Knowledge Discovery in Patent Databases

Konstantinos Markellos[1,2]
kmarkel@cti.gr

Penelope Markellou[1,2]
markel@ceid.upatras.gr

George Mayritsakis[1,2]
mayritsa@ceid.upatras.gr

Katerina Perdikuri[1,2]
perdikur@ceid.upatras.gr

Spiros Sirmakessis[1,2]
syrma@cti.gr

Athanasios Tsakalidis[1,2]
tsak@cti.gr

[1] Multimedia, Graphics and GIS Lab, Computer Engineering and Informatics Department, University of Patras, 26500 Patras, Greece

[2] Research Academic Computer Technology Institute, Internet and Multimedia Technologies Research Unit, 61 Riga Feraiou Str., 26221 Patras, Greece

ABSTRACT
In our days scientific, business and personal databases are growing in an exponential rate. What is truly valuable in large databases is the knowledge that can be extracted from the stored data. Knowledge discovery in Patent Databases was traditionally based on manual analysis carried out from statistical experts. Nowadays the increasing interest of many actors have led to the development of new tools for discovering and exploiting information related to technological activities and innovation, "hidden" in Patent Databases. In this paper we present a system that combines efficient and innovative methodologies and tools for the analysis of patent data stored in many international databases and the production of scientific and technological indicators.

Keywords
Knowledge Discovery, Text Mining, Patent Databases, Linguistic Preprocessing, Correspondence Analysis, Cluster Analysis.

1. INTRODUCTION
The general purpose of Knowledge Discovery is to "extract implicit, previously unknown and potentially useful information from data" [10]. Due to the continuous growth of the volume of data stored in scientific, business or personal databases, automated knowledge discovery techniques become more and more necessary. In addition, as the usual Data Mining techniques [7] are essentially designed to operate on structured databases, specific techniques, called Text Mining techniques, have been developed to process the available information that can be found in unstructured textual data. Text Mining therefore corresponds to the extension of the more traditional Data Mining approach to unstructured textual data and is primarily concerned with the extraction of information implicitly contained in collections of documents.

Patent Databases consists of text documents that describe a patent technology and its applications [16]. The traditional method of extracting knowledge from Patent Databases was based on manual analysis carried out from experts. Nowadays this method is impractical as Patent Databases grow exponentially. In addition, more and more R&D planners, business analysts, patents analysts, national and international patent offices, economic organizations, national statistical offices, venture capitalists and industrial bodies with high scientific and technological activity have increased their interest in discovering and exploiting information related to technological activities and innovation, "hidden" in Patent Databases.

In this paper we present the methodology used from a system that combines efficient and innovative tools for the analysis of textual data related to patents. The system uses existing Patent Databases (input), supports multidimensional analysis and produces new indicators (output). These indicators express information concerning the scientific and technological progress and can help the active actors (individuals or organizations) to understand the on-going changes and their effects.

The structure of this paper is the following. In section 2 we discuss the data related in the analysis of patents, while in section 3 we present the proposed methodology. Section 4 refers to the system architecture, while a case study with produced indicators is presented in section 5. Finally, in section 6 some conclusions are drawn with respect to the exploitation of the experience from data mining systems to improve the design and effectiveness of Patent Databases and how we can better interpret and visualize the extracted knowledge even for non-experts users.

2. MINING PATENTS
The analysis of the information "hidden" in patents, which are stored in many international databases, can provide a very clear view of the current trends regarding technological and scientific innovation [6], [13], [14]. However, in order to exploit information stored in Patent Databases more effectively, it is

Conference '00, Month 1-2, 2000, City, State.

necessary to develop efficient and innovative methodologies and tools for the analysis of existing information, as well as for the visualization [17], [9] of the extracted knowledge related to the technological innovation. In this section we give a brief description of patent data and Patent Databases. Consequently we introduce the main idea of the proposed methodology that will be presented in more detail in the following section.

Generally a patent is a legal title granting its holder the exclusive right to make use of an invention for a limited area and time by stopping others from, amongst other things, making, using or selling it without authorization [8]. The patent applicant has to provide a detailed technical description of its invention but also mention the points that render it an original application with innovative elements. A patent can be decomposed and described by several fields. Each field contains specific information while each patent is described by a code depicting its technical characteristics. These codes are given to patents due to the International Patents Classification system (IPC) or other classification systems. The fields contained in each Patent Database may differ between them. In the table below we present the fields describing a patent as listed in the Database of Patents: ESPACE ACCESS. We should also mention that patent documents can be either retrieved from on-line Patent Databases, or Patent Databases available on CD-ROMs.

Table 1. Fields describing a patent

PN: Priority Number (Number of the patent)
AN: Application Number
PR: Priority Year
DS: Designated States
MC: Main Classification Codes
IC: All Classification
ET: English Title
FT: French Title
IN: Inventor
PA: Applicant (Name of the company depositor)
AB: English Abstract
AF: French Abstract

Although all patent applications and granted patents are published, there is no uniform classification on a worldwide level. According to the International Patents Classification system (IPC), the patents are classified according to the technology related to the invention. IPC is designed so that each technical object to which a patent relates can be classified as a whole. In fact, inventions are classified by one or more symbols so that patents belonging to a technological field can be filed and retrieved. This classification follows a hierarchical structure. The sections themselves are then subdivided into sub-sections, classes, sub-classes, groups and sub-groups. Each subgroup may be further subdivided.

In a simple analysis approach of patent data we can use several of the fields mentioned above and produce reports for the patenting activity per field. In a more sophisticated approach we should be able to combine the above fields and produce indicators that measure the scientific and technological innovation. For this reason in our methodology we propose a multidimensional analysis of patent data, which could support multidimensional comparisons between the countries or the sectors but also, allow identifying competition within the same technological sectors.

For this reason we create homogeneous clusters of patents based on data analysis and classification techniques. The production of homogeneous clusters is based on textual analysis methods of the text describing a patent. More specifically textual analysis techniques are applied on the titles and abstracts describing a patent. The output of textual analysis is a frequency table, which associates the occurrences of words in a set of patents. Using the word frequencies of a patent as a vector we represent patents as points in a high dimensional space in which each word represents a dimension. Representing the patents, allows visualizing their relative positions and groupings, which are indicative of various implicit relationships and can serve as a basis for the analysis of technological innovation. However because of the intrinsic complexity of any interpretation within high dimensional vector spaces, we use factor analysis (more specifically the methodology of correspondence analysis), in order to find lower- dimensional subspaces, which approximate the original distributions of points, without a certain loss of information. Moreover the use of correspondence analysis ensures the robustness of the cluster analysis, which follows.

The basic objective in cluster analysis is to discover natural groupings of the patents according to a predefined similarity criterion, which expresses the distance measure among patents. In our case the aim is to look for families of patents characterized by their similarities in terms of shared technologies [15]. Each cluster represents a technological field and the relationships between the different clusters provide useful information concerning the interactions that exist between various domains of technological activity and the poles of innovation that exist inside these domains.

Although fields of patents, such as the companies submitting the patents, the inventors, the countries in which the patent is submitted etc., are not involved as variables in the main part of the analysis they could be used as complementary analysis variables enriching the results of the classification. This approach enables comparisons between countries, companies and specific sectors, as well as the extraction of other useful conclusions. Competitive indicators regarding the competitive level of each country are extracted, as well as countries that are actives in specific technological domain, etc.

3. PROPOSED METHODOLOGY

As already mentioned the main objectives of the proposed methodology are:

- to support multi-dimensional comparisons between countries, sectors or companies;
- to identify competition in a given sector;
- to capture interactions that might exist between domains of technological activity and poles of innovation inside these domains.

In the following sub-sections the used statistical techniques are described.

3.1 Linguistic Preprocessing

Linguistic processing refers to the problem of understanding the meaning of a text string [5], [2]. Its main aim is to transform the raw textual data of patents into a format suitable for statistical analysis (e.g. correspondence analysis) or for clustering procedures. So, this step is of high importance for further analysis of patents.

Linguistic preprocessing is an automatic procedure of our system. The output of the linguistic preprocessing gives useful information about the textual data by summarizing it in frequency tables (contingency tables). The user has the opportunity to see different views of these frequency tables (e.g. ordered by frequencies or alphabetically, etc.) and intervene by defining different parameters. For example he can define a threshold in the frequency of the used words or take into account only words with frequency greater or less than a predefined number. So, he has the ability to involve in the analysis frequent or rare words depending on the information he wants to extract, define synonyms and equivalent words or exclude specific words from the analysis. This permits to handle even data with very specific words that are not used in usual dictionaries. In this way the system can be considered as a powerful tool for textual analysis.

These modifications do not affect the original data. Therefore the user is able to reanalyze them by defining from the beginning new rules without loosing any part of the information.

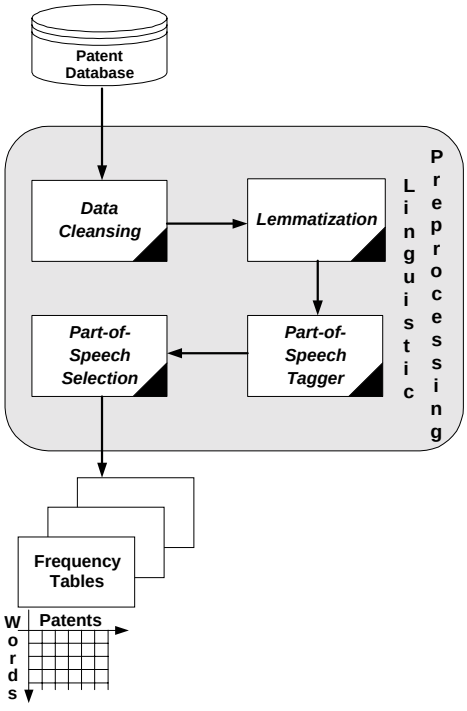

Figure 1. Linguistic preprocessing

As previously mentioned the output of this linguistic preprocessing is a frequency table with the selected patents in rows and the corresponding frequency of each selected word/lemma in columns.

Specifically, the steps of the linguistic preprocessing are:

- **Data cleansing:** cleans the input data by removing irrelevant html characters and punctuation characters.
- **Lemmatization:** focuses in restricting the morphologic variation of the textual data by reducing each of the different inflections of a given word form to a unique canonical representation (or lemma).
- **Part-of-speech tagging:** automatically identifies the morpho-syntactic categories (noun, verb, adjective) of words in the documents. The non-significant words can be filtered on the basis of their morpho-syntactic category. The part-of-speech tagging runs on the textual content i.e. on titles and abstracts of the patents.
- **Part-of-speech selection:** to further reduce the vocabulary size, the user has the ability to select the word categories as identified by the assigned parts-of-speech and restrict the analysis to specific word categories (i.e. nouns, verbs, adjectives). Moreover, the user can select words that will not be involved in the subsequent steps of the analysis, and create synonyms.

3.2 Correspondence Analysis

Correspondence analysis is a descriptive/exploratory technique for analyzing simple two-way or multi-way tables containing some measure of correspondence between the rows and columns [3]. This multivariate method explores cross-classified data by finding low dimensional geometrical representations and related numerical statistics [11]. This technique is adapted in the contingency tables and produces results in order to study eventual relations between the incorporated variables. This way features in the data are revealed without assuming any underlying distributions of the data.

Specifically, this method allows analyzing and describing graphically and synthetically big contingency tables in which can be found the number of individuals who share the same characteristic (intersection of the row and the column).

From a geometrical point of view correspondence analysis provides a method for representing data in a Euclidian space so that the results can be visually examined for structure. For data in a typical two-way contingency table both the row variables and the column variables are represented in the same space. This means that one can examine relations not only among row or column variables but also between row and column variables. Correspondence analysis is used for finding subspaces to represent proximities among the rows and columns of the contingency table.

Given that we have a frequency table the rows/patents can be viewed as n points in a p-dimensional space. When points are neighboring, this means that the two patents with the corresponding row-profiles are similar. Respectively, for the column/word profile we have a representation of p points in a n-dimensional space.

In particular, it is of central concern to be able to relate the observed relative position with the content of the corresponding patent, in other words with the underlying vocabulary used in the title and abstracts.

In other words, it becomes obvious that the profiles, which are a series of n or p points, depending on whether we are dealing with

rows or columns are used to define points in spaces of p or n dimensions. In order to identify the points that are close to each other we have to compute the distances between them. As it is quite difficult to actually perform any visualizing in more than two dimensions, it is very important to decide how the dimensions (i.e. factors) of the resulting factor space should be grouped.

3.3 Cluster Analysis

Cluster analysis is a multivariate procedure for detecting natural groupings in data [1]. Cluster analysis classification is based upon the placing of objects into more or less homogeneous groups, in a manner such that the relationship between groups is revealed [12].

The particularity of this clustering procedure consists in applying a hierarchical cluster analysis for the totality of elements, which however are characterized by the factorial coordinates created by the factor analysis that has previously been performed. Specifically a hierarchical clustering is performed in the factorial coordinates of the patents and a hierarchical tree is generated. Using an automatic procedure for the selection of the optimal cut of the hierarchical tree the user determines the number of cluster and consequently which patents are grouped under the same cluster.

In more detail we have n patents to be clustered as points in a Euclidean space with p dimensions. Each patent is represented by a set of coordinates that correspond to the factorial axes used. As a consequence each point x is a vector with p components. We start the clustering process by assigning each point to each own cluster (if we have n points, now we have n clusters). Furthermore we compute the distances between the points. Basically we need to calculate n(n-1)/2 distances between the n points. We find the closest (or most similar) pair of points and merge them into a single cluster (that is we have now n-1 points to be clustered). Each new cluster represents a node in the produced hierarchical tree, while the new cluster is represented as a point in a Euclidean space with p dimensions. Its coordinates correspond to the center of gravity of the merged points.

The above procedure is repeated until we regroup all the points in one single partition. The graphical representation through the hierarchical tree illustrates the iterative process more powerfully, while the cut of the tree determines the final number of clusters kept for the production of the indicators.

4. SYSTEM ARCHITECTURE

The system is based on a three-module architecture. The system's functions integrates the following modules for its operation:

- **Textual analysis.**

  The first module reads the patents data from the selected Patent Database and transforms them into the appropriate format, in order to be ready for further processing. The system database accepts textual and numerical data, in our case i.e. patents described by a set of fields. There is no restriction in the database from which will be taken the patent data. According to the user requirement analysis ESPACE ACCESS and ESPACE Bulletin are among the databases most frequently used. Therefore we used these databases for explaining the statistical methodology through application to real data. The flexibility of the system to different input databases is achieved through the adoption of specific input formats. Although data from different databases may be used depending on the problem and the information one wants to extract, they are each time standardized to a uniform representation in the system database. The parser is used for reading the textual data and consequently linguistic preprocessing is applied on them (this includes data cleansing, lemmatization, part-of-speech tagging and part-of-speech selection). The use of a dictionary and of a grammar is necessary for the linguistic processing

- **Statistical analysis.**

  The second module focuses in the available tools for analysis and processing. In particular, it applies textual analysis methods on the pre-formatted data, in order to extract valuable information and create the first groups of patents. The user can select complicated statistical methods such as the correspondence analysis in order to analyze the patent data. For this kind of analysis it is necessary, as already mentioned, to have previously performed the linguistic preprocessing of the textual data. The input data is the contingency table where the rows contain the patents, the columns represent the words, and each cell gives the frequency of the specific word in the corresponding patent. This analysis enables to explore the non-random dependencies between the variables involved in it and the vocabulary obtained from the title and abstracts of the patents. More precisely, the correspondence analysis produces a new vector space in which similarities between the rows and the columns of the input contingency table (as measured by the χ^2-distance) can be visualized as geometric proximities. Moreover, bootstrap techniques test the stability of the results obtained from the correspondence analysis. The user has also the ability to select different parameters in order to adapt the results in his needs. Of course different statistical measures are produced in order to help the user interpret the results. The clustering analysis is considered one of the important features of the system since it is basis for the derivation of technology indicators. It allows obtaining a clear view of the technology on-goings. The clustering procedure depending on the number of the data is distinguished in hierarchical clustering or k-means and hierarchical clustering. In the second case the hierarchical clustering is performed in the k clusters obtained from the k-means procedure. In addition in both cases the cluster analysis is performed in the factorial axes derived from the correspondence analysis. In the case of textual data, clustering techniques are used for representing the proximities between the elements of the lexical tables. In the general case, cluster analysis operates on contingency tables to identify relationships between the two different nominal variables. In the case of patent data, the aim of the procedure is to identify groups of patents that share common vocabulary and groups of patents that share common technologies in order to derive conclusions about technological trends and innovation. Especially for the clustering procedure we should mention the

production of the relationship map that demonstrates the relationships between the clusters or in other words the relationships between the different areas of technology. The technology indicators [4] are also based on the clustering procedure and constitute an important characteristic of the system. These are produced for each cluster separately and permit to identify the technology on-goings in different areas of technology. Furthermore these indicators are categorized in four different levels depending on whether they refer to the sector of technology (through IPC codes), the country or the continent, the assignees or the inventors and finally time (due to the priority year or other).

- **Visualization.**

  The third module is responsible for the extraction of the results and their visualization. This procedure is very important for the user in order to fully understand their meaning. Therefore many options from graphs and tables to ready-made reports are available. The interactivity is an important feature of the system giving to the user the opportunity to intervene in the outputs and adapt those to his real needs. Changes in the colors, types of lines, types of graphs (2-dimensional, 3-dimensional, etc.), fonts are supported so that one elaborates the results. Furthermore, the positioning of the graphs in the space is supported by the system, giving to the user the opportunity to have the optical view he considers appropriate for the visualization of the results.

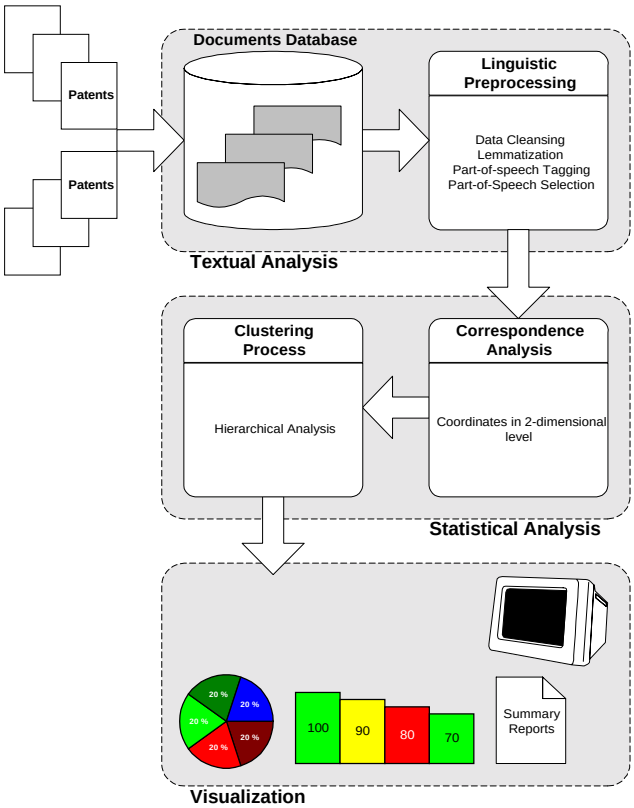

Figure 2. System architecture

Fig. 2 depicts the system architecture and its methodological scheme.

The modular approach ensures system's flexibility and openness to future changes and modifications e.g. addition of new statistical methods or techniques. Moreover the system can handle dynamically all the available information, regarding patents and indicators. Special consideration was given to user friendliness, interactivity and interoperability. In the following sub-sections the basic functionalities for each module are described.

The different flows are also depicted in the figure, defining the different stages of the analyses and the connections between them. It is also common sense that the natural sequence followed by the flows should be respected for the correct operation and for the robustness of the results.

5. CASE STUDY

In this section we present some of the indicators, our system produced from patents classified in the category "Wind Motors". Patent data were retrieved from the ESPACE ACCESS Database using the MIMOSA search engine.

Based on the percentages of patents that each country (either in Europe or worldwide) applied for, we can identify the more active countries or even continents (Figure 3). Another indicator refers to the evolution of patents over time for different areas of technology. More specifically Figure 4, permits to identify areas of technology that nowadays are still at peak or in the contrary, find out those areas of technology that over years have presented a course that declines more and more over the last years.

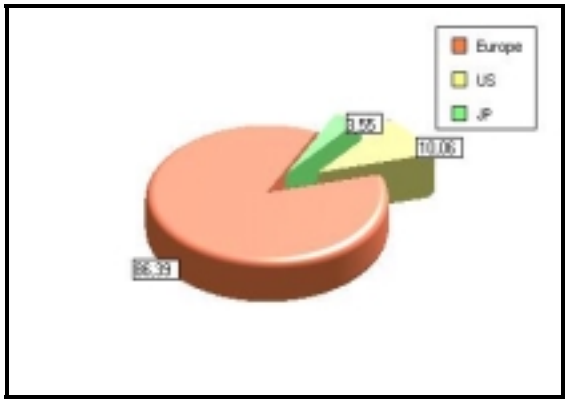

Figure 3. Distribution of applied patents per continent

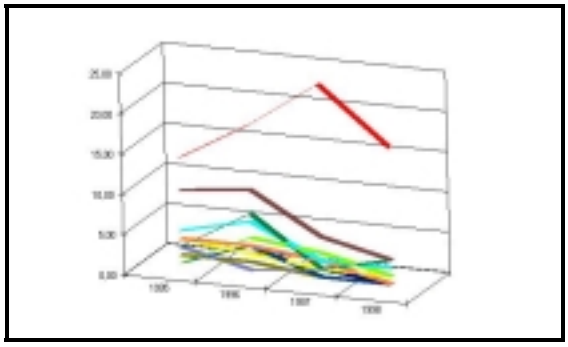

Figure 4. Distribution of patents technology per year

Furthermore another indicator allowing comparisons between countries patenting activity per year and per cluster is presented graphically in Figure 5.

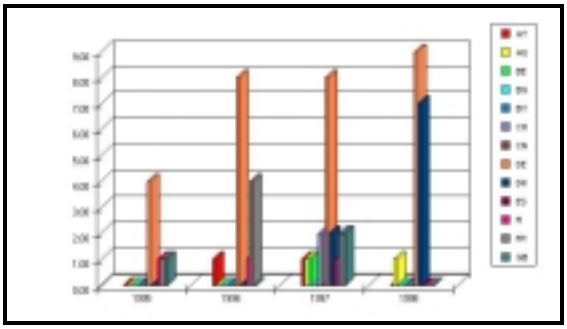

Figure 5. Evolution of countries patenting activity per year in cluster 1

6. CONCLUSIONS

The developed system permits to analyze patent data based on multidimensional analysis techniques that make use of all the information describing a patent. This approach enables to capture information not only at the level of a technological sector but also perform comparisons at the level of a country or at the level of a set of sectors. In addition it permits to identify technological trends and innovation.

Taking into consideration the produced indicators and the knowledge extracted from Patent Databases we could propose some improvements in the organization and management of Patent Databases, as a future work.

7. REFERENCES

[1] Alderferer, M.S., Blashfield, R.K. Cluster Analysis, Beverly Hills, CA., Sage Publications, Inc., 1986.

[2] Beaugrande, R., Dressler, W. Introduction to Text Linguistics, London Longman, 1981.

[3] Benzecri, J.P. Correspondence Analysis Handbook, New York: Marcel Dekker, 1992.

[4] Chappelier, J., Peristera, V., Rajman, M., Seydoux, F. Evaluation of Statistical and Technological Innovation Using Statistical Analysis of Patents, JADT 2002.

[5] Ciravegna, F., Lavelli, A., Pianesi, F. Linguistic Processing of Texts Using Geppetto, Technical Report 9602-06, IRST, Povo TN, Italy, 1996.

[6] Comanor, W.S., Scherer, F.M. Patent Statistics as a Measure of Technical Change. Journal of Political Economy, pp. 392-398, 1969.

[7] Edelstein, H. Introduction to Data Mining and Knowledge Discovery. Two Crows Corporation, 1999.

[8] EPO - European Patent Office. http://www.european-patent-office.org/index.htm

[9] Fayyad, U., Grinstein, G., Wierse, A. Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann Publishers, 2001.

[10] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communications of the ACM, November 1996, Vol. 39, No. 11, pp. 27-34.

[11] Hill, M.O. Correspondence Analysis: a Neglected Multivariate Method. Journal of Applied Statistics, Vol. 23, No. 3, pp. 340-354, 1974.

[12] Lewis, S. Cluster Analysis as a Technique to Guide Interface Design, International Journal of Man-Machine Studies, 35, pp. 251-265, 1991.

[13] Narin, F. Patents as Indicators for the Evaluation of Industrial Research Output. Scientometrics, 34, 3, pp. 489-496, 1995.

[14] OECD. The Measurement of Scientific and Technological Activities Using Patent Data as Science and Technology Indicators. Patent Manual, 1994.

[15] Rajman, M., Lebart, L. Similarities for Textual Data. In 4th International Conference on Statistical Analysis of Textual Data (JADT'98), Nice, 1998.

[16] Schmoch, U., Bierhals, R., Rangnow, R. Impact of International Patent Applications on Patent Indicators. JOINT NESTI/TIP/GSS WORKSHOP, Room Document No. 1, 1998.

[17] Thearling, K., Becker, B., DeCoste, D., Mawby, B., Pilote, M., Sommerfield D. Visualizing Data Mining Models. Published in Information Visualization in Data Mining and Knowledge Discovery, edited by Usama Fayyad, Georges Grinstein, and Andreas Wierse. Morgan Kaufman, 2001.