

The Data Mining

Alliaume Erwan

Tetzlaff Franck

03 / 26 / 2001

Contents

1	Introduction	3
2	Process of Data-Mining	4
2.1	To pose the problem	4
2.1.1	Formulation of the problem	4
2.1.2	The typology of the problem	4
2.1.3	The awaiting result	5
2.2	Data retrieval	5
2.2.1	Investigation	5
2.2.2	the reduction of dimensions	5
2.3	Selection of the pertinent data	6
2.3.1	Sample or exhaustiveness	6
2.3.2	Mode of creation of the sample	6
2.4	Cleaning the data	7
2.4.1	The origin of the data	7
2.4.2	The aberrant valor	7
2.4.3	The miss valor	8
2.4.4	null valor	8
2.4.5	Prevent the bad quality of data	8
2.5	Action on the variables	9
2.5.1	The monovisible transformation	9
2.5.2	The multivariable transformation	10
2.5.3	Tendencies	10
2.5.4	Linear combination	10
2.5.5	No-linear combination	11
2.6	Seek model	11
2.6.1	Introduction	11
2.6.2	Equations models	11
2.6.3	logic analysis	11
2.6.4	Production technic	12
2.7	Evaluation of the result	12
2.7.1	the qualitative evaluation	12
2.7.2	the quantitative evaluation	12
2.8	Integration of the knowledge	13
3	Technics of Data-Mining	14
3.1	Reasoning containing case	14
3.1.1	Construction of CBR	14
3.1.2	Applicability	16
3.1.3	Limits and Advantages	17
3.2	The Knowbots	17
3.2.1	What is a Knowbots	17
3.2.2	Use	18
3.2.3	First example	18
3.2.4	Second example	18

3.2.5	Conclusion	18
3.3	The associations	19
3.3.1	Definition and stakes	19
3.3.2	The applicability	19
3.4	The decision trees	20
3.4.1	Definition of chi 2	20
3.4.2	Introduction	20
3.4.3	Analogy with the trees	20
3.4.4	Stakes	20
3.4.5	principles of calculation	21
3.4.6	The descriptor is qualitative	21
3.4.7	The descriptor is quantitative	22
3.4.8	Application domain	22
3.5	Bayesian networks	23
3.5.1	Presentation	23
3.5.2	Recall on the theory of the graphes	23
3.5.3	Use	23
3.5.4	Setting	23
3.5.5	Variable initialization	23
3.5.6	Variable selection	23
3.5.7	Dependance identification	24
3.5.8	Matrix Probability	24
3.5.9	Advantage	24
3.5.10	Inconvenient	24
3.5.11	Conclusion	24
4	Conclusion	25
5	Bibliography	26
5.1	the web site	26
5.2	the books and documents	26
6	Appendice	27
6.1	Our Application 1	27
6.1.1	description	27
6.1.2	the source code	27
6.2	Our Application 2	34
6.2.1	description	34
6.2.2	the source code	34

Chapter 1

Introduction

What is Data Mining ? The term Data-Mining indicates the set of the extraction process of knowledge starting from data contained in a base of data. Because Data Base is more and more important with the augmentation of the capacity of storage, it is very difficult for people to study them and it take quite a while . So Data-Mining accelerate this process. Furthermore it can work on very big Data-Base in a minimum of time. The companies undergo an intensification of the concurrence. What pushes them to give an attention increasingly larger to the customers, to constantly improve quality of their products and to accelerate their process of new products marketing and services. Generally, Data-Mining has a reason to be everywhere where there are a lot of informations.

Here some examples of use of Data-Mining :

It can be used in the large-scale distribution: to analyze the behaviors of the consumers, seek similarities of the consumers according to geographical criteria, cross sale and selective activation with the discount cards, optimization of restocking. But it is also used in the pharmaceutical laboratories for the identifications (choice) of the best therapies, in the banks to have a search for frauds or the authorization of credit, but also in the insurances, aeronautics, cars, industry, transport, telecommunications, energy and in a lot of other domain. The companies are more and more interested in Data-mining. Why this interest? Because it's quite common that the first applications of Data-Mining generate more than ten times the investment which they will have required without it, especially with a return on investment of approximately a month!

Chapter 2

Process of Data-Mining

There is often a confusion between software of Data-Mining and the process of Data-Mining (or Knowledge Discovery in Database). Tools are only a component of the transformation of the data in knowledge. They are in a process with eight stages, that we are going to study. Most of times, to increase knowledge collected, it's necessary to repeat one or more phases. Here are the different stages.

2.1 To pose the problem

Here we go ! Let us start with the beginning : The first Phase. In the first phase we expose the problem and we define the objectives. ((The awaited result and the means of measuring successes of the phase of Data-Mining)). It will be necessary to create variables so has to extract the information. So we must understand the context of the research to give a logical signification of the variables.

2.1.1 Formulation of the problem

The first phase to resolve a problem is to understand it. If you don't understand the problem you can't resolve it. That is the same thing for the computer. You must pose the problem in a form that the tools of modelisation can treat. When we have a big and complex problem, like for the diagnostics of the breakdown, for the analyze of the defects of the production or for the target of customers, the best method is to cut the big problem in several sub-problems with a smaller complexity and collect the information necessary for all the subproblems.

2.1.2 The typology of the problem

Here is the primordial question of the problem's typology :

affectation or structuration ?

What is the difference ?

- When we know the membership of an element in one or several classes, we must identify the factors of affectation.
- If the objective is to highlight some classes or factors of differentiation, we will have to identify the factor

We have to highlight that the process of Data-Mining is better in a problem of affectation than in a problem of structuration.

2.1.3 The awaiting result

Before enter in a process of Data-Mining, we have to know what we are waiting for and what we will do of the knowledge. The launching of a Data-Mining's project must be accompanied by a critical analysis of process bound to the exploitation of the results (logistic, marketing, commercial force).

The identification of the person who uses the results of a process of Data-Mining and the identification of the decisions that they may take have a heavy influence on the choice of the algorithm because they don't offer the same lisibility. When the result must be understand by the final users, some technics, too hard to understand for someone who isn't an expert, couldn't be used.

2.2 Data retrieval

The objective in this second section is to determine the general structure of the Data and the rules used to make them. So, we must identify the exploitable information and check their quality and their accessibility.

2.2.1 Investigation

The search of an optimal selection of the Data is the center point of a process of Data-Mining. For this selection, we need almost of time help from experts for determinate the best attributes for the problematic. This experts are able to indicate the variables useful for the problem to result. It's important, in this phase, to know the elements of the context we can make a representation of the problem with. Without expert, we can search the factors the most useful by analysis'technical : we make Data-Mining in Data-Mining.

This phase of structuration of the data must clarified the associations that exist enter them, their semantic contents, the regrouping more used for some of them (like the different classes of age), the aberrant valor (like strike-day) to kill the results too current and improve the prediction. The structuration of the variables is useful to decrease the dimension of the problem by isolating the most pertinent elements.

2.2.2 the reduction of dimensions

Intergrading all the variables with a lot of details induce an over-sizing of the problem, that harm the capacity of the generalization. This capacity of generalization allow a model to conserve a level of performance between the base of training and the base of test. If there are too many variables by ratio of the number of examples, it's will be very difficult to find two examples in close parts. So we must reduce the number of variables (like replacing dates in interval).

To reduce the number of variables we must know which one are enough pertinent without changing the problem and which assumptions on the knowledge are taken.

+	Optimal zone	Long calculating time
-	Multiplication of training	Not enough examples
Nb ex / Nb var	-	+

Table1 : relation between dimension and examples.

2.3 Selection of the pertinent data

The best mean to create a model is to search in the past similar events. So we must make a base of information who allows to build the training. The recovery of the data can be helped by the technology, with data-base for examples. This phase of collect and selection is, when informatics systems are very hermetic (little documented, too old), a considerable workload. It can be 80 % of the workload of the process of Data-Mining. Moreover, some studies require the organization of a plan of data-gathering : qualitative talks, creation of programs to intercept data which do nothing but forward by the information system, With the feeling of uselessness and waste of time proved by the customers of Data-Mining during this stop, the person in charge of the project must implement a policy of animation and follow-up of the collection, with returns of intermediary information, to motivate the speakers on the importance of their mission. It is necessary to avoid the trap of GIGO : "Garbage In - Garbage Out", because errors of input induce errors of output.

2.3.1 Sample or exhaustiveness

The analysis must choose between a study on the exhaustiveness of the data-base and a work on a sample. This choice depends of the used tools, of the machine-power available, of the budget allowed and of the level of fiability seeked. To detect the general tendencies without a need to differentiate with a strong level of precision some sub-populations, a representative sample will be sufficient. The extraction by a representative sample will be sufficient by quota will be preferred when it is a question of undertaking an analysis on a specific subpopulation whose objectives are relatively restricted. For example, the search of the four or five segments the more representative of a bargain to engage a marketing reflexion can be based on a study starting from a sample. On the other hand, the implementation of a direct marketing campaign on narrow bargain segments requires a more significant volume of examples, even the exhaustiveness of the data available.

2.3.2 Mode of creation of the sample

It is necessary determiner if the sample must be representative of the population (with a random draught) or must allow to laminate the population according to some sub-populations. The process of stratification sticks to increase some populations little represented, but that constitute significant stakes : the rule of 20/80 of Pareto (20 % of clients contribute 80 % of the sales turnover) can be applied here. The sample size must be determinated in order to ensure a representativeness of the results, verifiable by statistical tests.

Analysis on an exhaustive base presents a better quality of the results, but the price of investment is sometimes too much compared to an analysis relating on a sampled base. In a general way, exhaustiveness is reserved to some " large holders of data " while the recourse to the samples is appropriate for the majority of the operations and present of the unquestionable advantages in terms of handiness and response time. This is more appreciable as Data-Mining is often iterative.

2.4 Cleaning the data

The definition of the size of the base of examples and the choice on the manner to constitute it pass by a diagnostic of the quality of data. A low quality of the data (errors of data entries, fields null, aberrant values) generally imposes a phase of cleaning of the data. This one aims to correct or circumvent the inaccuracies or the data errors.

2.4.1 The origin of the data

According to the size and the mode of constitution of the data base, the methods of assessment differ :

- The base of example is restricted (less than 300 recordings or less than 30 variables approximately) and its automatic feeding. It is easy to control in a manual and visual way each recording to detect the anomalies. The construction of histograms or scatter plots for the various variables makes it possible to isolate the aberrant examples.

- The base of examples is restricted and, its feeding being manual, the risks of data entry exist. It is necessary to supplement the process of preceding control by consistency check at the time of the data entry and to help to the maximum work of the data entry by lists with choice.

- The base of example is significant and its feeding is manual. The risks of data entry remain identical but the cost of collection of information and the time of implementation become such as they can be higher for the discounted benefits.

- The base of example is significant and its feeding is automatic. The risk of not-quality of the data increases by as much when some data only were little exploited even never. It is then necessary to envisage the development of the data-processing procedures controlling quality by tests of distribution and measurements of contribution of some examples.

Nb examples	Automatic seizure	Manual seizure
+	Zone with a lot of risk process of analysis of normality	price of collect important check in the seizure and tests of coherence
-	Good reliability of data visual check	Risk of seizure must be checked

Table2 : The reliability of the data.

2.4.2 The aberrant valor

The first stage of the cleaning of data consists in seeking the aberrant values. For that, there are three principal techniques :

- Simplest consists in isolating the peaks from some values in a statistical distribution (for example, a on-representation of the dates of birth forced at January 1, 1901 or November 11, 1911, which are values easy to seize, or telephone numbers forced to 01.00.00.00.00).

- Most usual consists in defining a space ranging between the average and a number of standard deviation, and excluding or has to put a ceiling to all the values higher than a certain threshold (of the type + 3 standard deviations). The procedure of exclusion is making safe because it makes it possible to reduce the original variance of a problem, but it can result in excluding too much from examples. If the operation of filtering results in excluding much from examples, this test can reveal a more total problem on the reliability of the data base.

- The last approach, more complex, consists in building a first score, then thanks to statistical indicators, examining all the examples which contribute in a too strong way to the constitution of this score. An abnormal level of contribution is often revealing of data aberrant or an example pertaining to a specific class (for example, incur them of a multinational in a sample of particular customers of a bank).

2.4.3 The miss valor

The second stage aims at managing the missing data. Indeed, the absence of value is not appropriate for all the tools of Data-Mining and the statistical techniques, support, they also, rather badly the existence of nonwell informed data; it is necessary to manage in a specific way these value missing according to one of the following methods.

delete the incomplete record

This first method, very restrictive, consists in excluding all the recording whose value is missing. This choice is penalizing because it reduces the base of training and conduit to be excluded from the examples with nonwell informed data whereas these same data can not bring, ultimately, no information.

replace the miss data

The second method supported by some software, replaces the data absent by a value which is chosen by the user (replaced by the average), is calculated (replaced by the result of a formula of score) or inherited (95 % have 4 wheels, therefore all the car of which the number of wheels is missing have four wheels)

Manage the miss valor

When the absence of data is acceptable from the point of view of the performance of the model, the algorithms generally make it possible to manage separately the missing value by distinguishing it from the indicated values, or that to regard the value missing as a factor of indecision and to duplicate the examples in as many subbranches as of possible values.

2.4.4 null valor

The third stage is interested in the zero values : the cleaning of the data must integrate an analysis specific of the examples to zero. The data storage over a long period induces sometimes a significant sum of ave examples of many zero values. The analysis of the existence of these completely null recordings must be carried out in order to identify the possible external causes of them (breakdown of sensors.), before launching the algorithms of training. The strong representativeness of recordings almost exclusively well informed by zero can involve a perverse behavior of some tools which will learn initially has to model the anomalies and will treat the examples inform like exceptions.

2.4.5 Prevent the bad quality of data

The bad quality of the data complexes the training and harms the performance of the model. To face this problem has, some tool integrate noise (random variation of a data) or vague processes (variation parameter) in the phase of training. For that, the software simulates the noise while varying the input data, and measures the stability of the model on samples of tests.

2.5 Action on the variables

Now that the variables are relevant and that given it are reliable, they should be transformed to prepare the work of analysis. It is a question of intervening on the variables, so that they are better exploitable by the tools of modelling. These transformations can be of two types, according to they modify one or more variables.

2.5.1 The monovariate transformation

The modification of measurement unit

In order to avoid some disproportions in the systems of units of the variables, it is recommended to carry out a standardization of the distributions. The interest of the standardization is to have variables with comparable value.

$$X' = \frac{X - Mv}{Ec}$$

with

X = value to convert

X' = new value

Mv = moyenne

Ec = ecart type

	variables before		variables after	
	age	income	age	income
Example 1	23	175	- 1,625	- 0,653
Example 2	55	235	- 2,375	0,147
Example 3	48	224	1,500	-
...
Example N	41	312	0,625	1,173
Moyenne	36	224		
Ecart Type	8	75		

Table3 : standardization of the variables

Another method consists in carrying out a transformation logarithmic curve of the variable in order to limit the impact of some exponential values. The following table shows the effects of such a transformation on the exponential value.

Examples	sales turnover	log(sales turnover)
1	23	3,135
2	78	4,357
3	123	4,812
4	131	4,875
5	2345	7,760

Table4 : transformation in log.

The transformation of date in last

The systems of production generally store dates. However, these absolute dates have generally much less value, from the point of view of a work of modelling, that frequencies or differences between dates. So, one will calculate, for example, the seniority of the customer starting from his date of first purchase, or the reactivity of a customer by the difference between a date of sends of a catalogue and a date of command. This type of calculation adds variables to the analysis and generally contributes to make denser the classes of populations similar than would not make a simple use of the dates.

Modification of geographic data in coordinate

The techniques of Data-Mining generally have evil to apprehend postal codes or departments. That holds, on the one hand, with the multiplicity of the codes and, on the other hand, with the randomness of codings (two bordering cities can be very well in different departments from which the numbers and the postal codes are very distant). An approach consists in associating the co-ordinates of longitude and latitude, in order to integrate the constraints of proximity in the reasoning.

2.5.2 The multivariable transformation

It relates to the combination of several elementary variables in a new aggregate variable. Indeed, the raw data are sometimes insufficient to bring a predictive capacity to a model. The types of transformation are multiple.

The ratios

The comparison of two indicators in the form of ratio makes it possible to circumvent the weakness of some software or some techniques of modelling. The amount of the relative purchases a family will be brought back to the total amount of the purchases to appreciate the degree of implication of the customer for this type of articles.

The frequencies

The follow-up of the data in time makes it possible to measure the repetitivity of the exchanges : a number of commands on X last periods.

	Time 1	Time 2	Time 3	Time 4	frequencies
Example 1	No	Yes	No	No	25 %
Example 2	Yes	Yes	Yes	Yes	100 %
Example 3	Yes	Yes	Yes	No	75 %
Example 4	Yes	No	Yes	Yes	75 %
Example 5	No	No	No	No	0 %

Table5 : examples of frequencies

2.5.3 Tendencies

The pattern of trade in time makes it possible to follow the progression on behalf of market of the sign in the budget of the customer. It is expressed by a growth in a number or sales turnover observed between the last periods and can be written in the form of or not linear equations.

	Time 1	Time 2	Progression	Tendencies
Example 1	235	536	128 %	++
Example 2	214	210	- 2 %	=
Example 3	345	100	- 71 %	-
Example 4	200	200	0 %	=
Example 5	110	4200	3718 %	++

Table6 : examples of tendencies

2.5.4 Linear combination

The expression of some concepts is built with the experts by the installation of indicators combining of the primary data. So, in the field of the credit, minimum with living, i.e. the share of the residual income after deduction of all the recurring loads, will be expressed in the following form :

$$\text{Income} - (\text{Charges} + \text{Number Adults} \times X F + \text{Nb children} \times Y F)$$

The combinations between variables also make it possible to calculate moving average or to measure phenomena of seasonality.

2.5.5 No-linear combination

The stock-brokers accustomed us to the calculation of complex composite indicators containing nonlinear formulas. It is indeed in the field of the prediction of course that one will generally find aggregations of variables by nonlinear formulas. The oscillator %R, indicator used in Chartism, will be calculated on a time series by :

$$100 * \frac{Hn - C}{Hn - Bn}$$

With :

- C for cloture of the day.
- Hn : high highest of the period considered.
- Bn : low of the period considered.

2.6 Seek model

2.6.1 Introduction

The choice of the calculation algorithms is determining for the performance of the model. It is necessary initially, to position the new tools of Data Mining compared to the statistics. There is not clear border between the tools statistical and the new tools of the inductive type, Bayesian or neuronal. The theory would like that Data-Mining it is exploratory, while the statistics would be confirmatory. In the facts, the algorithms of Data Mining are based, for whole or part, on the work completed by the statistical community. The new techniques of Data Mining seem more one extension of the statistical methods that like a revolution.

To position the various techniques of modelling, one uses a typology of the problems around three large poles :

- the search of the models based on equations, where the decision maker is based on a more or less complex function which combines the variables.
- logical analysis where the decomposition of the problem in successive subsets makes it possible to build a structured reasoning .
- techniques of projection where the initial complexity of the problem is reduced thanks to the description of the principal factors of explanation .

2.6.2 Equations models

They break up into two branches :

- the branch resulting from the statistics with the techniques of linear or logistic regression, the discriminating analysis.
- the branch resulting from the neuronal techniques with a distinction enters the networks of neurons, according to the technique of training (retropropagation, softmax, etc..).

The statistics remain relatively dominating in the models of equations with, in particular, the regression analysis, and the discriminating analysis more known under the name of scoring.

2.6.3 logic analysis

It also breaks up into three branches which represent three methods of inference:

- The analytical method consists in drawing a series from conclusions starting from the facts. All the conclusions will not be true to 100%, but the distribution of the facts within a conclusion (97% without defect and 3% with defect) makes it possible to build a diagnosis :

```
Florence is nice,  
Sylvie is nice,  
Aurelie is nice,  
All the women are nice (100% true)
```

The analytical method consists in drawing a series from conclusions starting from the facts. All the conclusions will not be true to 100%, but the distribution of the facts within a conclusion (97% without defect and 3% with defect) makes it possible to build a diagnosis :

- The abductive method seeks, starting from a list of deduction, to build diagnostic :

All the pretty women are perfect,
Florence is perfect,
Therefore Florence is a pretty woman.

The abductive methods are still relatively emergent, they tend to limit the size of the decision trees by seeking the elements more determining to synthesize information. This effort of synthesis is found in the techniques containing blur, some approaches based on genetic algorithms and the tools for associations.

- the last method of inference, the deductive method, seek, starting from a list of facts (premises), with lead a reasoning. It is used in the development of the expert systems to apply a reasoning thanks to the instantiation of rules of productions :

All the perfect women are pretty,
Florence is perfect,
Therefore Florence is pretty.

2.6.4 Production technic

They seek to restore an overall vision of a problem. The examples are to position on more or less structured levels. One generally distinguishes the factorial techniques, which associate the axes (called factors) the point to build an interpretation of the points, and them analysis from typology which position the examples compared to concepts of proximity. The techniques of projections are very clearly dominated by the statistics. It should be noted that the choice of the model has consequences, not only on the performance of the model, but also on the type of restitution of the results (tree...), and on its adequacy with the required objectives.

Knowledge will be more easily accessible by the combination from the various techniques which often contribute to a significant increase in the result.

2.7 Evaluation of the result

The evaluation of the result makes it possible to estimate the quality of the model, i.e. capacity to be correctly determined the values which it is supposed to have learned on new cases. This evaluation generally takes a qualitative form and a quantitative form.

2.7.1 the qualitative evaluation

The restitution of knowledge, in graphic or textual form strongly contributes to improve comprehension of the results and facilitates the sharing of knowledge. The restitution in an interpretable form contributes to improve the appreciation of the result.

2.7.2 the quantitative evaluation

The techniques of restitution in the form of rules contribute to the work of communication between the people implied in the project of Data-Mining. It is accompanied by indicators which measure the capacity of relevance of the rules and the threshold of confidence according to the sample size. The precision of a survey does not depend on the relationship between the sample size and the size of the population mother, but only of the sample size. The precision of a survey of 1000 people is identical, whether the population mother is 10 or 20 million people. This revision is evaluated by a threshold of confidence and a confidence interval. So, for a threshold of confidence of 95 %, the confidence interval is given by the formula :

$$i = \pm 1,96 \cdot \sqrt{\frac{p \cdot (1 - p)}{n}}$$

n : effective on the sample

p : frequency observed

This interval measures confidence that have can grant it to a survey. For example, if on a sample of 30 individuals, one notes the appearance with 65% of a phenomenon, we will be able to affirm that there is 95% of chances so that the percentage on a population mother amounts to 65%, more or less the confidence interval, equal to 17%. The percentage on the population mother lies between 47 and 82%! If a sample of 300 people is taken, the confidence interval varies from 5%. The percentage on the population mother then lies between 60 and 70%.

The increase in the sample size, as this example shows it and as one suspected it, makes it possible to make reliable the conclusions.

Validation by the test

With the exit of the construction of the model, it is theoretically possible to test the relevance of it on the basis of training evoked previously. It is frequent that some tools learn the data rather than the models. For example, the fact of forgetting to brew the data can result in obtaining a model which learned that the first 1000 recordings belong to class A and the 300 following with the class B! The best remedy to counter this risk consists in brewing by chance the data before any training and especially envisaging a distinct base of test. To validate models, it is preferable to constitute as a preliminary a base of test being used only for the test : the model discovers the examples which appear in it. The data of test subjected to the model make it possible to check if it is able to classify in a correct way of the data which it never met before. Stability between the results observed on the file of training and the file test is known under 'capacity of generalisation'.

In general, the performance of the model is appreciated through matrix of confusion which compares the real situation and the situation envisaged by the model.

2.8 Integration of the knowledge

Knowledge is nothing as long as it is not converted into decision then in action. This phase of integration of knowledge consists in establishing the model or its results in the information processing systems or the processes of the company. It is essential since it acts of the transition from the field from the studies to the operational field.

In some cases, data-processing integration is not necessary and the writing of a report/ratio or a book of procedure proves to be sufficient. But, most of the time, the model will find all its utility while being established in the information system either in the form of a data (the result of the model), or in the form of a processing (the algorithm of the model).

With the occasion of this final phase, it is also convenient to draw up an assessment of the course of the preceding stages. This assessment is used to improve existing it in terms of data and collection of these data :

- the low noted quality of the data results in re-examining the processes of feeding of Data Warehouse.
- the detection of the strong predictive capacity of a data pushes to modify the diagram of the data base and the rate/rhythm of feeding.
- the aggregates built in the process of analysis prove to be dimensions interesting for the piloting of the company and contribute to the extension of the existing management reports.
- extracted knowledge is in contradiction with existing knowledge, in which case a communication and explanations will be necessary.

Chapter 3

Technics of Data-Mining

3.1 Reasoning containing case

The systems of reasoning containing case (CBR) solve problems by the comparison of examples from a lot of cases stored before. With this method of resolution, if a last experiment and a new situation are sufficiently " similar ", all the conclusions applied to the last experiment remain valid and can be applied to the new situation.

The CBR follow a procedure of search to compare the descriptions of the case to be treated with those of the existing cases in their internal base. For this reason, the capacity of resolution increases with measurement of the addition of new examples in the reference index. The more significant the number of examples will be and the more the CBR will be likely to find a close example, even similar.

However, the growth of the base complexes the bringing together of a new case with N cases present. To mitigate this explosion combinative which appears when the base of case is packed, the CBR propose particular techniques intended to improve the capacities of search and speed. These optimizations require the addition of human expertise to enrich and guide search towards the most relevant criteria.

Moreover, the use of a tool based on decision trees, for example, facilitates the identification of the most significant criteria for the measurement of similarity. The combination of the techniques of Data Mining is frequent for the implementation of a CBR. Contrary to the expert systems, which distinguish base from knowledge and bases case, the CBR maintain in constant relationship the training and the reasoning. This amalgam avoids the collection of expertise, often expensive and difficult operation.

Moreover, the addition of the new cases comes to regularly enrich the capacities by deduction by the system. It acts of a significant advantage compared to the expert systems. The latter present a fixed vision of the reasoning, which requires regular installations of the base of knowledge. For this reason, the CBR seem a good response (pragmatic and evolutionary) to much of problems of diagnosis of breakdowns and assistance to the users. The recourse to the CBR does not exempt a certain structuring of the problem to facilitate the search of the similar cases. To illustrate that, let us take the index of similarity between the three following examples. It is of 75% (3 criteria of 4).

However, the common direction pushes us to note that the similarity between the two printers laser is stronger than that which exists between a laser and a screen. Also, to improve quality and the time the search, it is necessary to build a hierarchical structure of the variables. Dimension structuring is used as key to index the criteria and to avoid an exhaustive search for a similarity between a case, and the n-1 other cases of the base.

3.1.1 Construction of CBR

There are four stages in the construction of an CBR

1. The collect of the Data.
2. The search of the relevant factors.
3. The indexing of the data.
4. Tests and improvement of the performance.

Collect of the data

The data base of a CBR consists of case. A case represents a situation characteristic of an applicability. It gathers two types of information : a collect of facts which describe a state particular and coherent field, and a whole of deductions or interpretations applicable to the collection of facts. The input data are structured in the form of variables, defined by a finished list of methods, or in free textual form. This last type of format complexes the work of analysis. Indeed, it is more difficult, in this case, to identify the relevant factors and to isolate the context.

The data-gathering can take two forms : if the data exist in the information systems, the collection consists in building interfaces starting from the existing files. In the contrary case, the data-gathering passes by an effort of data entry to constitute a first whole of relevant cases. It is obvious that the number of examples is in relation to the number of variables and with the diversity of the possible values for each variable. To establish a parallel with the physical world, the addition of variables amounts increasing the number of parts in a dwelling, and the addition of methods is equivalent increasing the number of the cupboards in each part. The definition of a too large universe, with a cover in a too low number of examples, will result in a weak similarity between a new case and an existing case. It will be consequently difficult to obtain a good diagnosis.

Pertinent factor seeking

Most of the time, to provide a base of case to a CBR is not enough to solve a new problem. It is necessary to build a mode of representation of the data, more structured possible, according to the goals of the expert. This structuring of the data makes it possible to define the level of detail necessary to solve the whole of the cases. It passes by the construction of the hierarchy of the data items and leads to an indexing of the criteria. This one aims to accelerate the search and the selection of the cases. There are several techniques to build the hierarchy of the data items.

Keyword seeking The first method consists in making an analysis starting from the key words which describe an example. The measurement of similarity consists in counting, among the examples of the base, those which present the most common key words. One builds a distance between the new example and the cases present in the base according to the following formula :

$$d(a, b) = 1 - \frac{nbc}{nb}$$

$d(a,b)$: distance between a and b

nbc : number of common keywords between a and b

nb : number of keywords a or b

Key-Word case 1	Key-Word case 2
oil	smoke
odor	odor
noise	noise

Table7 : an example.

In this case, $d(\text{case 1, case 2})$ is :

$$d(a, b) = 1 - \frac{2}{4} = 0,5$$

Hierarchically concept The second method, less commonplace, consists in describing, at the time of the construction of the CBR a hierarchy of the concepts to describe a problem. The structuring of the problem, in the form of a tree structure, makes it possible to limit the measurement of the distance to the only relevant cases. For example, if a breakdown is localized on the screen of the computer, it is not very probable that the format of the diskettes takes into consideration the diagnosis. The creation of a classification of the topics makes it possible to determine contexts of analysis and to list the relevant factors for each specific context. The base of case being structured, it remains to associate a new case with others which present a similar context, and to present at the user the possible diagnoses. The search

for the most probable solution is based on the number of times where the case was presented and on the distribution of the diagnoses on this subset of case.

Data indexation

The indexing of the CBR consists in balancing the various criteria used for the calculation of the similarity between the new case and the existing cases. It aims to improve the performance of the diagnosis when the similarity between an existing case and the case to be analyzed are not strict. The indexing limits the number of the cases to those which are potentially similar with the new case, while identifying closest. Filtering improves the precision and the reliability of the diagnosis and decreases the search time.

The method of the most frequent indexing consists in seeking the closest neighbors of the new case with a function of similarity. It calculates a distance of the new case compared to the cases having the same context, then selects the shortest distances and presents at the user the most frequent diagnoses. As it is seen, the choice of the function of similarity is crucial. It will have a direct influence over the response times of the CBR. The presentation of the case nearest will take place starting from a simple counting or will require the determination of a function of similarity.

Case enumeration A first simple approach consists in counting the number of diagnoses present and carrying out a simple work of frequency. The answer is in this case the most frequent value. This technique of counting can be powerful if all the examples belong to the same class, or if the field is really restricted. In the other cases, the relevance of the response of the CBR rests on the level of precision and definition of the classes, which returns us at the preceding stage. It is frequent that the construction of this classification is carried out while being based on statistical or inductive techniques of classification.

The weighting of the criteria The one second measurement of similarity introduces a weighting of the criteria to define a total function. The algorithm of analysis of the CBR then selects the cases which present a minimal threshold of similarity, and seeks then the various types of diagnoses present in this subset of examples. This double processing makes it possible to present the various possible diagnoses at the user, with at the same time a frequency (percentage of time where this diagnosis is met) and a distance (method of the closest neighbors).

The structuring of the inputs makes it possible to very quickly identify the subset which contains the most interesting examples, with a minimum of questions. The search of the closest neighbors, allied a technique of weighting, makes it possible to sort the possible answers and to present at the user the whole of the possible diagnoses. The most probable answers will be presented at the head of list, and the least probable will be in bottoms of list or eliminated if the list is already long. As a CBR provides explanations to its proposals, they can be used to correct and improve the parameters of indexing.

Tests and improvement of the performance

Performance measure The last stage of realization of a CBR consists in measuring its level of performance. It is a question of launching a phase of diagnosis of the system of CBR on the whole of the cases. As several answers are possible, only that which presents the strongest similarity is retained. The comparison between the real diagnosis and the diagnosis predicted by the CBR makes it possible to build a matrix of confusion.

The comfort of use We saw how to build the " engine " of a CBR. It is necessary also a " interface for him " to dialogue with the user. The user interfaces are generally designed from a point of view of productivity in the data entry of the cases, and especially with a maximum of assistance to the data entry to limit the risks of errors of data entry. The menus with predetermined choices or the recourse to hypertext links are current solutions. They make it possible to improve quality of the descriptors introduced into the CBR.

3.1.2 Applicability

The applications of systems CBR are multiple. The majority of successes of this technique relate to the service after sale where the diagnosis of breakdown. The CBR contribute to improve the total

performance of the centers of call and to homogenize the consulting, even apart from the business hours where the experts are rare. The CBR can be directly integrated, in the form of a microprocessor or of a connected PC, in the product (a computer, an autopilot, a machine tool etc). In this case, one speaks about embarked application.

3.1.3 Limits and Advantages

The difficulty in integrating the textual data One of the problems of optimization of the CBR is related to the textual data management not structured. In this case, the search of the similarities is built starting from the identification of the key words; each case of the base being indexed with key words, this can lead to two types of problems : the case is indexed with a multitude of key words, in which case it will too often seem a possible diagnosis or, on the contrary, it is indexed with few key words, in which case it can never not be extracted. The choice of the key words is determining.

Problems of evolution

The problem of evolution arises initially at the time of the arrival of a new description, not envisaged in the phase of creation. Thus in a bearing diagnosis on computer equipment, a new type of hard disk must be able to be added in the list of the objects not to lose the concept of hard disk in the search of the case. The phase of indexing is longest in the construction of a CBR. The resulting structure is generally fixed. It poses problems of flexibility of the CBR which can find efficiently only the equivalent examples or sufficiently close relations. The successive addition of new keys of indexing, too many, fatally results in reducing the effectiveness of the CBR. One of the ways of current search consists in rebuilding in an automatic way the indices of indexing with techniques of decision trees.

Growth of the base

The performance tends to degrade with the growth of the base of case, when this one reached several thousands of examples. It is then necessary to re-examine the processes of classification and indexing to optimize the diagnoses suggested as well as the response times.

Weak costs of maintenance

Times of development of a CBR are about 3 to 6 months for a " normal " problem. These orders of magnitude are comparable with those of an expert system, except if the data base is already structured. On the other hand, taking into account their capacity of evolution, the CBR present weaker loads of maintenance. They for this reason offer recourse on often significant investments.

3.2 The Knowbots

3.2.1 What is a Knowbots

The term of Knowbots is a digest of Knowledge and Robot. Into French Knowbots were translated by intelligent agents . An agent is a physical or abstract entity able to act on itself and its environment. It has a representation partial of this environment and can communicate with other agents. It pursues an individual goal and its behavior is the consequence of its observations, its competencies, and the interactions which it can have with other agents and its environment. An agent has a personal objective which strongly distinguishes it from the traditional data-processing programs very collectivists. It is a software entity which shows the following characteristics :

- .Gerable (it takes its instructions of a man or an agent)
- .Autonomous (it preserves his own interests)
- .Persistent (it can nothing make over long periods)
- .Reliable (it meets the user's needs)
- .Far-sighted (it can anticipate the needs)
- .Active (it can take initiatives)
- .Communicating (it interacts to solve the problems or conflicts)
- .Adaptive (it can change environment)

3.2.2 Use

The explosion of Internet increased accessible volumes of information considerably. To be convinced some, it is enough to launch a search on a key word to find a few tens of thousands of sites in report. Such a quantity of data represents more one handicap than an advantage. The intelligent agents found in this field a sector completely adapted to their functionalities. With the service of the user, they are able to generate and carry out a research plan, to solve the problems in the execution of this plan and by interaction with the user, to improve their behaviors. These types of agents do not raise of data mining since they make only reproduce one manual process. However, with the development of the electronic trade on the Web, of new commercial agents (electronic advisers), are set up and one can completely compare their functions to the tasks of data mining.

3.2.3 First example

The electronic advisers the opening electronic commercial (on the Web) offers a new prospect for the intelligent agents. They can carry out the two facets of a negotiation : there are salesmen agents and negotiators agents.

NEGOTIATORS AGENTS : an negotiator agent traverses a list of potential salesmen. It diffuses a request for tariffs on the visited sites. By the same occasion, it fixes a time at its request. It manages the answers and takes care of the revivals. Lastly, it draws up a report for the applicant. The user selects the salesman and the agent takes care of the sending of the purchase order.

SALESMEN AGENTS : an agent salesman learns how to know a customer by examining his purchases and by supplementing its knowledge by complementary questions. The proposal for some offers and the answer of the customer make it possible the agent to build a precise profile of the purchaser. The agent is endowed with a capacity of training which enables him to know the potential customer better and better. Consequently, and thanks to the tools of data mining, it is able to make commercial offers of type one to one.

3.2.4 Second example

Currently experiments are in place in the United States. On subscription, a user initializes the process by filling out a questionnaire on what he likes and hates. Thereafter, all the electronic purchases are recorded and re-installed towards the company holder of the system. In same time, by analogy of tastes and behaviors, the system makes proposals individualized with the customer. This type of service is consultable for example on <http://www.firefly.com> or <http://www.hotmail.com>. at the beginning of 1998, the Microsoft company repurchased the company hotmail (which proposes this type of services) and has 9 million subscribers. In May 1998, the company firefly at summer also repurchased by Microsoft. This last plans to integrate this principle in the future versions of its navigator (Internet Explorer). The interest of Microsoft for these technologies shows its strategic and commercial stake well.

3.2.5 Conclusion

The intelligent agents or Knowbots are autonomous software entities whose most recent versions are integrated completely in the process of data mining. Some will go until regarding them as tools of data mining. Some of them, most elaborate, are able to follow and memorize the movements, visits and purchases on Internet and make it possible to work out user profiles to make them to commercial offers one to one. the user can, as for him, throw of the invitations to tender and competition automatically managed by these agents. This evolution (just like that of data mining) lead us to ask us questions of ethics and respect of the private life.

3.3 The associations

3.3.1 Definition and stakes

The search for associations aims at building a model based on conditional rules starting from a data file. A conditional rule is defined in the form of a continuation "if Conditions then Result". It is possible of mixer several conditions to reach a result : if A and B then C the combination of several logical operators introduced between the conditions makes it possible to extract from associations of conditions in elaborate formats : if A and not D then C. The search of associations can take place on the whole of the data (all the conclusions are tested) or on a target data (the conclusion is fixed by the user). The principal uses of the search for associations currently touch the diagnosis of credit as well as the analysis of the sales slips, that of the operation of the credit or discount cards.

Analyze sales slips

The analysis of associations (also called, in this case, basket analyzes) finds the its application most immediate in the analysis of the data of the points of sale. It is a question of identifying affinities existing between bought products and services. The analysis of associations leaves the finest data which make a transaction : elementary sales of articles. The search of associations aims at finding the connection which exists between two or N produced (80% of the purchasers of breeches layers buy beer; the tomato and salad purchasers buy oil in 80% of the cases), but also between behaviors of products (when the sales of X increase the sales of Y then increase in 80% of the cases).

Analyze sequences of purchase

The analysis of associations can function either into instantaneous to seek all associations of the same transaction or on the same sales slip, or in time to detect associations of sales, on the same customer, during two or three years. In this case, dimension time is obtained by using either the number of a card of payment, or that of a discount card. The search for associations in time adds a temporal dimension to the analysis and a concept of anteriority. A hypermarket can discover that 35% of the subscribers to a proprietary card bought an article electric household appliances during 6 months previous.

Stakes

The applications of the search for associations are multiple. They go from a better knowledge of the customer, of its basket, until the optimization of stocks or the "marchandising" (??).

- Optimization of stocks. The discovery of a logical sequence of the transactions allows the optimization of the procedures of provisioning of a store.
- "Marchandising". The discovery of associations between products can involve a reorganization of the surface of sale. For example, the observation of associations between food articles, clothing and pieces of furniture for the "toddlers" can result in defining a new space "child welfare" in a catalogue.
- Cross sales. The discovery of associations allows the realization of promotional campaigns personalized with the edition of reduction vouchers according to the purchases : if one notes the presence of coffee X in the transaction, then one publishes a reduction voucher for sugar Z because it is generally associated coffee X This personalized edition is carried out at output of case or is joined to the statement of the proprietary card.

This form of marketing of intimacy is essential to facilitate the purchases of the customer and to optimize the policy of restocking of the store. But the analysis of associations appears before just like the means of building the differentiation of a sign.

3.3.2 The applicability

This presentation of the operating mode of the search for associations makes it possible to understand that all the commercial transactions can be analyzed by means of an engine of associations. Consequently, the applicability is numerous and the most frequent uses touch the analysis of the purchases in the great distribution, the analysis of the movements in the bank, the analysis of the incidents in the insurance, or analyzes it communications in telecommunications. More generally, the analysis of

associations applies successfully in all the problems where the appearance of an event is conditioned by last events : analyze breakdowns in industry or study of the decisions in sociology.

3.4 The decision trees

3.4.1 Definition of chi 2

Il s'agit d'une technique qui tablit l'existence d'une relation entre deux variables qualitatives. Le test du chi 2 repose sur une comparaison de la frquence de distribution de ces deux variables une distribution thorique. Il consiste calculer la somme des carts entre la distribution observe et la distribution thorique et comparer ce rsultat une valeur prdtermine en fonction de la complexit du tableau.

Le test du chi 2 prsente cependant des limites qui doivent tre prises en compte avant qu'il ne soit utilis aveuglement : le test d'indpendance du chi 2 ne peut tre employ que si les effectifs totaux sont suprieurs 30 et si les croisements des modalits ont toujours des effectifs suprieurs 5.

3.4.2 Introduction

A decision tree is a hierarchical sequence of built logical rules in an automatic way starting from a base of examples. A logical rule includes a premise (the first part of the rule) and a conclusion (the second part of the rule). The premise expresses a logical condition built on tests relating to variables combined by logical operators (and, or, not...). The conclusion is supplemented by a frequency of membership (for a qualitative variable) or by an average (for a continuous variable).

3.4.3 Analogy with the trees

The arborescent shape of the decision trees is obtained by the next base division of examples using a sequence of decisions. The whole of origin, where all the examples of the base are present, is called the node root. This one is cut out in a successive way, in subsets called intermediate nodes. On each node, a new evaluation is made for a cutting in subsets. The final nodes are called sheets. From the decision tree obtained, it is hardly difficult to deduce from the rules. They describe, in the shape of a logical system, the path of reasoning. The connection between two levels can be compared with one " and " logical, and thus be read in the following way:

```
If (age  $\geq$  65) years and (sex = female)
then no_sale := 87%;
```

ID3 Algorithm

These systems of inductive training are based, for the majority, on system ID3, presented by Ross Quinlan in 1979. Its guiding principle rests on the manufacture of a tree of classification, starting from an experimental whole of examples. Technique ID3 calculates the minimal decision tree, by seeking, on each level, the parameter more discriminating to classify an example. It determines for that the sequence of attributes which leads as soon as possible to a correct classification. The visualization of the decision tree makes it possible to interpret the whole of successive cuttings immediately. One measures the quality of the model generated by his capacity to affect the examples in their good classes.

3.4.4 Stakes

The analysis of a tra-byte will require several years of work with a statistician. The possibility of extracting, in manner automatic, certain rules is the means of facing exponential growth of the data bases. Moreover, analysis automatic makes it possible to multiply the number of analyses. It is, with this title, a significant factor of competitiveness for companies which process data. Thus, a company who wishes to improve his production process can seek causes of failure of the whole of the components by one iterative method.

Detection of important variables

The very explicit formalism of the decision trees puts in obviousness the most significant variables. The construction of logical bonds between the variables makes it possible to structure very quickly the studied phenomenon. This structuring of the problem is a first stage to set up solutions correct. An engineer, who discovers that the combination of one temperature of more than 65 , on sensor 34, and of a pressure lower than 2 Bars, on press 3, a growth involves of 25 % of the rejects, can set up correct measurements targeted.

Information system construction

The possibility of locating the most relevant variables is also significant to build the information system. When it is a question of controlling a system, or of anticipating them evolutions of the systems, it is of primary importance to have data reliable and relevant. Analyses by decision tree, in helping to include/understand the key variables, will be able, for example, to improve the rules and the methods of food of one Dated Warehouse, or to refine the processes of historisation and of safeguard.

Data Mining of mass

The decision trees have a simple formalism. Restitution of a decision tree is easy to read. After a formation from one half-day, at one day, it is possible to entrust one software based on decision trees to a user trade. market very quickly included/understood the complementarity which these tools have with the traditional products of requests (association of Object business and of Alice, Impromptu and Scenario) and with spreadsheves. Forecasts of the number of users of the trees of decision are estimated (in a future that the editors want near!) to 10 % of the market of the spreadsheves. Awaited growth market of Dated Mining will necessarily pass by this type tools.

3.4.5 principles of calculation

The algorithm of determination of the significant variable is the base of the technique of construction of the decision trees.

The search of the order in the disorder The algorithm seeks to decrease the apparent " disorder " which exist in the data, while being based on a function of evaluation. There are many alternatives of this algorithm. Nevertheless, the common principle consists in choosing, with each level, the variable which makes it possible to extract the maximum of information. A good decision tree makes it possible to classify it better possible and by posing the minimum of questions (i.e. comprising a minimum of depth). We will illustrate it operation of an algorithm of decision trees with one simple function.

3.4.6 The descriptor is qualitative

The measurement of uncertainty, in the case of a variable qualitative, a formalism different borrows from pseudo-metric the Hamming one. Indeed, one uses, for this type of variable, probability of membership of 1a variable to one classify. For example, if a variable can take the values " large ", " average ", " small ", and if, among 100 observations, one 20 times the value has j large ", one associates 20 % this value.

Algorithms derived from information theory

The measurement of uncertainty can be appreciated by means of theorem of Shannon on information: $\sum P_i \cdot \log P_i$ with P_i who represents the percentage of membership of a class. This indicator is minimal when the probability of a class is equal to 1 (all the examples belong to only one class). If four classes out of four are represented in manner equiprobable, uncertainty maximum like is translated the indicator of Shannon who is worth two in this case.

This indicator is a good measurement of uncertainty or disorder. Principal technique developped at the point by J.R Quinlan compare the evolution of this indicator, during the test of one variable, to detect most discriminating. For each descriptor, one calculates the disorder which remains after its use. That which leaves less disorder is selected as being the next node of the tree of decision.

Algorithms from chi 2 statistic

Another approach of creation of the decision trees results from algorithm CHAID. Here, the definition of the variable more significant is based on the test of Chi 2. As we have it considering previously, the test of Chi 2 makes it possible to check conformity of a random phenomenon to a law of probability posed like assumption. The principle of Chi 2 is based on the comparison, enters frequencies observed, for each of the classes, and them theoretical frequencies. These last materialize the situation of independence enters the variables. Various methods (AID, XAID, Theta Chi 2, etc.) allow to circumvent the limits or skew of certain indicators. It acts, for example, to correct insufficiencies of Chi 2 when manpower are too weak.

3.4.7 The descriptor is quantitative

The objective is identical. However, the method changes because them values of the concept can be in infinite quantity.

The method of grape

A first method, known under the name of bunch, consists with to cut out the variable continues in ordered subsets. It cutting is built from traditional indicators such as the average, the median (for a partition in two classes) or deciles (for several classes). Thus, on a population of 1 000 individuals, there variable age is cut out in ten classes by a sorting on deciles.

Cutting by decile makes it possible to define the limits of each one classes. The method of bunch with several classes calculates, according to the formulas applicable to the discrete variables, the profit of information brought by each variable. The number of classes for the whole of the quantitative variables being equal (by example, 10 classes), the calculation of the profit of information is identical for all the variables. The function (for example ID3) allows to select the most discriminating variable. However, a cutting in 10 classes on each level is too fine it creates a tree quickly illegible: 10 nodes with the first level, 100 on the second level, 1000 with the third. In order to avoid this tree structure with multiple nodes (the bush of decision), a test is carried out between the various nodes adjacent to gather the methods with the tiny differences.

The method of bunch presents the disadvantage of not guaranteeing one optimal threshold of division variable. Indeed, if there the clearest difference is between the people of less 30 years and those of more than 30 years, the class created by decile of 27 to 34 years loses this threshold. Nevertheless, this method is fast in computing time and approaches the good value.

The exhaustive method

This method determines the optimal threshold of cutting of variable. This threshold is selected so that the partitions of explanatory variable make it possible to discriminate as well as possible the attribute. It acts, in the exhaustive method, to evaluate all possible thresholds and to retain the best. To choose it optimal threshold, all the possible values of the attribute are traversed by ascending order. With each value, one carries out one partition of the attribute and one calculates the discriminating capacity of the variable. When all the field of the values was traversed, the threshold retained for the binary partitions is that to which the best discriminating capacity corresponds.

The exhaustive technique is very expensive in computing times if it a many numerical attributes are significant and if each variable numerical many possible values present. On the other hand, it ensures a better attribute division.

3.4.8 Application domain

The applications of the decision trees are of two types there construction of an algorithm of segmentation of a population of which the groups of assignment are known, and the assignment of one classify with an individual, starting from certain descriptive elements. Taking into account the simple formalism of restitution, fields of application are numerous; the list below reflects them principal applications but does not want to be exhaustive:

the studies marketing, to include/understand the dominating criteria in the purchase of a product or the impact of the publicity expenses,

direct marketing, to isolate the best criteria explanatory of a behavior of purchase, sales, to analyze the performances by area, by teach or by salesman,

3.5 Bayesian networks

3.5.1 Presentation

The Bayesian networks aim at the discovery of the relations. They indeed allow the comprehension of some relations. The principle is that of a system which creates an inter-connected network of probabilities of concomitance between several actions and conditions. For example the small companion of Microsoft Office which is an intuitive help. Their operation is based on the graph theory.

3.5.2 Recall on the theory of the graphes

A graph is made up :

Of nodes which represent the objects.

Of edges which connect the objects.

Of paths which are an ordered succession of nodes connected by edges. A graph can related, completely related, be balanced or directed. Related graph : there is a path between each node. Completely related graph : there is an edge between each pair of node.

Balanced graph : each edge has a weight.

Directed graph : an edge is representative of a direction.

3.5.3 Use

A Bayesian network is a directed graph in which the nodes represent the variables and in which the edges symbolize the dependence between the variables. It measures the probability of appearance of an event knowing the result observed on other variables.

3.5.4 Setting

The installation of a Bayesian network passes by the following stages :

- preparation of the variables
- selection of the variables
- identification des dependence
- matrix probabilities.

3.5.5 Variable initialization

This stage consists in distinguishing the discrete and continuous variables, then to define intervals for the continuous variables.

3.5.6 Variable selection

It is the determination of the variables of input and output, knowing that a variable of output cannot be an input for another variable The selection is carried out by a classification according to the measurement of entropy whose formula is as follows.

$$H(X) = - \sum P(x) \log P(x)$$

with X a variable and P(x) its probability of appearance.

3.5.7 Dependence identification

That consists in measuring the dependence between the node then to classify them by descending order

The measurement of the dependence between two variables is carried out by calculating a factor of dependence.

$$I(X/Y) = H(X) - H(X/Y)$$

This factor makes it possible to determine the incidence of the variable Y on variable X, if this result is null, that means that the variables are independent. The difficulty remains to be determined which threshold is significant.

3.5.8 Matrix Probability

It is the counting of the occurrences between the node in order to establish the probabilities.

3.5.9 Advantage

Good resolution :

This technique allows a very good resolution of the problems based on the links It can be used in the fields of transport, telecommunications, .

Good visibility :

From its with dimensions graph, it gives a good visualization of the results and highlights the relations between the various elements.

Good discovery of relation :

Once the analyzed links, new decision criteria can be set up.

3.5.10 Inconvenient

Bad adaptation, Few tools, Bad performance.

Bad adaptation :

One of the principal disadvantages of this technique is the difficulty of adaptation to many types of data.

Few tools :

Moreover, there exists at the present time little of tools on the market.

Bad performance :

This technique generates many readings and many calculations, due to the multiple possible combinations what makes it very consuming power machine.

3.5.11 Conclusion

In conclusion, the Bayesian networks are not well adapted to the prediction or the classification of the data, but the relations discovered are a good entrance point for other techniques, the such networks of neurons or the decision trees.

Chapter 4

Conclusion

There are a lot of methods to practice Data-Mining, we saw just a part of the iceberg, we can use genetic algorithms or neuron networks too. Genetic algorithms copy on the nature. They have the same mechanism of natural selection : the selection, the reproduction, the mutation and 'crossing-over' like in real life. They describe the evolution during successive generations in function of the environment. The neuron network imitates the functioning of a human brain with stimulus, time of rest ... In these two new methods, Data-Mining takes example on the nature in order to have the best results as possible, but to have good results it's more difficult.

Data mining is very useful for the majority of the enterprise. With it, they, in most of the time, increase easily their performance and so make more money !

Chapter 5

Bibliography

To do this report, we use these supports :

- internet
- books

5.1 the web site

<http://www.datamining.org/>
<http://www.pcc.qub.ac.uk/tec/courses/datamining>
<http://www3.shore.net/~kht/text/dmwhite/dmwhite.shtml>
http://www.cs.bham.ac.uk/~anp/dm_docs/dm_intro.html
[http://www.math.chalmers.se/Cs/education/ ...](http://www.math.chalmers.se/Cs/education/)
<http://www.cisia.com/formations/>
<http://www.qucis.queensu.ca/home/cisc835/>
<http://home.nordnet.fr/~dnakache/valeurc/>
<http://santos.doc.ic.ac.uk/~yg/course/dmml/>
<http://scanner-group.mit.edu/DATAMINING/Datamining/datamining.html>
<http://www.iastate.edu/~CYBERSTACKS/4T9R.htm>
<http://www.pvv.org/~hgs/project/report/node8.html>
<http://open.cineca.it/datamining/articles/zanasi.htm>

5.2 the books and documents

- a report of a student about Data-Warehouse,1998
- "introduction au Data Mining" by Jambu Michel (Eyrolles),1998
- "Le Data Mining" by Lefbure Rene and Venturi Gilles (editor : Eyrolles),1999
- "Case-Based Reasoning" by Kolodner,1994

Chapter 6

Appendice

6.1 Our Application 1

6.1.1 description

This is the same example we use to explain associations. It shows the intermediaire step for an etud of sale slips.

From sales slips, the program first of all will count the number of presence of each articles, then it will calculate the frequencies of appearance. If this frequency is lower than a certain threshold (fixed in constant, but easily transformable into a variable), the answer below this threshold are deleted. Then, it will detect associations of two articles, present on the same sales slip and to make similar. This can to allow a "great surface" to make a reduction on a product, which will make it possible to increase sales of its associated object, or to bring them closer in the rays, or to even create a new ray with these articles inside.

6.1.2 the source code

```
program association;
uses
  crt;
const
  max = 5;
  max2 = 4*max;
  seuil = 30;

type
  ticket = record
    elt1 : string;
    elt2 : string;
    elt3 : string;
    elt4 : string;
  end;

  tablo = array[1..max] of ticket;

  result = record
    nom : string;
    nombre : integer;
  end;

  tablo2 = array[1..max2] of result;

{init des ticket}
```

```

procedure init(var tab : tablo);
var
  i : integer;
begin
  for i := 1 to max do
  begin
    tab[i].elt1 := '';
    tab[i].elt2 := '';
    tab[i].elt3 := '';
    tab[i].elt4 := '';
  end;
end;

{init du tableau de comptage}
procedure init2(var res : tablo2);
var
  i : integer;
begin
  for i := 1 to max2 do
  begin
    res[i].nom := '';
    res[i].nombre := 0;
  end;
end;

{remplir les tickets}
procedure remplir(var tab : tablo);
begin
  tab[1].elt1 := 'farine';
  tab[1].elt2 := 'sucre';
  tab[1].elt3 := 'lait';

  tab[2].elt1 := 'oeuf';
  tab[2].elt2 := 'sucre';
  tab[2].elt3 := 'chocolat';

  tab[3].elt1 := 'farine';
  tab[3].elt2 := 'oeuf';
  tab[3].elt3 := 'sucre';
  tab[3].elt4 := 'chocolat';

  tab[4].elt1 := 'oeuf';
  tab[4].elt2 := 'chocolat';
  tab[4].elt3 := 'beurre';
end;

{affiche les tickets}
procedure affiche(tab : tablo);
var
  i : integer;
begin
  i := 1;
  textcolor(yellow);
  writeln;
  writeln('Liste des tickets :');
  writeln('-----');
  textcolor(white);

```

```

writeln;
while (i <= max) and (tab[i].elt1 <> '') do
begin
  textcolor(yellow);
  writeln('TICKET ',i);
  textcolor(white);
  if (tab[i].elt1 <> '') then writeln(tab[i].elt1);
  if (tab[i].elt2 <> '') then writeln(tab[i].elt2);
  if (tab[i].elt3 <> '') then writeln(tab[i].elt3);
  if (tab[i].elt4 <> '') then writeln(tab[i].elt4);
  writeln;
  i := i + 1;
end;
readln;
end;

{affiche le vecteur resultat}
procedure affiche2(res : tablo2);
var
  i : integer;
begin
  i := 1;
  while (i <= max2) and (res[i].nom <> '') do
  begin
    writeln('NOM(S)      : ',res[i].nom);
    writeln('FREQUENCE : ',res[i].nombre);
    writeln;
    i := i + 1;
  end;
  readln;
end;

{recherche dans le vecteur resultat}
procedure recherche(var res : tablo2;x : string);
var
  i,j : integer;
begin
  i := 1;
  while (i <= max2) and (x <> res[i].nom) do
  begin
    i := i + 1;
  end;
  if (x = res[i].nom)
  then res[i].nombre := res[i].nombre + 1
  else begin
    j := 1;
    while (j <= max2) and (res[j].nom <> '') do
    begin
      j := j + 1;
    end;
    res[j].nom := x;
    res[j].nombre :=1;
  end;
end;

{compte le nombre d'elements}

```

```

procedure compter(tab : tablo;var res : tablo2);
var
  i : integer;
begin
  i := 1;
  while (i < max) and (tab[i].elt1 <> '') do
  begin
    recherche(res,tab[i].elt1);
    if (tab[i].elt2 <> '')
      then recherche(res,tab[i].elt2);
    if (tab[i].elt3 <> '')
      then recherche(res,tab[i].elt3);
    if (tab[i].elt4 <> '')
      then recherche(res,tab[i].elt4);
    i := i + 1;
  end;
end;

{creation du niveau 2}
procedure ajout(var res2:tablo2;ch1:string;ch2:string);
var
  i,j : integer;
  x : string;
begin
  x := ch1+' et '+ch2;
  i := 1;
  while (i <= max2) and (x <> res2[i].nom) do
  begin
    i := i + 1;
  end;
  if (x = res2[i].nom)
  then res2[i].nombre := res2[i].nombre + 1
  else begin
    j := 1;
    while (j <= max2) and (res2[j].nom <> '') do
    begin
      j := j + 1;
    end;
    res2[j].nom := x;
    res2[j].nombre := 1;
  end;
end;

{recherche de niveau 2}
procedure rechercher2(tab:tablo;ch1:string;ch2:string;var res2:tablo2);
var
  i,j : integer;
begin
  i := 1;
  while (i <= max) and (tab[i].elt1 <> '') do
  begin
    if (tab[i].elt1 = ch1) then
    begin
      if (tab[i].elt2 = ch2) or (tab[i].elt3 = ch2) or (tab[i].elt4 = ch2)
      then ajout(res2,ch1,ch2);
    end;
  end;
end;

```

```

    if (tab[i].elt2 = ch1) then
    begin
        if (tab[i].elt1 = ch2) or (tab[i].elt3 = ch2) or (tab[i].elt4 = ch2)
            then ajout(res2,ch1,ch2);
        end;

    if (tab[i].elt3 = ch1) then
    begin
        if (tab[i].elt2 = ch2) or (tab[i].elt1 = ch2) or (tab[i].elt4 = ch2)
            then ajout(res2,ch1,ch2);
        end;

    if (tab[i].elt4 = ch1) then
    begin
        if (tab[i].elt2 = ch2) or (tab[i].elt3 = ch2) or (tab[i].elt1 = ch2)
            then ajout(res2,ch1,ch2);
        end;
        i := i + 1;
    end;
end;

{compte le nom d'elements niveau 2}
procedure compteur2(tab : tablo;res : tablo2;var res2 : tablo2);
var
    i : integer;
    ch1,ch2 : string;
begin
    ch1 := res[1].nom;
    ch2 := res[2].nom;
    rechercher2(tab,ch1,ch2,res2);

    ch1 := res[1].nom;
    ch2 := res[3].nom;
    rechercher2(tab,ch1,ch2,res2);

    ch1 := res[1].nom;
    ch2 := res[4].nom;
    rechercher2(tab,ch1,ch2,res2);

    ch1 := res[2].nom;
    ch2 := res[3].nom;
    rechercher2(tab,ch1,ch2,res2);

    ch1 := res[2].nom;
    ch2 := res[4].nom;
    rechercher2(tab,ch1,ch2,res2);

    ch1 := res[3].nom;
    ch2 := res[4].nom;
    rechercher2(tab,ch1,ch2,res2);
end;

{virer les cas < seuil}
procedure virer(var res : tablo2);
var
    i,j : integer;

```

```

begin
  i := 1;
  while (i <= max2) do
  begin
    if (((res[i].nombre * 100) div 4) <= seuil)
    then begin
      j := i;
      while (j < max2) and (res[i].nom <> '') do
      begin
        res[j].nom := res[j+1].nom;
        res[j].nombre := res[j+1].nombre;
        j := j + 1;
      end;
    end;
    i := i +1;
  end;
end;

var
  tab : tablo;
  res,res2 : tablo2;

begin
  clrscr;
  textbackground(blue);
  textcolor(yellow);
  writeln;
  writeln;
  writeln;
  writeln('-----');
  writeln(' | DATA MINING | ');
  writeln('-----');
  writeln;
  writeln('      -< les associations >-');
  writeln;
  writeln;
  writeln;
  writeln;
  textcolor(white);
  writeln('      Ce programme montre les differentes');
  writeln('      etapes pour une etude de tickets de caisses. ');
  writeln('      Il met en evidences les doublets d''associations');
  writeln('      de produits. Cela pourra permettre, par exemple, ');
  writeln('      aux grandes distributions, de faire des bons de ');
  writeln('      reduction sur des produits cibles en esperant ');
  writeln('      une augmentation d''achat de son produit associe ');
  writeln;
  writeln;
  init(tab);
  init2(res);
  remplir(tab);
  readln;

  affiche(tab);
  compter(tab,res);

```

```

clrscr;
textcolor(yellow);
writeln;
writeln('Tableau de frequences d'apparition :');
writeln;
textcolor(white);
affiche2(res);

clrscr;
textcolor(yellow);
writeln;
writeln('Elimination des frequences trop basses !!');
writeln('(inferieures a ',seuil,'%')');
writeln;
textcolor(white);
virer(res);
affiche2(res);

init2(res2);
compter2(tab,res,res2);
clrscr;
textcolor(yellow);
writeln;
writeln('Tableau de frequences d'apparition niveau 2 :');
writeln('==<  differentes combinaison possibles  >=-');
writeln;
textcolor(white);
affiche2(res2);

clrscr;
textcolor(yellow);
writeln;
writeln('Elimination des frequences trop basses niveau 2 !!');
writeln('(inferieures a ',seuil,'%')');
writeln;
textcolor(white);
virer(res2);
affiche2(res2);

writeln;
writeln;
textcolor(red+123);
writeln('fin du programme');
textcolor(white);
readkey;
end.

```

6.2 Our Application 2

6.2.1 description

The Hamming's metric.

The aim of the Hamming's metric is to find the more pertinent question to obtain the next node of the tree. In a first time it calculates the distance between every questions and the response. The smallest value is the more pertinent, so we keep it. After this the program eliminates the positive similarities (in our exemple it destroy companies). And we found the next node in the same way until the value of the Hamming's metric is nul.

6.2.2 the source code

```
program hamming;
uses newdelay, crt;
var
attribut : array[1..10,1..12] of string;
hd, pm : array[1..10] of integer;
vlpd, nbe, nbet, nbq, i, j : integer;

procedure saisie;
var
i, j : integer;
begin
clrscr;
writeln('L optimisation de la prise de rendez-vous');
writeln('Nombre de questions ?');
readln(nbq);
writeln('Nombre d entreprises ?');
readln(nbe);
nbet := nbe;
for i := 1 to nbq do
  for j := 1 to nbe do
    readln(attribut[i,j]);
end;
procedure affichage;
var
i, j : integer;
temp : string;
begin
for i := 1 to nbq do
begin
  writeln;
  str(i,temp);
  write('Q' + temp + ' ');
  for j := 1 to nbe do
    write(attribut[i,j] + ' ');
  write(hd[i] , ' ' , pm[i]);
end;
end;
procedure affhamm;
var
i, j : integer;
temp : string;
begin
for i := 1 to (nbq - 1) do
begin
```

```

    hd[i] := 0;
    pm[i] := 0;
end;
for i := 1 to (nbq - 1) do
begin
    for j := 1 to nbe do
        if attribut[i,j] <> attribut[nbq,j] then
            hd[i] := hd[i] + 1;
        if hd[i] < (nbe - hd[i]) then
            pm[i] := hd[i]
        else
            pm[i] := nbet - hd[i];
        end;
    end;
end;
procedure vlpdiscr;
var
    temp : integer;
begin
    temp := pm[1];
    vlpd := 1;
    for i := 2 to (nbe-1) do
        if temp > pm[i] then
            begin
                temp := pm[i];
                vlpd := i;
            end;
        writeln;
        write('La question la plus pertinente est Q',vlpd);
    end;
end;
procedure nettoyage;
begin
    for i := 1 to nbe do
        if (attribut[nbq,i] = 'o') and (attribut[vlpd,i] = 'o') then
            begin
                nbet := nbet - 1;
                for j := 1 to nbq do
                    attribut[j,i] := 'o';
                end;
            end;
    end;
begin
    saisie;
    clrscr;
    affichage;
    readln;
    pm[vlpd] := 1;
    while pm[vlpd] <> 0 do
        begin
            readln;
            affhamm;
            affichage;
            vlpdiscr;
            nettoyage;
            end;
        readln;
    end.
end.

```

