Using Hierarchy to Extract Innovation; The use of Patents in Clarifying Innovation

Penelope MARKELLOU
Computer Technology Institute
61 Riga Feraiou Str.,26221 Patras, Greece
e-mail: markel@cti.gr

Angeliki PANAYIOTAKI
Computer Technology Institute
61 Riga Feraiou Str.,26221 Patras, Greece
e-mail: panayiot@ceid.upatras.gr

Spyros SIRMAKESSIS
Computer Technology Institute
61 Riga Feraiou Str.,26221 Patras, Greece
e-mail: syrma@cti.gr

Antonis SPINAKIS
QUANTOS S.A..R..L.
68, Boulevard De Port Royal, 75005 Paris, France
e-mail: aspi@quantos-stat.com

Athanasios TSAKALIDIS
Computer Technology Institute
61 Riga Feraiou Str.,26221 Patras, Greece
e-mail: tsak@cti.gr

Abstract: In this paper, we try to tackle the problem of the lack of a multidimensional analysis of the patents. The study is based on the principles and methods of textual analysis. The aim is to show effectively how the use of methods, like the hierarchical classification, can answer questions applied by industry concerning the interactions existing between the different fields of activity and the poles of innovation that are being created in these fields. We also describe in a detailed way the different step-by-step treatments applied to patents database.

Keywords: textual analysis, data mining, hierarchical classification, patents, indicators

1 Introduction

Nowadays, the increase of the volume of data stocked in the different computer systems is such that only one extremely reduced proportion of these data (typically between 5 and 10%) can be effectively analysed. The use of techniques of automatic analysis allows us to

valorise in a more efficient way the potential wealth of information that the textual databases represent, not only reply to a scientific and technically passionate challenge, but also to a real economic evolution, particularly crucial in fields such as old technology for instance. The continuous progresses realised in fields like the document research, the analysis of data and the automatic treatment of natural languages have lead us to the realisation of systems proposing functionalities relatively simple but operational in conditions of real use (important volume of data, textual data extremely varied).

The old technology must be today considered as the main equipment for the description of the scientific and technological evolution. The existing systems of database interrogation allow users to completely understand a query and provide them with references concerning the subject that he is interested to. These references used in our study are the patent references.

In general, the analysis and comparison of the scientific and technological activity between countries or between enterprises, with the help of patents, is made through a priori classifications and different types of indicators. Describing them in this way allows scientists to identify the tendencies and the essential points in the peak fields and to place them according to the industrial activity worldwide. On the other hand it does not allow a multi-dimensional comparison, neither of the countries and activity fields nor of the concurrence within the same activity field.

We shall thus try through this document to answer the problem of the lack of a multidimensional analysis of the patents. All the study will be based on the principles and methods of textual analysis**.** This type of approach allows us to manage in an elegant way data that are difficult to use, such as the patents. The final aim is to demonstrate how it is possible to answer questions of the industries, concerning the consideration not only of their environment as a whole but also the interactions existing between the different fields of activity and the poles of innovation that are being created in these fields, by the use of methods like the hierarchical classification. We will also describe in a detailed way the different step-by-step procedures applied in patents database in order to get all necessary information in order to place European industry in the concurrence environment of its evolution.

Of course, many problems (both algorithmic and conceptual) that we have met, remain internally complicated and still require the discovery of satisfactory theoretical solutions. However, the work already done in the different scientific disciplines has reached today a sufficient critical volume, in order to permit the realisation of techniques performing effectively enough so as to introduce the applications.

The methodology that will be described at the following sections is the one that is expected to be followed by STING[1] project (Evaluation of Scientific & Technological Innovation and Progress in Europe Through Patents – IST-1999-20847), which is partially funded by EC IST Programme. Patents data describe technological innovation in quantitative as well as qualitative terms and are used for the production of indicators for measuring the effects of S&T activities. However, data derived from patent documents require special treatment.

[1] More information is available at http://sting.cti.gr.

Therefore, it is required to extend the used technologies and tools and to develop the required abilities and means to exploit the information available in patent documents in order to produce accurate and reliable patent indicators. The main objectives of STING project are:

- Development of enhanced methodologies for the analysis and processing of patents data.
- Development of a reliable, accurate way to measure technological innovation and to produce indicators on a regular basis.
- Improve the quality and timeliness of the produced indicators.
- Exploitation of latest developments in IT in order to gain fast access to databases.
- Development of a computer-assisted system for the analysis of patents data.

2 General presentation of data

A reference of patents can be analysed and described by different aspects (through different fields). Every field contains specific information described by a code. The fields describing a patent in the Space Access[2] patent database are: AN (Number of the patent in the base), TI (Title of the patent), PA (Name of the deposing company), IN (Names of the inventors), PN (Number of the patent), DS (Country of extensions), IC (Codes of International Classification of the Patents-CIB), MC (Code Manual) and FT (Resume in French).

The different fields that illustrate the subjects treated by the invention influence a patent reference. All these fields handled together or separately can be the object of statistic treatments. Our approach is based on the analysis of CIB[3] codes (International Classification of Patents). The other fields will describe the result obtained by the use of the codes.

2.1 The system of patent classification

Every patent reference in the base is qualified, using computer terminology, by the code of classification CIB. A documentary classification separates the scientific sectors in sections. This classification is constructed according to an hierarchical principle, these sections are themselves separated in sub-sections, class, sub-class, group, sub-groups, etc. Every level in this hierarchical structure is represented by a codification.

3 Selection of patents, constitution of references' corpus

3.1 Consulted bases

This first step, of the construction of the references' corpus is certainly the one that influences most the quality of the results. This quality is of major importance as the results enter in the decision process on one side and in the description of the scientific and technological activity on the other side. Let us not forget that the principal aim is to present a method that allows considering the diversity of the meaning brought by the classification

[2] This is a CD-ROM series containing bibliography data about the patents for the period between 1978 and 1998.
[3] Classification established by the official institutes of deposition of patents.

used on the patents. But before concluding on the result coherence, it is necessary to create a total of references as homogeneous as possible.

The corpus that will be created has to be wide enough in order to cover the studied subject and restricted enough in order to contain a "noise" as weak as possible. This is the reason for which the total of references, collected throughout the study, is not probably exhaustive but has an element of homogeneity[4]. For the construction of the work corpus we have consulted the Space Access database.

3.2 Constitution of the reference corpus

Through our interrogation strategy 1056 patents were selected au hazard among the 7855 patents published in Europe in 1995 in the sectors of H04 (TECHNIQUE OF THE ELECTRIC COMMUNICATION) and H05 (ELECTRIC TECHNIQUES NOT PREVIEWED ELSEWHERE).

Not every patent is included in the CIB codes. In our corpus of references only 619 patents among the 1056 selected dispose of one or more CIB codes. On the opposite of the CIB codes, the texts describing the functionality of a patent exist on the totality of the corpus but their analysis creates a certain number of problems not always evident to solve.

4 Analysis of the patents according to EUROSTAT approach

EUROSTAT collects the data necessary for the analysis of the scientific and technological activity per country and per scientific sector from the office of European patents in Munich (EPAT patent database).

From these data are constructed matrices crossing the countries or the companies that make the deposition with the 3 or 4 first digits of the first code[5]. If a patent belongs to more than one companies, it is distributed by fractioning to the different depositing companies. These matrices are made or for one given (specific) year either for the totality of the years.

But this approach is a macroscopic one and for this reason is insufficient for the analysis of the scientific and technological activity, because it cannot consider the hierarchical level of necessary codes for the description and optimal analysis of the activity of every technological sector and the totality of the proposed codes in order to describe a patent. In this way there are created a priori classifications that do not exist in reality, as a patent can be valid for many activity sectors in which it can have many functionalities.

The fact that a patent is categorised, by this system of 3 or 4 first digits of the first code, in the sector of the specific activity, can create artificial phenomena: misjudgement of the countries or of the companies concerning their scientific and technological capacities, or valorise in an artificial way sectors of activity that in reality are not so important.

[4] All methodological steps that we are going to describe have to be conducted in presence of experts of the sector in order to guarantee not only the validity of the constructed base, but also the results of the analysis.
[5] A patent can be described by more than one codes.

5 Essay on the analysis of patents of the European electronic industry; comparison with the American and Japanese industries of the same sector

5.1 The problematic

The groupings of patents are multi-variable, but it seems that we must privilege the ones that are related to hierarchical methods as these ones provide us with a complete view of the grouping of corpus to homogeneous groups of patents, with their links inter and within the groups.

Obviously, we are still far from an easy to realise analysis, since the diversity of information sources, - CIB codes in different hierarchical levels, different databases etc. – the data formats, the redundancy of references, the multiplicity of treatments and the possible methods require an intelligent expertise for the overall treatment of the information.

5.2 The concept

The concept that was developed for the creation of homogeneous groups of patents will be the one that is used in textual analysis. According to this approach, two patents belong to the same category not only because they cover the same sectors, but also because these sectors are just a little bit shared by other patents. We will see the apparition of interesting and hard to reveal phenomena concerning the databases such as the existence of a priori patents with no direct relation combining corresponding technologies.

5.3 The statistic analysis

A code can be assimilated to a path among the possible branches of the hierarchy. In this hierarchy the branches are shared from a very large level of signification till levels of signification more and more detailed. The more down we are going to the hierarchical branches the more complicated representation a code has and the more precise its meaning is. But what hierarchical level must we consider given the previous difficulties? If we take the most precise hierarchical levels in order to have the finest descriptions, this is disputable. In fact, not all the codes are forcibly updated till the last hierarchical branch. For example, in our corpus of references we have 978 CIB codes updated till level 6, instead of 861 updated to level 8 and only 9 to level 9.

The choice of hierarchical level imposes thus a compromise between the loss of information and the gain of signification (for a very fine choice, there is loss of certain codes, but remaining codes are more precise). A pre-study was conducted in order to determine what hierarchical levels should satisfy the best statistic solution. The studied criteria for every hierarchical level were the number of remaining codes and the number of patents still updated.

Let us notice that the codes having a frequency equal to 1 have not been included in the analysis (hierarchical classification), as they do not establish any link between the patents.

The result of the analysis of the corpus is that level 6 is the best compromise among the

different classifications proposed. In fact, the number of patents, after elimination of patents at frequency 1 still remains important, whereas in case of 8 digits it begins to diminish sensibly; the mean (average) number per patent decreases from 1.04 to 0.42.

The case of 4 digits was excluded from the beginning, as the number of distinct codes after elimination of codes at frequency 1 remains very low (68 codes) relatively to the case of 6 and 8 digits (respectively 115 and 158 codes). The chosen hierarchical level for the study has been the codes CIB of 6 characters.

5.4 Factorial analysis corresponding to CIB codes

Many types of tables can be used by the methods of statistic analysis. Concerning the factor analysis, the entry tables are tables of frequency of the apparition of codes. These matrices cross a total of individuals and a total of descriptive variables. Naturally, the individuals here are the 577 patents and the total of variables is consisted of the 115 distinct codes after elimination of codes at frequency 1.

It is a first obligatory and necessary passage for the suite of the operations of the factor analysis of correspondence. It was applied to table crossing 577 patents (after elimination of patents at frequency 1) and the 115 distinct codes with 6 digits.

At this stage of analysis we present only the projection of the unique supplementary variable which we dispose of. This one was constructed by grouping the patents into groups according to the depositing countries.

So, we have constructed three big groups:

- European companies
- American companies (included Canada) and
- Asian companies (only Japan and Korea).

This choice, maybe prone to discussion[6], has provided us with quite interesting results. A first indisputable observation is the opposition between the industries coming from the United States on the one hand and those of Europe and of Asia on the other. Moreover, the European Industries are opposed to the Asian Industries.

5.5 Ascendant hierarchical classification

Our aim here is to reveal the classes of patents that are related by the descriptive codes that they have in common, otherwise look for the families of patents characterised by their similarities in terms of shared technologies. Thus, a certain number of codes can be found in the large majority of references; others appear in a less systematic way and others finally figure only in too weak number of references.

The classification criterion that will allow us to count similarities among patents estimates these phenomena. Our choice is based on Ward's criterion, which exposes very good

[6] In fact we could project all the countries separately or by grouping them in countries of North and South etc. but our aim is first of all to show a new way of research offering results of an indisputable clarity.

elements. This one guarantees the division of the corpus in homogeneous and discriminating patent families.

The analysis of the result allows also to explain each of these classes according to the codes or groups of codes that have presided their creation, in other words, to attach at each one of them a label resuming the technological subjects that it covers.

The classification was calculated on the base of factorial elements of 577 individuals on the first 50 factorial axes (80% of the total variance of the points cloud).

The hierarchical tree was divided in 43 classes. This is the optimal number of classes. On first sight the patent distribution seems to be quite balanced. In fact only 20 classes among the 43 contain more than 10 patents from which only one contains 56 (9% of the total). This ascertainment is quite encouraging as we avoid the construction of "heavy" classes (frequent phenomenon in big concave matrices).

	Number of patents	% of patents
Class 1	7	1.21%
Class 2	16	2.77%
Class 3	5	0.87%
Class 4	7	1.21%
Class 5	23	3.99%
Class 6	13	2.25%
Class 7	5	0.87%
Class 8	5	0.87%
Class 9	31	5.37%
Class 10	42	7.28%
…….	…….	……
Class 43	14	2.43%
	577	

By a first reading we find out that:

- The average number of codes per class is between 1 and 2.
- Only classes 3 and 31 contain 5.6 and 3 codes respectively in average.
- The classes with the biggest diversity are classes 2 and 14 each one with 12 different codes.
- Finally class 11 contains 93 codes of which 5 are distinct.

The examination of the tables given beneath provides the following description of classes:

- Very homogeneous classes (majority code at more than 70%) are 32 classes.
- Relatively homogeneous classes (majority code between 55 and 75%) are 4 (6%) classes.
- Non homogeneous classes (there is no majority code – between 35 and 45%) are 11 (24%) classes.

For 70% of the classes there is a clear technological tendency that comes out. For the rest we describe the class relying only to the first majority code.

Maybe the chosen rule for the description of classes, relatively and less homogeneous is not the most recommended. But in absence of a specialist of the sector, in order to propose another procedure, the solution of the majority code has a statistic meaning.

Let us note here, that, if we want the analysis to deliver trustworthy information with strategic aim, it is necessary to validate every step, from the selection of the corpus till the interpretation of the results, through different and adapted levels of expertise:

- Technical sector expert
- Patent expert
- Information expert
- Statistics expert

A system of supervision (alert) has to respect the difficulties in order to allow treating the subjects, the mass and the knowledge complexity of which cannot be apprehended by simple treatments, in acceptable time limits.

Having defined the classes the next step should be their description through the following illustrative variables:

- Depositing continents, an artificial variable, created by grouping the European industries on one hand, the Asian (Japan, Korea) and United States (plus Canada) industries on the other hand.
- Depositing companies
- Countries of deposition
- Inventories

Each one of these variables plays a specific role in the comprehension of the scientific and technological activity description process.

5.5.1 Depositing continents

From the 43 created sectors the European industry seems to get active in a satisfactory way in 10 among them. It is in the multiplex systems with time division (class11) that we find the biggest number of deposited patents. Other sectors with strong European tendency are the following:

- Automatic or semi-automatic centrals and
- Apparatus of selection in which the subscribers are linked through radioelectric or inductive links.

The U.S. is more attracted by the development of communication data webs (interconnection or transfer of information or other signs between memories, apparatus of entry/exit (class 10, 10.6%).

Asia is more preoccupied by the sector Details of transmission systems not characterized by the environment used for the transmission (agreement of the co-ordinated circuits) and by details of television systems (details of sweeping or their combination with the production of alimentation tensions; colour TVs). It is about classes 12 and 17 with 12 and

19% respectively.

5.5.2 Depositing companies

A reference patent can be deposited by different companies something that explains the raised number of companies in comparison with the number of patents of the class; i.e. class 2 is constituted from 16 patents and contains 24 depositing companies.

According to the study, the company, which has deposited the biggest number of patents in sectors H04 and H05, is Philips Electronics. The companies *AT&T*, *IBM* and *NEC* follow closely Philips in number of deposited patents.

5.5.3 Countries of deposition

We can also see the share per country of deposition something that could allow reveal the countries the markets of which are interesting. The countries, in which companies seem to want to protect their inventions, is first of all Germany (97% of the patents in a total of 691), then Great Britain, France, Italy etc.

5.5.4 Inventories

We can also distinct the share per inventory on the patents, fact that can reveal the activity of each inventor on specific areas.

6 Conclusions

This type of analysis and decision helping tool will enormously facilitate the experts' web, and at the same time decrease the risks of making mistakes or eventually forgetting, something that could be issued by not estimating in their entirety the links and relations between patents.

A classificatory approach is therefore the best adapted in order to respond to problems like those posed by the technological alert. With this approach two patents belong to the same class, not only because they cover the same sectors, but also because they are seldom shared from other patents.

We can then mention the appearance of interesting and not easy to reveal phenomena on the databases such as the existence of a priori patents with no direct relation combining equivalent technologies. The textual analysis seems to be, because of that, perfectly appropriate to the bibliometric analysis. It does not neglect any information, even if its weak presence gives to it an a-priori minor character.

Obviously, many problems (algorithmic or conceptual ones) that we face still remain complicated in them and still need the discovery of satisfactory theoretical solutions.

References

[1] Title of reference paper by name of authors, if relevant: where presented, publisher and publishing year (ISBN or other reference)