Mining information in order to extract hidden and strategic information

Taoufiq Dkaki[1], Bernard Dousset[1], Josiane Mothe[1,2]

[1] Institut de Recherche en Informatique de Toulouse
Université P. Sabatier, 118 Rte de Narbonne,
31062 Toulouse Cedex

tel : (33 / 0) 5 61 55 63 22 or (33 / 0) 5 55 67 81

{dkaki/dousset/mothe}@irit.fr

[2] Institut Universitaire de Formation des Maîtres
56 av. de l'URSS,
31400 Toulouse

Fax : (33 / 0) 5 61 55 58 62

http://www.irit.fr

Abstract

The amount of information available throughout the Internet or through specific collections is so huge that more and more sophisticated information handling systems are necessary to exploit it. In addition to efficient retrieval engines, the users need some tools to be able to analyse the relevant information without having to read all of it. The main objectives of a Knowledge Discovery System is to turn some selected pieces of raw information into knowledge or generalized patterns. Such a process includes a lot of problems to solve all over the three knowledge discovery phases: information harvesting and selection, information mining, displaying results. In this paper we present an interactive method of achieving knowledge discovery. It is based on information harvesting, homogenization and filtering. The discovery process itself is achieved by making several modules cooperate : different mining functions and visualization modules dynamically interact as directed by the user. The operational software we developed is also presented in this paper through some screen copies.

Key-words

Knowledge discovery, information mining, information analysis, hidden information, multidimensional analysis, multidimensional visualization

1. Introduction

The Data Base Management Systems (DBMS) and the Information Retrieval Systems (IRS) manage either factual data or information. In any case, these systems organize the information in order to retrieve some information or pieces of it. For example a query in the form of « SELECT X FROM B WHERE P » in a DBMS retrieves the part X of the tuples in B where P is true. In the same way a free text query or a boolean query in a IRS retrieves some documents according to the similarity between the query and the document indexing elements or the documents which are indexed by the term combination expressed in the query. In addition, some IRS handling long documents allow only the most supposed interesting passages to be retrieved [17]. Thus, in both DBMS and IRS, the aim of the system is to retrieve parts of relevant pieces of stored information. The processing performed when retrieving such information is make a choice according to a given query. The transformation of the retrieved information into knowledge does not fall within the system functionalities: it falls within the competence of the user. The user has to read the raw information and to appropriate it according to his objectives.

Knowledge Discovery Systems strive to achieve different goals and are based on different mechanisms. Their main objective is to turn the raw data or information into knowledge or into understandable, generalized patterns. In such a process, the data sources can still be the data bases from an information system DBMS or the information sources handled by an IRS. Nevertheless in such a process, the selection of some pieces of information is not an end in itself. It is followed by a data mining mechanism in order to extract strategic, hidden information or patterns. Knowledge Discovery processes are generally used in factual data bases [1],[8],[9],[10],[12]; we argue that this kind of process is also very useful using information collections particularly using the Internet information wide area. Because of the huge amounts of information available in specific data collections such as CACM, INSPEC, MEDLINE, ... and through Internet (Web, News, email, ...) it becomes less and less possible to search for and to read so much information. In addition, the knowledge of the whole document contents is not always the user's goal. He may want to find some strategic information that will used to make decisions, to predict some features, to verify some hypothesis or simply to have an overview of the information. Thus, it is necessary to propose some tools in order to allow the user to appropriate the information without having to read all of it. These tools have to handle and transform the information in order to extract some strategic elements from the raw information. It can then be used to make some strategic decisions. For example, from scientific bibliographic collections, one can be interested by having an overview of the main areas of one field and by their evolution in time, or one can may wish to discover the main teams that work in a given field and the collaborations they have made, or to discover the most appropriate journal for a publication on a given topic. This kind of information mining can be applied in a wide range of applications : from science monitoring to knowledge acquisition.

In addition, with regard to the Internet wide area, network has become more and more time-consuming to find relevant information sources -which correspond to any knowledge discovery or acquisition process inputs. Even though some information search engines exist, they are not satisfactory. Some tools are needed to make the Internet research engines more autonomous.

Our work has been done in that framework. The approach we propose allows knowledge discovery from specific data bases (as INSPEC, MEDLINE for example) or from the Internet (e.g. Web, email, News groups). To proceed, the harvesting of the information can be done either using a query on an IRS or through tools at the level we have developed on top of the web retrieval engines. All this information is then translated on an intelligible and synthetic view. With regard to the mining stage itself, we propose several functions based on the statistics. These functions are complementary and cooperate to achieve the knowledge discovering. We also propose an interface which is not only a way to display the results, but also interactively conduct the KD process.

To validate our research, we have developed software (called TETRALOGIE). This system is used in several French organizations for monitoring scientific progress.

In this paper, we first indicate the several stages of a knowledge discovery process (part 2) pointing out the problems to be solved in each of them. In particular, we indicate the interest in using a data/information warehouse in order to store the information and the difficulties to be overcome. Then (part 3), we develop our approach and describe more precisely the method we use to harvest and to reduce information from Internet or specific bibliographic sources. In part 4, we indicate the analysis methods we propose in order to perform the information mining. The next part (part 5) presents the interface of TETRALOGIE system, showing the interactive information discovery process. We end the paper describing the related works and we summarize the originality of our approach.

2. Knowledge discovery stages - problems to be solved

The amount of information available throughout the Internet or through specific collections is so huge that more and more sophisticated information handling systems are necessary to exploit it. In addition to efficient retrieval engine, the users need some tools to be able to analyse the relevant information without having to read all of it. Thus, the principles of Knowledge Discovery used in Data Bases will probably be more and more useful when handling information such as documents.

The main goal of a knowledge discovery process is to extract or infer global and unknown patterns from a raw data set [1][9]. This goal can be reached by discovering existing and previously unknown relationships among the data.

Generally speaking, a Knowledge Discovery (KD) process can be divided into three stages [6], see also [9].

1. Data selection: This consists in harvesting information, homogenizing, cleaning and reducing it.

2. Data analysis: Here the objective is to mine the cleaned information.

3. Results presentation: The objective is to answer to the knowledge user's needs and to allow him to take the relevant decisions.

These three steps have been defined for factual data bases but can easily be transposed to information collections.

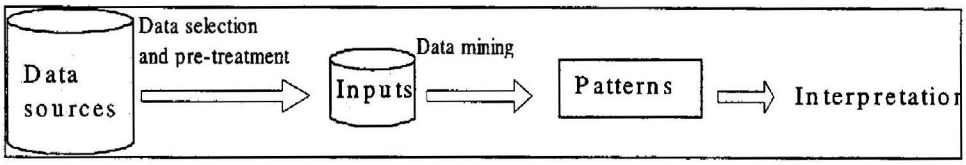


Figure 1 : Overview of a global knowledge discovery process

2.1. Input information

The data mining process inputs comprise cleaned data which has previously been harvested from **data *sources***. Generally speaking, the data sources are the databases of the information system. For our part, we are interested more specifically in information sources such as the information handle by IRS, the specific data collections as CACM, INSPEC, MEDLINE, ... and the information from the Internet (News Groups, Word Wide Web, emails).

In any case, one of the important features of the information is that it is *spread about* inside or outside the organization in *heterogeneous* forms. An other important point to notice is that the information changes with time ; that evolutionary information is not often stored. For example an information can be no longer available on the Web because the HTTP address is no longer accurate (the server location had changed, the web page name had been modified,...). The analysis of the information evolution is a very important aspect of knowledge discovery. The evolution of the information reflects the evolution of the field being considered. For example, the study of the evolution of the topics developed in scientific papers gives information about the evolution of the technologies, of the laboratories fields of interest,... Indeed it is important to keep all the information that has been valid at different times.

The information scattering in addition to the information volatility form the problem of information availability (because of the heterogeneousness and the lost of data values). That problem is much more important using Internet sources as the information contents are not well controlled. These features lead one to store the relevant information for mining purposes into a specific structure which could be called an information warehouse referred to as the data warehouse [14].

The design of such a warehouse brings up several problems (Figure 2):

- the **information sources** have to be chosen (their validity, their representativeness),
- the **source heterogeneousness** has to be solved before integrating all the information into the warehouse,

• the **amount of available information** can lead to storage and process efficiency problems,

• the **update frequency** has to be determined in order to take the source evolution into account.

2.1.1. The information source problem

The quality of the discovered information depends on the information source quality. This quality is very difficult to assess. It can be estimated using a wide range of collections, and by evaluating the induced models generalization power. That kind of problems has not yet been studied in the literature.

2.1.2. The heterogeneousness problem

The heterogeneousness problem is similar to the one encountered when querying heterogeneous databases [16] and the same kinds of solutions have to be brought with regard to semi-structured data. The information incompatibilities can be either semantic (naming conflicts, missing attributes, differences in abstraction levels) or syntactic, specially for non atomic attributes. The same kind of problems occurs when handling non structured data such as free text : synonymy, hierarchical relationships between the information elements.

2.1.3. The information quantity problem

It would be space consuming to store all the information in the warehouse. In fact it is important to choose only the really relevant information. Generally speaking the information, once it has being harvested is "cleaned". The cleaning process is generally based on a *filtering process* and on a *reducing process*. In the former process the relevant information or the important features or values to be kept are chosen. For example, using a bibliographic collection one can choose to keep the papers themes, the authors and the publication date for the papers related to computer science. In the reducing process the information is summarized to transform all the detailed information into global information. In the previous example, the only information retained could be the number of the publications written by an author on a given theme, and for a given period, instead of all the papers contents.

2.1.4. The update frequency problem

It is necessary to query the sources again from time to time because the sources are themselves updated. A recurrent querying is necessary. The querying frequency can be defined according to the source update frequency which is well known for specialized data collections but much more difficult to know for Internet sources. At the present time, that kind of problems have not been studied a lot in the literature.

In an information warehouse, some high level data is necessary. That meta data gives indications about the initial information sources, about the accurate querying frequency, about the pre-treatment to do, ...

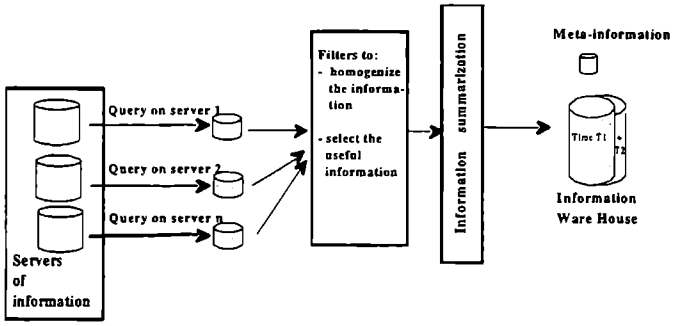


Figure 2 : Overview of a data warehouse creation and updating

2.2. Data mining or analysis

The data mining process it-self tries to find useful and unknown patterns from raw information or from an information / data warehouse.

[9] explains the main mining model functions using data as inputs. These functions can be grouped together into three groups :

* *classification*: mapping the data into predefined classes or into clusters constructed according to the data features similarities,

* *dependencies research*: (weighted) dependencies between variables, relations between fields, temporal dependencies, sequences or regression,

* *transformation:* summarization.

Each data mining model function is efficient for a specific goal nevertheless only a few systems make them coexist all together.

Notice that these functions have been defined using factual data from databases but remain the same when using information as input instead of data.

To achieve the classification, establishing dependencies goals, ... two main kinds of processes are used.

Machine learning provides some models for learning general patterns from examples, using induction mechanisms. They used either supervised learning (e.g. some neuron networks models, Quilian ID3) or unsupervised learning (e.g. auto-organizer neural network models).

Statistics proposes some models to find the links or the correlations that may exist between the data. Using statistics, it is also possible to assess the comprehensiveness of the found relationships.

Note that the quantity of available hidden information is really large ; the user has an important part to play in the mining process. His knowledge can help to overcome this combinatorial

problem and to lead the search toward the directions that fulfill his goals. Some knowledge can also be stored and automatically used during the process, making it more autonomous.

2.3. Results presentation

The last stage of a knowledge discovery process is to display the result and allows the user to interpret the data analysis or mining. Thus, the result presentation has to be intelligible. Graphical representations are among the most powerful (easily understandable, making possible to propose some interactive computing). Other intelligible representations can be obtained using rules or decision trees (see figure 3).

Finally, the figure 3 illustrates the global data mining process showing the several mining functions and the result representations possibilities:

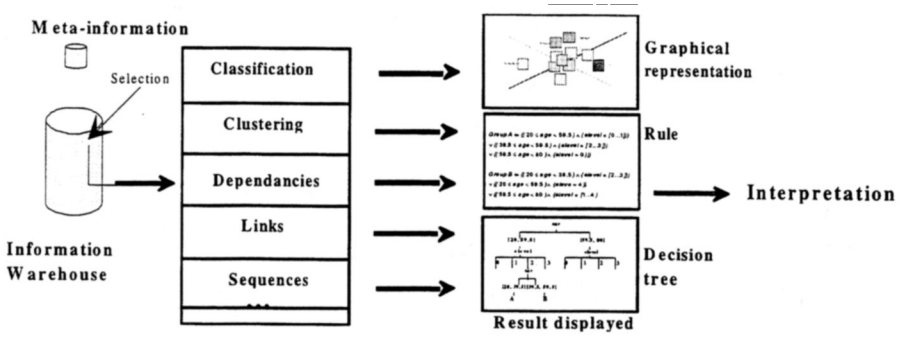


Figure 3 : Complete alternative information mining process overview

2.4. Interactive process

In fact a knowledge discovery process (selection / analysis-mining / display) is more flexible if the three stages are not sequential but are included within the framework of a cyclic process in order to refine step by step the mined and discovered knowledge.

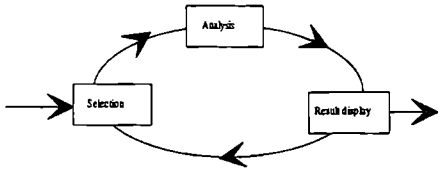


Figure 4 : Cyclic knowledge discovery process

Information mining is becoming a research field with increasing interest [7],[18]. It involves the solution of a lot of problems, and needs to take advantage of advances in other research areas:

- The *databases* field offers some solutions to the information homogenization and to the reduced information storage,

• The *statistics* and *machine learning* provides some models to implement information mining strategies,

• The *Human-Machine Interface* field gives elements which improve the result display and the interactivity between the user and the system.

Indeed, all these points are not developed in this paper, we stress the following points:

• with regard to the first stage we present our information harvesting choices and the way we represent the filtered and reduced information. Our homogenization propositions are not detailed in this paper (for more details on this step see [6]),

• with regard to the other stages, we describe the different information mining methods we use and the interactive process we proposed through the interface of the system we developed to validate our global approach.

3. Data mining inputs

In this section, we describe our approach to select the information from several sources types (specific data collections or the Internet as well). Our final goal -for the selection stage- is to obtain a comprehensible and synthetic view of the harvested information.

3.1. Information harvesting

We use two different kinds of process depending if the existing retrieval engines return raw information (such as Questel server, News or mail boxes) or addresses where the information can be picked (such as Alta Vista, Web Crawler, ...on the Web).

3.1.1. Direct harvesting raw information

With regard to the specific databases, we use the existing servers (Questel, ...). These systems retrieve the raw -supposed- relevant information according to a query. We use generic queries because we give a greater importance to the recall rate than to the precision rate. In fact, it is important to collect as much information as possible given the fact that it can be filtered at any stage. Generally speaking, increasing the retrieved documents exhaustiveness requires a librarian who is used to querying these servers.

With regard to the News group or mailbox information, we implement a harvesting tool which simply reformats the information according to their defined structure (e.g. *from, to, subject, message* attributes for an email file). This reformatting allows us to apply the filters we have defined (see § 3.2).

3.1.2. Automatic engine for Web information harvesting

Because some systems already exist, we tried to use them as much as possible. Nevertheless, systems such as Alta Vista, Yahoo, ... are not completely satisfactory, either in terms of usefulness nor in terms of efficiency. Two main drawbacks exist in the way we use the

collected information. On one hand it is necessary to sequentially select the different returned URLs, and on the other hand invalid URLs are often returned. That is why we implement a level on top of the Alta-Vista retrieval engine we called WHaT? (stand for Web Harvesting Tool). Its aim is to automatically harvest the information located at the valid URLs retrieved by Alta Vista. The filters can then be applied on the collected free text information (§3.2.2).

3.2. Information filtering

Different kinds of filters can be used depending on the needs. To summarize we define homogenization and selection filters as it is described in the following paragraphs (further details are available in [6]).

3.2.1. Homogenization filters

We define several kinds of homogenization filters, each of which has a specific function:

- *local rule extraction filters:* such a filter extracts the chosen values from a complex attribute (e.g. extracts the date of the publication from an attribute *SOURCE* composed of the elementary attribute DATE, JOURNAL, PAGES ; extracts the several authors from an attribute AUTHOR or extracts a key-word from a free text). One filter is composed of several rules : there is one such a filter per information source and one rule per atomic attribute. The rules are based on rewriting principles [13]
- *semantic function filter*: operates as a thesaurus, solving the problem of synonymy, hierarchical or inclusion relationships. It allows one to conflate several terms in a single form.

In addition, global extraction rules make the links between the local rule extraction filters.

3.2.2. Selection filters

We define two kinds of selection filters:

- *positive filters*: are used to select some information (e.g. the documents written by X, the documents where the term *(information retrieval) or (document retrieval)* occurs. These filters take advantages of IR mechanisms.
- *negative filter* operate in the reverse way (non selection of some pieces of information).

These filters are completed by external knowledge bases which can be compared to *strategic indicators* or *success factors*.

Homogenization and selection filters are complementary and are used together when filtering the information.

3.3. Information reduction

We wanted the result of the information reduction be a view of the information which correspond to:

• an homogeneous view of all the information whatever its source,

• a synthetic view of the information that keeps only the useful information in order to facilitate and speed up the analysis treatments.

• a single view over all the phases of the discovery process : in order to facilitate an interactive discovery process. This view has to correspond to each module inputs and to its outputs as well.

We choose to represent the information in the form of disjunctive tables and contingency tables. Contingency table is a powerful basic 2D knowledge representation [21]. In fact, we choose to use generalized contingency tables in order to model 3 dimensional knowledge (the third dimension is mostly related to time).

We defined a graphical language to allow the user to choose the information he wants for the three dimensions (see figure 7b).

An overview of the way we chose to achieve the selection and the pre-treatment on the information is shown on the figure 5.

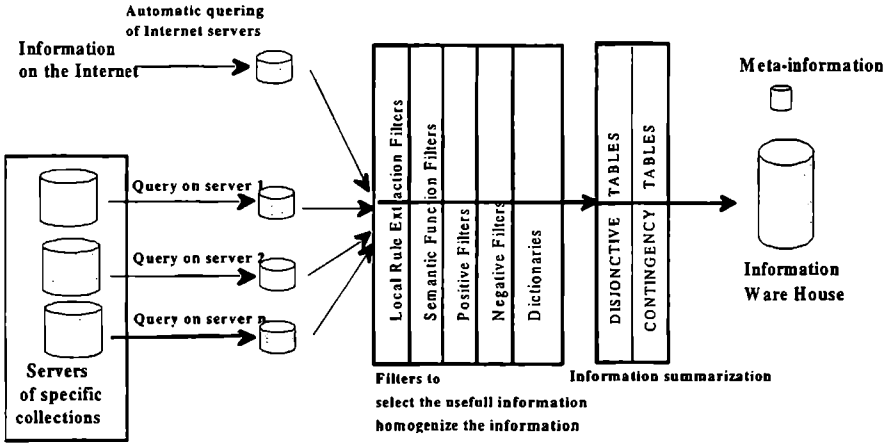


Figure 5: Overview of our approach to harvest and represent the information for KD purpose

4. Interactive process for hidden information and knowledge discovery

The precomputed and reformatted information resulting from the previous phase and stored in the warehouse is directly usable by the knowledge or hidden information discovery processes. To achieve we propose an interactive process including : mining/visualization/sub-selection functions. Indeed, the user participation to the discovery process is necessary because of the huge quantity of potential hidden information. Another point is that the user's need can change over the process becoming more and more precise. The discovered information can lead the user to define new needs.

Using our approach, the discovery process is done step by step : the user can refine his analysis, in addition at any time the user can verify the interpretations he does by returning back to the raw information. All the selection, mining or visualization functions behave as cooperative modules.

In this part of the paper, we present the different mining and visualization modules.

4.1. Information mining methods

4.1.1. Objectives

The mining methods we used answer the objectives of classification (either using predefined classes or not), search of data correlation and temporal data relationships. At present, we use methods coming from the statistics field ; all the methods used have a common point : they use our warehouse structure (contingency tables) as their inputs.

4.1.2. Statistical data analysis methods

The data analysis methods are used either to discover generic patterns or to make the data specificity appear clearly. More precisely, we used several methods that have been defined by Benzecri (see [2] for more details).

- *Principal Component Analysis (PCA)*

The aim is to obtained a representation of a set of points from a data chart, in a reduced space, that is to say in a smaller dimension that the initial one. The initial chart is a matrix n x p where n is the number of observed elements and p is the criterion number. The aim of a PCA is to compute the principal axes of the data set which maximize the inertia of the data set representation.

- *Correspondence Factorial Analysis (CFA)*

Using a CFA, the data are transformed in order to show the relative distribution of the observed elements among the criteria. It reveals the typology of the inputs appear more clearly.

- *Hierarchical Ascendant Classification*

A HAC is used to classify the elements according to their similarities. Two classes are grouped together if they are close enough to one an other. At the begining of the process, each element corresponds to a class. The process ends when the target number of classes is reached.

- *Classification by Partition (CbyP)*

A CbyP is used to classify the elements according to predefined classes.

- *Multiple Factorial Analysis (MFA)*

This is used to compare the discovered relationships which exist between the data according to an additional criterion. For example, if the additional criterion is time then, we will have to analyse several charts of data representing the same observed elements according to the same first criteria but at different times.

- *Procustean Analysis*

This is used to give prominence to the relative evolution of the inputs by erasing the average evolution.

4.2. Synthesis and display of the discovered information

A variety of different ways exist to present the mining results to the user. We choose a graphical visualization because we think that in addition to being a powerful representation, it is also a method that users are used to; it can make the processes more interactive and the interpretation more intuitive. In addition to the graphical representation, we also propose some principles to derive rules [5].

With regard to the graphical visualization, we propose two different kinds of visualization (see figure 7c) :

- a visualization in a four dimensional space in a graphic form,
- a two or three dimensional view of the warehouse (in a chart form).

These visualizations are associated with some dynamics as described in the next paragraph.

4.3. Interactive process

However powerful the mining functions are, an interactive process is still the most flexible. On one hand, the quantity of unknown information is so large that the user has to indicate to the system what he wants to discover, on the other hand, he may do not know exactly what he wants. Finally, when the system finds "some thing" the user would like to know its degree of reliability or simply to verify if it is right. Indeed, in our opinion, some interactivity is necessary to make a knowledge discovery system more flexible and such a process can not be totally automatic.

The relevance of a mining function, for example, is strongly related to the user's objectives. On the other hand, when a user observes interesting information, it is important for a user to be able to intervene in order to complete and to refine the knowledge discovery results. Finally, the user is also the best able to choose the different actions to perform on the data according to what he visualizes. So in our approach, the user can help the knowledge discovery process to be performed and he is helped in that process as well.

4.3.1. The dynamic associated to the visualizations

The dynamic we associate with the visualization have several goals. First, the global behavior of the data can be viewed: which data are close (correlated), which ones far away. Then, it is possible for the user to modify his center of interest in particular if he sees that some information is singular.

• Selection of a data subset

The user can graphically select a subset of data which wishes to analyse more deeply. In that case, that selected data will be used as an input in any mining function.

• Elimination of data

In the same way, the user can suppress some items of information by graphically selecting them. The remaining data will then be used as an input of a mining function.

These operations can be done either on the inputs themselves (using some functions associated to the charts) or on their graphical representation.

In addition, the following mechanisms can be applied to a graphic representation :

• Scanning of the data space

The scanning allows the user to change the point of view he has on the data. Indeed, two points can seem to be close each other according to one point of view whereas in fact they are far away. A scanning by rotation can make the correlations between the data more visual. The rotation in the space we propose is continuous and animated; the user can stop it when an interesting view is obtained. It is possible to find the optimum angle for the view, that is to say the plan which shows the associations, the similarities or the grouping of the data which are relevant for the user. All these transformations are an integral part of the knowledge discovery process.

• Modification of the representation space

To increase the amount of the visualized information, it can be relevant to use other axes in the space. We propose an action that produces "sliding" axes. This axis sliding can be repeated until the whole information is visualized. In fact, any combination of axes can be shown to the user.

4.3.2. The module combinations and communications

A simple module combination is the graphical visualization of some analysis function results. In fact, more complex combinations can be used for a step by step discovery process.

• Importation and exportation of rotation

When a user obtains a relevant point of view on the data, the angle made by the axes of that representation can be exported to another analysis module visualization. Notice that this exportation is only authorized when it has a meaning (CFA-PCA, CbyP-CFA).

• Interaction from a distant site

The same kind of importation can be done between two distant users. This communication is useful when different kinds of expert want to cooperate: for example a domain expert and a statistician expert. In that case, one can chose the different processes to do and the other visualizes dynamically the obtained results and direct the study. This kind of interactions can also be used for collective expertise.

• Linking of analysis modules (macro)

Several linking of analysis modules are relevant to proceed a knowledge discovery process. With regard to the study of the temporal data evolution for example, a scenario as Principal component analysis - Procustean analysis - Visualization can be used.

Finally, even if the knowledge discovery process is automated, the user remains the master of the choices made in the process (sub-set of data, mining function, function combinations, ...).

Figure 6 shows the interactive knowledge discovery process we propose.

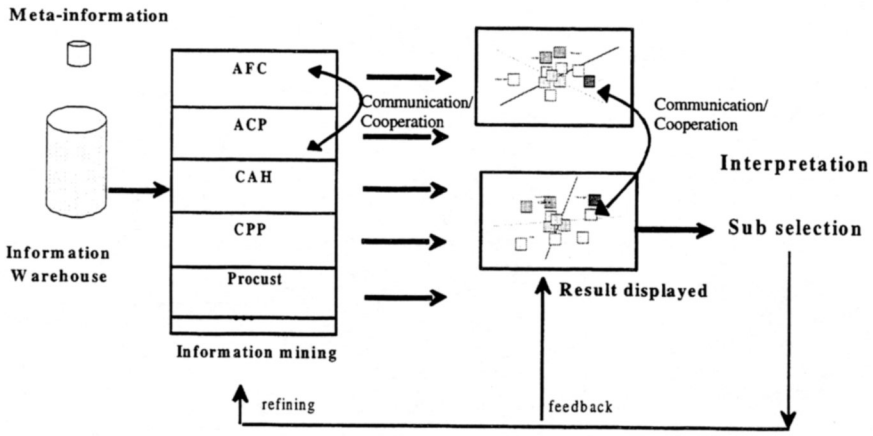


Figure 6: The interactive knowledge discovery process

5. Example of a KD result using Tétralogie system

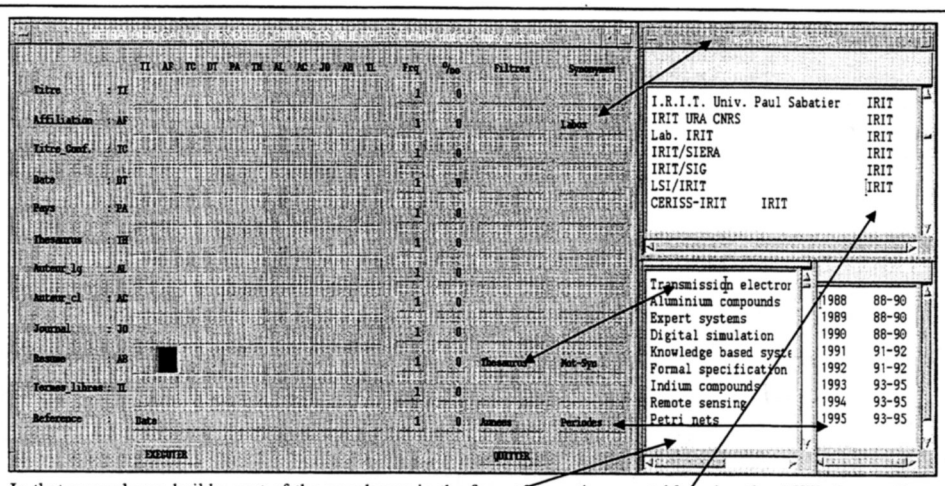
TETRALOGIE is the name of some software we have developed to validate our approach. In this section, we show the results obtained at some steps of a knowledge discovery process. The example we choose is an extract of a experiment we did on the "computer science research in Toulouse". The initial source we use is the INSPEC bibliographic database.

In that example we show how data specificity and strategic information can be discovered in an interactive way.

The INSPEC data collection has been queried through a dedicated server using a classification based query. We were interested by the publications related to "computer science in Toulouse". More than 2500 notices were retrieved. The INPEC data collection schema is used in order to identify some elementary components, using extraction functions (see § 3.2.1). Then some relevant pieces of information are kept using the other kinds of filters.

INSPEC 4241022 C9211-1160-002
Doc Type: Journal Paper
Title: When upper probabilities are possibility measures
Authors: Dubois, D.; Prade, H.
Affiliation: Inst. de Rech. en Inf. de Toulouse, Univ. Paul Sabatier, France
Journal: Fuzzy Sets and Systems Vol: 49 Iss: 1 p. 65-74
Date: 10 July 1992
Country of Publication: Netherlands
ISSN: 0165-0114 CODEN: FSSYD8 CCC: 0165-0114/92/$05.00
Language: English
Treatment: Theoretical/Mathematical
Resume: A characteristic property is given for a pair of upper and lower probabilities (induced by lower probability bounds on a finite set of events) to coincide with possibility and necessity measures. Approximations of upper probabilities by possibility measures are discussed. The problem of combining possibility distributions viewed as upper probabilities is investigated, and the basic fuzzy set intersections are justified in this framework. (23 Refs.)
Classification: C1160 (Combinatorial mathematics); C1140Z (Other and miscellaneous)
Thesaurus: Fuzzy set theory; Probability
Free Terms: Upper probabilities; Possibility measures; Characteristic property; Lower probabilities; Necessity measures; Possibility distributions; Fuzzy set intersections

Figure 7a: Example of INSPEC notice



In that example we build a part of the warehouse in the form of a contingency table using the **Affiliation** attribute (giving the main affiliation of the authors), the **Resume** attribute (giving a free text abstract of the paper) and the year of publication through the attribute **Date**.

A positive filter called THESAURUS is used to indicate the relevant terms we want to keep. A part of it is displayed in the right part. An other filter called MOT-SYN is used to homogenize the synonym values and transform the different synonyms into a single representative item. In the same way LABOS is a homogenization filter which lists the laboratories synonym names. The reference attribute is used to be able to analyse the information temporal evolution. Three periods have been defined (see the PERIOD filter): 1988-90, 1991-1992, 1993-1995.

Then, several parts like the one constructed here compose the warehouse. In the following , two examples of knowledge discovery process are presented.

Figure 7b: Example of a graphic query used to build the warehouse

We show an example of the module cooperation between the chart representing a part of the warehouse and a graphical visualization of discovered knowledge.

In that example, we study the correlations between the themes and the authors (one color is used for the themes, another for the authors). Notice that the fields of interest of an author are graphically near from his name. They can be clearly listed selecting them graphically. The author "LAUMOND" has an interesting behavior as he is between two closely but distinct groups. He may do the interface between two fields or had changed his fields of interest. These hypothesis can be validated by returning back to the warehouse information. Here for example, the chosen author has been graphically selected (in the frame on the right) and automatically retrieved in the frame on the left. The user can then work on the chart (ordering, selecting, ...some information). In the example, a column ordering has been done: the author's interests are on the top of the list. The full corresponding terms can also be displayed.

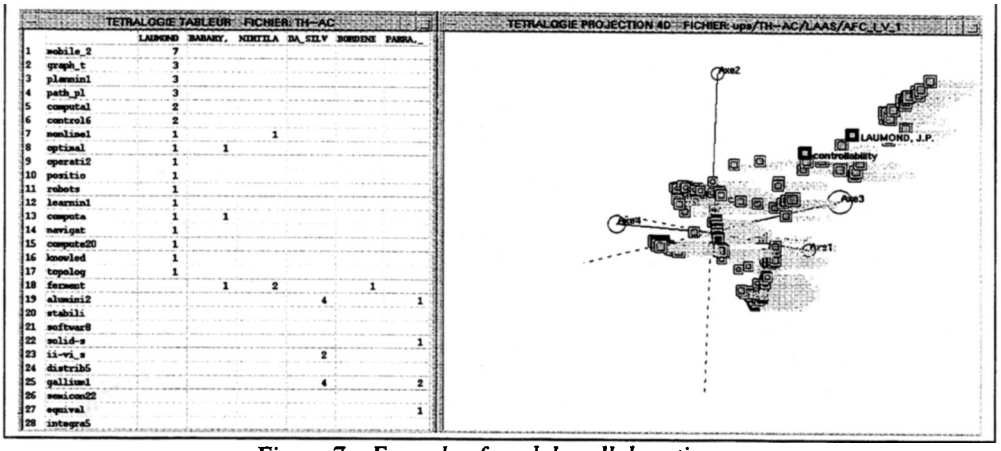


Figure 7c: Example of module collaboration

In this other example, we show some results obtained using a (Authors / Authors / Date) table.

After a scanning by rotation and an axis sliding in the representation space, one can see some emergent information as for example the "Segur" evolution. We can induce that in the first period he wrote with the team represented on the top-left part of the screen, then works in collaboration with the top-right team and finally comes to the center (probably no more publication) according to the selection done. This kind of interpretation can be verified coming back to the raw information if it is available or to the warehouse information.

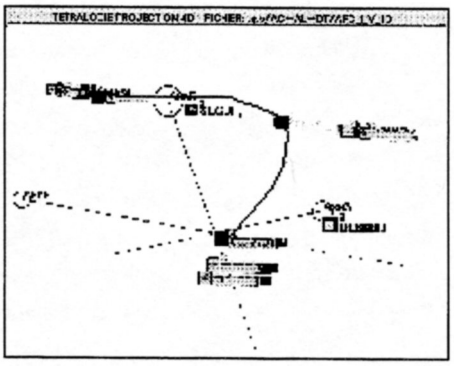


Figure 7d: example of temporal information analysis

/e show some examples of the functionalities of this system. Another example of its inctionalities is classification (either with predefined classes or not). In either case, the lassification results can be imported in a 4D view (one color per class). This is an other xample of how the mining functions can collaborate in order to allow the user to deeper nalyse the information.

6. Related work

Most of the work in knowledge discovery has been carried out on factual data and more precisely on relational databases. In that case the most suitable information sources are the ones given by the internal organization DBMS. The importance of the studies done on factual data can be explained by the hope that that kind of systems could help decision makers and analysts to become more productive. Knowledge discovery applications have been developed for a wide range of domains such as finance, marketing, production, ...

With regard to relational databases, [10] gives an interesting mechanism to induce generic rules from the database tuples. The rules are obtained by filtering the interesting attributes and by performing a generalization of the attribute values so that two different tuples can be represented by a single summarized tuple. A measure of the rule strength has been given. This strength is representative of the rule generalization power. [4] adds the possibility to aggregate two attributes in order to obtain rules which more closely fit the user's needs.

Some operational OLAP (On Line Analytical Processing), [12] for example, have been developed as an upper layer for relational databases. Their aim is to analyse multivariable data; and they allow one to rapidly exploit two or three dimensions of a given dataset. Generally speaking, these systems hold up data warehouse functionalities, multidimensional visualizations and a drill-down discovery process. Even with a systems emphasis on the interactive visualization of summarized data, these kinds of systems do not allow advanced mining treatments. Simoudis & al. (*Integrating Inductive and Deductive Reasoning for Data Mining* in [8] pp 353-373) insist on the fact that a cooperative use of several mining techniques is the most efficient. The authors propose the cooperation of rule induction, deductive databases and data visualization in order to create rule-based classification models. Histograms and bar charts are used to view the distribution of values. The discovered relationships can then be encoded as new concepts using the deductive database. Finally the generated data can be sent to the visualization module.

Our approach is more complete in the sense that in addition to an interactive discovery process and an interactive result visualization, a wide set of mining functions are proposed which can cooperate with each other. The visualization module includes dynamic functionalities, helping the user to deeper analyse the automatically generated knowledge.

The development of international networks such as the Internet increases the availability of a wide range of information types. Using documents as knowledge discovery sources is becoming more and more attractive. In [11], a multi-layered database is proposed in order to handle both information sources and to discovered knowledge. Layer-0 corresponds to the raw information; and the further you go up the layers, the more the knowledge information is generalized. As the upper layers can be stored into a classical DBMS, this model takes advantages of the SQL language. In addition, some other operators are defined in order to make the language -associated to the multiple layered database- fit the user's needs.

One of the advantages of our approach is that the data sources can be either factual data or textual data. When handling free text information, our filtering and reformatting methods (including the solving of structure and value heterogeneousness) transform it in order to handle it by our mining and visualization modules. Temporal data can also be mined to discover the evolution in term of vocabulary, of fields of interest, etc., ...

With regard to free text information, some of the issues proposed seems not to be new such as document classification [19] ; nevertheless the classification we propose is "filter" directed (the selection filters indicate the relevant items for the study), that means that it can also be user directed.

7. Summary

In this paper, we explain our approach to achieve a knowledge discovery process from heterogeneous information. We decompose such a process into three phases: each of the phases have been described first in a general point of view -pointing out the problems to be solved, then according to our approach.

With regard to the first phase (information selection and pre-treatment) we propose a filter oriented approach. The selection from specialized collections uses a query sent to existing servers, whereas the selection from the Web uses our own upper level automatic information harvesting based on Web retrieval engines. Some filters are used to homogenize the information (both in term of schema and in term of value), other filters are used to extract the information to be kept (both in term of information kind and in term of value). The selection filters are completed by external knowledge (as synonymy, hierarchical dictionaries). Notice that some selection filters can be view themselves as external knowledge. The selected information is then reduced in the form of 3D contingency tables which compose our warehouse.

The knowledge discovery process itself is done by making several modules cooperate. We define mining function modules -based on statistics functions- and visualization modules -a graphical visualization to represent the discovered knowledge and a visualization in a chart form for the warehouse data. These different modules can interact with each other and can be included into a scenario (succession of mining and visualization modules). An interactive discovery process is then possible because of the dynamics associated to the visualization modules.

We also present a part of the interface of the TETRALOGIE system that was developed in order to validate our approach. This system is used by several french organizations for science progress monitoring purposes.

Our future works consist in automatically finding a user's profile or in automatically formulating queries according to some relevant information analysis results. Those profiles will be used as the input to the automatic harvesting engine on Web information. Other complementary work consists in automating the creations of some filters. Some encouraging

results have already been obtained. An other perspective concerns the possibility of representing some knowledge in a neural network form, especially the selection and the user's profile filters. Some interesting results have been obtained using that kind of approach in order to reformulate a query (which is a notion very close to user's profile) [15][3]. We would like to make advantages of the neural network learning capabilities to adapt the information harvesting to the users' profiles. It is particularly interesting for uncontrolled information such as the information on the Web.

Even if the knowledge discovery area takes advantages of other research fields, it is in it own right still truly considered as a full research area. Some advances have been made in the last five years but there are wide evolution perspectives. Because of international networks development and because of the increasing amount of available information, it is more and more important to be able to retrieve the relevant information according to one's needs and to be able to automatically analyse it. In that sense, Information Retrieval Systems and Knowledge Discovery Systems are complementary. In our opinion, IRS have to be used to select the information that can be transformed into knowledge. A part of the discovered knowledge can in turn be used to improve IR performance through query reformulation.

8. References

[1] R. Agrawal, T. Imielinski, A. Swami, *Database Mining: A Performance Perspective*, IEEE transactions on knowledge and data engineering, pp 914-925, Vol.5, N.6, 1993.

[2] J.P. Benzecri, L'analyse de données, Tome 1 et 2, Dunod Edition, 1973.

[3] M. Boughanem, C. Soulé-Dupuy, *MercureO2: Adhoc and Routing Tasks*, in the 5th Text Retrieval Conference TREC-5, NIST SP 500, 1996.

[4] V. Dhar, A. Tuzhilin, *Abstract-Driven Pattern Discovery in Databases*, IEEE transactions on knowledge and data engineering, pp 926-938, Vol.5, N.6, 1993.

[5] T. Dkaki, B. Dousset, S. Koussoube, *Génération de règles pour l'explication des résultats des classifications,* Actes des Premières Journées de Mathématiques Appliquées, pp 1-8, Rabat, 1992.

[6] T. Dkaki, J. Mothe, *Extraction et synthèse de connaissances à partir de bases de données hétérogènes,* XIVème congrès INFormatique des Organisations et Systèmes d'Information et de Décision, pp 287-308, June 1996.

[7] O. Etzioni, *The Word-Wide Web:Quagmire or Gold Mine?*, in Communications of the ACM, Vol.39, N°11, pp 65-68, Nov.1996.

[8] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, AAAI Press, IBSN 0-262-56097-6, 1996.

[9] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *The KDD process for extracting useful knowledge from volumes of data,* in Communications of the ACM, Vol.39, N°11, pp 27-34, Nov.1996.

[10] J. Han, Y. Cai, N. Cercone, *Knowledge Discovery in Databases: An Attribute-Oriented Approach*, VLDB-92, pp 547-559, Vancouver, Canada, 1992.

[11] J. Han, O.R. Zaïane, Y. Fu, *Ressource and Knowledge Discovery in Global Information Systems: A multiple Layred Database Approach*, Proc. Of a Forum on Research and Technology Advances in Digital Libraries (ADL'95), Mc Lean, Virgina, May 1995 (*ftp://ftp.fas.sfu.ca/pub/cs/han/kdd/Ignis94.ps*)

[12] IDIS, Information Discovery Inc. http://datamine.inter.net

[13] P.Y. Lambolez, J.P. Queille, F.F. Voidrot, C. Chrisment, *EXREP : a generic rewriting tool for textual information extraction*, Engineering of Information Systems, Vol. 3, N°4, pp 471-488, 1995.

[14] M.E. Meredith, A. Khader, *Designing Large Warehouse*, 25-30, Database Programming & Design, Vol. 9, N°6, June 1996.

[15] J. Mothe, *Search mechanisms in a neural network model: comparison with the vector space model*, Intelligent Multimedia Information System and Management, RIAO'94, New-York, 1994.

[16] M.P. Reddy, B.E. Prasad, P.G. Reddy, A. Gupta, *A Methodology for Integration of Heterogeneous Databases*, IEEE transactions on knowledge and data engineering, pp 920-933, Vol.6, N.6, 1994.

[17] G. Salton, J. Allan, C. Buckley, *Approaches to Passage Retrieval in Full Text Information Systems*, Proc. of the 16th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 49-58, 1993.

[18] A. Silberschatz, M. Stonebraker, J. Ullman, *Database Research: Achievements and Opportunities Into the 21st Century*, Report of an NSF Workshop on the Future of Database Systems Research, May 1995.

[19] H. Schutze, J. Pederson, D.A. Hull, *A Comparison of classifiers and Document Representations for the Routing Problem*, Proc. of the 18th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 229-237, 1995.

[20] J.A. Yochum, *Research in Automatic Profile Generation and Passage-Level Routing with LMDS*, in The Fourth Text Retrieval Conference TREC-4, pp 153-191, NIST SP 500-236, 1996.

[21] R. Zembowicz, J. M. Zytkow, *Form Contingency Tables to Various Forms of Knowledge in Databases*, chapter 13 pp 329-349 in [8].