RESEARCH OF THE INTELLECTUAL SYSTEM OF KNOWLEDGE SEARCH IN DATABASES

Oleksii Vasilenko, Oleksandr Kuzomin, Bohdan Maliar

***Abstract**: The aim of the work is to research a modification for Rete pattern matching algorithm for medical expert rule-based systems. Developed modification uses the power of cloud system in order to shift the load from user's CPU and RAM.*

The work use Java programming language for implemented algorithms and support software. To provide high computing power used Amazon Elastic Compute Cloud – Amazon Linux AMI which included Java OpenJDK Runtime Environment.

***Keywords**: data mining, knowledge search, intellectual system, databases.*

***ITHEA Keywords**: E.2 Data Storage Representations, F.2 Analysis of algorithms and problem complexity.*

Introduction

The task of the differential medical diagnostic system is to determine the diseases that the patient may be ill, based on observations of his symptoms. Depending on the type of medical data there are two main approaches to medical diagnostics: diagnostics using the methods of probability theory and mathematical statistics based on objective statistical information; diagnostics using artificial intelligence based on subject information, i.e. knowledge and experience of a group of doctors.

One of the many areas of artificial intelligence is an expert system. The expert system has the following advantages: high efficiency; performance; reliability; accessibility to understand. To improve the efficiency of the system, knowledge base flexibility is required.

The system is focused on the possible expansion of the list of diseases, symptoms and interrelations between them by the criterion of the expediency of their use. A rule-based system is used to store and manipulate knowledge to interpret information in a useful way. A medical expert rule-based system is a software that is designed to help doctors around the world to choose the correct diagnosis based on a cluster of symptoms based on the list of knowledges provided by contact with a patient [Giarratano, 1994].

The common approach for rule-based system search is to analyze these knowledges and compare them with the knowledge base. The knowledge base consists of human-crafted rule sets.

Therefore, the goal of the work is to improve Rete algorithm for medical expert system by using the cloud as computing power.

Methods and models for problem resolution

Semantic Networks is the knowledge representation model most closely related to natural language. A semantic network is an oriented graph with vertices that correspond to symptoms, syndromes, etc. regarding to illness, concept or diagnostic situation, with other situations that can be identified by different methods that characterize the relationship between the objects of the study. The advantages of semantic networks include great expressive possibilities; visibility of the knowledge system, which is presented graphically; proximity to natural language, which is used when filling in the medical history (ontologies and precedents of “close” ones by the signs of the disease) compliance with the modern electronic representation of patients and their diseases; easy adjustment [Jackson, 1990].

The negative points of using the network model are the following facts:

- the model does not give a clear idea about the structure of the subject area, which corresponds to it, therefore, the formation and modification of such a model is very difficult;
- network models are passive structures, for the processing of which requires a special apparatus of formal withdrawal and planning;
- the complexity of the search and conclusion on semantic networks; the presence of multiple relationships between network elements;
- the presence of multiple relationships between network elements.

A few words according to the usage of frame structures for diagnostic modeling. In its organization, the frame model is a lot like a semantic network. A frame is a collection of nodes and relationships organized hierarchically, where the upper nodes are general concepts, and the lower nodes are more particular cases of these concepts. Frames reflect a typical pattern of action in a real situation. Each node is an attribute - a slot containing a certain concept and value of this concept, or how the value of a slot can be the name of another frame, or a slot can contain the name of a procedure with which a specific value can be calculated, or a range of values can be found in a slot. A set of frames can be networked, while the frame properties are inherited from top to bottom, the top frame contains more general information that is valid for the entire hierarchy of frames. The use of frames in the formalization of the medical subject area gives quite effective results.

Negative aspects when using frames of knowledge representation models include the difficulty of adapting the model when introducing new knowledge, the need to use sufficiently large amounts of memory that are necessary for storing model elements, as well as the lack of simple mechanisms for managing knowledge output.

The basic idea of building logical models of knowledge representation is that all information necessary for solving applied problems is considered as a set of facts and statements, which are represented as formulas in a certain logic. Knowledge is reflected by a combination of such formulas, and the acquisition of new knowledge is reduced to the implementation of inference procedures [Kuzomin and Vasylenko, 2010].

The positive qualities of logical models include the fact that the basis should be the classical apparatus of mathematical logic, whose methods are well studied and have a formal justification; the presence of sufficiently effective procedures for withdrawal; knowledge bases can only be stored using a variety of axioms, and all other knowledge can be obtained from them by the rules of inference. The main drawbacks of the logical method of knowledge representation are the considerable time spent on building a chain output; the inability to effectively describe the rules on exceptions; the need to describe a large number of rules and statements when modeling a real medical disease, which may be contradictory.

Ontologies are another way of describing the subject area, the basic concepts of this area, their properties and the connections between them. Practically all models of ontologies contain concepts (concepts, classes, entities, categories), properties of concepts (slots, attributes, roles), relationships between concepts (connections, dependencies, functions) and additional restrictions. Ontology, together with many individual instances, constitute the knowledge base, describes the facts based on the generally accepted meaning of the dictionary used. The advantages of ontologies are determined by the possibility of sharing by people or software agents for a common understanding of the structure of information; the ability to reuse knowledge in the subject area; the ability to separate knowledge in the subject area from operational data; ability to analyze knowledge in the subject area.

As an example of using the semantic network model for solving medical diagnostics problems, we can use the CASNET system (Causal-Associational NETwork). On the basis of logical models, such systems as MYCIN - a system for diagnosing bacterial infections or a system INTERNIST - a diagnostic system in the field of general therapy are built. Given the level of development of information computer technologies, ontologies have found their application in many subject areas, in particular in medicine, the SNOMED dictionary or the Unified Medical Language System can be noted, which are used to formalize common terminology and annotation in the field of medicine.

Formally, the task of medical diagnosis can be presented as a classification task, which consists in the fact that in order to match the set of input parameters a specific disease.

The basic approaches that are used to solve the problem of medical diagnosis can be grouped as follows:

1. Logical approach.
2. Statistical approach.
3. Bionic approach.

The logical approach to making decisions in medicine is quite common, as it is a direct reflection of the doctor's considerations. The reasoning of the physician during the diagnostic process must be assured, consistent and evidence-based. That is, to determine a violation, which is characterized by a complex of symptoms, it is necessary to use the laws of formal logic. You can make a diagnosis using inductive reasoning associated with the prediction of the results of observations based on past experiences. But the inductive method in the diagnosis of the state, since the logic of knowledge and considerations is deductive in nature. Therefore, the deductive method is usually used in determining the diagnosis, that is, the process of deductive inference from a certain set of suspected diseases of the required diagnosis.

A variation of the logical approach is the method of decision trees or classification trees. To use this approach, the rule base of the form is composed: "If> Then" in the form of a hierarchical structure, which is a tree. To determine the class of the disease, it is necessary to answer the questions in the nodes of this tree, starting from its root, and thus make the transition to the next question. The positive aspects of this approach are the clarity of the method and clarity. But in practice, a number of questions arise that limit the use of this approach in solving the problem of classification. One of the problems associated with the choice of the next node. To solve it, various algorithms are used, which often give a too detailed tree structure and can lead to errors. Despite the various features of this approach, the logical scheme of questions and answers of the diagnostic process in the form of decision trees has been successfully applied in practice.

The group of statistical methods includes the Bayesian approach, methods of discriminant analysis, and conclusion based on precedents. The use of the Bayes theorem in determining the class of the disease is a common approach due to its simplicity, clarity and simple mathematical calculations, but it has a clear disadvantage - a large database of archival data is needed for the likelihood of a diagnosis to meet reality. Also, a negative point is the fact that a rare disease or non-standard symptoms are difficult to correctly identify.

Discriminant analysis is characterized by the presence of many calculations, the emergence of various kinds of connections between symptoms, the elucidation of the effect of the relationship of signs on the result of diagnosis, usually pulls in difficulties in solving a medical problem.

The case-based deduction method (Case Based Reasoning) is an approach based on the use of previous cases and experience. To make a decision, the search for similar precedents that took place in the past is first carried out, and then a measure of proximity between the new and all found cases for which solutions are determined is calculated. This approach has a number of limitations: a system based on precedents, built on knowledge, which is “drawn out” of experts, which means that there is a measure of subjectivism; knowledge gained from people needs to be assessed, verified and assessed for their credibility; lack of solving such problems for which there are no precedents; arbitrariness in the choice of a measure of proximity; a sharp decline in performance with a large set of input parameters.

The bionic approach is a process of artificially reproducing those structures and processes characteristic of the human brain. The advantages inherent in this approach are quite large: the ability to adapt (learning and self-learning) the parallelism of information processing; robustness (resistance to individual failures). Neural networks (NN) are used in solving problems of medical diagnostics, problems of classification (clustering), approximation, forecasting. Sometimes the applicability of the neural network approach is limited due to some drawbacks: a large amount of archival data is needed to train the created neural network; The factor of subjectivity in creating the NN structure is very important - the number of layers, the number of neurons in each layer and the activation function are selected by an expert, which, in turn, introduces additional uncertainty; Also, when training an NN, there is a possibility that the selected structure will not give adequate results when using some input data; Sufficient sensitivity to the input data - a training pattern on which NN learns must be properly prepared to eliminate the additional weight of individual input parameters, leading to unreliable learning outcomes.

Thus, the task of medical diagnostics is a rather complicated task due to the fact that patient data is poorly structured and have a different character. Some of the necessary information relating to the patient is usually missing, which introduces additional difficulties in the processing of medical data; some of the information is qualitative, since it is determined by the doctor, that is, there is a share of subjectivity; some of the information reflects the results of the analyzes, which means that it is necessary to take into account the randomness factor due to measurement errors. To date, there are several approaches to work with uncertainty in the tasks of medical diagnostics.

The probabilistic approach is an approach where unknown factors are statistically stable and therefore represent random variables or random events. In this case, all the necessary statistical characteristics must be determined: the laws of distribution and their parameters, functions, or probability density distributions, which, in turn, introduces additional difficulties.

Another approach to accounting for uncertainty is the use of the theory of fuzzy sets, which allows you to take into account the inaccessible or no data, or the data are of an exclusively qualitative nature. However, the application of this approach is associated with a number of difficulties, for

example, with determining the type of membership function, as well as difficulties with the subsequent processing of fuzzy data.

Resolution of the problem

The medical expert system of differential diagnostics is a system for determining diagnostic hypotheses based on the medical knowledge of a group of doctors and the facts of the patient's symptoms found. Diagnostic hypotheses are possible diseases (expert judgment) that a patient suffers from [Kuzomin and Vasylenko, 2010]. Based on diagnostic hypotheses, it is possible to determine the possible specialty of a doctor, according to which the patient should be treated [Kuzomin and Vasylenko, 2014].

The formal model of the system can be represented as a tuple:

$$MESDD = \langle WM, KB, UI, IE, EM, KA \rangle, \quad (1)$$

where WM - the working memory; KB - medical knowledge base; UI - user interface; IE - control output diagnostic solutions; EM - explanation of performance information; KA - the acquisition of medical knowledge.

In the mode of acquiring knowledge, expert doctors fill the system with medical knowledge, which allow it to independently solve the problems of finding a diagnostic solution in the mode of medical consultation. In the mode of medical consultation, the user-patient is involved in communicating with the system, who is interested in the effective diagnostic information and explanatory information of the result.

The key concept of the system is the knowledge base. For the presentation of knowledge in the system, a combination of frame and fuzzy knowledge bases was chosen. The frame knowledge base is presented to describe the current state of the field of diagnostics, that is, a quantitative assessment of each disease based on knowledge from the knowledge base and evidence of symptoms. A fuzzy knowledge base is presented to describe the dynamic known in the transitions between the states of the diagnostic area, that is, a cause-effect relationship that relates a disease to symptoms in its symptom complex. Using procedural knowledge and the inheritance of frame properties, you can implement a mechanism for controlling the output of a diagnostic solution.

Frame knowledge base can be represented as a tuple:

$$KB = \langle FC, FSM, FSD, FSS, FSC, \{FIM_i\}, \{FID_j\}, \{FIS_k\}, \{FIC_h\} \rangle \quad (2)$$

where FC - a frame class; FSM - specialty frame prototype; FSD - a prototype of the disease; FSS - symptom frame prototype; FSC - prototype frame of the symptom complex; $\{FIM_i\}$ - a set of frames-instances of specialties; $\{FID_j\}$- a set of frames-instances of diseases; $\{FIS_k\}$ - a set of frame-instances of symptoms FSC_h - a set of frame-symptom complexes.

Formally, a fuzzy rule can be represented as a tuple:

$$FR = \langle NFR, \{\langle FSMS_i, SF_i \rangle\} \rightarrow FSMD, CF \rangle, \quad (3)$$

where WFR is the name of a fuzzy rule; $FSMS_i$ is a fuzzy statement of a symptom variable; SF_i is the coefficient of symptom specificity in a symptom complex; $FSMD$ is a fuzzy statement of a variable disease; CF is the coefficient of confidence of the likelihood of the disease.

Formally, a fuzzy statement of a single variable can be represented as a tuple:

$$FSM = \langle LV, LT, M \rangle, \quad (4)$$

where LV is a linguistic variable; LT is the linguistic term of the variable M - modifier, which correspond to the words “very”, “more or less”, “no”, etc.

Formally, a linguistic variable can be represented as a tuple:

$$LV = \langle NLV, TSLV, ULV, GLV, MLV, TLV \rangle \quad (5)$$

where NLV is the name of the linguistic variable; $TSLV$ - term set of linguistic variable; ULV - the domain of definition of each $TSLV$ element; GLV - syntactic rules, often in the form of a formal grammar, generating the name of linguistic terms; MLV - semantic rules that define the membership functions of linguistic terms generated by the syntactic rules of GLV ; TV is a type of linguistic variable (symptom or disease).

Formally, the linguistic variable term can be represented as a tuple:

$$LT = \langle NLT, MF \rangle, \quad (6)$$

where NLT is the name of the linguistic term; MF - a variable membership function of a linguistic term. As the membership function, the following functions are used:

$$\mu_{LT}(u) = \frac{1}{1 + \left(\frac{u - b}{c} \right)^2}, \quad (7)$$

where b and c are the settings: b is the coordinate of the maximum of the function; c is the coefficient of concentration-stretching function.

In a system based on a combination of frame and fuzzy models, symptom slots are treated as weekends, and disease slots are targeted. When generating additional questions, procedures-methods are initiated that implement the opposite conclusion to determine the initial values of possible symptoms. When assigning the initial values of the slots, the demons procedures that are responsible for the direct inference and perform a fuzzy inference to obtain the target values of the disease slots will work.

To highlight the final diagnosis for the disease, 3 main criteria are used:

- the most minimal range of reliable solutions;
- the most maximum current integrated assessment of the disease;
- the maximum area of unreliable decision.

When implemented, the system checks the suitability of the fuzzy-production rules of each disease for each fact of symptoms in the working memory, if necessary, performs them and proceeds to the next disease, returning to the beginning when all diseases are exhausted. To ensure speed with a large knowledge base and a large number of facts in the working memory, the Rete algorithm is used. This algorithm sacrifices the amount of memory for speed, so the calculations must be carried out in cloud systems, it will simplify the load on the doctor's working machine.

When using the Rete algorithm, the medical knowledge base translates into a network of Rete (or a prefixal tree), in the end nodes of which there are, on the one hand, procedures-demons attached to output slots, and on the other, procedures-methods for obtaining values of target slots with the truth of the premise of fuzzy-production rules, information about which is stored in the intermediate nodes (α and β-memory). When symptoms enter the working memory, output slots are assigned a value, and only a small part of the network connected to it is updated.

At the time of assignment, not all rules are known under conditions of uncertainty. Therefore, it is impossible to build a single network for all the rules. Such a modification of the Rete algorithm is called the fast Rete algorithm.

The following components should be stored in the modified Rete network:

- an activation list in which parent slots are stored, i.e. slots of prototype frames;
- the context of the activity, which stores references to the current frames that caused the activation, that is, instance frames.

When changing the value of some source slot, it is in the premise that all the associated demons are activated, which directly try to calculate the value of the target slot in the conclusion. When calculating using the fuzzy inference algorithm, β-memory stores intermediate results, and the slots are used as α-memory. Unification of the rule with values in the working memory is not carried out as such, but is replaced by implicit inheritance unification, which is achieved by calling the daemon procedures of all parent frames with passing the current frame (triggering activation) as the call context. Thus, the network is implicitly formed by demons attached to the slots, rules associated with them, and a fuzzy conclusion, in the nodes of which intermediate results of calculations are stored.

Comparison of the symptom complexes with the available facts from the working memory is carried out after the approval of the incoming symptoms. As a result, at the stage a conflict set consists of potential diseases according to the following criteria:

- a potential disease corresponds to a symptom complex, in which the symptoms coincide with the symptoms entering the working memory;
- a potential disease corresponds to the age and sex of the patient, possibly suffering from this disease.

In case all the symptoms of a disease are present in the working memory, then such a disease will occur in a variety of diseases under consideration.

Conflict resolution is performed to select one or more of the most appropriate diseases from the conflict set. The result of this phase is the set of active diseases and determines the order of their implementation.

Conflict resolution is based on the principle “first come, first served”, i.e. priority over the choice of the first of the active diseases to trigger. In addition, the following criteria are used to select appropriate diseases:

- novelty. Active diseases correspond to the manifestations of symptoms that enter the working memory as the most. For this, it is necessary to provide the facts with a special attribute of time spawning;
- specificity. Active disease corresponds to a symptom complex with many facts of manifestations coming into the working memory of symptoms.

You can select a conflict resolution criterion or define a queue of several criteria. In addition, worn fuzzy-production rules should not be applied to existing facts.

To improve the effectiveness of teaching a fuzzy-production model of knowledge representation, it solves the problem of forming a knowledge base, the use of a genetic algorithm is proposed. To do this, the first is to put a method of encoding / decoding a fuzzy-production knowledge representation model, which defines some parameters (membership function parameters; coefficients of specificity and confidence), which are combined into a single vector. The value of one parameter lies in a certain neighborhood, which can be divided into 2-16 intervals. Then, to encode the slot number, you can use a 16-bit value in the gray code, in which the neighboring numbers have fewer positions.

To create the initial chromosome population, 100 chromosomes are randomly generated from the initial initialization of gene values in each neighborhood using the Gaussian method. Then, using the composition operation, combine a set of genes into a single chromosome to assess the fitness of chromosomes.

Each chromosome from the population is mapped to an assessment of its fitness of chromosomes in the population, the calculation of which is performed on the basis of training samples and vectors of model parameters. The learning process is considered complete if the condition that the resulting estimate is greater than the threshold value is satisfied. Selection of chromosomes. In a selection procedure based on the principle of a roulette wheel, the larger the sector on the roulette wheel (that is, the corresponding assessment of chromosome fitness), the higher the chance that this particular chromosome is chosen, which later on after performing the decomposition operation, genetic operators are used to create populations.

Application of genetic operators to chromosomes. In genetic algorithms, the crossover operator (90% probability) is responsible for the transfer of parents' genes to descendants. The mutation and inversion operators (probability 10%) are designed to maintain the diversity of chromosomes in a population.

The formation of a new population. Productive chromosomes must be placed in the population after performing the composition operation. To reduce the population to the original number of chromosomes, a reduction operator is used.After stopping the work of the genetic algorithm, a trained model comes out, approximating data from a training sample with a given accuracy and forms a knowledge base consisting of a system of fuzzy-production rules.

Creating a knowledge base.

At (Fig. 1)–Disease stores general information about acute renusitis (GDS).

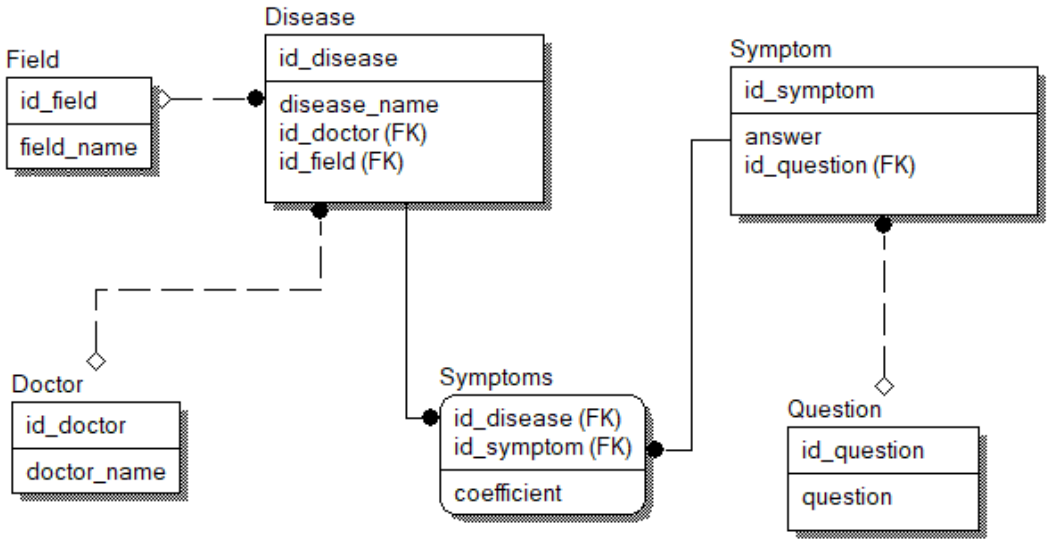

Figure 1–Knowledge Base MES

Field is a grouping of diseases by region in accordance with the International Classification of Diseases (ISS). Symptom (Fig. 1) stores the symptoms, the separate table contains the questions that the MES forms during the analysis of fuzzy knowledge. For example, the question “The

nature of nasal breathing" and in the Symptom table are stored the diagnoses referring to this question with the answers "free", "moderately difficult", "difficult", which are stored in the answer field. Each option is treated as a separate diagnosis. If the question can be answered yes / no, then the answer is stored null. The Symptoms table connects the symptoms with the diagnosis, there is a coefficient field that takes into account the weights — an estimate for different diagnoses; if one is more important than the other, put more weight.

Conclusions

The aim of the work is to begin research for possible improvements of the Rete algorithm for medical expert system.

During the work was implemented modification for Rete algorithm. The key idea of the work is that it can be used as for daily routine operations in every medical institution. Unfortunately, we do not have enough time to present them in the work.

In the work, we finished only theoretical part of our goal. Right now, we implemented this modification with the help of Java language code and Amazon Cloud Services. The next step will be to implement this modification to newer variants of Rete algorithm and to find similar algorithms which can be modified in this way also.

Bibliography

[Jackson, 1990] Jackson Peter. Introduction to Expert Systems. Addison-Wesley Pub, Vol.2, 1990.

[Giarratano, 1994] Giarratano Joseph, Riley Gary. Expert Systems: Principles and Programming. PWS Publishing Co, Vol.2, 1994.

[Forgy, 1982] Forgy Charles. Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem. Artificial Intelligence, pp 17-37, 1982.Authors' Information

[Kuzomin and Vasylenko, 2014] Kuzomin, O.Ya., Vasylenko, O., Obespechenie bezopasnosti ispolzovaniia baz dannyh v usloviiah chrezvychainyh situazii. International Journal "Information Technologies Knowledge", Vol. 8, Num. 2. 2014. pp. 173-187.

[Kuzomin and Vasylenko, 2010] Kuzomin, O.Ya., Vasylenko, O., Analiz estestvenno iazykovyh obiektov I predstavlenie znanii. Vostochno-Evropeiskii zhurnal peredovyh technologii, Vol. 6/2(48). 2010.

Authors' Information

Oleksii Vasylenko *– Aspirant of Kharkiv National University of Radioelectronics; Kharkiv, Ukraine;*

e-mail*: ichbierste@gmail.com tel.: +380 63 841 66 23*

Major Fields of Scientific Research*: General theoretical information research, Knowledge Discovery and Engineering, Business Informatics.*

Prof. Dr.-hab. Oleksandr Kuzomin *– Informatics chair of Kharkiv National University of Radio Electronics; Kharkiv, Ukraine Ukraine;*

e-mail*: kuzy@daad-alumni.de tel.: +38(057)7021515*

Major Fields of Scientific Research*: General theoretical information research, Decision Making, Emergency Prevention, Data Mining, Business Informatics.*

Bohdan Maliar *– Master student in Information and Communication technologies of Kharkiv National University of Radio Electronics; Kharkiv, Ukraine Ukraine;*

e-mail*: bohdan.maliar@gmail.com*

Major Fields of Scientific Research*: Big Data, Data Mining, Data Analyses.*