Using Nosql Databases and Machine Learning for Implementation of Intelligent Decision System in Complex Vision Patalogies

Eremeev A.P., Ivliev S.A., Vagin V.N.
National Research University "Moscow Power Engineering Institute" (NRU "MPEI")
Moscow, Russia
eremeev@appmat.ru, siriusfrk@gmail.com, vagin@appmat.ru

***Abstract*—One of the most important fields of decision support system development is processing medical data for helping experts to make decision in the case of complex pathologies. In generally, a system for storing data and a decision module is main parts of these systems, what that is the reason why is very important to create systems, which can handle medical and expert information, that can be presented in various types and forms. One of the decision in this case is combining methods of machine learning and NoSQL databases.**

Keywords—NoSQL; machine learning; decision systems; artificial intellegence; ontology

I. INTRODUCTION

The work in a heterogeneous environment is a frequent phenomenon during the process of complex systems' development. The reason for this is constant interaction among the developers and the experts. Moreover, the subject area itself vary as a result of both technical and scientific progress or the replacement of equipment, that can lead to a change in data format with which the system works. It is also worth to remember that in different medical institutions, various reporting formats can be adopted as well as to different levels of personnel qualification and other features. As a result, a situation may arise when the developed system will not have the flexibility to work with all the new changes made in this institution. One method of solving this problem is a development of a more abstract ontology of the subject area description, however it increases the overheads and complicates system modifications.

It is worth noting that since about 2007, when the volume of data and its complexity was quite high, the extensive development of data storage technology in non-relational databases (NRDBs) of the type NoSQL DBs has begun. These DBs provide greater flexibility when working in complex environments and ease of modification at the lowest level– storage level [1, 2].

Another important task is to process incoming data. Due to the fact that data can be presented in a different form, various methods and algorithms can be applied, that involve data mining, its analysis and knowledge discovery. In order to work effectively in this case, one has to storage both data and its ontologies, this process can be performed through a NRDBs implementation.

Neural networks [4], genetic algorithms, SVM (Support Vector Machine) algorithm, and other learning methods of Bottom-Up AI (Artificial Intelligence) approach can be used for data processing as well as Top-Down AI methods, such as a method of Bayesian networks and the Dempster-Shafer method [3].

II. DATABASES FOR MEDICAL DATA STORAGE

The storage system is required for any system related to the data [5]. Due to the fact that the designed expert system is a Decision Support System (DSS) in problem situations, we should expect the presence of high heterogeneity in the data submitted. In this regard, the comparison of conventional relational and non-relational model of data storage is the following.

As an example, one can use medical reports, containing the following information:

1) the results of specialists' examinations, written in the natural language;
2) the results of the analyses that can be:
- data series;
- tuples (name, value);
- photographs;

3) the results of the surveys, written questionnaires;
4) prescribed drugs and procedures.

In addition to the data distribution over time (that is always difficult to proceed in conventional (relational) databases (RDBs) and requires the implementation of additional structures to work with temporal data [6,7]), one can see that there is data represented in the natural language, in the form of time series and images. On the other hand, in medical reports there are knowledge that may have been obtained in the framework of another model.

In case the chosen implementation of storage is RDBs, you must implement a set of tables and relationships between them. And the closer a data model for a storage layer, the more it

978-1-5386-7531-1/18/$31.00 ©2018 IEEE

loses the flexibility. Moreover, the further the data model, the more complicated is the work on the presentation layer.

This can be illustrated by the following examples:

1. Let the DB contain the following table names: “Patients”, “Tests”, “Medical tests results”. The first table stores general information about the patients, in the second one there are interactions between patients and test results, and information about the medical tests (time, location, type), the third table stores the actual research data. We can immediately notice that the third table will be overwhelmed with data, regardless of the method of placing it. In addition to the degradation of a work speed because of the large number of entries, the creation of intermediate tables and entities for continuous data mining is required, what will lead, in turn, to the slow work speed of the system as a whole.
2. Let the DBs store the knowledge gained as a result of processing data that was presented, id est we have the knowledge base (KB). However, in this case, the opportunity to re-explore the data when you receive any new information is lost, since actual data are not processed for a permanent use.
3. Let the DBs store both data and derived knowledge. Then the number of tables will grow with the growth of the new knowledge forms. The quality of access to raw data will remain low, since there will be proceeded a division of the primordial essence (of a medical test) on the artificial sub-entities.

In addition to described complexities in the examples, it is also worth to highlight the difficulties with scalability, portability and processing of large amounts of data in a conventional RDBs [1,8].

As an alternative to the traditional RDBs, it is suggested to use NoSQL DBs. Unlike ACID (Atomicity, Consistency, Isolation, Durability) approach, NoSQL DBs use the BASE concept [9]:

- BA (Basic Availability) – basic availability, i.e. any query to the DB is completed (successfully or unsuccessfully) without errors;
- S (Soft state) –the system state can change over time even without introducing new data to achieve data consistency in the DB.
- E (Eventual consistency) – the data can be inconsistent for some time, but eventually they come to a consistent state.

Even these "physical" properties of the NoSQL DBs allow us to understand that they can be used for solving problems, connected with the heterogeneous environment, since they do not require the storage only of crisp date schemas.

The most important property of NoSQL DBs is the ability to store documents as a single entity that allows you to effectively organize different types of data processing and search in knowledge (for example, using the MapReduce algorithm [10]), as well as to create new forms of knowledge storage and retrieve data by supplementing the existing documents and establishing links between them.

Thus, each item of history can be interpreted as some kind of a document. One can also form new derivative documents, conduct sampling among them in order to select those which best correspond with the results obtained in future data, i.e., to form certain ontology.

III. THE ONTOLOGY REPRESENTATION OF A MEDICAL REPORT IN A NON-RELATIONAL DATABASE

In [2] there is an example of using ontologies for the representation of genomic data for storage in the NoSQL DBs for example, graph-oriented DBs. In our case, it is proposed to use a document DBs and to build an ontology. As the basic ontology, one can consider the following.

Concepts (Classes):

- *Observation (O)* is some type of structed or expressed based on a structured or unstructured document. The observation can contain different instances, such as "a study of ERG (Electroretinogram)", "retinal photography", etc.;
- *appointment (N)* is a final result issued by the expert based on the consideration of data in one or more observations. The appointment may contain different instances, such as "prescriptions", "physician from another doctor", etc.;
- *knowledge (Z)* is a result of learning from data by automatic methods or by forming with the help of questionnaires and other methods for directly obtaining knowledge from experts. Knowledge consists of instances of different kinds, for example "the ophthalmologist questionnaire", "questionnaire of neurologist", "the result of the analysis by the neural network," "the inference results by the Dempster-Shafer method", etc.
- *document generation (DR)* is a document generated in the application to the generation of new documents and associated with one of the relations of generation.

Relationships:

- *the generation of* (R_{gen}) is a generation of *Z* from the sets *O, Z, N*;
- *generalization* (R_{int}) is a generalization of *Z*;
- *conclusion* (R_{sum}) is a receiving (reasoning) of *N* from the set of *Z* on the basis of an expert opinion.

This ontology allows to describe the following process of introducing a new appointment *N* available according to the following formula:

$$Z(\tau+1) = P\iota\nu\tau(P\gamma\varepsilon\nu((O(\tau)\times N(\tau)\times Z(\tau))2)$$
$$N(\tau+1) = P\sigma\upsilon\mu(Z(\tau+1))$$

where:

- *N(t+1)* is a new conclusion;

- $Z(t+1)$ is a generalization of all knowledge at the time of building an output;
- $O(t)$ is the set of all surveys at the time t;
- $N(t)$ is the set of all assignments at the time t;
- $Z(t)$ is the set of all knowledge at the time t;
- $(O(t)\times N(t)\times Z(t))^2$ is the set of all subsets of the Cartesian product received at the time t of examinations, assignments and knowledge, i.e., the original data for the re-operation from data and knowledge obtained at the given moment.

This ontology is a simplified basic model for storing knowledge. Since each generation occurs through the implementation of methods and algorithms embodied in the relations of the generation R_{gen}, then such operation is generated by some document DR, what allows to proceed a retrospective study of conclusion sequences.

It should be noted that all concepts can be stored in NoSQL DBs in an open and extensible forms both formalized or not. Also these DBs are "open" to populate new concepts.

In the conclusion, we can tell that using the ontology allows to formalize the process of working with NoSQL DBs, expand possibilities of applications in DSS for complex problem situations, as well as the opportunities for NoSQL DBs to integrate ontology with DB, making it expandable at the lowest level.

Such balanced performance opens great opportunities for data processing through machine learning.

IV. Machine learning methods for data mining and knovledge extraction

The proposed model allows to integrate different models of data mining and to obtain final solutions described in [3,4,11]. At the same time for various observations, one can use different methods of knowledge acquiring.

In particular, during the analysis of electroretinogram (ERG) [4], a neural network (NN) can be used, that solves the classification problem formally presented in the form of:

RD (classification based on NN) ← (O(ERG), R_{gen}, Z(a possible disease based on NN)).

The classification problem can be used on the basis of simulation models (SM), described in [11]:

RD(classification simulation model) ← *(O(ERG),* R_{gen}*, Z(a possible disease based on SM))*.

In the formation of questionnaires that allow to gain knowledge (Z), the generation of new knowledge that will be better and understandable to an expert [3] can be expected:

RD (output-based on the Dempster-Shafer method) ← *(Z(questionnaire 1), Z(questionnaire 2)), Rint, Z(possible disease questionnaire))*.

The formation of general conclusions can be made based on the integrated output:

RD (the final output appointment) ← *Z(a possible disease based on NN), Z(possible disease-based on SM), Z(a possible disease based on the Dempster-Shafer method), N* R_{sum}*(appointments))*.

As can be seen from the dependencies, they do not take into account the possibility of using data obtained in time, and change their dynamics as the result of certain assignments. For these purposes, the forecasting methods can be applied, in particular:

- based on neural networks when a neural network simply analyzes N-parameters [12];
- based on the clustering and feature extraction (SVM, random forest [13], etc.).

It should be noted that the model does not exclude, but even promotes the use of methods of semantic search and analysis among the doctors (experts, decision makers) that will allow to allocate additional knowledge [14]).

V. Practical realization

Prototype of knowledge aggregation and research system was developed by using proposed model and framework. Methods of knowledge extraction of this system was described previously in this article.

System based on Flask framework, No-SQL DB MongoDB, Python programming language version 3.0, and numerous python's packages, such as numpy, scipy and others. The main part of system contains next functionality:

1. Extracting data from storage files of biophysical researches, such as EDF (European Data Format) or simply csv file, in case of therapy vision pathologies.
2. Adding some additional information for this data, which connected with process of research, such as patient name, stimulation type etc.
3. Storage data on natural language (for example examination results in free form) for next semantic search.
4. Process data using methods, described in [4].
5. Generation summary of results of processing.

Next major points for developing system is:

1. Realization of methods, which are described in [3, 11]
2. Using method of semantic search for extraction knowledge from records on natural language.

It is worth nothing of using No-SQL DB and described model allowed develop wide-range system, which can work not only with eye pathologies, but also it was tested on ECG data with methods was described in [4].

VI. Conclusion

This paper presents a model based on NoSQL database that can store and process data by machine learning methods obtained during medical observations. The difference from conventional data representation in a traditional RDB, the

storage capabilities of the ontologies together with the data themselves, as well as the flexibility of the proposed model are shown. The description of data processing methods is given, the possibility of their transformation into knowledge is shown, as well as the analysis methods and their combinations (integration) with the aim of obtaining the final conclusions are produced. The research is conducted at the Department of applied mathematics, National Research University "Moscow power engineering Institute" together with the Moscow Helmholtz research Institute of eye diseases for creating an intelligent DSS under the diagnosis of complex pathologies [3,4].

REFERENCES

[1] Zohreh Goli-malekabady, Mohammad kazem Akbari-fatidahi, Morteza Sargozaei-javan. An effective model for store and retrieve big health data in cloud computing // Computer Methods and Programs in Biomedicine, Volume 132, August 2016. — Pp. 75-82.

[2] Naresh Kumar Gundla, Zhengxin Chen. Creating NoSQL Biological Databases with Ontologies for Query Relaxation // Procedia Computer Science, Volume 91, 2016. — Pp. 460-469.

[3] Aleksandr Eremeev, Ruslan Khasiev, Irina Tcapenko, Marina Zueva. The Intelligent Decision Support System for Diagnostics of Difficult Diseases of Vision // International Journal “Information Content & Processing”, Volume 1, Number 3, 2014. — Pp. 269-279.

[4] A.P. Eremeev, S.A. Ivliev. Analysis and Diagnosis of Complex Pathologies of Vision Based on Wavelet Transformations and the Neural Network Approach // Collection of Scientific Works of the VIII International Scientific and Technical Conference "Integrated Models and Soft Computing in Artificial Intelligence" (Kolomna, May 18-20, 2015). T. 2. — Moscow: Fizmatlit, 2015. — Pp. 589-595 (in Russian).

[5] Gary D. Riley; Joseph C. Giarratano Expert Systems: Principles and Programming, Fourth Edition. Publisher: Course Technology, 2004.

[6] A.P. Eremeev, E.I. Kurilenko, P.R. Varshavskiy. Temporal Case-Based Reasoning Systems for Automatic Parking Complex. // Proc. of the 17th International Con-ference on Intelligent Systems and Tecnologies, ICIST 2015), May 25 - 31, 2015, Tokyo, Japan. International Science Index, eISSN: 1307-6892. — Pp. 3129-3136.

[7] A.P. Eremeev. The Logic of Branching Time and Its Application in Intellectual Decision Support Systems // Sb. Tr. The 10th National. Conf. On artificial intelligence with the international membership. KII-2006. In 3 volumes, T.3. - Moscow: Fizmatlit, 2006. — Pp. 746-754 (in Russian).

[8] Ken Ka-Yin Leea, Wai-Choi Tangb,1, Kup-Sze Choia. Alternatives to relational database: Comparison of NoSQL and XML approaches for clinical data storage // Computer methods and programs in biomedicine, 2013. — Pp. 99-109.

[9] I.A. Kozlov. Analysis and classification of non-relational databases // Youth Scientific and Technical Herald, No. 02, February 2013. — Pp. 1-23 (in Russian).

[10] Ralf Lämmel. Google’s MapReduce programming model — Revisited // Science of Computer Programming, Volume 70, Issue 1, 1 January 2008. — Pp. 1-30.

[11] D.N. Anisimov, D.V. Vershinin, O.S. Kolosov, M.V. Zueva, I.V. Tsapenko IV Diagnostics of the current state of dynamic objects and systems of complex structure by fuzzy logic methods using imitation models // Artificial intelligence and decision-making. — 2012. — No. 3. — Pp. 39-50 (in Russian).

[12] L.N. Yasnitsky, A.A. Dumler, K.V. Bogdanov, A.N. Poleshchuk, F.M. Cherepanov, T.V. Makurina, S.V. Chugainov. Diagnostics and prognostication of cardiovascular system diseases on the basis of neural networks / / Medical technology. — 2013. — No 3. — Pp. 42-44 (in Russian).

[13] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda. Machine Learning and Data Mining Methods in Diabetes Research // Computational and Structural Biotechnology Journal, Volume 15, 2017. — Pp. 104-116.

[14] Jay Urbain. Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models // Journal of Biomedical Informatics, Volume 58, Supplement, December 2015. — Pp. S143-S149.