Extraction of structured data from unstructured medical records using text data mining technologies: process automation

I V Moskalev¹, O S Krotova¹, L A Khvorova¹ and D G Bobkova
Altai State University, 61 Lenin Avenue, Barnaul, 656049, Russia
E-mail: moskalev.igor.v@gmail.com, khvorovala@gmail.com

Abstract. The paper discusses technologies for processing text-based medical data stored in the Microsoft Word text format. Processing such data is aimed at data mining the text for new, potentially useful knowledge that can later be used to study various diseases and to form a personalized approach to diagnosis and treatment. During the study, 3244 depersonalized medical records of children and adolescents in Altai Krai suffering from diabetes mellitus were processed. Information is stored in the records in both structured and unstructured forms. Most of the valuable data, such as the dynamics of the disease course, patient complaints, patient’s life history, etc. are kept in natural language. The difficulty of processing text medical records is associated with a great number of abbreviations, synonyms and misprints, which makes it impossible to use a unified template. Therefore, the study is aimed at minimizing information losses while extracting knowledge by means of applying various text data mining methods. The practical outcome of this study is a database containing a large amount of valuable information on diabetes mellitus, various types of its clinical course and complications. The obtained data will be further used to build mining models for diagnosing and predicting the disease and its complications. To reach the goal of the research, we used the PostgreSQL DBMS and modern linguistically oriented software created within the framework of the Python programming language and its libraries: python-docx, natasha, Natural Language Toolkit (NLTK).

1. Introduction
In recent years, electronic data management systems have been widely applied in medicine. With the help of such systems, healthcare institutions have accumulated a large amount of data, a significant part of which is represented by natural language texts [1]. The purpose of this work is to create a structured data set from unstructured texts [2] in the medical records of children and adolescents with diabetes mellitus and to data mine for additional information in order to solve the task of predicting complications.

The authors see the novelty, relevance and practical significance of this work in obtaining valuable and necessary data that can be collected from the case records of children and adolescents suffering from diabetes mellitus with the purpose of further predicting any possible disease complications.

When working with machine learning algorithms, as a rule, large volumes of heterogeneous data are used. Therefore, one of the tasks that scientists in the field seek to solve is the implementation of a convenient way to store upload and download information. A handy tool to solve this problem is relational databases.
We used anonymized medical case records of children and adolescents with diabetes to create the database. A total of 3244 documents containing data on the hospital care of 865 patients within the period of 2011 to 2018 were processed.

A medical case record is an electronic document drawn up in the Microsoft Word text editor. Part of the data is the information on insulin therapy and test results stored in a chart form [3], another part is in the form of a doctor’s notes in natural language. This document is filled in manually by the doctor and, in the case of diabetes mellitus patients, has the following sections: patient information, information on the diagnosis and complications, medical case and patient’s life history, physical examination, diagnostic results, given treatment, results and recommendations in natural language. Such records contain a significant amount of information and carry a great semantic meaning [4]. This fact and a large number of documents make the task of obtaining structured data inaccessible without pre-processing [5].

Thus, to form a data set from medical case records it is necessary to solve the following tasks:

- extraction of numeric and line data from charts;
- identification of the medical concept references in the text (diseases, symptoms, medications, medical procedures);
- extraction of numerical characteristics from the text.

2. Database

We chose a free and open source PostgreSQL as a database management system (DBMS) [6]. The database contains 15 tables; figure 1 displays the tables chart.

According to data ownership, the tables are divided into 2 types:

- general patient data, which do not depend on the period of treatment (gender, the disease onset date, patient’s life history data);
- data related to the period of hospital stay (laboratory test results, physical examination, drug treatment, etc.).

The patient identifier is used to link the tables of the first type, and the record identifier is used for the tables of the second type.

3. Record data processing

The quality of the models is greatly influenced by the attributes that are used for their learning, that is why the most important task is the quality origination and transformation of the source data. The first phase of data origination is the exclusion of non-informative parameters. Classical methods for attribute selection are statistical data analysis and the application of sequential attribute selection algorithms.

The records data processing was carried out with the use of the Python high-level programming language [7]. We used the python-docx library to work with docx-documents. The advantages of this library are a convenient object representation of the document and comprehensive documentation.

All the textual data were pre-processed with the purpose of increasing their informational value and rigorizing. The following transformations were performed to prepare the texts for data mining:

- tokenization;
- conversion to lowercase;
- stop-words deletion;
- lemmatization.

We used the Natural Language Toolkit (NLTK) library for basic text transformations (tokenization, stop-words deletion) [8]. Lemmatization was carried out using the pymorothy2 library. These operations allowed reducing the total number of tokens from 4353462 to 2700843, and the number of unique tokens from 26920 to 16937. Figure 2 displays the frequency distribution of 30 most used Russian words from anonymized medical case records of children and adolescents of the Altai krai (Russia) with diabetes before and after pre-processing.
Each record has a number of tables characterized by different header structures. Based on the statistical analysis of the column heads, 6 main types of tables containing the most valuable data were identified. First of all, the table handling module determines which type this or that table belongs to. Since there is no strict template for filling out records, the structure of tables belonging to the same type may vary from document to document. Therefore, at the next stage, the matching search mechanism between the table columns in the document and the database fields was implemented. Then the interline data transfer from the record to the database is performed. At this stage, the basic processing of the text values in the source table is also carried out.
Using the NLTK library [8], we obtained lists of word combinations, grouped according to the sections of the record where they are encountered. Experts in the field identified the entities set, singled out their synonyms and abbreviations used to indicate them. This allowed creating a dictionary of concepts, which serves as a basis to form templates applied to search for these concepts in text documents.

The extraction of medical concepts was carried out involving the search for the template values of the concepts in the corresponding sections of the document. Levenshtein distance [9] was used as a criterion for evaluating the similarity of the identified elements.

Medical texts contain a large amount of relevant quantitative information. Examples of such data are the quantitative characteristics of diagnostic results. In order to be able to analyze these data, it is necessary to present them in a structured form. In a general case, quantitative data in the text can be represented by a set of the following elements: object name, attribute name, quantitative value, unit of measurement, value modifier. A text block containing these elements is called a quantitative group.
(the presence of all the elements from the list is not necessary) [10]. To implement the task of the quantitative data identification and mining from the medical record texts, we used the Natasha library where context-free grammars for the description of constructions requiring extraction were compiled. Based on these grammars, the parser identifies the entities and their corresponding values in the text.

4. Results and discussion
Due to the medicine digitalization and the introduction of digital technologies in the healthcare sector, medical institutions of the state healthcare system have accumulated a large amount of the electronic medical data relevant to statistics, modeling and prediction. A significant part of these data is stored in the form of natural language texts.

The medical texts set – medical records – falls into the Big Data category, which the tools and methods for processing structured and unstructured data of significant volumes were applied to in order to obtain important results. To achieve the objectives of the study, the authors developed a set of software modules that process medical information stored in the Microsoft Word text format and contained in medical case records. The software package was tested and endorsed on anonymized data (3244 anonymized medical records) of children and adolescents in Altai krai suffering from diabetes mellitus. The difficulty of processing text medical records was associated with a great number of abbreviations, synonyms and misprints, which made it impossible to use a unified template. Therefore, the study was aimed at minimizing information losses while extracting knowledge by means of applying various text data mining methods.

The practical outcome of this study is a database of medical case records of children and adolescents suffering from diabetes mellitus, which contains a large amount of valuable information on diabetes mellitus, various types of its clinical course and complications. For the database development, we used the PostgreSQL DBMS. We applied the text processing tools and methods that enabled extracting knowledge from unstructured data in medical records and presenting them in a structured form using modern linguistically oriented software created in the framework of the Python programming language and its libraries: python-docx, natasha, Natural Language Toolkit (NLTK). A total of 38 attributes characterizing the laboratory test results were obtained from the tables of medical records. From the text records, we managed to obtain additional 41 attributes of the patient’s medical case history, self-monitoring level and complaints at admission, which can be used to improve the quality of diagnosis and prognosis models of the disease and its complications.

References
[1] Shelmanov A O, Smirnov I V and Vishneva E A 2015 Computational Linguistics and Intellectual Technologies: Annual International Conference “Dialogue” 14(21) pp 537–49
[2] Ingersoll G S, Morton T S and Faris E L 2015 Unstructured text processing. Search, organization and manipulation (Moscow: DMK Press) p 414
[3] Abhyankar S, Demner-Fushman D and McDonald C J 2012 J. Biomed. Inform 45 3–8
[4] Jensen P B, Jensen L J and Brunak S 2012 Nat Rev Genet 13(6) 395–405
[5] Velupillai S, Mowery D, South B R, Kvist M and Dalianis H 2015 Year book of Medical Informatics 10 pp 183–93
[6] PostgreSQL: Documentation. URL: https://www.postgresql.org/docs/
[7] Bengfort B, Bilbro R and Ojeda T 2019 Applied analysis of text data in Python. Machine learning and building natural language processing applications (St. Petersburg: Peter) p 368
[8] Bird S, Klein E and Loper E 2009 Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit (USA, California, Sebastopol: O’ReillyMedia) p 504
[9] Levenshtein V I 1965 Reports of the USSR Academy of Sciences 163(4) 845–48
[10] RubashkinV Sh and Chuprin B Yu 2006 Computer Linguistics and Intelligent Technologies: Proceedings of the international conference "Dialog 2006" (Moscow: Publishing House of the Russian State Humanitarian University) pp 456–58