Medical Knowledge Management since the Integration Heterogeneous Data until the Knowledge Exploitation in a Decision-making System

Nadjat ZERF BOUDJETTOU, Fahima NADER and Rachid CHALAL

Abstract— Knowledge management is to acquire and represent knowledge relevant to a domain, a task or a specific organization in order to facilitate access, reuse and evolution. This usually means build, maintain and evolve an explicit representation of knowledge. The next step is to provide access to that knowledge, that is to say, the spread in order to enable effective use. Knowledge management in the medical field aims to improve the performance of the medical organization by allowing individuals in the care facility (doctors, nurses, paramedics, etc..) to capture, share and apply collective knowledge in order to make optimal decisions in real time. In this paper we propose a knowledge management approach based on integration technique of heterogeneous data in real time. In this paper we propose a knowledge management approach based on integration technique of heterogeneous data in the medical field by creating a data warehouse, a technique of extracting knowledge from medical data by choosing a technique of data mining, and finally an exploitation technique of that knowledge in a case-based reasoning system.

Index Terms— Data warehouse, Data Mining, Knowledge Discovery in Database (KDD), Medical Knowledge Management, Bayesian networks.

I. INTRODUCTION

The orientation of patients toward different services and different specialties is a major problem in major health institutions such as center university hospitals. If the patient is unable to recognize the first signs of a medical problem or injury, he can make a two or three months long appointment with a specialist who may not be the right one.

The relevant medical decision-making both diagnostic and therapeutic needs a good knowledge of the case to be treated. The doctor needs to know its clinical data, which implies a strong collaboration between health professionals and interoperability between the systems being used. Given the complexity of the medical field, we encounter several problems such as [1]:

• The diversity of the distributed information sources and their heterogeneous.
• The problems of access to relevant information for the care are related to the dispersion of medical information on different health information systems which are often autonomous and heterogeneous.
• The formalization of the information does not allow mining, sharing, dissemination and use of medical knowledge.
• The difficulty to understand the mechanisms of interpretation and medical reasoning.

II. OBJECTIVE OF WORK

Taking into account the problems cited is one of the keys to the establishment of a system that allows knowledge management to intervene in the medical field, we propose our system that can:

• Provide a solution to integrate heterogeneous data,
• Propose an approach to extract relevant knowledge from data,
• Exploit the result (knowledge) in case-based reasoning system “Orientation of patients to medical services”.

Our goal is to design a system of knowledge management in the medical field. We will show in this paper how we could meet this goal while trying to take into account the synergy between different approaches such as data warehousing, knowledge discovery from data and case-based reasoning.

III. STATE OF THE ART

The management of medical knowledge is developing and evaluating methods and systems for the acquisition [1], processing and interpretation of knowledge extracted from data "patient" with the help of knowledge from scientific research. To achieve these objectives, medical uses scientific methods that inherit from computer science, artificial intelligence, mathematics and management science [2].

The first intelligent systems or diagnostic support systems in medicine were case-based reasoning systems. Then, as and when time and as required, data mining methods have been integrated into these systems. These methods have filled gaps in the systems of Case-Base Reasoning [3].

IV. PROPOSED APPROACH

In this section we will describe the proposed approach, present the process of building a medical data warehouse to solve the problem of heterogeneous data sources, then we will choose a data mining technique for extracting knowledge from data, then the result of the proposed approach is a knowledge base used for the case-based reasoning system.

A. Integration of Medical heterogeneous data

This part of the work is to extract, in advance, the relevant medical data for use by users (nurses, doctor, radiologist, biologist ...), to filter, transform and store. To meet our needs we will build a data warehouse that stores medical data from physically repartees sources, these generally correspond to the experimental data sources from the
health facility. To be usable, all data from distributed systems must be organized, coordinated, integrated, and finally stored to give the user an overview of information.

1) The design of the medical data warehouse

In our work, we have chosen the real architecture, it is generally used for decision maker systems, data storage is performed in a DBMS separate from the operational system, and these systems are designed in an autonomous and distributed way. The complete system includes all systems of production such as: management consultation, management of medical appointment, the management of admissions, management of hospital accommodation, management and medical analysis laboratory, management of radiology, gynecology and obstetrics unit management, financial management and accounting, management of pharmacy and medicine stocks, the management of hemodialysis, human resource management (HRM).

2) Building the medical data warehouse

To build the medical data warehouse, we followed the process of ETL (Extraction, Transformation, Loading) [4][5]:

a) The data extraction

This phase collects useful data of operational data sources from various disparate sources. The extracted data can be: Interaction of a doctor, a biological test result, image file, Time series, Voice comment and clinical data.

b) The data transformation

Sources can also present problems of semantic heterogeneity, these problems are addressed by specifying a set of transformation rules to achieve a uniform representation, for example, replace the word "gender" with "sex" or "Women" by "F" and "Men" with "M".

During the transformation phase of data, we encountered several types of conflicts that are each treated separately: conflict of classification, descriptive conflict, structural conflict, data / metadata conflict and data conflicts.

c) The loading of data contained in the target system

The cooling process in our system is incremental, the update uses changes in the sources each time a source changes or periodically with a period that depends on the needs of users and the access charge to the warehouse. The following figure summarizes the sequence of these processing steps:

3) Construction of the warehouse schema

The result of processing of the data warehouse building is to define a global schema providing an integrated view of the sources that will be exploited later in the process of extracting knowledge from data [6].

We have identified in the global schema a representation of the different packages and classes that comprise the medical data warehouse.

Fig. 1. Steps of processing building of medical data warehouse

Fig. 2. Global schema of data warehouse.
4) Data Mart Consulting

The data mart "Medical Consultation" is focused and driven by the needs of our system. It has the same purpose as the medical data warehouse (provide architectural decisions), but it aims to solve our problems with a smaller number of users [7]. The following figure illustrates the decomposition of the data warehouse in multiple data marts:

![Fig. 3. Decomposition of the data warehouse in multiple data marts.](image)

Our work is based on the data mart "Medical Consultation"; it aims to solve our need for medical orientation toward services. It includes three packages:

- **Consultation Package**: contains classes that detail the data such as consultation code, consultation date, functional signs, pain, symptoms, diagnosis and treatment.
- **Patient Package**: contains classes that detail patient data such as name, sex, age, weight, date of birth, address...
- **Assignment Package**: contains the consultation service.

5) Multidimensional modelling warehouse

The modelling that we adopted for our data mart is the star schema [8]. This pattern consists of a fact relationship "Consult" and ten dimensional relations (Patient Record, Service, Age, Type weight, fever, pain, functional Sign 1, functional Sign 2, functional sign 3, functional sign 4).

![Fig. 4. Star schema of the Data Mart "medical Consultation".](image)

B. Extracting Knowledge from Data

In this part of the system, we'll create a process for extracting medical knowledge from the data warehouse (data mart "Medical Consultation") constructed in the previous section. The process we have used is divided into several phases [9]: data selection, data preparation, use of an intelligent data mining method applied to the processed data and finally the evaluation and validation of models.

![Fig. 5. Process of extracting knowledge from medical data.](image)

1) Data selection

This phase can be summarized in the selection of data which allows the operation to respond to our problem, it requires:
Data specification: the whole data contained in the medical data warehouse, and that will be used for the resolution of our problem is the data mart "Medical Consultation".

Data mining methods specification: the Bayesian Network is the appropriated method to reach our target, because it's based on the causality notion (cause for purpose).

Specification of measurement: from all existing algorithm, we prefer to use the k2 algorithm, for its speed and his capacity to restrict the search space (seek interesting arcs).

Representing the results of data mining and representation of knowledge extracted: in our system the extracted knowledge will be represented by a Bayesian Network Graph. The quantitative part is the graph that has the links indicate causality between the different variable, the qualitative part is the whole of the conditional probabilities tables which represents the degree of reliability.

2) Data Preparation
In this step, several procedures are required [6]:

- The procedure of data cleaning: we carried out several methods to supplement missing data in the medical data mart, example: ignore incomplete instances, complete the data manually and complete the incomplete data using a function or a global constant.

- The transformation procedure: we realize that the pretreatment is already made by the phase transformation of ETL, but depending on our needs we were obliged to reprocess the data so that they are adequate for the k2 algorithm. The k2 algorithm work on discrete data then we have to discretize the attributes continues.

- The reduction procedure: in cooperation whit an expert (a doctor), we reduce the number of variables to the maximum, the expert proposed that most important variables for the orientation of a patient (from 100 to 40 variables), the remains variables doesn’t had a significant effect on the results, but a very important impact on the execution time of the algorithm.

3) The data mining
It is the heart of the process of KDD [10]. It is at this level to find knowledge from data. The work consists of applying intelligent methods in order to extract this knowledge.

a) Choice of method
The whole problem of data mining lies in choosing the appropriate method to a given problem. The method we chose in our study and that meets our needs is that of Bayesian Networks. This technique can be applied to any sector of activity: banking, telecommunications services, for example. It can also be used in medical research [11].

The Bayesian Networks [12] are based on the conditional probabilities. They allow somehow predicting the future from the past, assuming reproducibility of probabilities. This technique derives from Bayes theorem, which is used to calculate the probability of an event, A knowing that E was observed. Therefore, the construction of a Bayesian network allows finding rules with transitional probabilities. These techniques are also widely used for fault detection.

The characteristics of this method [13]:

- The ability to gather and merge knowledge from different sources in a single model: data feedback from experience and expertise (expressed as logical rules, equations or subjective probabilities), or observations.

- The accessibility: a Bayesian network is graph model who is understandable and easy to handle by a non-specialist.

- The Universality: we can use the same model to asses, plan, troubleshoot, and optimize the decisions.

b) Choice of algorithm: K2 [14]
Cooper & Hersovits proposed a method widely used in learning structure of BN. This method is called K2. Its purpose is to maximize the probability of the structure given the data in the space of ADG (space acrylics directed graph) satisfying the constraint in the enumeration order. This constraint can restrict the search space and limited it on interesting arcs only.

Indeed, the k2 algorithm test the addition of parents with respect of this order in the following manner: if Xi is the first on the list he can’t have parent. If not, if Xi is quoted before Xj then there can be no bow from Xj to Xi. This technique used the Bayesian Dirichlet score, calculated by the following formula:

\[
f(i, \pi_i) = \prod_{j=1}^{n} \left( \frac{(x_i - 1) + \alpha_{ij}}{(x_i + \alpha_{ij})} \right) \prod_{k=1}^{m} \alpha_{jk}!
\]

With:

- \( \pi_i \): set of parent of node Xi
- \( q_i : \{\phi_i \} \): List of all possible instantiations of the parents of Xi in the database D.
- \( r_i [V_i] \): List of all possible values of the attribute Xi.
- \( a_{ijk} \): Number of cases in DB in which xi attribute is instantiated with its value of kth, and the parents of Xi in \( \pi_i \) are instantiated with instantiation j in \( \phi_i \).
- \( N_{ij}^{\pi_i} = \sum_{k=1}^{r_i} a_{ijk} \): Number of cases in DB with which parents of Xi in \( \pi_i \) are instantiated with instantiation j in \( \phi_i \).

Our data mart contains a considerable mass of data, and our goal is to discover relevant and useful knowledge in order to be able to guide patients to the different services. Each record stored in the data mart "Consultation" contains data from a patient:

- Identification: name, age, sex, weight, address...
- Functional signs of the patient: fever, diabetes, blood pressure, the list of pain, the list of other signs...
- Administrative data: consultation code, file number, the doctor, date of consultation, service of consultation ...

List of variables used: we used more than forty variables which represent the score part (sex, age type, weight, fever, diabetes, blood pressure, smoking, pregnancy, pain, burn, burning micturition, colic, constipation, diarrhea, palpitation, cramp...) And "Service" with its overall criteria represent the result part.

The data mining methods specification: the Bayesian Network is the appropriated method to reach our target, because it’s based on the causality notion (cause for purpose).

Specification of measurement: from all existing algorithm, we prefer to use the k2 algorithm, for its speed and his capacity to restrict the search space (seek interesting arcs).

Representing the results of data mining and representation of knowledge extracted: in our system the extracted knowledge will be represented by a Bayesian Network Graph. The quantitative part is the graph that has the links indicate causality between the different variable, the qualitative part is the whole of the conditional probabilities tables which represents the degree of reliability.
For application we considered more than 10000 records.

- **Structural learning:**
  Initially our job is to determine a graph taking into account different conditional independence between variables of this graph, we will calculate the score of each variable by the formula (1) to have list of parents of each node:

  \[
  \rho(X_i = x_k | Pa(X_i = x_j)) = \delta_{ij,k} = \frac{N_{ijk}}{\sum_k N_{ijk}}. \tag{2}
  \]

  With: \(N_{ijk}\) is the number of events in the database for which the \(X_i\) variable is in the \(x_k\) State and her parents are in the configuration \(x_j\).

  Result is illustrated in the following figure:

- **Settings learning:** once the structural part of the graph is created, it remains to fill the table of conditional probabilities of each node in the network, learning parameters is to estimate these probabilities.

  - In our case all variables are observed, the simplest method and most used method is «maximum of likelihood» calculated by:

  \[
  \rho(X_i = x_k | Pa(X_i = x_j)) = \delta_{ij,k} = \frac{N_{ijk}}{\sum_k N_{ijk}}. \tag{2}
  \]

  Tables of probabilities are defined by statistics on the problem to be solved (can also be determined by experts). Each prepared a table of conditional probabilities variables relating to the causal variables which they are extracted from data mart, result is illustrated in the following figure:

- **Evaluation and presentation of results**
  - The interaction with the domain expert is privileged in our system, the validation whit expertise is provided by a domain expert (doctor) who will judge the relevance of the outputs (graph relationship and conditional probability table).
  - The presentation phase is to interpret the results to the user through visualization techniques. The results of data mining are shown as Bayesian graph.
3) Process of case-based reasoning

The operation of this part of our work is based on the four parties that comprise the RBC systems [15]: part research, part adaptation, part revision and the part of memory:

a) Party Research

When presenting a new case, this phase determines the cases that are most similar to the problem given the patient, after extraction of known indices that will be used to make the search for similar cases.

- Overall similarity calculation [16]: the overall similarity (distance) of two cases C1 and C2 is calculated by the following formula:

$$d(x,y) = \sum_{i=1}^{n} \frac{|x_i - y_i|^2}{W_i}$$  

With: $W_i$ represents the weight of each functional sign in the CBR interpreted by the results of Bayesian Networks, it is the conditional probability of each node.

- Selection of similar cases

The search procedure is implemented by a selection of the nearest neighbor (k-nearest neighbors). Similar cases are classified according to the similarity measures (weight) and if the highest ranked and a weight of reliability above the threshold is proposed as a solution. In our work, the part of case-based reasoning is based on the previous data mining and the problem of defining the weight of similarity of cases is resolved, we consider that the degree of similarity in each case based reasoning is the weight of reliability that extracted by the previous part.

In the case before it, we have four such similar with different degrees of similarity, where the service is offered "Gastroenterology", with a degree of similarity of 62.5% (considering threshold = 50%).

C. Reasons to Base Case

In this section we present our reasoning system based on a case by the medical guidance which helps the user to direct the patient to the appropriate department. The knowledge base is built by the process described in the previous part.

1) Construction of Base Case

The base case is constructed from the causal graph (structure learning)

Each path in the graph represents a case in the base case since $(x_1 = \text{sex})$ until $(x_{39} = \text{cloudy})$ and $x_40$ is the solution of the CBR. Conditional probability tables (settings learning) can define the weight of reliability of each solution of similar cases.

2) Case Representation

The description of the case under the care of medical referral is clinical cases already stored in our database of cases. Three main elements appear in the content of the case:

- The goal: the goal is trying to achieve the solution; it is the orientation of a patient to a medical service.

- Characteristics: whatever the representation of knowledge at "symbol chosen, the characterization is described by a finite set of couples <attribute, value>. It is about giving the serial number value.

- The solution to the problem: In our work the solution is the "medical service", "weight of reliability" of the service compared to the case.
Recent Advances in Environmental and Biological Engineering

Our approach is broken down into three basic phases: and case-based reasoning. combining the techniques of data warehousing, data mining the management of knowledge in the medical field, by the expert has access to the system to insert the new case. that there is no similar case or fairly close to this new case, case and those of the base, the system generates a message If after calculating the similarity measures between the new by human intervention (doctor). a real environment [17]. In our system, the revision is made The review is to evaluate the proposed solution by testing in part storage If after calculating the similarity measures between the new case and those of the base, the system generates a message that there is no similar case or fairly close to this new case, the expert has access to the system to insert the new case.

V. CONCLUSION
In this paper, we showed the importance of combining the management of knowledge in the medical field, by combining the techniques of data warehousing, data mining and case-based reasoning. Our approach is broken down into three basic phases:

- First, we built a data warehouse for medical solve the problem of heterogeneity of sources of medical data. To do this we followed the ETL (Extraction, Transformation, Loading).
- Second, we conducted a process for extracting medical knowledge from the data warehouse is already built, the process we have chosen is divided into several steps: data selection, preparation of selected data, using the technique of Bayesian networks of data mining applied to data processed, and finally evaluating and reporting results.
- Finally, we presented our system of case-based reasoning for medical guidance which helps the user to direct the patient to the appropriate service; we have exploited the knowledge extracted from data mining.

References
[14] Prof. Carolina Ruiz Department of Computer Science, WPI “Illustration of the K2 Algorithm for Learning Bayes Net Structures".