Competitive Intelligence through Data Mining Public Sources

Alessandro Zanasi

IBM Corporation

EXECUTIVE SUMMARY

A new paradigm has appeared in computer science in the recent years: data mining. The success of this new approach comes from recent advances in three different, yet connected fields: mathematics—new efficient and quick algorithms have been developed; database technologies—new research has allowed the improvement of databases; computer power—new powerful computers and architectures have made possible the elaboration of huge data volumes. Data mining's applications to the business world are several and include database marketing, basket analysis, and crime detection. In this article, the author discusses the application of data mining to competitive intelligence analysis. © 1998 John Wiley & Sons, Inc.

"If you know your enemy and you know yourself, you need not fear the result of a hundred of battles"
*—**SunTzu,** 500 B.C.*

The Cost of Knowledge

As knowledge is becoming more and more important to the creation of wealth, we must look at companies as knowledge operators. Within this data-intensive environment, seeing that the fight to produce, control, and rapidly exploit "know-how" is heating up, many businesses have decided they need more information about the plans, products, and strategies of their competitors.

This need has always existed. Why only now do we see this dramatic rise? Different reasons have been the cause:

- *Increased competition all over the world.*
- *Ease of "invading" a competitor's market even if located on the other side of the world.*
- *Product lifetime is becoming shorter and shorter.*

Competitive Intelligence Review, Vol. 9(1) 44–54 (1998)
 CCC 1058-0247/97/01044-11

So, before investing many years in the development of products that may remain on the market for only months, accurate analyses must be made.

And surely, before implementing a decision, it would be useful to know what the competitor is planning to produce in the future. But one could ask, "Can this activity be undertaken legally?" Of course, it depends on the means we use to obtain the data. Very often important, strategic information is "open" to everyone. One can obtain additional information without the risk of being accused of industrial espionage in a cheap, quick, legal, and secure way.

Some Definitions

Data Mining

By data mining we mean all the techniques that allow us to "discover" knowledge otherwise hidden in huge databases. The computer algorithms that make this possible are based on sophisticated mathematics and on new, intelligent ways of utilizing computer power. For a simple explanation of these techniques and of their applications see Zanasi (1997).

Competitive Intelligence

Timely and fact-based data on which management may rely in decision making and strategy development. It is obtained through industry analysis, which means understanding all the players in an industry; competitive analysis, which is understanding the strengths and weaknesses of competitors (Society of Competitive Intelligence Professionals).

It is also information that tells us how competitive the firm is. It's understanding the competitive arena, being able to predict competitors' and customers' intentions, government actions, and so forth.

Competitive Intelligence (through Data Mining)

The process of discovering or predicting the competitors' strategic decisions and/or understanding the characteristics of the business using quantitative analysis techniques applied to open sources (for example, online databanks).

Benchmarking

The continuous process of researching new methods, practices, and processes. Companies either adopt or adapt what works for them in order to become the best of the best (Camp, 1989; Balm, 1992).

Scouting

The process of finding alternative technologies outside a company's research, development, and engineering laboratories, which will enhance its marketing and business goals. It searches out processing or technical breakthroughs unknown to company researchers due to cost or time constraints.

In this article, we use the term CI to mean all the activities directed to study the strategy of competitors and the trends of the market, utilizing techniques borrowed from bibli-, scientо-, and info-metrics, which are techniques used primarily to analyze scientific output.

Bibliometrics

Measures the output of a research system. It was born after the establishment of databanks (that allowed the creation of new indicators) and resulted from the need of research managers for a technique to evaluate the effectiveness of their efforts.

Scientometrics

Establishes a relationship between the results and the resources given to a research system.

Infometrics

Studies the relationship between the research system and other systems (economic and social).

In this article we will use the term competitive intelligence as inclusive of the above techniques (often utilized in CI activities). For an analysis of this matter see Huot (1993).

Competitive Intelligence in the Advanced Information Age

In today's competitive business environment, it is vital for any industrial company to watch what its competitors are doing, and in what direction the market is moving. The most important questions that the managers of a modern company ask themselves are:

- *What are the specific areas of R&D that need to be developed?*
- *What are the market trends?*
- *In what sector are our competitors preparing new products to be put on the market?*

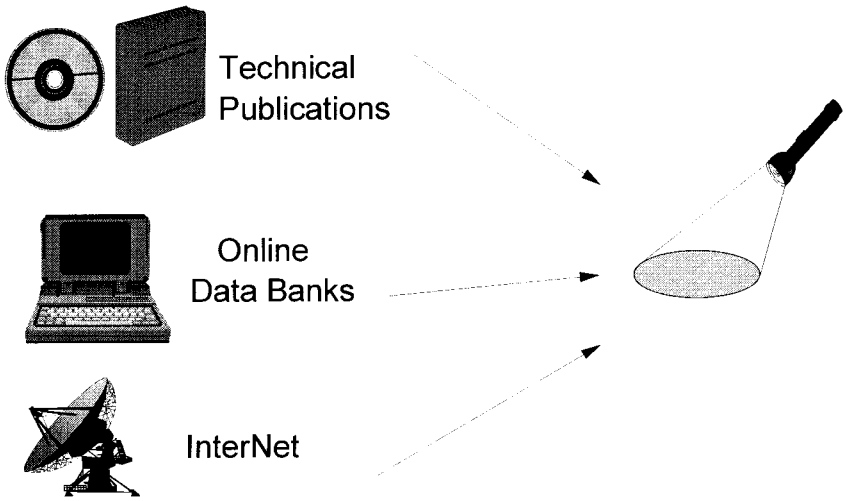


Figure 1.

Data sources for competitive intelligence.

- *When they will probably do so?*
- *What sector will be abandoned by them in the next years?*
- *Are we in a strong position, given our expertise, to enter a new market?*
- *What are the technological "bricks" that are key to a certain sector?*

Where can we look for the answers?

We can say that in our advanced Information Age, if something has been written it probably exists in electronic format in an online databank. Whether it is an article published in a Seoul newspaper or proceedings from a scientific conference in Johannesburg, it will be available in electronic form.

However, the principal disadvantage of these databanks is that they are very large and the information contained within them is very complex.

For example, one of the most reliable sources of technical information are international patent databases; they, as well as other technical and research publications, are currently available in online databanks. It has been calculated that, more or less, 80% of the significant technical information is patented. So no one doubts of the importance of patents to perform a useful CI analysis. But there is a problem: The available patents or technical reports are many, often written in so specialized a language that they can't be read easily. And they can't be easily put in relation to each other. So, how can we "extract" this knowledge?

Data mining provides an innovative solution for extracting useful information from these databases, automatically, quickly and easily, for targeting CI objectives (Zanasi, 1995).

> While 80% of all significant technical information is patented, the available patents and technical reports are often written in so specialized a language that they can't be easily read, nor easily placed in relation to one another. Data mining provides an innovative solution.

Users, Skills, Objectives

End Users

The results of competitive intelligence is often delivered directly to the top decision makers (CEOs, Presidents or General Managers, Division or Business Unit Managers).

The interested sectors are principally:

- *Production Companies (to understand the market and the competitors).*
- *Commercial Banks (to prepare Study Office reports, to evaluate specific clients).*
- *Investment Banks (to evaluate the innovative companies and the dynamics of their industries for venture capital investments or mergers and acquisitions).*
- *Insurance Companies (to evaluate any potential risks especially within technical companies).*

Performers
Who performs and which people are directly involved in competitive intelligence? This is such a new activity that not all companies are structured with an office dedicated to it (this would be the best solution). Sometimes, professionals in direct contact with top management, the documentation office professionals, or marketing employees are in charge of performing this role.

Skills
The minimum requested skills for such performers are:

- *Broad cultural awareness*
- *Documentation retrieval experience*
- *Computer literacy (UNIX and programming capacity is often useful)*
- *Knowledge of databanks peculiarities as:*
  - *access languages*
  - *data organization*
  - *specific codes*
  - *knowledge of particular techniques (as data mining)*

More often than not this activity is outsourced. In fact, in competitive intelligence cases the problems of confidentiality are absent: Nothing about the company client is given outside (except the request), and the company is free and capable of utilizing (or not) the results that have been obtained.

Objectives/Why
The current and most important target of CI activity is to understand, as far in advance as possible, the competitor's technological strategy and/or the trends of the market. After all, years of R&D have to be planned, and millions of dollars have to be invested.

Given the strong need for CI analysis, CI specialists can now be found in a number of different positions inside corporations, and academic and government agencies. Their titles include Market Research Director; Strategy, Industry or R&D Analyst; Informations Specialist.

The process

Steps to be followed to perform the CI process correctly include:

- *Problem Understanding. To understand, working with the end user, the business problem.*
- *Data Sources Definition. To define the strategic data sources where the information is obtainable.*
- *Strategy Research. To define the research strategy to best identify the most important data set.*
- *Automatic Analysis. To choose the correct tool for an automatic analysis of this data set to extract the "hidden" knowledge ("hidden" because the data to be analyzed takes up hundreds of thousands of pages). This is the reason data mining is a useful tool for CI.*
- *Results Analysis and Interpretation. To have a simple, possibly automatic interpretation of the results.*

We now discuss these steps in detail.

Problem Understanding

At times, it is only at the end of an engagement that the client and the internal or external consultant realize that a misunderstanding on the content of the research has occurred, but by then, it is too late to rectify it.

The client and the consultant should spend some time together planning their research, to be sure that the problem has been well targeted.

Data Sources Definition

Today, the problem for people who need to work with information is no longer about obtaining data: On the contrary, the volume of data available is often too extensive. The problem is finding *useful* information within the data.

The Net
The Internet provides a way for all UNIX workstations to get worldwide information using TCP/IP communications protocol. Many search tools such as WAIS, Gophers, and FTP allow the end user to get information from various firms, research centers, and financial institutions. The recent availability of the most important hosts (DIALOG, ORBIT, ESA/IRS) on the Internet is a major step for online users in the commercial world. This evolution is characterized by a decrease in network prices, an important increase in speed transmission, a simplification of the connection process, and, above all, new innovative online services provided to end users.

Internet Data
Internet navigators know well how many documents, how much data and information can be obtained through

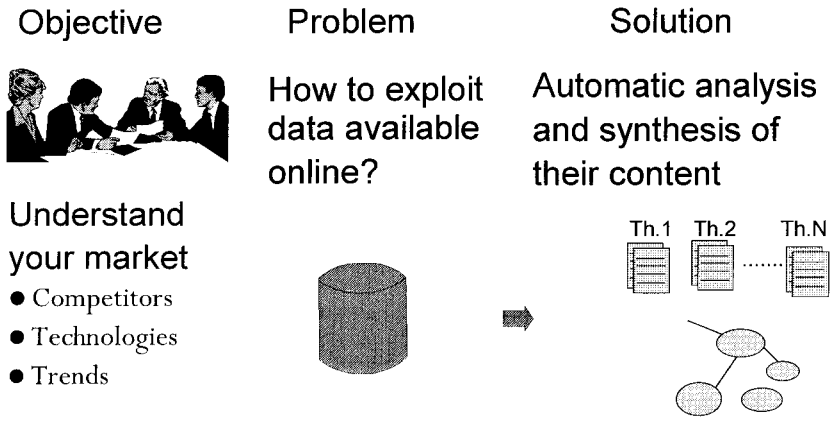


Figure 2.

The competitive intelligence problem and the data mining solution.

this means. The main advantage is that this "flood" of information is often free, however, the information is not certified (that is, we are not sure that this information is correct, complete or true; on the contrary, it could be completely false, put on the Net to lead the CI professionals astray). So, the CI professional doesn't usually utilize this data (or, at least, uses this source only as first-level information requiring re-confirmation).

Online Databanks

Online databanks are used more extensively. They contain well-organized documents, usually dedicated to specific, defined subjects. Databanks are dedicated to press surveys, patents or to scientific articles (chemical, physical, or mathematical).

Before being featured in these databanks, the articles need to be "certified" by the producers, thus giving credibility to the information.

Other important features of databanks are the keywords, codes, and classification tools that allow for better organization of the documents in the databank. Some of the most well known online databanks are:

- *WPIL—World Patent Index Latest (Derwent)*
- *Chemical Abstracts*
- *Compendex*
- *Dunsdata*
- *INSPEC*

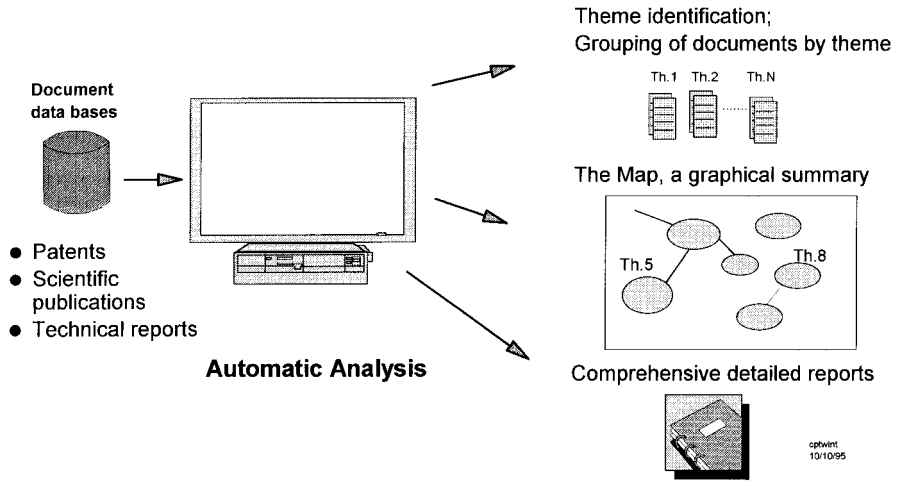


Figure 3.

How the raw data, through data mining, become real knowledge.

- *Medline*
- *Pascal*
- *Pharmaprojects*
- *Science Citation Index*
- *Tulsa*

Private Sources

A company may have a private databank that can be merged with other sources. This private databank can contain data completely different in format and the content from the data obtained through databanks.

Research Strategy

After having defined the data source, a strategy of document research has to be defined and implemented.

The difficulty in using databases is determining the keywords from which to extract citations and abstracts. This may be obvious in some cases. Often, it is a trial-and-error process. An extract could include more diverse information than might be useful. So several limiting descriptors will most likely be necessary. Also what is being excluded by the use of certain qualifying descriptors should be considered.

Because fees are paid on the basis of connection time and number of documents extracted, it is important to search efficiently to avoid paying an excessive sum of money. Searching documents is a specialized and difficult job; the success or the failure of competitive intelligence activity, both from an information point of view and from an economic point of view is based on database searching capability.

After we have located our information, we download a data set to which we apply data mining techniques.

Automatic Analysis

Various methods exist to analyze automatically a downloaded data set (Huot et al., 1992). A good approach is to utilize these methods in successive steps, grouping them under the titles of Statistical Techniques and Data Mining Techniques.

Statistical Techniques

These treatments (showing the distribution of data and of their variables), allow us to detect the phenomena that will lead the analysis.

- *Frequency Analysis. Counting the references we have information about:*
  - *The subjects about which we are working (very frequent)*
  - *New methodologies (frequent enough)*
  - *Noise (very rare)*
- *Pairing Techniques. They highlight the connections among the documents. They also show some drawbacks.*
- *Data Analysis Techniques. They were utilized to solve the previous drawbacks. We have Inertia, Classification, and Co-Word Methodologies.*

Data Mining Techniques

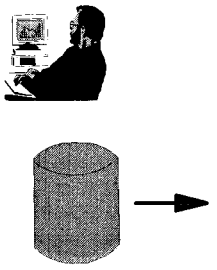
Techniques have been produced or improved in the last years: neural nets, genetic algorithms, association algorithms, cluster analysis, decision trees. These techniques, applied to competitive intelligence problems, have been studied extensively at IBM ECAM in Paris (Marcotorchino, 1986, 1991). Below we outline a particularly efficient approach (Huot, 1996).

The input is the downloaded data, which is then organized into a matrix format. The variables, chosen among the codified fields found in the references, will be used in the analysis as describers of documents. They have to be homogeneous and characteristic of phenomena to be studied.

A two-dimensional matrix is built with the downloaded references forming the rows, and the chosen variables plus additional variables that contribute to the results analysis step comprising the columns. Many different scientific and technical problems can be addressed through this data representation.

For example:

- *Rows = Patents*
  - *Columns = Codes (CIB, Derwent, Manuals, etc.).*
    For typology of activity domains (dominant technologies, innovation poles, etc).
  - *Columns = Codes (CIB, Derwent, Manuals, etc.).*
    Supplementary variables = companies.
    For analysis of competitors by theme.
  - *Columns = Extension Countries*
    Supplementary Variables = Companies and Years
    For evolution of the international commercial set-up.
- *Rows = Scientific Publications*
  - *Columns = Key words*
    Supplementary Variables = Years
    For research evolution.



Patent references

1/3881 - (C) Derwent Info 1994
AN : 94-364398 [45]
TI : Television with function for enlarging picture by variation of deflect ion frequency - has microprocessor for controlling system synchronous signal output, horizontal and vertical frequency drive circuit, sync. signal counter, signal detector.
DC : W03
PA : (GLDS) GOLDSTAR CO LTD
IN : HO J
NP : 1
PR : 88KR-011143 880831
IC : H04N-005/262; C08J-005/18; G11B-005/704
PN : KR940043 B1 940120 DW9445 HO4N-005/262 001/pp
AB : No abstract

Figure 4.

An example of a patent reference layout.

- *Columns = Authors Co-Citations Network. To discover connections between people.*

- *Rows = Manufacturing standards*

  - *Columns = Technical Codes*
    Supplementary Variables = Products
    For normative products segmentation.

- *Rows = Technical Cards Related to Special Phenomena (breakdowns, failures, operating systems, etc.)*

  - *Columns = Key Words*
    For specific features to be identified.

The aim of the process is to group together, within homogeneous clusters, the row elements with the most similar profiles with respect to the descriptive variables.

Results Analysis and Interpretation

The Graphical Output

Data mining provides a visual summary of the analysis, in the form of a map showing the different groups, the number of patents in each group, and the keywords characterizing each group. A visual indication of the relationships between the groups is given through colored lines between the groups—the colors indicating the strength of the relationship. Detailed information is provided for

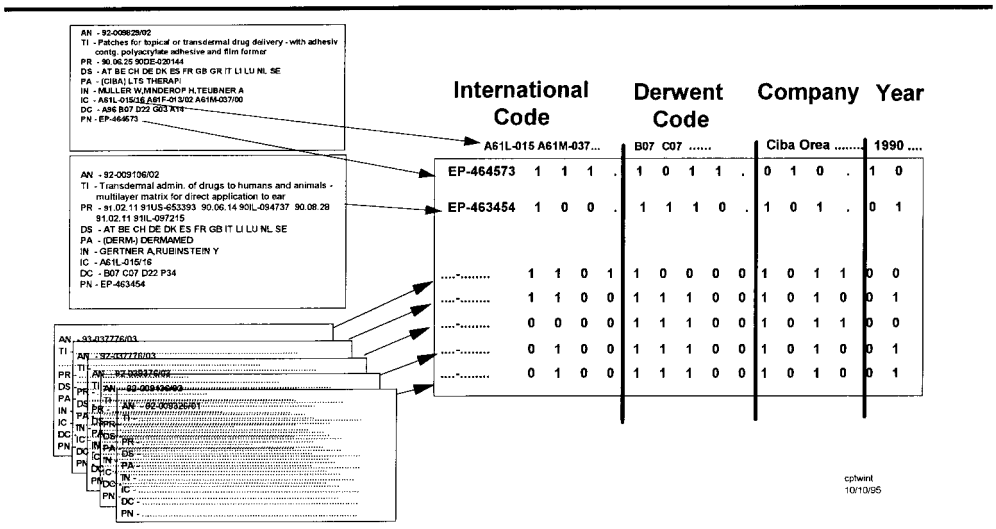


Figure 5.

How the matrix is built.

each patent or record in a group, based on the formatted fields selected.

DATA MINING PROVIDES A VISUAL SUMMARY OF THE ANALYSIS IN THE FORM OF A MAP SHOWING THE DIFFERENT GROUPS, THE NUMBER OF PATENTS IN EACH GROUP, AND THE KEYWORDS CHARACTERIZING EACH GROUP.

The Analysis Indicators

The partition obtained from the process has to be compared to the basic data. The standard analysis is based upon the following ideas:

- GLOBAL INDICATORS DESCRIBING THE WHOLE PARTITION:
  - *Global quality: It measures the ability of the objects to be classified.*
  - *Number/sizes of the clusters: They characterize the splitting of the objects.*
  - *Repartition of the objects within the clusters: It expresses the typology of the objects highlighting the objects with the strongest characteristics.*
- INDICATORS FOR EACH CLUSTER THAT MEASURE:
  - *Intrinsic quality of the cluster: Homogeneity and coherence of the cluster.*
  - *Links between clusters: Network basis of the linked clusters.*
  - *Support of the analysis variables: Explanation of the clusters construction (characteristic and discriminant variables).*
  - *Support of the supplementary variables: Description of complementary phenomena.*

The Reachable Objectives

After having collected the data and prepared them for data mining activity, the process begins.

The *first* objective is to reduce, for example, 10,000 documents to a map of 20 areas. These areas, automatically discovered, treat different business segments (or sectors of research). They have been discovered comparing each document with all the others: The most "similar" ones are put in the same group. But in what sense are they similar? For example:

- *They share the same, very uncommon, technology.*
- *The scientists involved usually work together.*
- *They make reference to the same body of previous knowledge.*

This analysis is humanly impossible, given the many variables by which a document may be defined, and the great number of documents that can be found. But it is very useful to understand the "real" segments that compose a business area.

The *second* objective is to find relationships between them. Two areas, which from a superficial point of view may seem completely different can actually be found to be very related, i.e., that their basic technologies are the same. Is this a sign that there is the possibility of finding a synergy? Is someone already trying to take profit from it? This mapping is very useful for discovering "opportunities," areas where the revealed relationships are sufficient to change our mind/our approach to the competition. These relationships are shown by colored bars.

The *third* objective is to define time trends. Utilizing the time variable, we can assess how over time, strategies change. We can discover that some areas are disappearing and that others are augmenting their importance or interest.

The *fourth* objective is, in cases where the target is a business area, to determine competitors, the nature of their focus on certain areas, and their movements in time. We can discover that our most dangerous competitor isn't the most well known (on the contrary, he may be leaving the sector) but a new one located far from our own business sector or geographic area, but is nonetheless, developing competitive products.

The *fifth* objective is to define the names of scientists or experts that work in the field and their particulars. If a scientist who has always worked in a certain area suddenly begins to work in another, is it a sign that he changed his interests, or that he found a new methodology to solve an old problem?

Other objectives are defined following the requirements of clients and their particular fields of activity.

An Applicative Experience: Bologna Data Mining Center (BDMC)

Cineca (Italian Universities Consortium) located in Bologna provides services to universities, government, public and private companies, inside and outside Italy in advanced information technology applications. It is one of the most important European Internet providers. It owns several databanks, and subscribes to the most well-known

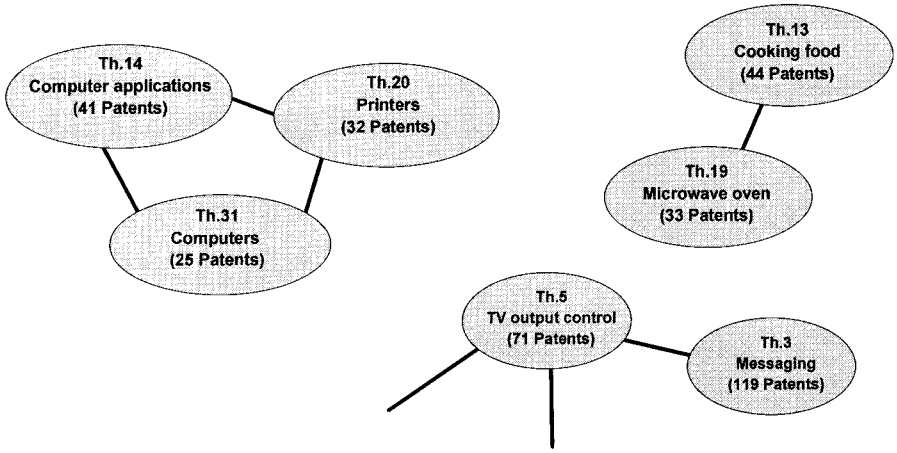


Figure 6.

Example of results presentation.

commercial and scientific databanks. Cineca and IBM were partners in creating the Bologna Data Mining Center to exploit the opportunities of applied data mining. Currently, different international companies and universities are working in connection with this Center. One of its more innovative offerings is a Competitive Intelligence Service.

Key Success Factors

- *Integration in only one provider of the different, necessary steps to fulfill a complete CI analysis based on exploiting electronic data.*
- *Competence in Information Retrieval and Document Analysis.*
- *Powerful elaboration power (32 nodes IBM SP2, the largest one installed in Italy).*
- *Ad hoc software.*
- *Special expertise in competitive intelligence through data mining (Zanasi, 1995).*

Business Process

- *The client explains its problem to Bologna DM Center experts.*
- *The best suited databanks are chosen and a correct strategy of research is defined and implemented.*
- *The databanks' output is processed through IBM data mining software.*

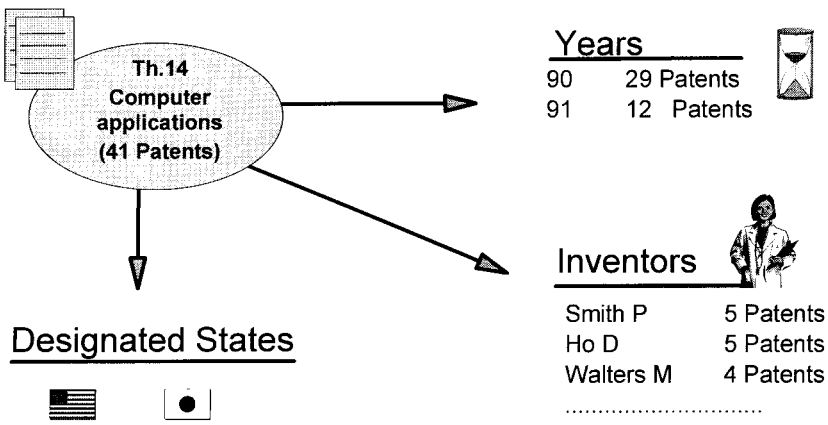


Figure 7.

Detail of CI analysis of only one competitor. If the analysis is on a market, there is a list of the involved companies.

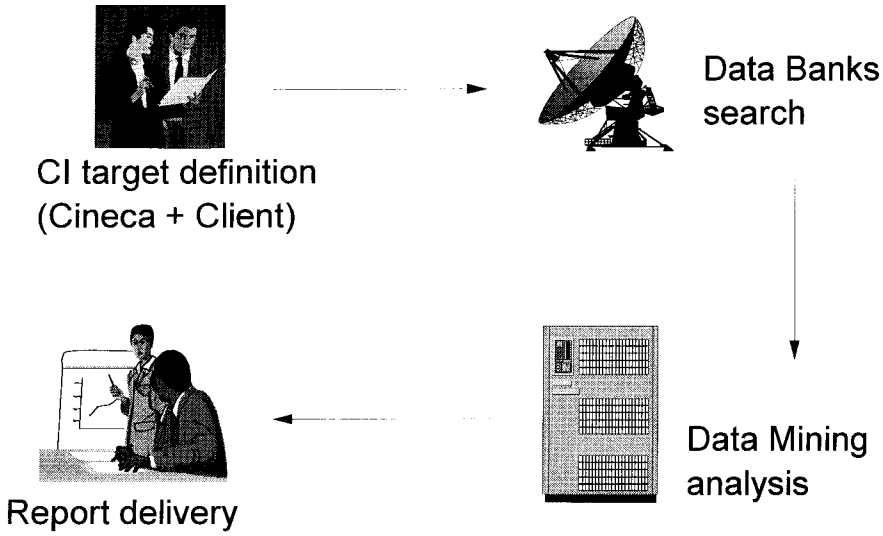


Figure 8.

The "CI by Data Mining" process implemented by BDMC.

- *The final report is presented to the client. Different formats are available: paper, diskette, CD, HTML for Intranet use.*

The report is always composed by:

- *Management summary*
- *Graphs*
- *Data mining complete technical output (usually about 100 pages)*

Timing/Scheduling

- *First project: From 20 days to 4 months.*
- *Following projects (for Subscription Service): two weeks.*

Bologna DM Customers

- *Geographic regions: Europe, Middle East, and South America.*
- *Clients are from Processing and Manufacturing, as well as from Media and Government sectors.*

Some Engagement Examples

A Chemical Company

A multinational company, working in chemical production, asked us to "discover" what one of its competitor was planning. We discovered that this competitor was working hard on a new, strange chemical molecule for use against a pest present only in Amazonian area. It was clear from our research that this competitor was planning new activities in Brazil which happened a year later.

A Media Company

Our client asked us to discover what new multimedia services through telecom were being studied worldwide, and to identify the most "dangerous" competitor in this business. We discovered that a company was particularly active in that area through patent registrations, seminar participation, business associations and alliances, even though they hadn't declared an interest in that business via related product development or selling. Only seven months later the situation changed: they transformed their logo, declaring that the new business in which they were entering was, indeed, multimedia services.

Visit our web site at http://open.cineca.it/datamining/

References and Related Readings

Balm, G.J. (1992) *Benchmarking,* QPMA Press, ISBN: 0963216716.

Camp, R.C. (1989) *Benchmarking,* ASQC Quality Press, ISBN: 0873890582.

Huot, C. (1993) *Analyse Relationelle pour la veille technologique: Vers l'analyse automatique des bases de donnees,* Thesis, Marseille University.

Huot, C. (1996) *IBM Technology Watch*–Paris, France, IBM Internal Report.

Huot, C., Quoniam, L., and Dou, H. (1992) "A New Method for Analyzing Downloaded Data for Strategic Decision," *Scientometrics,* 25(2):279-294.

Marcotorchino, J.F. (1986) *Maximal Associations as a Tool for Classification, Classification as a Tool for Research,* pp. 275–288, North Holland: W. Gaul & M. Schader.

Marcotorchino, J.F. (1991) *L'Analyse Factorielle Relationelle (partie I et II)*, Etude du CEMAP IBM France–N°MAP–003.

Zanasi, A. (1995) *Data Mining and Competitive Intelligence Thru Internet,* III NIR-IT-95, Third Network Information Retrieval Conference Proceedings, Milan, Italy.

Zanasi, A., Cabena, P., Verhees, J., Hadjinian P., and Stadler R. (1997) *Discovering Data Mining,* New York: Prentice Hall.

About the Author

Alessandro Zanasi is currently working for IBM in Bologna, Italy. He graduated in Nuclear Engineering at Bologna University, specialized in Probability Theory and Finance at Paris VI University, and in Business Administration at Modena University. He began his professional activity at Bologna University (Information Theory researcher) then, as Carabinieri Lieutenant, he worked at the Carabinieri Scientific Center in Rome. After a period as Information Broker, he entered IBM where he was assigned until 1995 to the European Applied Mathematics Center in Paris. He is now running the activities of Bologna Data Mining Center. He can be reached at Via G.B. Amici 29, 41100 Modena, Italy; Tel: +39-337-285470/+39-51 4136726; Fax: +39-51-406052; e-mail: zanasia@vnet.ibm.com.