STING: EVALUATION OF SCIENTIFIC & TECHNOLOGICAL INNOVATION AND PROGRESS IN EUROPE THROUGH PATENTS

Spiros SIRMAKESSIS, Konstantinos MARKELLOS, Penelope MARKELLOU, Giorgos MAYRITSAKIS, Katerina PERDIKOURI, Athanasios TSAKALIDIS

Computer Technology Institute, Research Unit 5: Internet and Multimedia Technologies,
61 Feraiou Str., 26221 Patras, Greece
Tel: +30 61 960335; Fax: +30 61 960322;
E-mail: {syrma, kmarkel, markel, mayritsa, perdikur, tsak}@cti.gr

Georgia PANAGOPOULOU

National Statistical Services of Greece, IT Division
Agisilaou 43-45, 10166, Athens, Greece
E-mail: panag@statistics.gr

Keywords: textual analysis, data mining, hierarchical classification, patents, S&T indicators

INTRODUCTION

Nowadays, the increase of the volume of data stored into different computer systems is so huge that only one extremely reduced portion of these data (typically between 5 to 10%) can be effectively analyzed. The use of techniques of automatic analysis allows us to valorize in a more efficient way the potential wealth of information that the textual databases represent.

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

In general, the analysis and comparison of the scientific and technological activity between countries and/or enterprises, with the help of brevets, is made through a priori classifications and different types of indicators. Describing them in this way allows scientists to free 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.

In this paper, we give an answer to the problem of the lack of the brevets' multidimensional analysis. All this study is based on the principles and methods of textual analysis. This type of approach allowed us to manage in an elegant way data that are difficult to use, such as the patents. The aim is to show effectively how the use of methods, like the hierarchic classification, makes possible to 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, in order

to get all necessary information required to find the position of the European industry in the international environment.

This type of analysis and decision-making assist 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 when not estimating in their entirety the links and relations between patents.

A classificatory approach is therefore the most suitable, 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 with other patents.

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

Of course, many problems (algorithmic or 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 disciplines has gained today a sufficient critical volume, in order to permit the realization of the performed techniques and effectively introduce the application of results. In this paper we present the way in which we use the international patent classification hierarchy, in order to produce meaningful results, as well as the methodology applied to clustering. The results of the study are also presented.

METHODOLOGY FOR THE ANALYSIS OF PATENTS

The proposed methodology provides a multidimensional analysis of patent data. This enables multidimensional comparisons between the countries and/or the sectors, while it also allows identification of competition within the same sector. Interactions that exist between various domains of technological activity and the poles of innovation that exist inside these domains are captured. The proposed technique of automatic analysis allows exploiting the information stored in patent databases, in a more effective way. The innovative character of the methodology proposed in this project consists in the fact that it does not only use the first digits of the IPC code of a patent, but also any other information that is stored in the classification record of a patent. This approach enables comparisons between countries, companies and specific sectors, as well as the extraction of other useful conclusions. Furthermore, through this approach we have the ability to analyze scientific and technological innovation and progress in Europe at three different levels.

- Firstly, analysis can be performed at the level of a sector. This means that a specific sector can be isolated in order to be analyzed and consequently identify technological evolution, which is not very evident. More specifically, for the analysis we use the IPC codes and/or the text describing a patent in order to create homogeneous

classes from which technological trends can be identified and derive useful information. Variables that are not involved in the main part of the analysis, such as the companies submitting the patents, the inventors, the countries in which the patent is submitted etc. would enrich the result of the classification as complementary analysis variables.

- The same analysis could be performed at the level of all sectors. This would enable us to obtain information about scientific and technological activity per selected sector with relation to the set of companies submitting patents in this specific sector, for a specific year or for several years.
- Furthermore, analysis at the level of a country is available. Through this analysis, homogeneous groups of countries are formed, which show the progress of the scientific and technological analysis in the different countries of Europe. Indicators regarding the competitive level of each country are extracted, as well as listings of the countries that are active in specific technological domain, etc.

The kind of information described above is produced by data analysis methods and more specifically by classification techniques. The production of homogeneous clusters is based on textual analysis methods, a brief presentation of which will be provided next in this document.

The basic steps of the analysis are presented below:

- Textual analysis of the IPC codes and the texts describing the patents.
- Multiple Factor analyses based on boostrap techniques for controlling the stability of the method.
- Creation of homogeneous classes of patents based on their codes and texts that describe the patents, using the methods of correspondence and cluster analysis. The objective of this stage to look for families of patents characterized by their similarities in terms of shared technologies.

Textual Analysis Techniques

As already mentioned, textual analysis techniques form the basis of the developed methodology for the analysis of patents. Textual analysis of the codes and titles or abstracts describing a patent is applied. Some basic concepts of textual analysis are presented, in order to have a complete view of the undertaken statistical methodology for the exploitation of patent data.

In every statistical problem the following procedure is followed: firstly, we identify the nature of the problem and formulate it according to specific statistical or probabilistic models. The type of data may lead to various types of analysis. A pre-processing of the data may be necessary before involving them in any kind of analysis. Also, the pre-processing phase may include the testing of hypotheses or models. Finally, for the interpretation of the results many activities may be necessary, such as a critical evaluation of the hypotheses and models etc. However, the need for analysing texts is increasing in many fields

of scientific research. Statistical methods rely on measurements and counts based on objects that are to be compared. In the case that textual units are under consideration, these should be analysed through the use of discrete, qualitative variables, such as the counts, rather than with variables of continuous nature. Generally, the statistical analysis of texts is quite complicated since every text possesses a sequential or syntagmatic dimension; therefore its formulation is too complex. In addition the relationships between words of a text should be taken into account and could be brought into light through counts. Counting elements and adding them together means that they are treated as identical occurrences. Before performing any kind of analysis, textual data should be decomposed in simpler lexical units. Several methods of lexicometric processing exist that help identifying specific units of the text upon which counts are carried out. In order to choose units from the text the procedures of segmentation, identification, lemmatisation, disambiguation can be applied. Segmentation consists in subdividing the text into minimum units i.e. units that are not to be subdivided further. Then follows the phase of identification i.e. the grouping of identical units. Once the segmentation unit is defined then textual statistics methods are applied. Furthermore, the elements resulting from segmentation can be lemmatised. In this case should be established identification rules so that words arising from the different inflections of a lemma being grouped together. The main steps of a lemmatisation are the following:

- Verb forms are out into the infinitive.
- Nouns are put into the singular.
- Elisions are removed.

Automatic lemmatisation requires a disambiguation, including a morpho-syntactic analysis and some times a pragmatic analysis. In some cases, in order to have a systematic determination of the lemma to which a form in a text belongs, a prior disambiguation is required. This need for disambiguation arises in many cases such as: when a given unit corresponds to inflectional forms of different lemmas, or when units of the same etymological source exist. In other cases ambiguities concerning the syntactic function of a word have to be removed, requiring a grammatical analysis of the sentence containing it. So, this procedure assists the automatic identification of the morpho-syntactic categories (noun, verb, adjective, etc.) of the words in the documents. In this way, we can filter non-significant words on the basis of their morpho-syntactic category.

In addition, there are cases where the meaning of words is closely related to the way they appear together. In these cases, the identification of repetitive segments may be another solution for identifying lexical units that could be used in the statistical analysis. Textual data analysis, when used for extracting information from documents relies on a cluster analysis based on the chi-square distance between the lexical profiles. For each of the resulting clusters characteristic words, i.e. words with a frequency in the cluster significantly higher than the one expected according to a predefined probabilistic model, are

extracted. Each of the clusters is then represented by a characteristic document, which is the document in the cluster that contains the most characteristic words. After the pre-processing of the data and in order to proceed with the analysis, several types of methods can be applied for obtaining useful results. Firstly, it should be mentioned that computer-based processing of textual data is greatly simplified by applying a technique called numeric coding. This technique consists in giving a numeric code to each word involved in the analysis. This code is associated with each occurrence of the word. Then a quantitative analysis of the vocabulary is performed. In fact, a table of frequencies of each word is obtained, on which the next steps of the analysis of data will be based. In addition, indexes and concordances of the data can be produced in order to provide a different perspective from the one obtained when having a sequential reading of the corpus. Consequently, an alphabetic index can be obtained, where words that participate in the analysis are arranged in alphabetical order. There is also the ability to obtain lexical tables, as well as to take into account repeated segments. It is important to have the ability to identify segments since the meaning of words is closely related to the way they appear in compound words or in phrases and expressions that can either inflect or completely change their meanings. So, in several cases it is useful to count larger units consisting of several words. These elements can be analysed in the same way as words. In addition in such analysis there is always the problem of misleading segments, which are identical because of the existence of punctuation marks. In order to face this situation the status of strong separator or sentence separator is assigned to some of the punctuation marks (such as period, exclamation point, question mark). Also, weak punctuation marks are defined (such as comma, semicolon, colon, hyphen, quotation marks and parentheses). These delimiters are called sequence delimiters. Apart from them, exist the word delimiters, which are treated as characters called blanks or spaces.

Correspondence and Cluster Analysis Techniques

The main idea behind this approach is to represent the patents in a high dimensional vector space, in which each of the available codes represents a dimension. Representing the patents in such a space, allows the visualization of their relative position 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 highly dimensional vector spaces specific data analysis tools need to be used such as factorial and clustering techniques for textual data.

The main method on which the derivation of good results depends is cluster analysis. However, in some cases, a factor analysis may precede the cluster analysis. In fact the use of factor analysis has a dual goal: Firstly, multiple factor analyses will be performed and through the use of bootstrap techniques the stability of the method will be presented to the user. Thus multiple ellipsoid graphs will be demonstrated enabling the user to draw conclusions about the stability of the method. Secondly, it is under consideration to use

correspondence analysis that will not be presented to the user with the aim to ensure the robustness of the cluster analysis. A schematic representation of the [illegible] the proposed statistical methodology [illegible]n in the following figure.

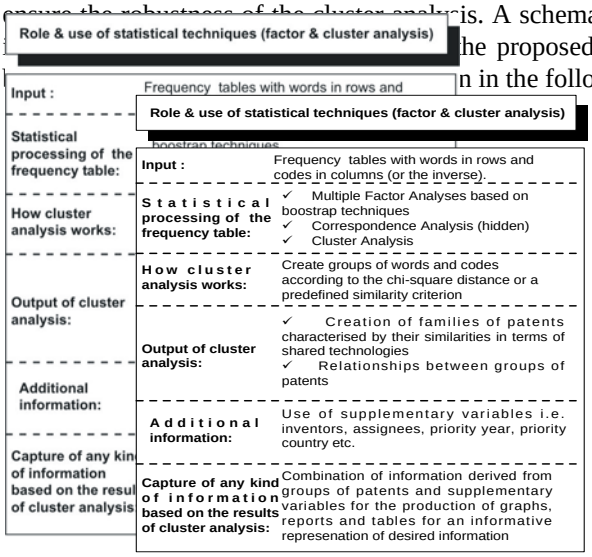

Figure 1: Process of cluster analysis: input data, processing of data, aim of the analysis, outputs

Factor Analysis (Correspondence Analysis)

Factor analysis (in this case correspondence analysis), allows to approximate (with minimal distortion) the high dimensional representations in the more manageable 2D or 3D spaces that can be directly visualized. So, the role of factor analysis is to find lower-dimensional subspaces, which more accurately approximate the original distributions of points. However, the reduction of dimensionality cannot be obtained without a certain loss of information. Factor analysis is applied to the contingency tables and permits to study the relationships that exist between nominal variables. More specifically the columns of this contingency table represent the modalities of the nominal variable while the rows represent the other variable. In the case of textual data a special contingency table is used where rows represent words or lemmas of words etc. while the columns are groupings of texts. In a typical correspondence analysis, a cross-tabulation table of frequencies is first standardized, so that the relative frequencies across all cells sum to 1.0. One way to state the goal of a typical analysis is to represent the entries in the table of relative frequencies in terms of the distances between individual rows and/or columns in a low-dimensional space. More analytically, in order to analyze a contingency table we use the repartitions in percentage in the interior of each line or column i.e. the line-profile or column-profile, which makes comparable the modalities of a variable. The proximities between points are interpreted in terms of similarities. In fact, a geometrical representation of similarities between different modalities of the same variable is performed.

The first step in the analysis is to compute the relative frequencies for the frequency table, so that the sum of all table entries is equal to 1.0 Then this table shows how one unit of mass is distributed across the cells. In the terminology of correspondence analysis, the row and column totals of the matrix of relative frequencies are called the row mass and column mass, respectively. If the rows and columns in a table are completely independent of each other, the entries in the table (distribution of mass) can be reproduced from the row and column totals alone, or row and column profiles. According to the formula for computing the Chi-square statistic, the expected frequencies in a table, where the column and rows are independent of each other, are equal to the respective column total times the row total, divided by the grand total. Any deviations from the expected values (expected under the hypothesis of complete independence of the row and column variables) will contribute to the overall Chi-square. Thus, another way of looking at correspondence analysis is to consider it as a method for decomposing the overall Chi-square statistic (or Inertia=Chi- square/Total N) by identifying a small number of dimensions in which the deviations from the expected values can be represented.

Cluster Analysis

The basic objective in cluster analysis is to discover natural groupings of the items. Grouping is done on the basis of similarities or distances. In our case the aim is to look for families of patents characterized by their similarities in terms of shared technologies i.e. eliminate classes of patents, which are similar in the codes, and words that they have in common. Several similarities measures can be adopted. Usually distances are used for clustering items and correlations to cluster variables. At the first step, when each object represents its own cluster, the distances between those objects are defined by the chosen distance measure. However, once several objects have been linked together, there is the need for defining a linkage rule in order to determine when two clusters are sufficiently similar to be linked together. Several linkage rules exist which are briefly described below. The single linkage method according to which two clusters could be linked together when any two objects in the two clusters are closer together than the respective linkage distance, i.e. we use the "nearest neighbors" across clusters to determine the distances between clusters. This rule produces "stringy" types of clusters, that is, clusters "chained together" by only single objects that happen to be close together. Other linkage rules are the complete linkage (furthest neighbor) method and the Ward's method**.** In the complete linkage method distances between clusters are determined by the greatest distance between any two objects in the different clusters (i.e., by the "furthest neighbors"). The Ward's method is distinct from all other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In short, this method attempts to minimize the Sum of Squares (SS) of any two (hypothetical) clusters that can be formed at each step.
Apart from the measures of similarity used for identifying relationships between objects, the grouping into clusters can also be performed through

different approaches. Therefore, different clustering algorithms have been created, among which are these of the hierarchical clustering, the k-means algorithm or a mixture of these algorithms.

Hierarchical clustering techniques proceed by either a set of successive mergers or a series of successive divisions. Agglomerative hierarchical methods start with the individual objects. Thus there are initially as many clusters as objects. The most similar objects are first grouped and these initial groups are merged according to their similarities, as the similarity decreases all subgroups are fused into a single cluster. On the contrary, divisive hierarchical methods start from an initial single group of objects, which is subdivided into two subgroups so that the objects in each subgroup are completely different between each other. These subgroups are then further subdivided into dissimilar subgroups and the process ends up when there are as many subgroups as objects i.e. until each object forms a group. The results of both methods can be displayed in a dendrogram.

Non-hierarchical clustering techniques (such as k-means method) are designed to group items rather than variables into a specific number of clusters, which can either be specified in advance or determined as a part of the clustering procedure. Non-hierarchical methods start either from an initial partition of items into groups or an initial set of seed points that will form the nuclei of clusters.

Cluster Analysis of Words and Texts

In the case of textual data, clustering techniques are used for representing proximities between the elements of a lexical table through groupings or clusters.

When methods of hierarchical clustering are used, then we obtain a hierarchy of groups partially nested in one another, starting with a set of elements that are characterised by variables. On the other hand, when direct clustering methods are applied, then simple segments or partitions of the population are produced without the intermediate step of hierarchical cluster analysis.

So, cluster analysis is applied to the column and row matrices of lexical tables. Either the entire set of columns (usually words and sometimes text parts) or the entire set of rows (most often the different text parts) is clustered. Hierarchical agglomeration as already described in the previous section is based on the agglomeration of elements that are close to each other according to the measure of distance used. In the general case, the distances among the elements subjected to cluster analysis are measured using the chi-square distance between the columns of the table. Each of the groupings created at each step, by following this method, constitutes a node. The set of terminal elements corresponding to a node creates a cluster. The representation as a dendrogram of a hierarchical cluster analysis shows that the groups created in the course of a cluster analysis constitute an indexed hierarchy of clusters that are partially nested in one another.

Use of supplementary variables

Supplementary variables can also be involved in the analysis in order to provide a complete description of the clusters and let us identify information about technological trends, competitors, inventors as well as technological evolution over time and information about priority countries. Below it is presented how additional information can be exploited in order to get a complete view of technological trends and innovation. As already mentioned in previous sections apart from the main variables used in the analysis i.e. the codes and the abstracts or titles that describe a patent, additional information can be obtained when considering supplementary variables such as inventors, assignees, countries that apply the patent and priority year.

If we take into account the structure of clusters relating to the codes that describe them, then we can derive any other kind of information. Thus, for each cluster considering which codes belong to it and consequently the technological trends it represents, we can have a global view about which companies are especially interested in specific technological sectors, which are the most prominent inventors, the technological activity of different countries and in which sectors as well as technological evolution over years.

In the following sections we attempt to describe how additional information taking into account supplementary variables can be exploited in order to get results that allow detecting different “situations” related to the technological development and innovation. For reasons of simplicity we present the way supplementary variables can be used for exploiting patent data information in two sections. The first one refers to the societies that deposit patents and what kind of information can be derived from it. The second one refers to the use of two other supplementary variables that describe a patent, those of inventors and priority year. Of course, these are not the only supplementary variables that describe a patent. Therefore, in the two sections mentioned above we describe how these variables can be used in combination with the other supplementary variables. The main aim is to present how the developed statistical methodology offers the ability to exploit all information related to patents and combine different variables in order to get the desirable results. But also an idea of what kind of analysis can be available through the use of this methodology. Furthermore, we should mention that the methodology allows exploiting information in each cluster separately as well as combined information for more than one cluster. As a consequence it is possible for each cluster to exploit information not only from the main variables we used but also from supplementary variables. Below we present how can be exploited information from supplementary variables.

Pre-processing Steps

Transformation of Data

It is important to present the data in the form of a table where each field describing a patent (i.e. assignees, inventors, etc.) will form a column while each patent will form a row. This is a more flexible format of the data that will

facilitate their use in the subsequent steps of the procedure but also in the selection of different fields. Furthermore, this will enable us to preserve the relationship that exists between patents and the different fields. It is of great importance to keep up the relationships between different fields since this will enable us to associate supplementary information such as inventors, assignees etc. to the results obtained from the main part of analysis and especially those of cluster analysis. In order to do this it is therefore necessary to have data sorted according to the number of patent. Therefore we should assign the number of each patent to all fields related to it and then sort according to field. Then for all patents according to their order of appearance in the original data we have a list with all the fields. Thus it is easy to create the table with patents in rows and fields in columns.

Selection of fields involved in the analysis

It has already been mentioned that the analysis will mainly be based on the use of codes and either titles or abstracts. However, although there is the ability to download selected fields from the patent database in the case that data are in other forms, we have to choose the fields that will mainly participate in the analysis. These are the main and IC codes as well as titles or abstracts.

Transformation of selected field in appropriate format for being used in the analysis

For continuing the procedures of the analysis it is necessary to appropriately form the data that will participate in the analysis. We should take into account that both the file with titles or abstracts as well as the file with the codes should be transformed. In the case of titles or abstracts the corresponding file should have the format described below. Titles or abstracts should be distinguished among them. Therefore separators should be defined and used. The separators among patents are lines that only contain the symbols "----" in columns 1 to 4.

Furthermore, the file with the codes that will be involved in the analysis will be treated as the case of numerical data that has to be associated with the file of textual data. The numerical variable is considered nominal with a number of modalities. Therefore, a file with the labels is created first and a file with the numerical data next. The last file consists of two columns, the first contains the identifiers of patents while in the second column the numerical data are represented.

Additionally, it should be mentioned that in the procedure described above it is possible to use not only the main codes but also IC codes. In this case, special consideration should be given to the way that the relationships between different codes are defined, as well as to which texts (titles or abstracts) these are related. In the development phase, this kind of problems should also be taken into account. Indices denoting which texts the codes are related to should therefore be used.

A breakdown of the developed statistical methodology is presented graphically in the figure that follows

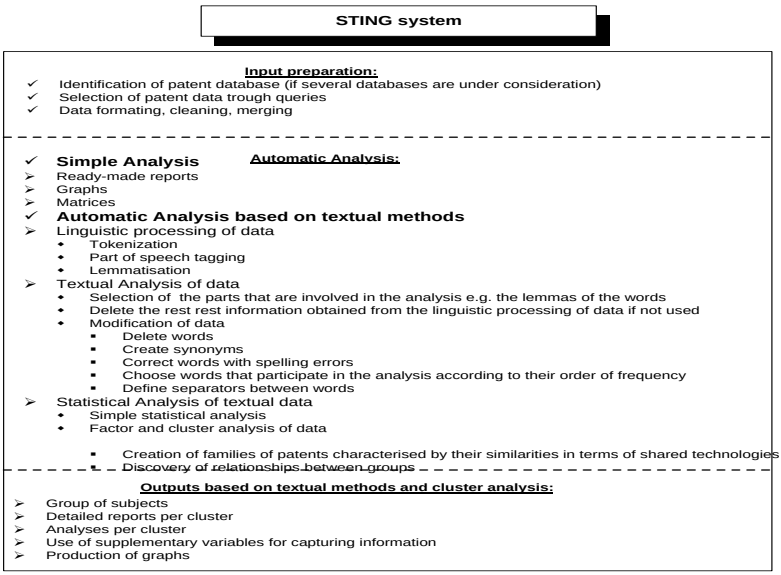

Figure 2: Basic Idea of the way STING works.

SYSTEM DESCRIPTION

One of the main objectives of STING project is the design and implementation of a system for the efficient analysis of existing information "hidden" in patent data, in order to produce indicators for the technological innovation in pan-European level [8], [9], [11], [16], [17], [18]. The system is based on the proposed statistical methodology described in the previous section and focuses on the maximization of end-user's productivity and satisfaction by offering him not only the correct guidance (wizards and help options) but also useful tools and functionalities (filters, charts, reports, etc.). This means on the one hand that the system enables the user to extract only necessary information and exploits it in an effective way for drawing useful conclusions. On the other hand, the system is capable of conforming to continuously changing user needs. Specifically, the target user-groups of the STING system are Science and Research Institutes, Educational Centers (Universities), Statistical Offices, industries, companies, or individuals, etc.

The system consists of three main modules, which integrate a full-operational patent data analysis environment. The functional architecture of STING system is presented in Figure 3. It shows the modules-tools of the system and the interactions between them. Each module corresponds to a basic system task, has a pre-specified input and output data format and executes a set of well-defined operations. The main system modules are:

- Database manager module.
- Statistical analysis module.
- Results presentation module.

All the above tools interact and allow easy navigation to the user from the data import level to the analysis and presentation level. It is highlighted the modular basis of the system, which enables the flexibility of the system to subsequent changes in the used statistical methodologies (addition of new methods, etc.). Furthermore, the structure of the system allows the use of different patent databases.

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. In the following sub-sections the modules of the system as well as the user interface are analytically described.

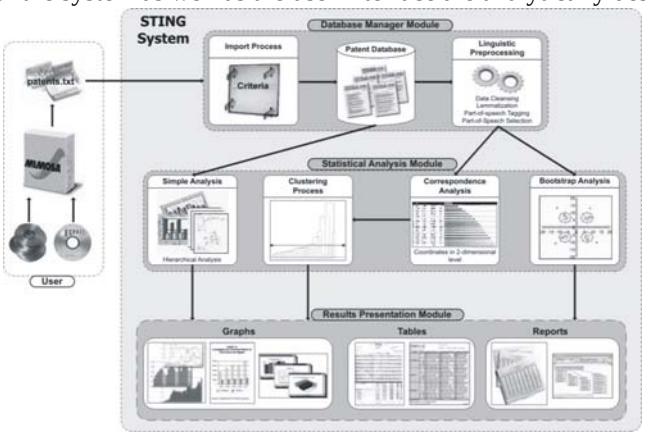

Figure 3: STING functional architecture.

Database manager module

The database manager module is mainly responsible for the patent data import, cleaning and preparation for further processing and analysis.

On a first level the module reads a text file from MIMOSA search engine and import it into the STING database (appropriate format). The database fields correspond to the fields describing a patent as presented in the ESPACE ACCESS database of patents. 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.

Before the data import the data are cleaned and filtered. The parser is used for reading the textual data and consequently linguistic preprocessing [2], [7] 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.
On a final level, the module encodes the textual information describing the patents in a lexical table, which associates the frequencies of the selected words/lemmas with the corresponding patents.

Statistical analysis module
The statistical analysis module is mainly responsible for the statistical analysis of data and the production of the technological and scientific indicators. 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.
There are four different analysis methodologies integrated into the system: simple analysis, correspondence analysis [3], [14], cluster analysis [1], [15], bootstrap analysis. Normally, this module starts after the linguistic processing of data and the creation of a lexical table, which associates the frequencies of the selected words/lemmas with the corresponding patents.
In the case of simple analysis the user can proceed directly to the production of indicators based on the selected database. Otherwise, the user can choose to proceed with a factor analysis on the lexical table and consequently perform the cluster analysis. At the end of cluster analysis the user can explore the homogeneous classes of patents and run the simple analysis in each cluster separately. 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.
It is also worthwhile to 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 [6] 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).

Results presentation module
The results presentation module is mainly responsible for the data export and the graphical representation of the analysis results.
It is common sense that the visualization of the results is a significant factor for their understanding [10], [19]. This module permits to visualize the desired information in a useful way and is very important for the user in order to fully comprehend their meaning.
Therefore, different 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 these 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.
Finally, 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.

User interface

Effective user interfaces generate positive feelings of success, competence, mastery, and clarity in the users. However, designing user interfaces is a complex and highly creative process that blends intuition, experience, and careful consideration of numerous technical issues. STING system's main aim was the development of a *flexible* and *user-friendly* environment that enables the effective analysis of information stored in patent databases. Especially, novice users have the opportunity to easy navigate the system through an effective graphical user interface and explore every field in patent data analysis.
Hence, the basic user interface specifications that the final system offers include: language/vocabulary, form of dialogue, data entry, windows, use of color, response times, help messages, error messages, system feedback, human-computer interface configuration, documentation, general qualification requirements, [4], [5], [12], [13].

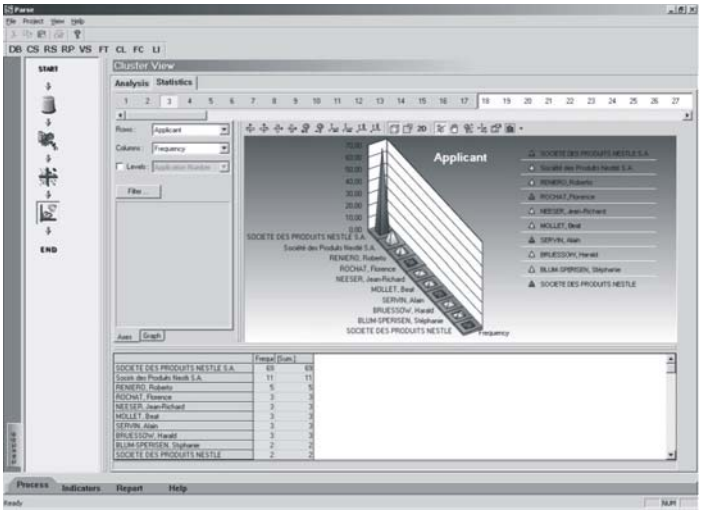

Figure 4: STING linguistic process.

Figure 4 provides a first view of the user interface of STING. The system is designed to convey information to users efficiently. For this reason different menus are available, as well as different windows and controls that visualize the functionalities of the software. Specifically, the main window is separated

into sub-windows each of which describes a different functionality. The basic parts of STING's workspace are the following:

- Main menu.
- STING's Toolbars.
- Tool Window.
- Tool Selector.

The system functionalities and interfaces were tested by groups of users during the validation phase. However, extensive testing and iterative refinement were performed early on to validate the design.

TECHNOLOGY INDICATORS

It has already been mentioned that STING will produce indicators either in an overall basis or within clusters. Here particular emphasis is given to the indicators within clusters since they are the result of a specific statistical methodology based on cluster analysis. In addition, these indicators can be categorized according to the patent field they are related. More specifically, these are categorized as follows:

- Indicators in the level of areas of technology.
- Indicators in the level of continents/countries/designated states.
- Indicators in the level of assignees/inventors.
- Indicators over time.

A schematic representation of the categorization of indicators is also given in Figure 5.

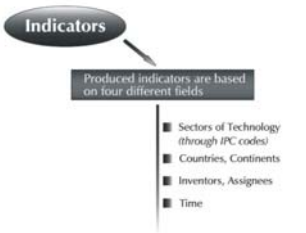

Figure 5: Categorization of indicators.

The simplest type of patent indicator is obtained when counting the number of patents satisfying some criteria. These criteria are imposed to other fields such as inventors, assignees, priority countries, priority years or set of criteria imposed to a combination of these fields. In addition counting the number of patents in each cluster either by imposing some criteria or not is an index of great importance since it let us identify technological tendencies.

Indicators based on the technological sector

The use of cluster analysis permits to identify areas of technology that share common technologies. Each created cluster express a specific area of

technology. Therefore, there is the ability for each area of technology to define several indicators.

A first indicator is the one that gives the number of patents that correspond to each area of technology. This permits to identify the maturity of a specific area of technology as well as innovate areas of technology. According to the information one is interested in, he can obtain either the top areas of technology or the less active technologies based on the number of patents.

A more sophisticated indicator about the maturity of each area of technology can be defined as follows. Categorize clusters in “very homogeneous” (this is the case that the major code exist in a percentage greater then 70%), in “ relatively homogeneous” (this is the case that the major code exist in a percentage between 55 and 70%) and finally in “non homogeneous” (this is the case that there not exist a major code, its percentage is between 35 and 55%) and thus draw conclusions about the homogeneity of clusters as well as about technological activity.

Another indicator may be the dominant codes of each cluster. Then we have at once the major technologies that characterize each area of technology. Percentages or additional information such as how many times they appear in other clusters may also be demonstrated.

Indicators in the level of continents/countries/designated states

The most frequently published indicators are counts of patents taken in a given country, broken down by country of the patentee (the inventor or the applicant) or by priority country (first country where the invention is filed before protection is extended to other countries). However when counting patents applied in a given country by patentees from various countries raises the issue of comparability i.e. to what extent do country share reflect their technological output. In fact residents in any country have a higher propensity to patent inventions at home than foreigners. This means that protection for smaller inventions is searched on the local market only. In the sequence we present some indicators that take into account the fields of country, inventor or assignee. These are very simple and are based on the calculation of some simple statistics such as percentages.

Level of continents

For each area of technology, each cluster identifies the patenting activity per continent based on the number of patents or the corresponding percentages. In addition another indicator is this that gives the top areas of technology of each continent or the less active areas of technology for each continent.

Level of countries

A first indicator is this that gives the active countries in each area of technology. Furthermore another indicator is this that specifies the hot areas of technology for specific countries.

Level of designated states

The top designated states (as well as these countries in which the fewest inventions are protected) are a very important element since it gives a first view about the strength of the market. Thus a country that is between the top designated states probably indicates the existence of a powerful market. Therefore the ability to identify "strong" markets (i.e. competitive markets) enhances the capabilities of the patent analyst to take good decisions.

Furthermore the countries that are not considered so important and therefore present a low place in the selected designated states can be identified.

- The evolution of each country as a designated state over years or for specific years should be presented.
- The selected designated states for USA patents/over years.
- The selected designated states for European patents/over years.
- The selected designated states for Japanese patents/over years.
- The selected designated states for specific European countries patents/over years.
- The average number of designated states per invention.
- Ratio of the selected designated states for USA patents over the totality of patents.
- Ratio of the selected designated states for European patents over the totality of patents.
- Ratio of the selected designated states for Japanese patents over the totality of patents.
- Ratio of the selected designated states for specific European countries patents over the totality of patents.
- For the dominant designated states:
    - Top areas of technology.
    - Top industries.
    - Top inventors.

Indicators for inventors/assignees

Inventors

In the sequence are presented some statistics based on the notion of inventors.

- Top inventors/less active inventors (also at the level of clusters).
- Average number of inventions per inventor.
- A comparative view of the patenting activity of inventors over years based on the average number of inventions or on patents counts.
- Resident and non-resident inventors.
- Inventors per country, per continent etc.

The subject of counting the patents relating to the inventors may be faced from different perspectives. According to the patent manual in the case that the inventors are of different nationalities we should share the patent among the various countries concerned. In measuring a country's patent output this results in fractional counting.

Assignees

The field of assignees can be used for different kind of analysis in order to catch points of interest. Firstly the top assignees as well as the less active is of importance in order to identify the leaders in specific areas of technology, for specific years etc. Then we should have in mind that these that apply a patent vary from individuals to companies public or private, universities, etc. Therefore the analyst should be able to define all these categories that will form the basis for different analyses.

Additionally, the distinction between assignees of the same nationality as these of the applicant or of the inventor may be of interest.

- Top assignees/Less active assignees.
- Distinction in companies, institutes, universities etc.
- Distinction in individuals and societies.
- Distinction in public societies and private societies.
- Distinction in small and big companies.
- Patenting activity of assignees for specific years/over years.
- Average number of assignees per invention.
- Average number of assignees with same nationality of applicant/inventor.
- Average number of assignees with foreigner applicant/inventors.

When patenting by type of assignee is investigated fractional counts can be used to assign patents to the different groups considered, such as firms, universities, government laboratories, individual inventors and so on.

Indicators over time

Taking into account the indicators mentioned above, we could derive statistical measures based on the number or percentages of patents for each technology area, as well as for the designated states or inventors. The evolution over time is another feature that helps tracing the technology on-goings and monitoring the novelty of an invention. The field of Priority Year can be used for different kind of analysis. Firstly one can depict the evolution of patenting activity over time independently of the country of filling. Additionally the evolution over time taking into account the country of filling can be demonstrated. More detailed information can be obtained by combining the field of priority year with this of the inventors. More specifically the inventions can be distinguished to the following categories:

- National applications: all applications filed in a national patent office.
- Resident applications: all applications filed in a national patent office by inventors resident in a country.
- Non-resident applications: all applications filed in a national patent office by persons resident abroad.

CONCLUSION

The described system refers to a significant scientific subject, the measurement and assessment of technological innovation through indicators. Specifically, it

specializes in the analysis of patents exploiting the totality of information related to them. The fact that we take into account all the information describing a patent i.e. the codes as well as the titles or abstracts allows to have more reliable and complete results. This is the feature that differentiates it from other existing approaches for the exploitation of patent data. Another innovative point is the ability to exploit information stored in various patent databases. Patent data can be reached using filters and downloaded in the appropriate format in order to being involved in the procedures of various supported analysis.

System's statistical methodologies are based on the use of textual and advanced statistical methods such as correspondence and cluster analysis. These statistical procedures permit the effective exploitation of patents enabling the user to capture the desired knowledge in an easy and informative way. Moreover, the automation of the patent data analysis does not require statistical knowledge from non-experts users. The use of the most appropriate statistical tests for ensuring the accuracy and reliability of the method are under consideration. 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 handles dynamically all the available information, regarding patents and indicators. Special consideration was given to user friendliness, interactivity and interoperability.

The produced results can be presented through different ways, which enhance the usability of our system in the domain of patent analysis. The information can be depicted graphically through different kind of graphs such as pie charts, bar charts, etc. In addition, several reports can be derived in order to present information summarizing important features. Another characteristic of the system is the production of ellipsoid graphs as a result of a correspondence analysis. This permits to identify information in an easy way and provide conclusions in a comparative basis.

Finally, it should be also mentioned that efficiency is a very important factor of the system. In order to achieve greater efficiency, in the design of the system was followed a modern and innovative philosophy. More specifically, the following parameters guided the development of the modules:

- Use of open system architecture, to allow the easy adaptation of the system on many different statistical databases;
- Interoperability;
- Development of an integrated processing environment;
- Implementation of a standardized interface.

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.

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. Brooks R. (1997). User Interface Design Activities. In: Helander et al. 1997, pp. 1461-1473.
5. Carroll J.M. (ed.) (1997). Human-Computer Interaction. Part VII of The Computer Science and Engineering Handbook, Allen B. Tucker Jr. (Ed.), Boca Raton: CRC Press, Inc.
6. Chappelier, J., Peristera, V., Rajman, M., Seydoux, F. Evaluation of Statistical and Technological Innovation Using Statistical Analysis of Patents, JADT 2002.
7. Ciravegna, F., Lavelli, A., Pianesi, F. Linguistic Processing of Texts Using Geppetto, Technical Report 9602-06, IRST, Povo TN, Italy, 1996.
8. Comanor, W.S., Scherer, F.M. Patent Statistics as a Measure of Technical Change. Journal of Political Economy, pp. 392-398, 1969.
9. Edelstein, H. Introduction to Data Mining and Knowledge Discovery. Two Crows Corporation, 1999.
10. Fayyad, U., Grinstein, G., Wierse, A. Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann Publishers, 2001.
11. 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.
12. Galitz W. O. (1989; 1993). User-interface Screen Design. Wellesley, MA: QED Information Sciences.
13. Helander M., Landauer T. K. & Prabhu P. V. (eds.) (1997). Handbook of Human-Computer Interaction. Amsterdam: North-Holland (Second, completely revised edition).
14. Hill, M.O. Correspondence Analysis: a Neglected Multivariate Method. Journal of Applied Statistics, Vol. 23, No. 3, pp. 340-354, 1974.
15. Lewis, S. Cluster Analysis as a Technique to Guide Interface Design, International Journal of Man-Machine Studies, 35, pp. 251-265, 1991.
16. Narin, F. Patents as Indicators for the Evaluation of Industrial Research Output. Scientometrics, 34, 3, pp. 489-496, 1995.
17. OECD. The Measurement of Scientific and Technological Activities Using Patent Data as Science and Technology Indicators. Patent Manual, 1994.
18. Schmoch, U., Bierhals, R., Rangnow, R. Impact of International Patent Applications on Patent Indicators. JOINT NESTI/TIP/GSS WORKSHOP, Room Document No. 1, 1998.
19. 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.