Using rules to enhance evolution of knowledge mapping: Application on Healthcare

Menaouer Brahami[1], Baghdad Atmani[2] and Nada Matta[3]

[1] Department of Computer Science, University of Oran, Laboratory LIO
Oran, BP:1524 El M'naouer 31000, Algeria

[2] Department of Computer Science, University of Oran, Laboratory LIO
Oran, BP:1524 El M'naouer 31000, Algeria

[3] Laboratory ICD/TechCICO, University of Technology of Troyes (UTT)
12 Rue Marie Curie CS 42060/10004, Troyes, France

Abstract

Preserve knowledge, retain knowledge, these are the objectives of a scalable enterprise. Knowledge mapping is graphical techniques which allows of preserving and visualizing the patrimony strategic and trades of the domains of knowledge acquired over the years. The approach presented in this paper draws, firstly, on the exploitation of different data sources for improving the process of acquisition of explicit knowledge on an organization. In this sense, our contribution is to produce, from one hand inductive Boolean rules that will feed knowledge base CASI, and from other hand, refinement of Boolean model of the knowledge mapping already achieved by the MASKII method (Critical Knowledge Mapping).

Keywords: *Knowledge management, knowledge mapping, Boolean modeling, Cellular machine, Data mining, Machine learning, Decision Tree, Decision Support System.*

1. Introduction

Currently most of companies are aware of the need to manage their wealth of knowledge [6]. Sharing and knowledge transfer between generations is a topical issue related to the expected retirement for years to come. The conclusion is that there is a risk of loss of knowledge that will keep [9], [12]. Among the method lies available to formalize this strategic heritage of knowledge, we are interested in our study, mapping of critical knowledge domains (MASK II). The method of knowledge mapping is used to represent and analyze the knowledge of a company grouped by domain and by viewing them as a map [8], [12], [21], [24].
Moreover, companies are faced with new ones problems due to the growth increase of the size of their data [13]. This flow, continuous and increasing information, can now be stored and prepared for study using new techniques of Data Warehouse. Among obstacles to successfully extract knowledge from data mining, we quote: the increasing amount of information generated and made available to the departments concerned the right information at right time. As a result of the arrival of these two fields of application (*Data Mining and Data Warehouse*), a new idea is obvious: « Why not combine all these techniques to create powerful methods for automated knowledge extraction to improve the mapping process, including all stages from data collection to evaluation of knowledge gained». Thus was born the idea of *mapping critical knowledge (MASK II)* through a process of *knowledge discovery in data (KDD).* Our contribution in this area consists of designing and experimenting on a new approach of critical knowledge mapping by automatic learning. We did so by exploiting the proven performances with the techniques of the mathematical formal of a cell machine CASI (Cellular Automata for Symbolic Induction) [2]. The result of the critical knowledge mapping, made by MASK II [11], [8], is refined by a symbolic automatic learning process graph-based induction. This refinement is done by the Cellular Inference Engine (CIE) who is attend of the symbolic induction to optimization of the induction graph and who going ensure, thereafter, the internal representation of the new map of critical knowledge domains.

2. Related work

In recent years, awareness about the strategic importance of knowledge of an organization that has its value is linked to its knowledge and its use. The potential damage caused by the loss of a key competence and volumes of departures, scheduled or not, most experienced staff alert, in a manner becoming stronger, the need to adopt

management strategy knowledge. Indeed, tacit/explicit knowledge management is extremely rich and dynamic: It has become necessary to model them. This modeling is to transform large amounts of data, from interviews with experts and searching documents in multiple repositories related to business activities [25], [23]. To this end, a multitude of tools and methods exist for knowledge discovery in data, expert interviews and/or reference materials. These methods are classified into two categories: explicit methods (*capitalization*) and methods for automatically extracting knowledge [23] (Fig 1).

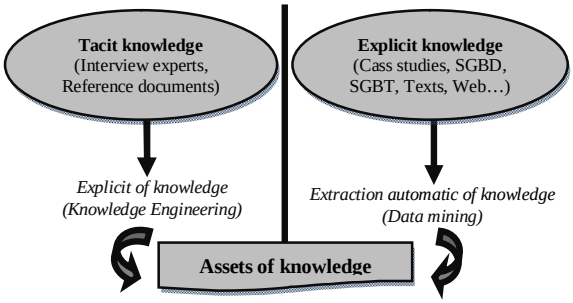

Fig. 1 Mining and explicit knowledge [23]

Knowledge mapping, which is considered as a method of knowledge explicitation, aims to showcase of the trade's critical knowledge of the company [9]. Similarly, it allows for the indication of the importance of knowledge that is at risk of being lost and that must be preserved [12].
Among the problems open from the knowledge mapping, include notably research dynamic of knowledge domains decisional in a knowledge map that is becoming increasingly complex, due to several parameters (*number of domains, criteria and degrees of criticality*) thereby the evolution of knowledge maps exploiting other sources of enterprise data. Then, it does not allow formalize the data and information in real time. For this, the generalization of the extraction process of knowledge on all available knowledge in the organization (Database, reference documents, enterprise portal) is desirable and possible by the use the data mining techniques.

In this field of knowledge extraction from data, several techniques of artificial intelligence are the basis of numerous studies. These studies include the work of Hai Wang et al. [30] on economic intelligence. Hai Wang et al. proposed a model of knowledge sharing based on various « *blogs* » to support the process of knowledge management. In biology, Fabien Jalabert et al. [5] recommended an integrated environment I2DEE (*Integrated and Interactive of Data Exploration Environment*) that was applied to two distinct application areas of engineering terminology and ontology, and to data analysis gene expression from ADN chips. The main objective was the integration and visualization of heterogeneous data in the design process of an RTO (*Terminological Resource or Ontological*) that is specific to a given application. This environment allows, through a chain of analysis and data processing, filter concepts and relations that are pertinent to this application and present it to the user through a knowledge map with which it can interact via a communication interface. Since the information system of each company contains a huge amount of data Jelena Mamcenko et al. [23] proposed a methodology to extract pertinent knowledge that has been previously hidden through correlations between data. Nhien Year Khac et al. [15] have created «*a distributed knowledge map*» that easily and effectively shows the new knowledge discovery in data sets stored in the distributed platforms (*data grids*). Emmanuel Blanchard et al. [4] proposed a method that is based on reasoning mechanisms to enhance the system of knowledge management and the reliability of the identification skills of an individual. The main objective is to propose a knowledge mining process that is defined by an analogy with the extraction of association rules, in order to induce a rule base from a knowledge base. Vladimir Kvassov et al. [29] used the extraction of knowledge from data to improve decision making in the knowledge management of two technology companies and telecommunication. Finally, Tipawan Silwattananusarn et al. [28] have explored the applications of data mining techniques which have been developed to support knowledge management process. The journal articles indexed in ScienceDirect Database from 2007 to 2012 are analyzed and classified.

In this context, and as we have already pointed out, we are interested to raise in how the knowledge discovery from a corpus of data centralized or distributed can improve critical knowledge mapping of the health SEMEP service «*Services of Epidemiology and Preventive Medicine*» of the city of Mostaganem in Algeria.

3. Proposed approach "BMKMDM"

The objective of our approach BMKMDM (**B**oolean **M**ethod of **K**nowledge **M**apping guided by **D**ata **M**ining), as illustrated in Fig. 2, is double: first, the data mining is made in a first step by using the algorithm ID3 (*Inductive Decision Decision Tree*) that generates the rules of induction from the practical cases, subsequently in the second stage produce Boolean mapping rules which will feed the knowledge base the machine cellular CASI (*Cellular Automata for Symbolic Induction*) [2]. On the other hand, refine the Boolean model of the critical knowledge mapping of already achieved by the method MASK II (*Method for Analyzing and Structuring Knowledge*) [11], [8] by the mapping rules obtained by this

process of symbolic automatic learning to basic the induction graph.

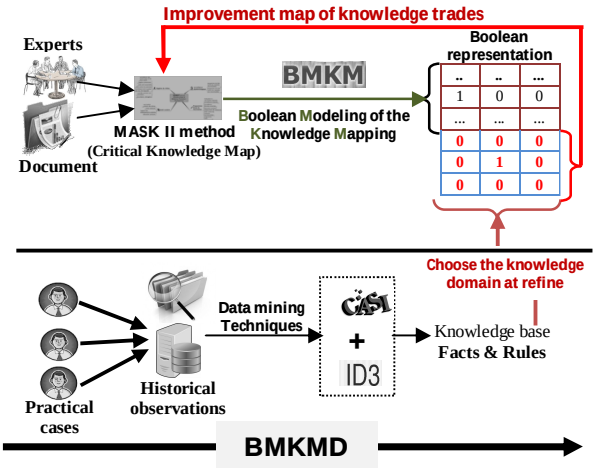

Fig. 2 General architecture of the proposed approach

3.1 Cellular Automaton "CASI"

CASI (*Cellular Automata for Symbolic Induction*) is a cellular method of generation representation and a means to optimize induction graphs generated from a set of learning examples [2]. This Cellular system is organized into cells where each cell is connected only with its neighbors (*subset of cells*). All cells obey in parallel to the same rule, which is called the "local transition function" This results in an overall transformation of the system. CASI is composed of three modules: COG (*Cellular Optimization and Generation*), CIE (*Cellular Inference Engine*), and CV (*Cellular validation*) (shown in Fig. 3).

- **COG (Cellular Optimization and Generation) module:** Using a cellular automaton and cooperating with an induction graph (*SIPINA* method), COG module will extract new knowledge from training data. Two finite layers of finite automata represent the knowledge that is generated.
- **CIE (Cellular Inference Engine) module**: The CIE module, heart of the cellular machine (*CASI*), simulates the functioning of basic cycle of an inference engine using two layers of finite automatons finite. The first layer, *CELFACT*, for the basis of facts and, the second layer, *CELRULE*, for the basis of rules. The states of the cells consist of three parts: EF, IF and SF, respectively ER, IR and SR are the input, internal state and output of a cell of CELFACT, respectively a cell of CELRULE. The internal state, IF of a cell of CELFACT indicates the role of fact: in our graph IF=0 corresponds to a fact type top (s_i), IF=1 corresponds to a fact type attribute= value (X_i = *value*) [2]. For defining the vicinity of cells, we are using the two incidence matrices of the (R_E) input and the (R_S) output of the automaton. R_E and R_S represents the input / output relationship of facts and are used in forward chaining: of the root to the leaves. We can also use (R_S) as input relation and (R_E) as the output relationship to launch an inference in backward chaining: the leaves to the roots. Finally, the dynamics of CASI for simulate the operation of CIE inference engine uses two transition functions δ_{fact} and δ_{rule}, where δ_{fact} is the phase of evaluation, selection and filtering, and δ_{rule} corresponds to the execution phase and that all the cells in parallel to obedient the same branch called local transition function, which a as consequence in a global transformation synchronous of the system [2].
- **CV (Cellular Validation) module**: After the rules have been generated by the *SIPINA* method, which has been coupled along with the CASI machine, validation of this knowledge could be done using the CV module.

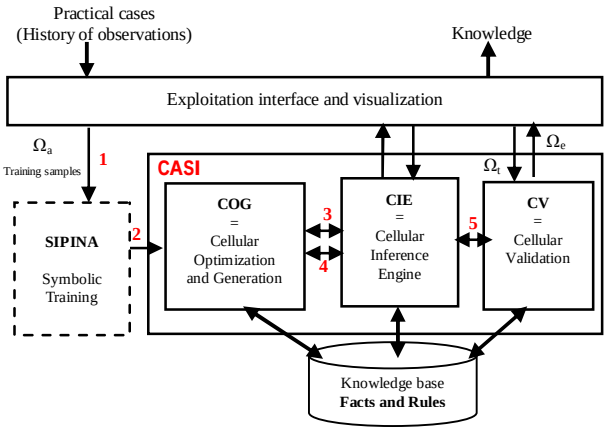

Fig. 3 General diagram of the cellular machine (CASI) [2]

3.2 Towards a mapping guided by data mining

The mapping process is guided by machine artificial learning, (shown in Fig. 2) of our project. It proceeds in the following two steps: data mining and automatic mapping. Data mining consists of initially to launching the symbolic induction from case studies using CASI [2]. The mapping rules obtained are used to automatically improve the Boolean model of critical knowledge mapping already carried by the MASKII Method [11], [8].

We recall, in this step, that the realization of the map was based on an analysis of references documents

(organizational chart, description of the distributions of activity services, directory of staff activities, plan medium term of the vaccination, studies, balance sheet of vaccination, etc.) and interviews with business experts (doctors, health technicians, psychologists, midwives, etc.) and others who are responsible in the health sector.
The adopted principle of knowledge mapping [16] is to group the different activities' knowledge domains, of them get in shape format via a representation vulnerable then the complete and validate the mapping produced with experts, in an iterative manner. These iterative validations allow having co-construction work. They also guarantee the maximum involvement and appropriation of the interviewees. The result is a map of knowledge domains or know-how map trades. This mapping, illustrated by our CARTOCEL system [19] (shown in Fig 4), is a description of a level of the meta-knowledge [30] of SEMEP know-how. It provides a system for being able to address know-how in order to facilitate access to knowledge domains.

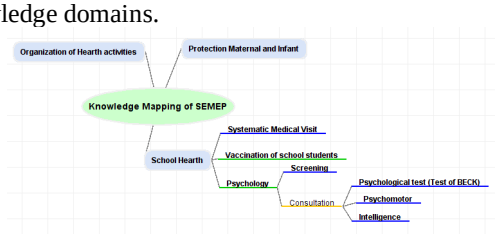

Fig. 4 Example of Knowledge mapping by our CARTOCEL [19]

The problem of Knowledge Discovery in Databases (KDD) uses the principles of machine learning and the methods of inductive or deductive supervised learning. Among the inductive methods, we were interested in decision trees, especially those that are based on induction graphs, because the classification function is expressed by a graph which can be transformed in the form of production rules.
Automatic learning is certainly, in terms of artificial intelligence, the scope of application the most fertile of recent years. It is generally known that the prerogative of artificial intelligence is to learn from past experience so that its behavior becomes adaptable. The machine learning is thus the field of study where one tries to reproduce human capacity to be learned. The pioneers of machine learning should be viewed as a set of changes in a system that allows it to accomplish the same task the best, or a similar task in the same population in the future.
Rakotomalala [26] confirms that Dietterich proposed a more functional approach to machine learning that can evaluate, by linking it to the notion of knowledge. Dietterich distinguishes three levels of description for a learning system, as listed below:

- A system that doesn't receive input and that performs the task best;
- A system that receives an input of knowledge, but doesn't perform induction;
- And finally, a system that receives inputs and extracts knowledge that is either known implicitly or explicitly is inductive learning.

It is the latter that interests us in our study, specifically, we are interested in empirical learning, which is aimed at generating new knowledge from case studies: examples, observations, etc… that have been baptized « symbolic induction by machine learning ».
For [27], inductive learning extracts a model from a set of examples of cases that have been resolved by domain experts. From this set, which is called the learning set, we generate a model that is used to study new examples in the same domain. Each example (object) of this set is represented by a vector of attributes and each attribute is a set of values.

Application of KDD process on psychological sheets: In [17], the process of Knowledge Discovery in Databases comprises 5 stages: selection, pre-processing, transformation, data mining and interpretation. In this framework, the data we used are *the sheets of the psychologist's* SEMEP. The service SEMEP has task the collection, analysis and interpretation regular and systematic the sanitary data for the description and observation contained of health the students schools a view to facilitating the planning, implementation of evaluation of interventions and school health programs. For example, psychologists of the SEMEP have making systematic visits in collaboration with their counterparts in secondary schools to study the psychological state of students. In addition, they are very diverse and are not necessarily all usable by the data mining techniques. Most techniques that are used deal only with the tables of data in traditional rows /columns. The objective is to prepare the tables rows/columns. In other words, to prepare the individual tables/variables those are obtained by the steps (shown in Fig. 5):

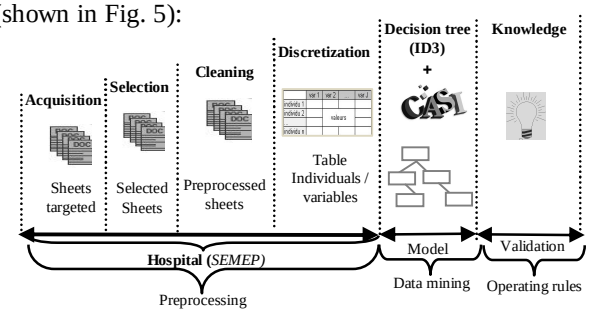

Fig. 5 The process of extracting knowledge from the psychological cards (Personality Test).

1. **Pre-processing:** Preprocessing is a key preliminary step in the process of knowledge discovery in databases (KDD). The results that are generated at the end of this phase depend in large part on the quality of data that is used.
The preprocessing steps include access to data in order to construct two-dimensional tables, which are called table individual variables. These include observations (*explicit data*). Based on the type of data (e.g. numeric, symbolic), methods of preprocessing format the data, *clean, handle missing data* and *select* the attributes or observations (individuals) when there are too many of the following choice of attributes that are most informative in the first case and samplings in the second case. This phase is important because it is the one that will determine the quality of the models developed in data mining. Indeed, these choices are intended to bring out the information contained within this mass of data [3].
   - **Selection of data**: in our case and in order to supplement, refine and accelerate the mapping process we searched the records of psychologists services (*School Health*) to collect case studies of students in all levels (*middle, secondary*). After selecting the plugs and cleaning considered relevant we looked at records of *psychological tests* that students have met, especially the *personality test (the Test of Beck or the Beck Depression Inventory «BDI-II»).*
   - **The discretization**: this is compromise, and from among the data reduction strategies we use the discretization that transforms continuous attributes by cutting the field of values of these attributes into intervals in order to obtain qualitative attributes. Indeed, there are many methods for the purpose of discretization that we can mention. These include, and quintiles discretization, the discretization as nested averages or that has been standardized according to the discretization etc.

   In our case and to complete this step we had to understand the principle of the tests (*the psychological tests «Personality Test- Beck-II»*) [1], and extract the variables (attributes) in order to launch the qualitative phase data mining and generation of knowledge.

2. **Data mining by decision tree:** Data mining, which is the heart of the process of KDD, is the analysis of observations of a data set in order to not identify the suspected relations and to summarize the knowledge included in this data in new forms that is both comprehensible and useful to experts [17], [14], [11], [23]. KDD, by means of data mining is then seen as engineering knowledge discovery in data. Data mining principally uses the disciplines of artificial intelligence, statistics, and data analysis [10], [20]. It is usually done on two-dimensional tables and essentially decomposes into three major families of methods: descriptive, explanatory and structuring.
The objective of the implementation of data mining techniques is to obtain operational knowledge. This knowledge is expressed in terms of models of varying complexity, which are a series of coefficients for a forecast model, and the type of logical rules for the *If Condition then Conclusion* or graphs. For these models to acquire a status of knowledge, they must be validated. Then a series of operations post-processing steps, which are used to validate the models to make them intelligible if they are to be used by humans or if they are to be formalized by being automatically processed by a machine, need to be implemented. Beyond statistical validation, the intelligibility of the models is often a criterion for their survival. Indeed, including a model will be used by the user is consequently critical and needs to be perfected [17], [14].

Construction of a decision tree and the generation of rules: In the literature, decision trees are among the classification techniques of data mining that are the most popular and that are the fastest and easiest to use. According to [3], decision trees are learning tools that produce rules such as (If condition then conclusion) in which 'condition' means a disjunction of conjunctions of logical propositions of a type «of attribute, and value» [10]. The set of rules is thus the prediction model. They use a set of individuals (n-uplets) as input that is described by variables (attributes). Each individual belongs to a class, with the classes being mutually exclusive. To construct a decision tree, it is necessary to have a learning population (table or view) that consists of individuals whose class is known. The learning process then consists of determining the class of any individual based on the value of its variables [3]. The construction method of decision trees consists of a segmentation of the learning population in order to obtain groups in which the class size is maximized. This segmentation is then reapplied recursively on the partitions that have been obtained. The search for the best partition for the segmentation of a node returns to find the most discriminating variable for the classes. Thus the tree (or more commonly the graph) is made. Finally, decision rules are obtained by following the paths from the root of the tree (the entire population) to its leaves.
To illustrate this notation, consider the problem of psychological tests in particular the personality tests

(BDI-II). Table 1 shows our learning sample of 14 patients (the students in secondary schools). Each example or patient is described by four descriptive variables (Sense of failure, Sense of guilt, Sadness, and Pessimistic) « in statistics called exogenous variables » and is associated with a particular attribute (Depression).

Table 1: Example of learning sample-patients

	X_1	X_2	X_3	X_4	X_5
Ω_α	Sense of failure	Sense of guilt	Sadness	Pessimistic	Depression
1	yes	yes	no	no	Major
2	no	yes	yes	no	Major
3	no	no	no	no	Dysthymia
4	no	can be	no	can be	Dysthymia
5	yes	can be	no	yes	Major
6	no	can be	can be	yes	Dysthymia
7	no	no	can be	yes	Dysthymia
8	yes	yes	yes	can be	Major
9	yes	no	yes	can be	Major
10	yes	no	no	can be	Major
11	no	can be	can be	no	Major
12	no	yes	no	no	Dysthymia
13	yes	yes	yes	no	Major
14	no	no	yes	no	Dysthymia

In the table shown below, we have summarized an example of exogenous variables out of our learning samples. The value taken by X_j *(w)* is called the term or value of the variable X_j for each individual (patient). We denote by L_j the number of different modalities assigned to the variable X_j.

Table 2: Example of learning sample-patients

Variables and (l_j)	*Signification*	*Values*
$X_1(l_1=2)$	Sense of failure	yes ; no
$X_2(l_2=3)$	Sense of guilt	yes ; no ; can be
$X_3(l_3=2)$	Sadness	yes ; no; can be
$X_4(l_4=3)$	Pessimistic	yes ; no ; can be
$X_5(l_5=2)$	Depression	Major ; Dysthymia

In this case, the patient population that is affected by the problem of learning is a set of tuples consisting of the four predictor variables of X_1, X_2, X_3, *and* X_4 and their classes (*Major depression and Dysthymia).* From these examples, we construct a tree decision as:

- Each *node* corresponds to a *test* of the value X_j *(w)* of an attribute X_j in which l_j has the possible values $\left(x_j^1, \ldots, x_j^{l_j}\right)$;
- Each *branch* leaving a node corresponds to a value of x_j^v for the test on x_j with $v=1,\ldots,l_j$;
- Each leaf is associated with the c_k value of the target attribute *Y*.

Suppose our learning sample is composed of 14 patients. The initial partition (s_0) has a single element that is denoted as s_0, which includes all of the learning samples with eight (8) individuals (patients) belonging to the class « *major depression*» and six (6) belonging to the class «*dysthymia*».

For the construction of the decision tree, we used the *ID3* algorithm method [14]. *ID3* (*Inductive Decision Tree*) is a heuristic tree that is used to construct a decision tree. Its principal consists in generating a succession of partitions by splitting nodes of the tree. Its objective is to optimize a criterion of *information gain*. From the sample of the learning method *ID3* is symbolic processing that begins the construction of the decision tree [3], [26]:

- Choose the measurement uncertainty (Shannon or quadratic) ;
- Initialize the parameters *of gain, info* and the initial partition S_0;
- Apply the method *ID3* to pass of the partition S_t to S_{t+1} and to generate the decision tree ;
- Finally, generate the prediction rules.
- Starting from the root of the tree (see Fig. 6) the partition $S_1 = \{s_1, s_2\}$ is generated by the variable X_1 with:
  - $s_1 = \{\omega \in \Omega\alpha \mid X_1(\omega) = \text{yes}\}$;
  - $s_2 = \{\omega \in \Omega\alpha \mid X_1(\omega) = \text{no}\}$;

Just as with the node s_0, we distinguish s_1, and s_2 the individual classes of 1 and 2 of the partition S_1. The process is then reiterated in search of partition S_2 which would be better according to the chosen gain (see Fig. 6). Thus, the decision tree can be then exploited to: extract the classification rules concerning our target attribute « *depression* ».

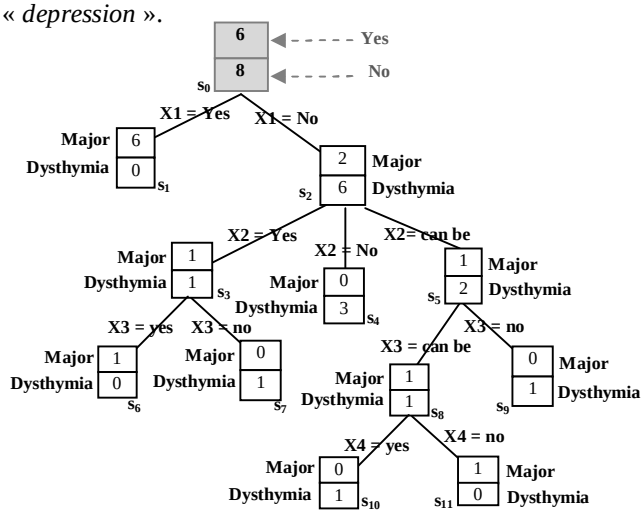

Fig. 6 Example of a figure caption decision tree obtained by ID3 on the example "Test by BECK"

Finally, we exploited the tree before (Fig. 6) to extract five rules R_1, R_2 … and R_7 of the psychological inductions (*Test by Beck*) also on the target attribute « *depression* ». Useful and critical knowledge have not been explicit before and are of the form: **If** *Condition* **then** *Conclusion.* Where ***condition*** is a logical expression that is composed of summits that will be called the ***premise*** and the

c***onclusion*** is the majority class in the summits described by the condition.

1. **If** (sense of failure = yes) **then** major depression.
2. **If** (sense of failure = no and sense of guilt = no) **then** dysthymia.
3. **If** (sense of failure = no and sense of guilt = yes and sadness = yes) **then** major depression.
4. **If** (sense of failure = no and sense of guilt = yes and sadness = no) **then** dysthymia.
5. **If** (sense of failure = no and sense of guilt = can be and sadness = no) **then** dysthymia.
6. **If** (sense of failure = no and sense of guilt = can be and sadness = can be and pessimistic = yes) **then** dysthymia.
7. **If** (sense of failure = no and sense of guilt = can be and sadness = can be and pessimistic = no) **then** major depression.

Exploitation of the rules of induction: In this section, we launch the validation phase across the CV module (*Cellular validation*) on the psychological induction rules (*Personality Tests «Test by BECK»*) presented in the previous section.

Table 3 shows how the Boolean knowledge base was extracted starting from the psychological induction rules (*personality tests - BDI-II*) and how it is modeled on the layers *CELFACT* and *CELRULE* (as shown in section 3.1). Note that in this step, the two incidence matrices of input (R_E) and output (R_S) are generated (see Table 4).

Table 3: The initial configuration of CELFACT and CELRULE

Facts	EF	IF	SF
sense of failure = yes	0	**1**	0
sense of failure = no **and** sense of guilt = no	0	**1**	0
sense of failure = no **and** sense of guilt = yes **and** sadness = yes	0	**1**	0
sense of failure = no **and** sense of guilt = yes **and** sadness = no	0	**1**	0
sense of failure = no **and** sense of guilt = can be **and** sadness = no	0	**1**	0
sense of failure = no **and** sense of guilt = can be **and** sadness = can be **and** pessimistic = yes	0	**1**	0
sense of failure = no **and** sense of guilt = can be **and** sadness = can be **and** pessimistic = no	0	**1**	0
Major depression	0	**1**	0
Dysthymia	0	**1**	0
CELFACT			

Rules	ER	IR	SR
R_1	0	**1**	**1**
R_2	0	**1**	**1**
R_3	0	**1**	**1**
R_4	0	**1**	**1**
R_5	0	**1**	**1**
R_6	0	**1**	**1**
R_7	0	**1**	**1**
CELRULE			

Table 4: Input / output of the incidences matrices

R_E	R_1	R_2	R_3	R_4	R_5	R_6	R_7
sense of failure = yes	**1**	0	0	0	0	0	0
sense of failure = no **and** sense of guilt = no	0	**1**	0	0	1	0	0
sense of failure = no **and** sense of guilt = yes **and** sadness = yes	0	0	**1**	0	0	0	0
sense of failure = no **and** sense of guilt = yes **and** sadness = no	0	0	0	**1**	0	0	0
sense of failure = no **and** sense of guilt = can be **and** sadness = no	0	0	0	0	**1**	0	0
sense of failure = no **and** sense of guilt = can be **and** sadness = can be **and** pessimistic = yes	0	0	0	0	0	**1**	0
sense of failure = no **and** sense of guilt = can be **and** sadness = can be **and** pessimistic = no	0	0	0	0	0	0	**1**
Major depression	0	0	0	0	0	0	0
Dysthymia	0	0	0	0	0	0	0

R_S	R_1	R_2	R_3	R_4	R_5	R_6	R_7
sense of failure = yes	0	0	0	0	0	0	0
sense of failure = no **and** sense of guilt = no	0	0	0	0	0	0	0
sense of failure = no **and** sense of guilt = yes **and** sadness = yes	0	0	0	0	0	0	0
sense of failure = no **and** sense of guilt = yes **and** sadness = no	0	0	0	0	0	0	0
sense of failure = no **and** sense of guilt = can be **and** sadness = no	0	0	0	0	0	0	0
sense of failure = no **and** sense of guilt = can be **and** sadness = can be **and** pessimistic = yes	0	0	0	0	0	0	0
sense of failure = no **and** sense of guilt = can be **and** sadness = can be **and** pessimistic = no	0	0	0	0	0	0	0
Major depression	**1**	**0**	**1**	**0**	**0**	**0**	**1**
Dysthymia	**0**	**1**	**0**	**1**	**1**	**1**	**0**

For the experimentation phase we used the platform WS4KDM [7] for extraction and the Boolean modeling determining the rules of psychological prediction. The objective and the automatic improvement of the Boolean model of knowledge mapping critical epidemiological guided by data mining. WS4KDM takes as its input the learning sample as a table of individuals /variables in order to supply a basis of classification rules during output, by applying the principle Boolean of the cell machine. The result of the cartography of knowledge domains of SEMEP obtained by CARTOCEL [18, 19, and 20] is illustrated in Fig. 7.

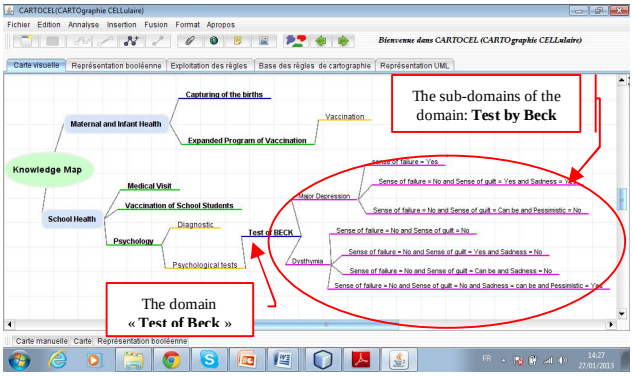

Fig. 7 Knowledge domains mapping of SEMEP as refined by a process of inducing rules « Test of Beck »

4. Results and discussion

Several methods of capitalization, operation and enrichment [11, 12] have been proposed in the literature of knowledge management. Among these methods the MASK method should be mentioned. According to [11], MASK is defined as a task of observing and mastering a knowledge system in its totality and in its complexity. It also benefits from various documentary sources, observation analysis and experience feedback that should all be continually refined and enriched.
However, MASK encounters limitations because it is not possible to set up MASK on more than a limited domain of knowledge patrimony of the company. This method is thus concerned with past practices or procedures. It does not allow for the formalizing of data and information in real time. Moreover, MASK essentially provides a tool for aiding in decisions, but it does not allow for conducting an automatic search, in the knowledge map, for having the rules of mapping decisional.

As we stated earlier (as shown in section 3.2), we are also interested in empirical learning, which aims to generate new knowledge from practical cases (observations) that are known to neither implicitly nor explicitly, in order to facilitate the work of the expert and that meet the needs of an organization.

For the experimental phase we used the psychologists' evaluation sheets from the SEMEP service. Through the evaluation educational psychologists will be able to issue recommendations which may serve the basis for follow-up strategies, intervention or orientation towards other professionals in education or health. In addition, the latter are highly diverse and are not necessarily valorized by the data mining techniques. We are being interested in the scale of BDI-II (*Beck Depression Inventory 2nd Edition Scale*) [1] to evaluate the depression of the *students attending schools*. The latter is a self- evaluation questionnaire that is designed to measure the depression severity. It contains 21 elements (items) describing many of the symptoms of depression. Each of its items are presented in the form of four propositions whose subject must make a choice from among them, while choosing the statement that best describes his state of mind during the last 2 weeks. Beck et al [1] have identified two types of depression are usually: ***major depression*** and ***dysthymia***. Major depression consists in one or more depressive episodes which slice with the usual operation of the person, while dysthymia is characterized by less severe depressive symptoms but chronic.
From a practical point of view we constructed, as a first step, a new questionnaire, which is composed of four items. We did so in collaboration with the school psychologists of SEMEP and with the fundamental schools. We retained the following categories: 1 (*Sadness*), 2 (*Pessimistic*), 3 (*Sense of failure*), 5 (*Sense of guilt*), in order to measure the depressed state of the students from secondary schools.
To refine the Boolean mapping of the critical psychological knowledge, we conducted an experimental study using the platform WS4KDM [7] for extracting and modeling the Boolean rules for predicting psychological. WS4KDM takes in the input of the learning sample (*psychological sheets - Test by BDI-II*) in a table format showing the individuals / variables and outputs a base of Boolean production rules by applying the principle of the cell machine CASI (see Fig. 8).

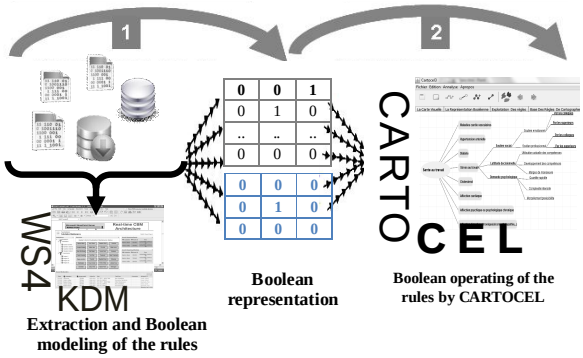

Fig. 8 Boolean exploitation of the rules extracted by WS4KDM.

Our objective, through using the CARTOCEL system [18, 19, and 20], is to bring about a mapping of knowledge domains of SEMEP that have been refined by the

[1] BDI (*Beck Depression Inventory*) was published for the first time in 1961 by psychiatrist Aaron T. Beck (Beck et al., 1979) and revised in 1996 (BDI-II) to account including current diagnostic criteria for depression (Beck et al., 1996).

classification rules that were extracted using case studies (*BDI-II test*) in an automatic way.
Figure 9, below, illustrates the exploitation window of rules induction and initialization of two layers of cellular automaton (CELFAIT & CELREGLE).

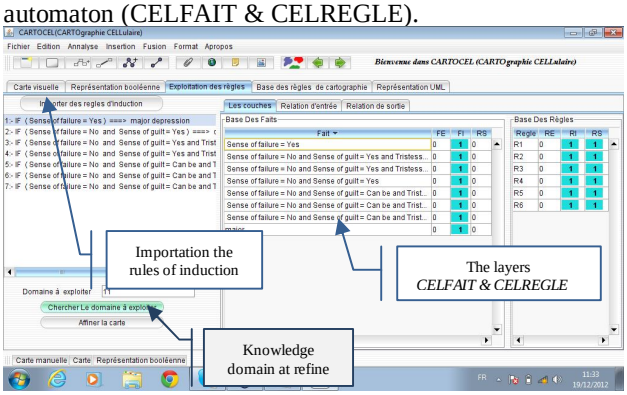

Fig. 9 Exploitation and knowledge base of rules in CARTOCEL

Finally, we choose the knowledge domain to exploit in the mapping performed (map of the service SEMEP "Axe: School Health") and finally refine this latter. Figure 7, above, illustrates the knowledge domain selected "Beck tests" with subdomains extract from the rules.
The experimental results improved the construction of the knowledge map by proposing a new process of knowledge mapping that is guided by data mining. The advantages of our approach, which is based on the principle of Boolean modeling to render mapping of critical knowledge that is more flexible and scalable may be recapitulated as follows:

- The representation of knowledge well as his control is simple, as they are in the form of binary matrices and require the minimum amount of preprocessing.

- The facilities of the implementation transition functions are complex, efficient and robust concerning of extreme values. Moreover, they are well suited with situations that have many attributes.

- The results of the mapping are simple for being reorganized and used by and for the data mining process.
- The mapping system is a cellular model that is composed of a simple set of transition functions and production rules, which allow ensure the evolution of knowledge maps by exploiting other sources of the company data;
- The ease of navigation and dynamic search of knowledge decisional by using the transition functions and production rules.

5. Conclusions

Competing motivations have led us to propose this principle cellular for optimization, generation, representation and use of Boolean knowledge mapping. Indeed, we have, not only wished to have a knowledge map optimal but also, we have, too, wanted to improve the construction and the visualization of this map by proposing a new process of knowledge mapping guided by data mining. The Boolean model of the critical knowledge mapping domains thus obtained is perfected by a process of the automatic learning symbolic graph-based induction. This improvement is made by the CASI cellular automaton that goes assist the ID3 method in the process of extracting new knowledge starting from past experience «*in the form a practical case*».
Finally, critical knowledge mapping domains that are guided by *data mining* have interesting properties and numerous advantages over other techniques of knowledge explicitation. This new principle of mapping the knowledge domains forms a very satisfactory tool for formalization (*the creation of new knowledge*), representation and the visualization of knowledge domains by adopting a Boolean modeling, which ensures, thereafter, a contribution in the general process of the creation, transfer and reuse of knowledge (*tacit or explicit*).
The results of our work offer many perspectives for further research at both the theoretical and practical level. Whence, we cite, the following, succinctly few of these perspectives:

- Apply our approach to large masses of heterogeneous data and hyper-documents (*text, web*) in order to reach the effective exploitation of this approach and validate its performance;
- Contribute to the visual data mining by the proposal a new approach to decisional knowledge mapping;

- Extending our work to other psychological tests and other knowledge domains of SEMEP service;
- Improving the CARTOCEL system by the integration of advanced techniques of computing as well to enable collaboration and more extensive of the community of SEMEP services and epidemiological researchers.

Acknowledgments

This project is registered in the context of National Program of Research (NPR) , launched in collaboration between the health sector SEMEP, our research team SIF «*Simulation, Integration and Data mining*» of the laboratory LIO «*Computer Science Laboratory of Oran*»,

and the laboratory Tech-CICO in University of Technology of Troyes (UTT). The authors also acknowledge the service team SEMEP the psychologists in secondary schools for her assistance on to finalize this project.

References

[1] A. T. Beck, and B. A. Alford. Depression: Causes and Treatment, Publisher: University of Pennsylvania Press; 2nd edition: February 25, 2009, ISBN-10: 0812219643, pp. 432.

[2] B. Atmani, and B. Beldjilali, "Knowledge Discovery in Database: Induction Graph and Cellular Automaton", Computing and Informatics Journal, Vol.26, N°2 (2007), pp.171-197, 2007.

[3] D.A. Zighed, and R. Rakotomalala, Graphs of induction, Training and Data Mining, Hermes Science Publication (Edition Hermès Sciences), 2000, pp. 21-23.

[4] E. Blanchard, M. Hazallah, and H. Brinad, "Reasoning in competence management", Workshop: Extraction and Knowledge Management- EGC 2005, Volume II, pp. 587, cépaduès-editions, ISBN : 2.85428.677.4

[5] F. Jalabert, "Knowledge mapping: biological data integration and visualization applied to knowledge engineering and gene expression data analysis", Doctoral thesis at the University of Montpellier, France, specialty : structures and systems, 2007.

[6] G. Aubertin, "Knowledge mapping: a strategic entry point to knowledge management". Trends in Enterprise Knowledge Management, ISTE, Londres, 2006.

[7] H. Kadem, and B. Atmani, "Designing a Cellular Platform Open Source of Extraction and Knowledge Management: WS4KDM". 7th National Seminar in Computer Science BISKRA (SNIB'2010), University of Mohamed Khider - Biskra, Algeriae, 02-04 Nov 2010.

[8] I. Boughzala, J.L. Ermine , Management des connaissances en entreprise, Collection technique et scientifique des télécommunications, Hermès, 2004.

[9] I. Nonaka, and H. Takeuchi, The Knowledge-Creating Company, Oxford University Press, Oxford, New York, 1995.

[10] I. H. Witten, Eibe Frank, and M. A. Hall. Data Mining: Practical Machine Learning Tools and Techniques, Editor: Morgan Kaufmann, Edition: 3 (3 February 2011), ISBN-10: 0123748569.

[11] J.L. Ermine, I. Boughzala, and T. Tounkara, "Critical Knowledge Map as a Decision Tool for Knowledge Transfer Actions", The Electronic Journal of Knowledge Management, Vol. 4, Issue 2, 2006, pp.129-140.

[12] J.L. Ermine, "A Theoretical and formal for Knowledge Management Systems", dans D. Remenyi, 2nd International Conference on Intellectual Capital and Knowledge Management (ICICKM'2005), Dubia, United Arab Emirates (U.A.E), 2005, pp. 187-199.

[13] J. Han, M. Kamber, Data Mining : Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, University of Illinois at Urbana-Champaign (Canada), 2nd Edition, Elsevier, 2006, ISBN : 10: 1-55860-901-6, Available online : http://www.cs.uiuc.edu/~hanj/bk2/

[14] J. R. Quinlan, Induction of decision trees, Machine Learning, Volume1: Issue 1, 1986, pp. 8-106.

[15] Le Khac. Nhien An, and M. A. Lamine "Distributed Knowledge Map for Mining Data on Grid Platforms", IJCSNS International Journal of Computer Science and Network Security, Vol.7, No.10, October 2007, pp. 98-107.

[16] T. Buzan, "A head well done - Use you intellectual resources", Editor: Organisation Eds, Novembre 2011, and ISBN: 2212552149, pp. 1-186.

[17] U. Fayyad, G.P. ShapirO, and P. Smyth, "The KDD process for extraction useful knowledge from volumes data", In: Communication of the ACM, Vol. 39, Nov.1996, pp. 27-34.

[18] M. Brahami, and B. Atmani, "Towards a knowledge mapping guided by data mining: first step Boolean modeling". 2nd Conference GECSO'09, Electronic Journal ISDM N° :36, 2009.

[19] M. Brahami, B. Atmani, and M. Mokaddem, "CARTOCEL : A tool for knowledge mapping guided by cellular machinery CASI", In 10th International Conference on Extraction and Knowledge Management, EGC'2010, RNTI (E-19), Edition Cépaduès, ISSN : 1764.1667.

[20] M. Bramer, Principles of data mining, Editor Springer 2007, ISBN1846287669, pp.15-40.

[21] M. Grundstein, "From capitalizing on company knowledge to knowledge management", Knowledge Management: Classic and Contemporary Works M. Press, Daryl Morey, Mark Maybury and Bhavani Thuraisingham, 2000, pp. 451.

[22] M. Jelena, and I. Beleviciute, "Data mining for knowledge management in technology enhanced learning", Proceedings of the 6th conference on Applications of electrical engineering, Istanbul, 2007, Turkey, ISBN: 0-9547096-5-9, pp.115-119.

[23] N. Matta, and J.L. Ermine, "knowledge capitalization with a knowledge engineering approach: the MASK method", IJCAT'2001, knowledge management and organisational memory workshop. International Joint Conference on Artificial Intelligence, seattle, Etats-Unis, 4-10 août 2001.

[24] P.H. Speel, N. Shadbolt, W. De Vies, and P.H. Van Dam, O'hara K, "Knowledge Mapping for industrial purpose", *Conference KAW'99*, Banff, Canada, 1999.

[25] P. Van Berten, J.L. Ermine, "Applied Knowledge Management: a set of well-tried tools". The Journal of Information and Knowledge Management Systems, vol.36, 4, 2006, pp.423-431.

[26] R. Rakotomalala, "Induction Graphs", Thesis for the obtaining of the Diploma PhD, University of Claude Bernard-Lyon 1, 1997, France.

[27] P. Flach. Machine Learning, Editor: Cambridge University Press; Edition: 1 (November 29, 2012), ASIN: B009ZRNT0C, pp.5-50.

[28] T. Silwattananusarnl, and K. Tuamsuk, "Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012, pp.13-24.

[29] V. Kvassov, and S. C. Madeira, "Using data mining techniques for knowledge management: an empirical study". 9th AIM (Association Information and Management) Conference – Information Systems: Critical Perspectives (AIM 04), May. 2004, PP.5-23.

[30] W. Hai, and S. Wang, "A knowledge management approach to data mining process for business intelligence", Journal: Industrial management & data systems, Volume: 108, Issue: 5, 2008, pp. 622-634.