rahm and Hai Do (2000), where the data is being taken from different sources in order to create a data

warehouse. In this Section the steps of the ECTL process are explained in detail. Preparation of data is not a process that can be carried out blindly. Hitherto, no fully-automatic toolbox for data cleaning purposes have

been developed, so that it could be pointed at a data set and the dirty data could be eliminated. Probably, when artificial intelligence techniques become more powerful that they are today, fully-automatic data

preparation toolboxes will become feasible. Until then, data cleaning will remain as much art as science, when talking of good data preparation tools.

See figure 1 in full PDF online

Extraction

The information used in this work was taken from the U.S. Patent Office website (USPTO). Since the information is

published in a web context, the available format is HTML. Therefore, the data must be extracted in a more manageable format, in order to analyze it and make some inferences.

The followed steps were: first, a query within the USPTO website of the technology that we want to analyze was performed; in this case MEMS. The results of this query were several HTML pages with a

total of 14191 patent files shown as links within the website. Second, for reading any specific patent file, the desired file must be click, in order to story the gathered information; but this manual

process was automated by using a Web Data Extraction technology. Third, some Wrappers were used in order to detect on each of the links the main data for each patent; this data included:

1) Patent Name,
2) USP Number,
3) Inventors names,
4) Assignment name,
5) Assignment city,
6) Assignment country,
7) Application number, and

8) Filed date.

Last, an XML file was created containing all the information from each patent, separating the important fields with specific tags.

XML is the new standard for information exchange and retrieval. An XML document has a schema that defines the data definition and structure of the XML document [9] An XML structure was selected, due to

the wide acceptance of XML. A number of techniques are required to retrieve and analyze the vast number of XML documents. Automatic deduction of the structure of XML documents for storing

semi-structured data has been an active subject among researchers. A number of query languages for retrieving data from various XML data sources also have been developed. However, the use of this query

language is limited (e.g., limited type of inputs and outputs, and users of this language should know exactly what kinds of information are to be accessed). Data mining in the other hand, allows the user to

search to unknown facts, the information hidden behind the data. It also enables users to pose more complex queries as described by Duhnham (2003).

Data Cleaning

The Data Cleaning process proposed by Muller and Freytag (2003) requires the analysis of the following steps to verify the quality of the data:

- *Parsing*. Cleaning of the data to detect syntax and grammatical errors in the tuples or in the values of the database is carried out. This process is analyzed by

following a language compiling approach.

- *Data Transformation*. Requires modifying the data from an original format to the requested one, affecting

the scheme of the tuples as well as their values.

- *Assurance of the integrity of the restrictions*. After all the changes are made to the database, the

original restrictions in fields and tuples must still comply with the new database.

- *Duplicate Elimination*. In this step, all those tuples or fields that are duplicated,

and therefore do not give additional information, are eliminated.

Anomalies

Anomalies in the data of an information system can be

classified, according to Muller and Freytag (2003), in three groups:

- *Syntactic Anomalies*. These are those that describe the format and

values used to represent an entity. In this field, it can be considered lexical errors, like the expected size of a tuple; format errors, like the absence of some union character; and

irregularities, like the use of different values to represent the same instance.

- *Semantic Anomalies*. Anomalies where the representation of the real

world is not correctly stored in the database. Some of these anomalies are due to integrity, like non-consistent schemas in some of the tuples; contradictions, like

information that remarks the opposite in some tuples or fields related among each other; duplicates, when there are two instances of the same element; invalid tuples; that even if it

complies the previous anomalies the information does not represent the real world.

- *Coverage Anomalies*. These anomalies exist when the stored data has an

absent value in some field or some tuples that are supposed to be within the database are missing.

Data Cleaning Rules

When the database structure is established, the next step is to

format the raw data, which is the data that comes from the extraction process, in this case from a dynamic context.
Since the original information contains several anomalies, the following criteria were

followed to clean and transform the data stored in the XML file towards mySQL instructions. For the completion of these steps, Java language was used to load,

parse, transform and clean the data.

The rules applied to the raw data were to extract the individual information from each patent, using the tags

<patent>, for the start and </patent> for the ending. This step gives a Vector of N individual patents to which the next rules were applied.

- Extract the name of the patent using the tags <patent-

name>, for the beginning and </patent-name> for the ending.

- Extract the USP number using the tags <USP-number>,

for the beginning and </USP-number> for the ending.

- Extract the information of the inventors using the tags <Inventor-name>, for the beginning and </Inventor-

name> for the ending, which gives a Vector of size I of each of the inventors associated with this patent. If there are punctuation marks at the beginning of the name,

like commas, these characters are eliminated.

- Extract the information of the company, to which the patent was assigned, using the tags <assigment-name>, for

the beginning and </assigment-name> for the ending. If this tag does not exist, the first name of the inventors list is taken as the company name. In case there is a value for this tag, but its

content is numeric, then it will be moved to the application number field. If there are parentheses in the name, or initial commas, these characters are eliminated.

- Extract the information of the city of the company, to which the patent was assigned, using the tags <assigment-city>, for the beginning and </assigment-city> for the ending. If there is

no tag, the default valued as not availabel <n.a.>

- Extract the information for the country of the company, using the tags <assigment-country>, for the begining and

</assigment-country> for the ending. If there is no defined country, the default value is USA.

- Extract the application number of the patent, using

the tags <assigment-appl.n>, for the beginning and </assigment-appl.n> for the ending.

- Extract the assignment date using the tags <assigment-

filed>, for the beginning and </assigment-filed> for the ending. The date is translated to a new format in order to comply with the SQL standard yyyy-mm-dd.

Transformation

The transformation process was executed at the same time that the data cleaning process. Due to their close relationship, the rules for transforming data

were performed simultaneously.
Some of the steps done for transforming the data are:

- *Translate values*. The date is something that was changed

from the original format to comply with the mySQL syntax.

- *Splitting*. Some columns were separated into two fields, like the number of inventors

for each patent that required a variable length.

- *Validation*. There were some specific values that did not correspond to the supposed field value.

- *Generating*. Some values that will be used for consistency in the database were generated, for example assigning a default value for the country.

Loading

When the data cleaning and transformation is concluded, the next step is to perform the mySQL instructions, in such a way that they can be stored in a file and loaded on a posterior

date. This process is done in parallel with the data cleaning one, creating dynamically patent per patent.
The file is constructed writing sequentially the information from the tables in a predefined

order, due to the related information, indexes and specific restrictions of the database. For example to relate the USP-number with the inventor_ID in the inventions table, it is required that both

fields are already created before they can be loaded together. The process is to load sequentially the tables in the following order: 1) Assignee, 2) Place, 3) Assignment, 4) Inventors, 5) Inventions.

On each INSERT update of the database it is added the word IGNORE for consistency of non duplicate instances and avoiding mySQL mistakes. At the end of the process, a SQL file is created, and it contains

the instructions to generate new elements of the repository.

Patent Information System

In order to develop a unified system of technology intelligence, certain steps are performed to transform the

original data into useful knowledge that helps in the decision making process.
The next steps describe this process:

- Identify the type of data that is required for the analysis

and join different sources in a single repository.

- Clean and format the data so that some relationships can be established between inventors, companies, places, etc.

- Construct a system that joins all the data and present visually the most important information.
- The final objective is to establish a model for a system that can be usable in

the future for the analysis of different technologies, not only for the one that originally was built.

System Design

Designing an adequate system for managing the corporate information related to patent research is very important in the technology intelligence process. The goal is to develop

a software tool that let the users to interact with it and perform an Extraction, Cleaning, Transformation, Load Process (ECTL). In this software coexists several technologies and it is used

across platforms since it was developed in Java. It generates views and reports of the historic patent assignments done in the USPTO. Figure 3 depicts the information system schema.

See figure 2 in full PDF online

Database

Within the U.S Patent Office database, related with the MEMS field, it was established a data structure shown in the

Figure 4, where the tables created for data storage are shown with the specific fields, primary and foreign keys, as well as the relationships among them.

See figure 3 in full PDF online

The Places table stores the information related to the location of a patent, in other words, it saves the city and country of the company or

person to which the patent was assigned to. The Assignees table stores the names of the companies or people to whom a certain patent is assigned. To each instance is created an identification number for

posterior use in other tables. The name of the authors that work in this technology is stored in the Inventors table, where a unique identifier is assigned to establish future relationships.

The Patents table is the main one, since it makes a connection between the patent number assigned by the U.S. Patent Office and the rest of the tables. It also stores the name of the patent, its date,

assignation number and the identifiers from other tables such as the place.

The Inventions table is an intermediate table, designed for storing the multiple

relationships between several patents and several inventors. This table is the result of normalizing the original model where the relationship between the inventors and

patents tables was of the type many-to-many.

Patent Visual Analysis

Patents are a wide field, where techniques, products,

applications, and legal considerations are highly mixed. Most of the time, this is also a field dedicated to the industrial users and, for example, the academic community do not cite patents

very often. Nevertheless, patents are a unique source of information since most of the data and information published in patents are not published elsewhere. However, using and managing a set of patents is

rather complicated because most of the tools available today are both expensive and complicated or need a strong expertise in the field of intellectual property. According to Dou (1997), the

cost of patent databases, in which users are allow to perform complete searches (involving a large number of patents) or to automatically establish relationships between patents is very high

and most of the time out of the reach of middle size enterprise, academic laboratories, or developing countries. We had the opportunity to use patents in many circumstances and then from these uses develop a

basic knowledge to design and develop various tools to integrate patent data in Competitive Intelligence or Competitive Technical Intelligence as well as in innovative thinking, as

presented by Quoniam et al (1993).

Furthermore, it is very important to analyze the potential of insights that patents will provide in a

graphical way. We can analyze multiple patents in different contexts (Figure 4) such as:

See figure 4 in full PDF online

1) The amount of patents that were generated over the world during a period of time with a specific technology. This information will be generated in a top-ten list

from different countries with a specific technology. In order to know how the scientific production over the world has been growing or decreasing (Figure 5).

2) We also can analyze the amount of patents in MEMS technologies per company during a period of time. This kind of analysis will show which companies are leading the research and

innovation with this technology (Figure 6). At the same we will know where these companies are located and who the scientists that develop such technology are.

3) The third kind of patent analysis is per country, some time the corporation want s to know which country is leading the MEMS technologies until today. It is important to

know the leaders or the follower of our company and where those leaders are located. The report also include an area to do some comments in the right side of the screen, this

comments will be saved and it will generate report in a ".pdf" format.

See figure 5 & 6 in full PDF online

4) The last kind of analysis that the Corporate Technology Intelligence Research System can generate is a mixed analysis, where we apply the technology of Tag

Clouds, to give a brief review of the relevant topics in a patent or the relevant topics in the MEMS technologies in a country (Figure 7).

Conclusion and Future Work

A large literature exist about using patents to build up various indices to R&D, to ascertain quality of

inventions, to compare patent production in various countries, to evaluate the R&D policy or firms within or outside a country, etc. We propose to extend the process of the patent analysis

with the introduction of new technologies such as: data mining (classification, clusterization, etc), business intelligence and patent market analysis.

At the same time, the Extraction, Cleaning, Transformation and Loading (ECTL) for the required data is an important analysis topic due to its contributions to any given information system. In

addition, due to the large amount of data stored on XML format, an automatic process is required to identify anomalies and solve some errors.

See figure 7 in full PDF online

The Data Cleaning process is formed by several steps that improve the quality of the stored data. It is very important to increase the quality of the results obtained from a technology intelligence

system, especially if it requires managing large amounts of complex information, which nowadays is becoming more familiar in normal applications.

Technology intelligence is becoming the most important trend for the intellectual property industry, especially for those companies interested on their long term survival. The ECTL process is the first step

for developing a software tool that helps in the decision making process, if it is done correctly the next steps will have better results.

Furthermore, in the short future, the purpose is to integrate data mining algorithms with XML documents to achieve knowledge discovery. For example, after identifying

similarities among various XML documents, a mining technique can analyze links between tags occurring together within the documents keeping the structure within and XML document. Note that in this

kind of data mining, knowledge is inferred from the internal structure of XML documents.

The further research on the Data Mining and Patent Landscaping processes will be

made towards the goal of creating a Technology Intelligence research system. This will be an important issue in the improvement of decisions regarding technology

investments.

References

Abiteboul, S., Buneman, P., and Sucie, D. (2000) Data on the Web: From relations to

semistructured data and XML. San Francisco,CA: Morgan Kaufmann.

Dou, H., (1997) ‘Hearing 97 - Patents in Europe - Usage and dissemination of Patents as a

tool to improve SME's strategies', Hearing, 1997. Proceedings on the future patent politique d'information brevets de l'Organisation *européenne des brevets* (pp.66-

68). Munich, Germany: European Patent Office.

Duhnham, M.H. (2003) ‘Data Mining: Introductory and advanced topics.” Upper Saddle River. NJ: Printice Hall. 2003.

Fisher, K., Walker, D., Zhu, K.Q., and White, P. (2008) ‘From dirt to shovels: fully automatic tool generation from ad hoc data’. *SIGPLAN* Not., 43(1):421-434.

Müller, H. and Freytag, J.C. (2003). ‘Problems, Methods and Challenges in Comprehensive Data Cleansing'. Technical Report HUB-IB-164, Humboldt-

Universität zu Berlin, Institut für Informatik. Berlin.

Oliveira, P., Rodrigues, F., and Henriques, P. (2006) ‘An ontology-based approach for data cleaning’. Proceeding of

the 11th International Conference on Information Quality, MIT, Boston, EUA, November 2006.

Quoniam, L., Hassanaly, P., Baldit, P., Rostaing, H. and Dou, H. (1993) ‘Bibliometric analysis of patent documents for R&D management’. *Research Evaluation,*(3), 13-18. April 1993.

Rahm, E. and Hai Do, H. (2000) ‘Data cleaning: Problems and current approaches’. *IEEE Data Engineering Bulletin*, 23:2000, 2000.

Simitsis, A. (2003) ‘Modeling and managing etl processes’. In: *VLDB PhD Workshop*, 2003.

Yeap,T., Hwa Loo, G. and Pang, S. (2003) ‘Computational patent mapping: Intelligent

agents for nanotechnology'. *MEMS, NANO and Smart Systems*, 2003. Proceedings. International Conference on, pages 274 {278, July 2003.