Pergamon

Information Processing & Management, Vol. 30, No. 3, pp. 407–418, 1994

0306-4573/94 $6.00 + .00

0306-4573(93)E0014-F

A NEW METHODOLOGY FOR SYSTEMATIC EXPLOITATION OF TECHNOLOGY DATABASES

CHANTAL BÉDÉCARRAX and CHARLES HUOT
European Center of Applied Mathematics, CEMAP IBM-France,
68,76 quai de la Rapée, 75592 Paris Cedex 12, France

(*Received* 28 *May* 1992; *accepted in final form* 29 *April* 1993)

Abstract—Nowadays technology watch must be considered as a strategic tool for business enterprises. The increase of database volume has forced a change in information management. The purpose of this article is to explain how a mathematical data analysis method can help to transform sequential raw data into valuable information.

Keywords: Technology watch, Relational analysis, Patents, Bibliometrics, Database, Automatic classification, Strategic information.

1. INTRODUCTION

In recent years, we have been faced with an extraordinary growth of information from public or private scientific or technological databases. Databases are physical structures devoted to load information, and they transfer it to final users. Nevertheless, databases should not simply be considered as information warehouses. We must see them as deposits to be exploited, the ore being the data they contain.

The information can be extracted from the transformation and the exploitation of this ore according to a given problem. Let us list some examples of systematic exploitation of database processes to transform data into valuable information:

Patents DB	→ Technological watch
Scientific DB	→ Research evaluation
Technical text DB	→ Technical thesaurus
Questionnaire	→ Market survey
Customers DB	→ Customers typology
Press DB	→ News synthesis

Information appears under different forms in a document or a reference: free text (titles, abstracts) or codified fields (index terms, inventor names, companies or laboratories, countries, classification codes, . . .).

The references are given by the downloaded data after a remote query. We have to extract as much information as possible from this data set.

At first we must have a good description of the basic data by using classical statistical analysis. This step is important to determine the relevant variables and the variable distributions. Afterwards we can analyze the corpus to detect an underlying data structure to reveal what we really call information. The full text analysis, which is the most informative, is still premature today (natural language processing has not yet reached the maturity stage). Thus we base our study upon the codified fields (including index terms), which are excellent data sources to analyze.

2. THE RELATIONAL ANALYSIS

In this paper we use classification methods based on the general methodology of the Relational Analysis (Marcotorchino, 1986, 1991b). This methodology has been used in

many applicative fields as diversified as computational lexicography (Warnesson & Bédécarrax, 1988; Bédécarrax & Warnesson, 1989), manufacturing (Marcotorchino, 1987), and epidemiology (Parisot, 1985). The purpose of this paper is to present a new extension of the relational analysis application's field, namely, Technology Watch (Huot *et al.*, 1992).

The relational analysis methodology is based upon the general following points:

- data management
  under a relational logic form (paired comparisons),
  of all types: qualitative, binary, contingency, quantitative;
- linear programming modelization
  unified methodology approach;
- non hierarchical classification
  simple classification (similarity aggregation, relational matching),
  cross classification (seriation, quadri-decomposition);
- detailed result analysis
  quality indicators,
  coherence indicators.

Moreover, within the database systematic exploitation framework, the data present specific properties, such as

- large scale matrices,
- sparse matrices, and
- specific distributions.

As we will see, the relational analysis approach allows for the management of these characteristics.

3. A NEW VISION OF TECHNOLOGY WATCH

A recent trend in business firms' needs shows that one of the most useful techniques for competitive intelligence is undoubtedly the statistical analysis of patents. This interest is confirmed by the growing scientific literature on this subject (Brockhoff, 1992; Mogee, 1991; Griliches, 1990; Albert *et al.*, 1991; Hamers *et al.*, 1989; Courtial & Callon, 1991; Dou *et al.*, 1991).

This phenomenon clearly expresses the importance of data analysis methods in the technological survey process. The development of these methods will radically modify working uses and open new horizons. But the key to success for this mutation will be the control of these concepts and the capability of an intelligent exploitation of this mine of raw material.

Broadly speaking, we can distinguish two environments: on one hand, the database contents, and on the other hand, the methods and tools that will allow for their detailed analysis. The technology watch transforms a group of punctual data into an elaborate information base. Thus it can provide strategic information on competitors, detect innovations, evaluate the research axis, and survey scientific and technical evolution.

The application of relational analysis to database analysis answers more specifically some fundamental questions in technology watch, such as:

- typology of activities domains (main technologies, innovation, cross application technology, . . .),
- knowledge of a technological environment for the creation of new markets,
- highlight of extension strategy and international coverage, and
- analysis of the matching between coding classification on an external database and company internal classification.

4. METHODOLOGY DESCRIPTION

4.1 *General data presentation*

This preliminary step consists of

1. reference extraction from the database and matrices generation, and
2. processing of the matrices.

4.1.1 *References extraction.* This step is the classical first step for any statistical analysis of databases. This work is very tricky and requires a good knowledge of international databases to isolate a coherent reference data set for the study. Our methods take into account specific matrices generated by the relationship of different fields in data set references.

For our study, the matrix generation is provided by the C.R.R.M. researchers (Centre de Recherche Rétrospective de Marseille, University of Aix-Marseille, France) using prototype softwares specialized in this kind of application.

Let us briefly describe the main fields contained in a patent reference. These fields are indexed with codes dedicated to a specific item of information. For example, in the Derwent database (Derwent is the server center of the World Patent International and World Patent International Latest databases), a patent record looks like the following:

AN 92-009829/02
TI Patches for topical or transdermal drug delivery—with adhesive layer contg. polyacrylate adhesive and film former
TT PATCH TOPICAL TRANSDERMAL DRUG DELIVER ADHESIVE LAYER CONTAIN POLYACRYLATE ADHESIVE FILM FORMER
PR 90.06.25 90DE-020144
PN EP-464573-A 92.01.08 (9202)
DE4020144-A 92.01.09 (9203)
AP 91.06.24 91EP-110409 90.06.25 90DE-020144
DS AT BE CH DE DK ES FR GB GR IT LI LU NL SE
PA (LOHM) LTS LOHMANN THERAPI
IN MULLER W,MINDEROP H,TEUBNER A
LA G
CT (G)DE3843238 DE3843239 EP-305758 EP-379933
IC A61L-015/16 A61F-013/02 A61M-037/00
DC A96 B07 D22 G03 A14 P34 P32
MC A04-F06E5 A08-P01 A12-V03A B04-C03B B12-M02F D09-C04B G03-B02D1 G03-B04
AB (EP-464573)
Topical or transdermal patches comprise a backing layer, an adhesive layer and a release liner. The adhesive layer comprises 100 pts.wt. of a polyacrylate adhesive (I), 5-150 pts.wt. of a polyacrylate-compatible film former (II), 0.250 pts.wt. of non-plasticising active agents and/or additives, and 10-250 pts.wt. of plasticising active agents and/or additives.

ADVANTAGE—Inclusion of (II) overcomes consistency problems associated with high levels of plasticising components. (10pp Dwg.No.0/0)

The definition of the main fields codes are:

- AN: Accession number (patent number in the database)
- TI: Title
- PN: Patent number
- DS: Extension countries
- PA: Patent assignee (the company that registers the patent)

- IN: Inventors
- CT: Citations
- IC: International Classification codes
- DC: Derwent classes
- MC: Manual codes
- AB: Abstract.

All this information can be processed separately or together, as they can be organized in different statistical tables. We focused on the study of the relations among three fields: patent assignee, patent number, and IC codes.

Before detailing the data processing, let us briefly explain what IC codes represent and what they look like (see Fig. 1). The International Classification of patents describes the whole range of domains that can give rise to invention patents. This classification is indexed by hierarchical codes divided into *sections*, *classes*, *subclasses*, *groups* and *subgroups*. According to this organization, the IC codes can be considered at different levels: from the less precise (sections: 1 digit), corresponding to 8 descriptors, to the most precise (the subgroups: 11 digits), corresponding to about 700,000 descriptors.

4.1.2 *Relational presentation.* Starting from these fields, we have defined three sets of objects:

- the set of IC codes, noted I with $\text{Card}(I) = n$; the value of n can vary according to the number of digits chosen for the study (i.e., the precision of the IC codes);
- the set of patent assignees, noted J with $\text{Card}(J) = m$;
- the set of patent numbers, noted L with $\text{Card}(L) = p$.

The data is extracted from the descriptors of the p patents that represent the studied data set. The different relations found in the data can be outlined with the two tables, T and T' as shown in Fig. 2. Table T describes the cross connection between the patents and the IC codes they contain. The table T' describes the cross relation between the patents and the companies they belong to. These two tables simply restate, under a relational form, the information extracted from the references. Their general terms are given by:

$$t_{li} = \begin{cases} 1 \text{ if patent } l \text{ is described by code } i \\ 0 \text{ otherwise} \end{cases}$$

$$t'_{lj} = \begin{cases} 1 \text{ if patent } l \text{ is registered by company } j \\ 0 \text{ otherwise.} \end{cases}$$

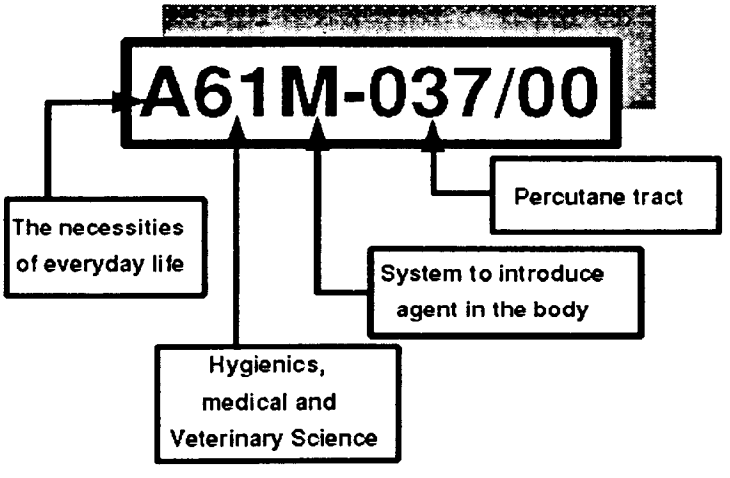

Fig. 1. Example of IC code (Derwent WPI database).

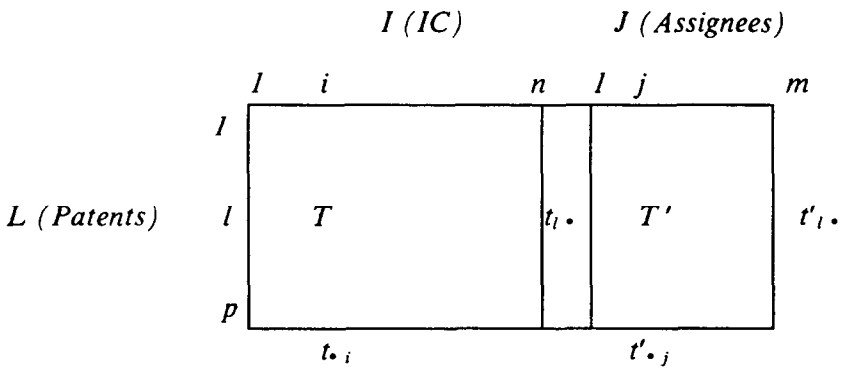

Fig. 2. Characteristics of the binary tables T and T'.

Row sums:

$$t_{l\cdot} = \sum_i t_{li} = \text{number of IC codes describing the patent } l$$

$$t'_{l\cdot} = \sum_j t'_{lj} = \text{number of assignees of the patent } l^*.$$

Column sums:

$$t_{\cdot i} = \sum_l t_{li} = \text{number of patents described by IC code } i$$

$$t'_{\cdot j} = \sum_l t'_{lj} = \text{number of patents registered by the company } j.$$

4.2 *Addressed problems*

We used automatic classification methods developed within the relational analysis framework to illustrate the structured information from the sequential data extracted from the database. Let us briefly recall the principles of the two methods we have implemented in that paper, namely, classification and seriation. The classification method deals with the relations among the objects within a single set. It groups together the objects that resemble each other in homogeneous classes. The structured relation we try to build is an equivalence relation, which means a partition of the objects set. The general form of such a relation can be drawn as shown in Fig. 3.

The seriation method deals with two different sets. The relations are given by the cross correspondence of a set of objects and a set of attributes (or descriptors). The seriation

*A patent is generally registered by only one company, that is the reason why, in our study, we always have $t'_{l\cdot} = 1$.

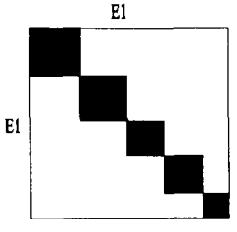

Fig. 3.

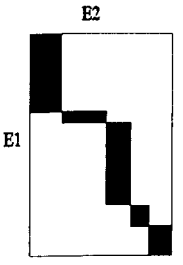

Fig. 4.

method aims to find the optimal correspondence linking these two sets. This rectangular correspondence is called a block correspondence, because it shows the connections between classes of the two sets. The general form of a block correspondence is shown in Fig. 4.

In this paper we will also refer to the notion of quasi-seriation, a special kind of block correspondence, in which there is no systematic correspondence between each class of one set with a class of the other set. Some row object (respectively, column object) may be isolated without any correspondence with a column object (respectively, a row object). The general form of such a relation is shown in Fig. 5.

All the problems can be modeled in the form of linear programs under constraints. The economical function of the mathematical program corresponds to the criterion that measures the adequacy between the solution and the data. The choice of this criterion is a fundamental point, because it induces the nature and the intensity of the resemblances we want to highlight. This point needs to be seriously discussed before starting any data analysis process (Hamers *et al.*, 1989; Courtial & Callon, 1991; Huot *et al.*, 1992).

With the relational approach it is possible to choose from a large range of criteria to meet the particularity of the problem and the available data. Some criteria deal with binary data, some others better suit frequency data; most of them are based upon majority rules that determine the level of the relation defining the threshold beyond which two objects are considered similar.*

Let us recall one of the main properties of the relational analysis methodology: Within non-hierarchical classification methods, it does not require a choice of the number of classes of the solution prior to the processing.

4.2.1 *Patent numbers* × *IC codes.* The table T describes the basic relation between these two fields under the form of a rectangular binary matrix $p \times n$.

Patents number × IC codes seriation: Each patent, $l \in L$, is described by a set of IC codes. This description defines the profile of the patent in the matrix T. Inversely, any IC code, $i \in I$, belongs to a certain amount of patent's references.

The goal of this first processing is to highlight the different trends of the studied data set, that is, to show the distribution of the patents according to families of domains along

*The most classical criteria are the Condorcet criterion (derived from the vote theory problem described in Michaud, 1982) and the weighted Condorcet criterion (which connects the relational analysis to the factorial analysis; Marcotorchino, 1991a).

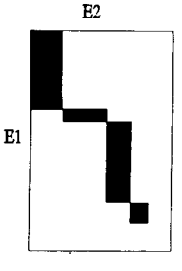

Fig. 5.

with their characteristic code groups. To answer this question, we have submitted the matrix T to the seriation process. As previously stated, the solution relies upon the criterion used. In this case, we chose the classical seriation criterion on binary data, defined as follows:

$$T(Z) = \sum_{l} \sum_{i} (2t_{li} - 1)z_{li},$$

where the binary relation Z satisfies the constraints of a block correspondence on $L \times I$, that is:

- affectation:

$$\sum_{l \in L} z_{li} \geq 1, \quad \forall\, i \in I$$

$$\sum_{i \in I} z_{li} \geq 1, \quad \forall\, l \in L;$$

- impossible triade:

$$z_{li} + z_{li'} + z_{l'i'} - z_{l'i} - 1 \leq 1 \quad \forall\, (l,l') \in L, \forall\, (i,i') \in I$$

$$z_{l'i'} + z_{l'i} + z_{li} - z_{li'} - 1 \leq 1 \quad \forall\, (l,l') \in L, \forall\, (i,i') \in I$$

$$z_{l'i} + z_{li} + z_{li'} - z_{l'i'} - 1 \leq 1 \quad \forall\, (l,l') \in L, \forall\, (i,i') \in I$$

$$z_{li'} + z_{l'i'} + z_{l'i} - z_{li} - 1 \leq 1 \quad \forall\, (l,l') \in L, \forall\, (i,i') \in I.$$

After the maximization of $T(Z)$ under these constraints, the solution Z shows the best correspondence between patent classes and IC code classes. The blocks of the seriation provide us with groups of patents mainly described by the IC codes pertaining to the same block. A classical descriptive statistical analysis would give "sequential" information on each patent or each IC code regardless of the other ones. The relational analysis provides us with a real data structure that leads to a global vision of the basic information.

The seriation method, through the block correspondence relation, *de facto* generates two partitions, one on the set L and the other on the set I. By construction, these two partitions do not unfortunately have any optimal property on their own set. In order to find the optimal equivalence relations on each set, we must have recourse to the classification process.

Patent classification: Starting from the table T, we can build the similarity matrix $\hat{B}$, $p \times p$, crossing the patents. The general term of this matrix is defined as

$$\hat{b}_{ll'} = \sum_{i} \frac{t_{li} t_{l'i}}{t_{\cdot i}}.$$

$\hat{b}_{ll'}$ is a presence-rarity index: Two patents are more similar (high value for $\hat{b}_{ll'}$) as they share IC codes ($t_{li} = t_{l'i} = 1$) that are seldom present in the data set (low value for $t_{\cdot i}$).

According to the relational analysis methodology, we build complementary similarities (i.e., dissimilarities) between objects $\bar{\hat{b}}_{ll'}$, in the following way:

$$\bar{\hat{b}}_{ll'} = \frac{\hat{b}_{ll} + \hat{b}_{l'l'}}{2} - \hat{b}_{ll'}.$$

The optimal partition on the set L is then given by the maximization of the criterion:

$$B(X) = \sum_{l} \sum_{l'} (\hat{b}_{ll'} - \bar{\hat{b}}_{ll'}) x_{ll'} \quad \text{or} \quad B(X) = \sum_{l} \sum_{l'} \left(2\hat{b}_{ll'} - \frac{\hat{b}_{ll} + \hat{b}_{l'l'}}{2}\right) x_{ll'},$$

where the binary relation X satisfies the constraints of an equivalence relation on $L \times L$, that is:

- reflexivity: $x_{ll} = 1 \quad \forall\, l \in L$
- symmetry: $x_{ll'} - x_{l'l} = 0 \quad \forall\, (l,l') \in L$
- transitivity: $x_{ll'} + x_{l'l''} - x_{ll''} \leq 1 \quad \forall\, (l,l',l'') \in L.$

In this case, we choose a similarity based upon a presence-rarity index. We could have used a similarity that is much more common for classification matter: $b_{ll'} = \sum_i t_{li} t_{l'i}$ = number of IC codes shared by the patents l and l'.

This processing would group together the patents sharing a majority of codes because, in this case, the similarity between patents is not simply measured by the number of descriptors they share. This kind of classification would be relevant if we wanted to highlight the basic similarity between profiles, which means the main trends of the data. In this case, the more subtle phenomena would be neglected.

However, it seems important, according to the given problem, not to neglect the rare configurations. We must keep in mind the fact that we are dealing with typical distributions such as Zipf or Bradford distributions. It is worth noting that 80% of patents are registered by only 20% of registering companies. The index $\hat{b}$ allows us to take into account the less frequent similarities and to restitute their contribution in the solution. Thus two patents will belong to the same class not only because they cover the same domains, but also because these domains are seldom shared by other patents in the data set. Then interesting phenomena appear that would not have been detectable in the base data, such as some patents that do not possess any link at first glance, but prove later to combine similar technologies. Moreover, the obtained classification makes it possible for a business enterprise to position its own patents according to its competitors and at the same time to detect the patents that very much resemble its own. This kind of analysis, as a decision-making support tool, can make the expert's work easier, as it would minimize the risks of error and the possible oversights due to the *nonglobal vision* of the whole links network.

IC codes classification: From the table T, we can also derive the similarity matrix $\hat{C}$, $n \times n$, crossing the IC codes. Its general term is defined as:

$$\hat{c}_{ii'} = \sum_l \frac{t_{li} t_{li'}}{t_{l\cdot}}.$$

$\hat{c}_{ii'}$ is a presence-rarity index: Two codes are more similar (high value for $\hat{c}_{ii'}$) as they simultaneously describe patents ($t_{li} = t_{li'} = 1$) having few codes (low value for $t_{l\cdot}$).

As previously mentioned, the code classification is given by the maximization of the function:

$$C(Y) = \sum_i \sum_{i'} (\hat{c}_{ii'} - \bar{\hat{c}}_{ii'}) y_{ii'} \quad \text{or} \quad C(Y) = \sum_i \sum_{i'} \left(2\hat{c}_{ii'} - \frac{\hat{c}_{ii} + \hat{c}_{i'i'}}{2} \right) y_{ii'}$$

under the constraints of an equivalence relation on $I \times I$ for Y.

The obtained partition groups together IC codes that possess similar code profiles. These codes simultaneously appear in patents with few descriptors. This classification provides us with a vision of the current technical situation: Big classes are characteristic of great activity. The innovation detection is also a strategic point for any company. It is also an absolute necessity for its long-term welfare. Innovation phenomena are obviously detected in the low frequencies. That means small classes, having few connections with the others, will reveal information concerning innovation. This processing methodology is very useful for business enterprises that search for appropriate information for decision making or for the state of the art in certain domains. Moreover, the comparison of such classifications, over the years, would give a good panoramic view of the trends in the application of technologies.

4.2.2 *Patent assignee and IC codes.* The simultaneous exploitation of the two tables T and T' makes it possible to build matrices dealing with cross relations between the sets I and J.

Seriation patent assignees × IC codes, in terms of weight: The first matrix that comes to mind is the rectangular matrix, noted S, $n \times m$, crossing the IC codes and the patent assignees. The general term of the matrix is defined by the number of patents described by the code i and registered by the company j:

$$s_{ij} = \sum_l t_{li} t'_{lj}.$$

Matrix S characteristics:

- Row sums: the number of patents described by the code i:

$$s_{i\cdot} = \sum_j s_{ij} = \sum_j \sum_l t_{li} t'_{lj} = \sum_l t_{li} \sum_j t'_{lj} = \sum_l t_{li} = t_{\cdot i}.$$

- Column sums: the number of codes (with redundancies) describing the patents registered by the company j:

$$s_{\cdot j} = \sum_i s_{ij} = \sum_i \sum_l t_{li} t'_{lj} = \sum_l t'_{lj} \sum_i t_{li} = \sum_l t'_{lj} t_{l\cdot}.$$

At this point, it is important to mention that we take into account the *weight* of a company within a domain as the index S counts the tokens of codes, and not only their presence. Let us explain this remark with the help of a little example: Let the company j_0 register 3 patents l_1, l_2, and l_3. These patents are described by six codes: i_1, i_2, i_3, i_4, i_5, and i_6, as shown in tables T and T' (see Fig. 6): We have $s_{\cdot j_0} = 4 + 3 + 5 = 12$. The codes l_1 and l_2 are counted three times, the codes l_5 and l_6 twice, and the codes l_3 and l_4 once.

We could choose to focus on the presence of the companies in the different domains by means of weighting the IC codes according to their global presence. In that case we would speak about a company *spectrum* within a domain. This approach will be presented in the following section. We will submit matrix S to the seriation process. The data are neither binary nor derived from the sum of a certain amount of relations with an underlying notion of majority; thus the problem of the criterion choice for the seriation arises.

We chose a criterion that deals with the profiles of the company's weight in the different domains:

$$S_1(Z) = \sum_i \sum_j \left(\frac{s_{ij}}{s_{i\cdot}} - \frac{1}{m} \right) z_{ij},$$

where $s_{ij}/s_{i\cdot}$ = part of the company j within the domain i (compared with other companies) and $1/m$ is the arithmetic average of the values $s_{ij}/s_{i\cdot}$.

As we can see:

$$\frac{1}{nm} \sum_i \sum_j \frac{s_{ij}}{s_{i\cdot}} = \frac{1}{nm} \sum_i \frac{1}{s_{i\cdot}} \sum_j s_{ij} = \frac{1}{nm} \sum_i \frac{s_{i\cdot}}{s_{i\cdot}} = \frac{n}{nm} = \frac{1}{m}.$$

This criterion can be seen as a deviation from the mean: the higher $s_{ij}/s_{i\cdot}$ from the mean $1/m$, the more important the part of the company j. The seriation process based upon this criterion creates blocks mixing IC codes and companies that are the most representative of each other. Thus we get company classes with similar weights in the domain groups they are in correspondence with. Within a given block, the weight of a company over the concerned domains may vary from an important part up to exclusivity. This kind of processing makes it possible to position a company in relation to its competitors. This information is provided by a more or less important presence of the other firms within the technological domains this company covers.

	i_1	i_2	i_3	i_4	i_5	i_6	$t_{l\cdot}$	j_0
l_1	1	1	1	0	0	1	4	1
l_2	1	1	0	0	1	0	3	1
l_3	1	1	0	1	1	1	5	1

Fig. 6.

The parallel of this criterion for the exploitation of the weight profiles of the different companies is built in the following way:

$$S_2(Z) = \sum_i \sum_j \left(\frac{s_{ij}}{s_{\cdot j}} - \frac{1}{n} \right) z_{ij},$$

where $s_{ij}/s_{\cdot j}$ = part of the code i in the patents belonging to the company j (compared to the other codes) and $1/n$ is the mean of the values $s_{ij}/s_{\cdot j}$.

As we can see:

$$\frac{1}{nm} \sum_i \sum_j \frac{s_{ij}}{s_{\cdot j}} = \frac{1}{nm} \sum_j \frac{1}{s_{\cdot j}} \sum_i s_{ij} = \frac{1}{nm} \sum_j \frac{s_{\cdot j}}{s_{\cdot j}} = \frac{m}{nm} = \frac{1}{n}.$$

Thus this criterion can be seen as a deviation from the mean: the higher $s_{ij}/s_{\cdot j}$ from the mean $1/n$, the more the part of the domain i towards the patents belonging to company j.

As explained above, the indicator $s_{\cdot j}$ counts the basic tokens of the codes. Thus, the criterion $S_2(Z)$ will highlight the most common codes and isolate those that appear more seldom.

According to the goal of this study, the use of such a criterion is not very interesting. Therefore we chose another similarity index $\hat{S}$.

Patent assignee × IC codes seriation, in terms of spectrum: The matrix $\hat{S}$ is derived from the table T and T' as follows:

$$\hat{s}_{ij} = \sum_l \frac{t_{li} t'_{lj}}{\sum_i t_{li}} = \sum_l \frac{t_{li} t'_{lj}}{t_{l\cdot}}.$$

$\hat{s}_{ij}$ is a presence-rarity index: a code i and a company j are more closely linked (high value for $\hat{s}_{ij}$) as they simultaneously appear in patents references ($t_{li} = t'_{lj} = 1$) having few codes (low value for $t_{l\cdot}$)

Now we consider the IC code frequency as a basic item of information. The result of the seriation will take into account the presence of a company in a domain, even if this presence is not prominent.

Characteristics of $\hat{S}$:

- Row sums

$$\hat{s}_{i\cdot} = \sum_j \hat{s}_{ij} = \sum_j \sum_l t_{li} \frac{t'_{lj}}{t_{l\cdot}} = \sum_l \frac{t_{li}}{t_{l\cdot}} \sum_j t'_{lj} = \sum_l \frac{t_{li}}{t_{l\cdot}}$$

  $\hat{s}_{i\cdot}$ is higher as the code i appears ($t_{li} = 1$) with few other codes in the patents references (low value for $t_{l\cdot}$). The most frequent codes will not "take up" the whole information item. Thus it will be possible to bring about an illustration of rare phenomena.

- Column sums: number of patents registered by company j.

$$\hat{s}_{\cdot j} = \sum_l \hat{s}_{lj} = \sum_i \sum_l \frac{t_{li} t'_{lj}}{t_{l\cdot}} = \sum_l \frac{t'_{lj}}{t_{l\cdot}} \sum_i t_{li} = \sum_l t'_{lj}.$$

The criterion can then be chosen as shown in the previous section. In any case, the solution will provide us with a correspondence between activity domains and companies in terms of presence, but not in terms of main coverage. This information is useful as far as variety is concerned, that is, when the range of technologies covered by a company is meaningful.

4.2.3 *Summary.* Figure 7 summarizes the different crossings we can build from the three sets and the structures of the relational matrices.

- Direct classification perspective:
  1. Patents classification (on B or $\hat{B}$)
  2. IC codes classification (on C or $\hat{C}$)
- Cross classification perspective:
  3. Patents seriation × IC (on T)
  4. IC codes seriation × patent assignees (on S or $\hat{S}$).

Each of these problems, presented in a data analysis perspective, corresponds to an information analysis concern, in the perspective of technology watch:

1. to define patents families,
2. to highlight the relationships between different domains of fundamental and application research,
3. to show the relationships between patents and application domains, and
4. to determine the common or specific research strategies of competitive companies.

5. FURTHER DEVELOPMENTS

In the previous parts we have studied the relations between several fields such as patent assignee, patent number, and international classification codes describing the patents. It is of course possible to implement the same methodology to analyze relations between other descriptive fields. The results will provide us with another kind of strategical information. At least, any conventional information restructured under a conventional form could be analyzed the same way.

Among the various concerns in a technology watch approach, we think that two kinds of problems are also very interesting: the choice of the extension countries and the comparative analysis between company internal classification and international patent classification.

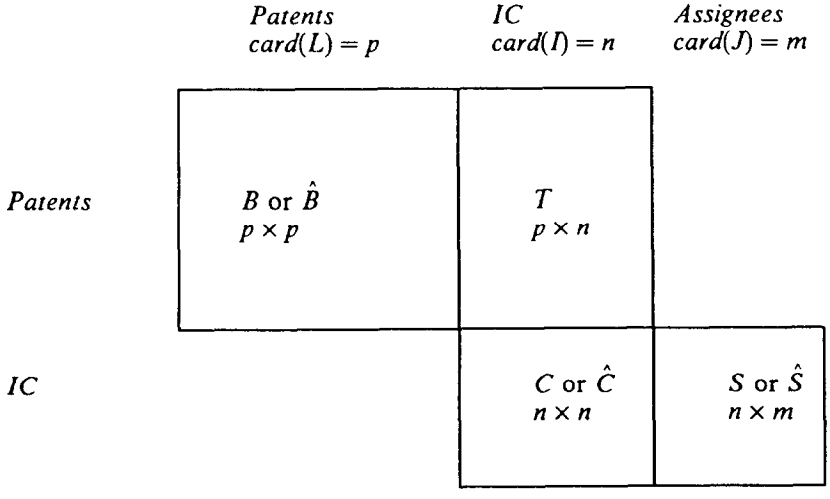

Fig. 7.

When an industrial firm registers a patent in its country, this firm can use, during a period of 12 months, its priority registered rights to extend its patent to other countries, members of The International Union for the Protection of the Industrial Property (Union of Paris). After this process, the invention is protected in the countries where the deliverance was issued.

This information, which is mentioned in the patent reference, is very interesting because it provides us with a good idea of the international coverage of the companies. Then we can imagine the analysis of the cross relationships between patent assignees and extension countries to highlight some foreign strategic axis of business enterprises.

Company or state laboratories working on leading fields or pathbreaking domains are frequently not satisfied with the IC codes. Thus they create their own internal classification of science or technologies trying to keep a certain correspondence with the official classification. This task is time consuming and prone to error when manually done. The classification methodology described in this paper could be of great help in this perspective.

6. CONCLUSION

In this paper we have presented the theoretical aspects of a data analysis methodology, specially seen from the database content-analysis point of view. We tried to show the main methodological points that allow for the use of these tools to extract information from raw data. The relational analysis approach has already successfully been applied in the context of technology watch surveys through patent record analysis, in different domains such as cosmetics, biology, and household appliances. One can find (Rostaing *et al.*, 1993) a study conducted on a set of patents in the medical field showing the different steps of the database mining process.

Acknowledgement – The authors are grateful to the referees for helpful comments on an earlier version of this paper.

REFERENCES

Albert, M., Avery, D., Narin, F., & Allister, P.M. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy*, *3*, 251–259.

Bédécarrax, C., & Warnesson, I. (1989). Relational analysis and dictionaries. *Applied Stochastic Models and Data Analysis*, 131–151.

Brockhoff, K. (1992). Instruments for patent data analysis in business firms. *Technovation*, *1*, 41–59.

Courtial, J., & Callon, M. (1991). Indicators for the identification of strategic themes within a research programme. *Scientometrics*, *3*, 447–458.

Dou, H., Quoniam, L., & Hassanaly, P. (1991). The scientific dynamics of a city: A study of chemistry in Marseille from 1981 to the present. *Scientometrics*, *1*, 83–93.

Griliches, Z. (1990). Patent statistics as economic indicators: A survey. *Journal of Economic Literature*, *4*, 1661–1707.

Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., & Vanhoutte, A. (1989). Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing & Management*, *3*, 315–318.

Huot, C., Quoniam, L., & Dou, H. (1992). A new method for analysing downloaded data for strategic decision. *Scientometrics*, *2*, 279–294.

Marcotorchino, F. (1986). Maximal association as a tool for classification. *Classification as a tool for Research*, 275–288.

Marcotorchino, F. (1987). A unified approach of the block-seriation problems. *Applied Stochastic Models and Data Analysis*, *2*.

Marcotorchino, F. (1991a). L'analyse factorielle-relationnelle: Parties I et II. *Etude du CEMAP IBM France.*

Marcotorchino, F. (1991b). Seriation problems: An overview. *Applied Stochastic Models and Data Analysis*, 139–151.

Michaud, P. (1982). Agrégation à la majorité I: Hommage à Condorcet. *Etude du Centre Scientifique IBM France.*

Mogee, M. (1991). Using patent data for technology analysis and planning. *Research Technology Management*, *4*, 43–49.

Parisot, P. (1985). Application of similarity aggregation techniques to a population of paraplegic patients. *Applied Stochastic Models and Data Analysis*, *1*, 35–54.

Rostaing, H., Nivol, W., Quoniam, L., Bédécarrax, C., & Huot, C. (1993). L'exploitation systématique des bases de données: Des analyses stratégiques pour les entreprises. *Les cahiers de l'ADEST.*

Warnesson, I., & Bédécarrax, C. (1988). Optimization of monolingual and bilingual dictionaries. *Proceedings of the ELS Conference on Computational Linguistics.*