Using NoSQL Databases and Machine Learning for Implementation of Intelligent Decision System in Complex Vision Pathologies

-Eremeev A.P., Ivliev S.A., Vagin V.N.

National Research University “Moscow Power Engineering Institute” (NRU “MPEI”)

Moscow, Russia

Arghyadeep Das

1711072

SY COMPS B

nosql maketable < research.tpl

```
< stdout
>Abstract
>Introduction
>Databases for medical data storage
>Ontology representation of medical reports in NRDBs
>Machine Learning for data mining and knowledge
 extraction
>Practical Realization
>Conclusion
>Bibliography
```

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 intelligence; ontology

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.

INTRODUCTION

- 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.
- 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.
- Neural networks, 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.

DATABASES FOR MEDICAL DATA STORAGE

- The storage system is required for any system related to the data. 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 specialist's examinations, written in the natural language

2) the results of the analyses that can be:

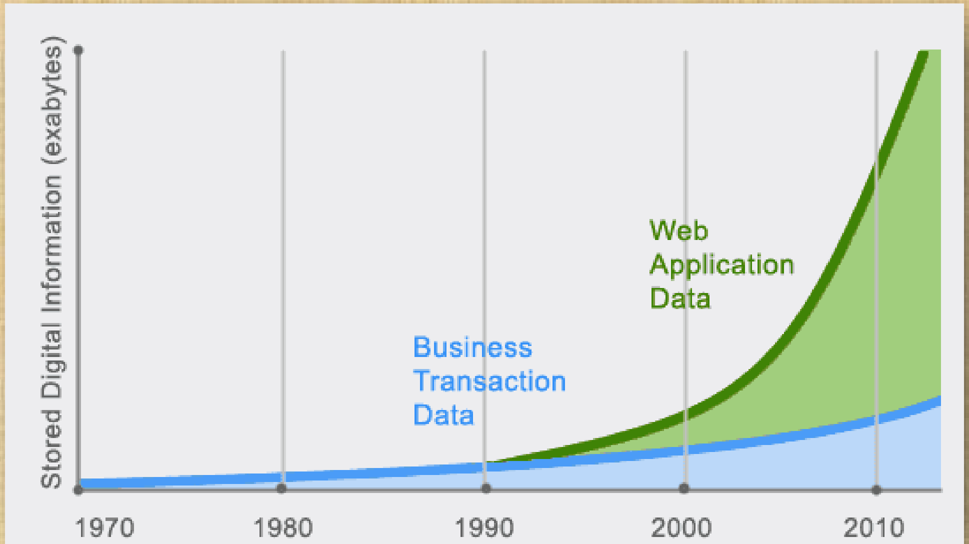
   a. data series

   b. tuples (name, value)

   c. photographs

3) the results of the surveys, written questionnaires

4) prescribed drugs and procedures



DATABASES FOR MEDICAL DATA STORAGE

- The data will not be only textual but can also be images, documents, etc. The scans of various body parts are huge in size and take a lot of space. The retrieval becomes an issue in RDBs.
- 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 loses the flexibility.
- Moreover, the further the data model, the more complicated is the work on the presentation layer.

DATABASES FOR MEDICAL DATA STORAGE

- Let’s understand with the help of an example:

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, i.e. 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.

DATABASES FOR MEDICAL DATA STORAGE

- In addition to previously mentioned complexities, it is also worthy to note the difficulties of scalability, portability and processing of huge chunks of data in traditional RDBs.
- So, it is proposed to use NoSQL DBs as an alternative to traditional RDBs.
- Unlike ACID(**A**tomicity, **C**onsistency, **I**solation and **D**urability), NoSQL DBs use the BASE(**Ba**sic availability, **S**oft State, E(**E**ventual consistency) approach.

Here is an excellent article on understanding difference between ACID and BASE.

DATABASES FOR MEDICAL DATA STORAGE

- 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), 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.

THE ONTOLOGY REPRESENTATION OF A MEDICAL REPORT IN NRDBs

- The ontology proposed for our use case:

Concepts (Classes):

- ***Observation (O)*** is some type of structure or expression 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.

THE ONTOLOGY REPRESENTATION OF A MEDICAL REPORT IN NRDBs

- *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:

$$Z(\tau+1) = P_1 V\tau(P\gamma\varepsilon\tau\big(O(\tau)\times N(\tau)\times Z(\tau)\big)^{\,2})$$

$$N(\tau+1) = P\sigma\vartheta\mu(Z(\tau+1))$$

THE ONTOLOGY REPRESENTATION OF A MEDICAL REPORT IN NRDBs

where,

i. $N(\tau + 1)$ is a new conclusion;

ii. $Z(\tau + 1)$ is a generalization of all knowledge at the time of building an output;

iii. O(τ) is the set of all surveys at time τ;

iv. N(τ) is the set of all assignments at time τ;

v. Z(τ) is the set of all knowledge at time τ;

vi. $\left(O(\tau) \times N(\tau) \times Z(\tau)\right)^{2}$ is the set of all subsets of Cartesian product received at time τ of examinations, assignments and knowledge, i.e. the original data for the re-operation from data and knowledge obtained at the given moment.

MACHINE LEARNING FOR DATA MINING AND KNOWLEDGE EXTRACTION

- We will understand the implementation of machine learning with the help of an example. During analysis of electroretinogram(ERG), we can use a neural network for classification problem.

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):

RD(classification simulation model) ← *(O(ERG),* $\mathrm{R}_{\mathrm{gen}}$*, 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 can be expected:

RD (output-based on the Dempster-Shafer method) ←*(Z(questionnaire 1), Z(questionnaire 2)),* $\mathrm{R}_{\mathrm{int}}$*, 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* $\mathrm{R}_{\mathrm{sum}}$ *(appointments)).*

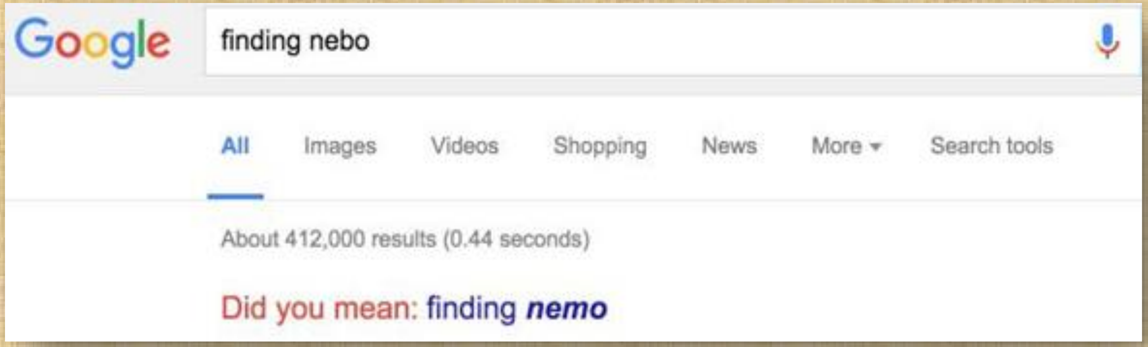
MACHINE LEARNING FOR DATA MINING AND KNOWLEDGE EXTRACTION

- 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:

a. based on neural networks when a neural network simply analyzes N-parameters;

b. based on the clustering and feature extraction (SVM, random forest, etc.);

c. semantic search.



Google finding nebo

All Images Videos Shopping News More ▾ Search tools

About 412,000 results (0.44 seconds)

Did you mean: finding *nemo*

PRACTICAL REALIZATION

- With all that’s proposed, now it’s time to practically realize it, or rather, implement it. A system based on Flask framework, MongoDB, Python3 and numerous Python packages such as NumPy, SciPy and others. The main part of system contains next functionality:

i. Extracting data from storage files of biophysical researches, such as EDF (European Data Format) or simply csv file, in case of therapy vision pathologies.

ii. Adding some additional information for this data, which connected with process of research, such as patient name, stimulation type etc.

iii. Storing data on natural language (for example examination results in free form) for next semantic search.

iv. Process data using methods.

v. Generation summary of results of processing.

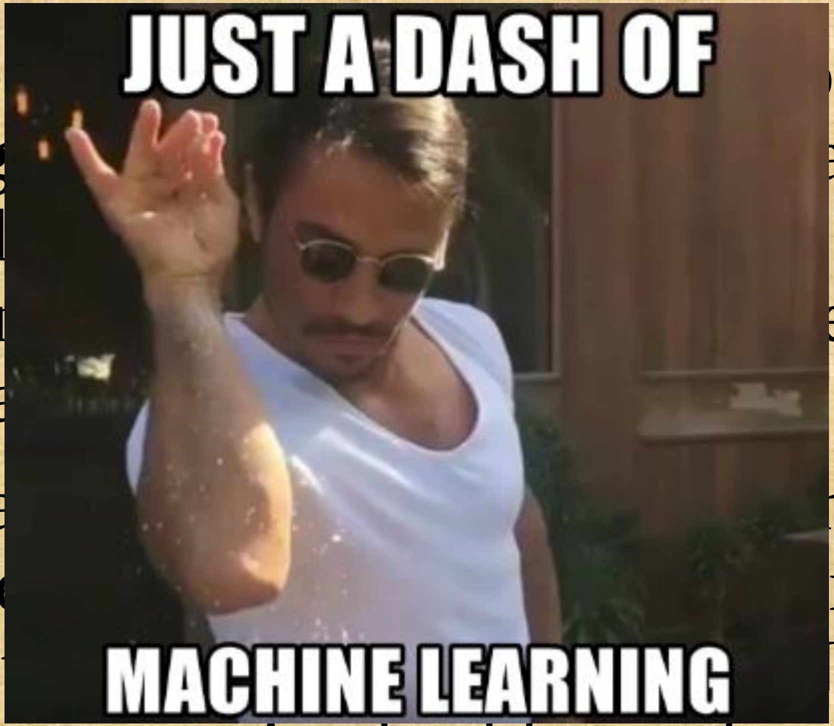
vi. Using method of semantic search for extraction knowledge from records on natural language.

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.

MY TAKEAWAYS FROM THIS PAPER

- I am a machine & de practitioner since my freshman year of eng arious algorithms in in machine learning l MapReduce and also done a bit of study or emantic Search, LDA, word2vec, lemmatiza
- This paper gave me a nine Learning algorithms can be use n healthcare domain. It gives a detailed on RBs to tackle the current issues faced in managing healthcare data when juxtaposed with a bit of Machine Learning.



BIBLIOGRAPHY

- Original paper: https://ieeexplore.ieee.org/document/8482230
- https://towardsdatascience.com/
- www.tutorialspoint.com
- https://www.couchbase.com/solutions/nosql-for-healthcare

Thank You!