Big Data & Health Care
Volume 1, Issue 1, No. 1
Published online August 20 2022

Sanderman
Publishing House

Similar medical record retrieval system based on medical big data platform

Hui Yang[1], Song Xue[1], Guangli Gu[1], Feng Huang[1], Saijuan Jin[1], Yongzhang Ji[2*]
[1] Information Section, No. 454 Hospital of the People's Liberation Army, Nanjing, China
[2] Medical Department, No. 454 Hospital of the People's Liberation Army, Nanjing, China

Abstract. **Objective** To realize similar medical records retrieval in medical record database on medical big data platform based on natural language processing technology. **Methods** The platform index retrieval technology was used for the structured part of medical records; for the unstructured natural language description part, features were extracted and similarity was calculated to retrieve similar medical records based on the constructed medical feature database. **Results** The system can retrieve the similar medical records in the medical record database, and the users can make auxiliary diagnosis or scientific study analysis based on the retrieval results. **Conclusion** It is proved that the similar medical record retrieval system based on natural language processing technology is feasible by judging the retrieval results, but its accuracy shall be further improved.

Keywords. Index Search, natural language processing, similar medical records

1. Introduction

Electronic medical record (EMR) has been widely used in major hospitals with the gradual deepening of hospital informatization. EMR system has collected much information and gradually entered the era of big data through years of accumulation. Much text information in these electronic medical records has become a valuable asset of various hospitals. However, the relatively simple statistical function of HIS system can no longer meet the growing needs of people [1]. How to use the massive text information of EMR system to serve doctors and patients has become a study topic. In this paper, a method based on semantic similarity calculation is proposed by using natural language processing technology, such as automatic word segmentation, building medical vocabulary ontology database and index retrieval technology based on open-source search engine Solr, to realize the similar medical record retrieval function, providing a reference for using electronic medical record text information and the quality monitoring of electronic medical record [2].

Electronic medical records are popularized in hospitals at all levels gradually. Except for the course records, more and more clinical system data, such as tests and examinations are integrated into the electronic medical record. Therefore, how to store, retrieve and reuse the data of the electronic medical record has become a study hotspot [3].

The standard of clinical data format has been studied both at home and abroad, such as HL7 CDA can be used as the design specification of electronic medical record. Except for complying with the overall clinical data standards, the large domestic electronic medical record manufacturers also strive to standardize each module [4]. For example, some EMRs provide symptom dictionaries and set several templates for some diseases in the current history input link.

All these study works aim at formatting and standardizing the data entry and storage of electronic cases. However, there is no national or industrial unified symptom dictionary and common term dictionary till now, and most descriptions of diseases cannot be input according to fixed templates.

There are also some difficulties in the retrieval and secondary utilization of electronic medical record data. For example, when doctors encounter difficult and complicated diseases or do medical study, they hope to customize some input conditions to search similar cases of history for reference. The existing system can rarely meet the needs of the above-mentioned doctors to search and analyze medical records.

Many hospitals are building internal medical data platforms to response to the national call to build a regional medical platform [5]. The 454 Hospital of the People's Liberation Army has explored and built a big data platform based on medical data storage. The platform integrates various formats of data from HIS, LIS, EMR, and PACS, etc., and realizes the basic rapid retrieval function. An important significance of establishing a big data platform is to analyze the data after collecting a large amount of data and mine the related information that cannot be found on a single system [6]. A similar medical record retrieval system is designed in this paper based on the above big data platform to make secondary use of the value of medical data. The relevant inspection data and image data can be further reviewed after the similar medical records are retrieved through text.

2. System design and implementation

2.1 Design idea

Use the data collection function of the big data platform to extract user-defined meta data (metadata) from the standard data or files in the form of DB, HTML, PDF, HL7, DIC0M, etc. of HIS, LIS, EMR, PACS, etc., and save the metadata and corresponding source data files in the form of objects on the content storage platform. This platform uses Hitachi storage [7]. The platform also uses Solr to establish a full-text index of these metadata and source files, which can quickly retrieve and display relevant files [8]. The function of similar medical record retrieval system based on this platform is designed to use the index retrieval technology of the platform for the structured part of medical records; for the unstructured natural language description part, features shall be extracted and similarity shall be calculated to retrieve similar medical records based on the constructed medical feature database. For system architecture diagram, see Figure l.

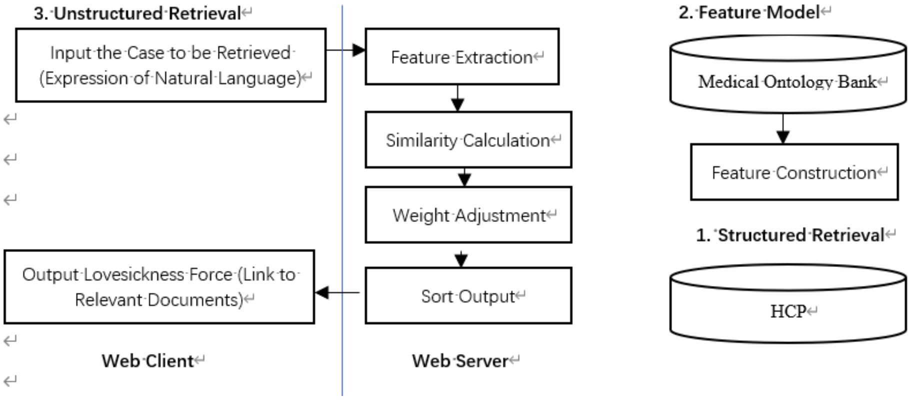

Figure 1. Architecture of similar medical record retrieval system

2.2 Structure analysis of electronic medical record

The electronic medical record source files collected on the platform are HTML files exported by the EMR system of our hospital. The key information contained in the XML file is analyzed, as shown in Fig. 2. For the structured part, Solr tool is directly adopted to establish the indexes, such as the patient's gender, age, and physical examination results, etc. [9], and the corresponding search input interface is provided on the search interface; for unstructured data, such as the description of current medical history, Solr has established a full-text index. On the search interface, the sentences and keywords contained therein can be inputted for query, but the search performance is general.

< fieldelem name = "patient’s name" > Zhang San < /fieldelem>

< fieldelem name = "patient’s gender]" >female< /fieldelem>

< fieldelemname = "age" > 63 < / fieldelem >

<fieldelemname ="Content of current medical history"> the patient had chills and fever after getting cold last night. The temperature was up to 39.0 °C, accompanied by headache, mild pharyngeal pain, nausea, no vomiting, no nasal obstruction, runny nose, smooth defecation, difficulty in urinating at night, and pain in urination. The patient came to our hospital this morning. The outpatient planned to check the positive and lateral position of the chest: no clear substantive lesions were found, and she was given oral Cefixime tablets and Composite Pseudoephedrin Hydrochlorid Tablets, but still suffering from high fever. She went to the outpatient clinic this afternoon and is planned to be admitted in our department for "fever to be examined". During the course of the disease, the patient's spirit was poor, her sleep and diet were general, her stool was smooth, and she urinated a little more at night, accompanied by urination pain. There was no significant change in body weight recently</fieldelem〉

< fieldelemname = "body temperature"> 42.0 < / fieldelem >

< fieklelemname = "pulse value" > 96 < / fieldelem >

Fig. 2 Electronic Medical Record Segment

2.3 Electronic medical record retrieval based on similarity

Based on the description part of the unstructured data in Fig. 1 that is similar to the current medical history, although Solr has established a full-text index and can input some words and sentences through the search interface, the users shall organize the key sentences independently, and Solr does not make special treatment for each word segmentation, and cannot distinguish the importance of symptom words from other words, so the search results are not easily to be controlled. And for the unstructured data retrieval in similar medical records, the similarity based on semantics is calculated.

2.3.1 Retrieval based on structured data

Some meaningful features of the medical records to be retrieved are analyzed first and the retrieval conditions are set. If the set search conditions (gender: female, age: 60 ~ 70, body temperature: 39 ~ 42, Department: Respiratory Medicine, etc.), a group of medical records can be roughly screened.

2.3.2 Building feature model for unstructured data

An ontology library was first prepared in the medical field to build a feature model, describing various features of electronic medical records. The symptom characteristics can be expressed in the common symptom dictionary, such as chills, fever, headache, sore throat, nausea, vomiting, nasal congestion, runny nose, and urinary pain. For the unstructured part of each screened medical record, such as the description of current medical history, a feature vector can be constructed through the symptom dictionary [10]. The specific design is as follows: the words with symptoms are represented by 1; the absence is indicated by 0; when it appears but is modified by "None", it is indicated by **-** 1. According to the rules, "None", "Deny", "Not Accompanied" and other words are similar. Due to the large number of words in the symptom dictionary, the initially constructed vector dimension is large, which shall be reduced in terms of operation speed and semantic meaning. In this system, singular value decomposition (SVD) is adopted to reduce each vector to ten or tens of dimensions. A matrix model of eigenvector is constructed for all screened medical records till now.

2.3.3 The similarity calculation between the original medical record and the matrix model is performed.

The unstructured part of the original medical record is executed in the same process as in "1.3.2" to obtain a feature vector, such as the description of the current medical history. The initial similarity between the medical record and each medical record in the medical record group is obtained by comparing the distance between the feature vector and each vector in the feature matrix. For example, the weight of each symptom is also provided in the ontology database, reflecting the importance or frequency of the disease. This knowledge can be used for further modifying the initial similarity and obtain the final similarity. The symptom weight can also be tested by using statistics based on word frequency and then confirmed by experts. After the similarity is calculated, it is displayed in the output interface in the order of similarity. Except for directly displaying similar text information, it also provides links to original medical records and related images and other files. Users can conduct more in-depth viewing and analysis according to their own needs.

3. Search results and discussion

A batch of electronic medical records of a certain department are used for preliminary test. Since there is no industry standard for judging the calculation results of similarity, and there is no unified test database in the industry, only the advantages and disadvantages of the calculation results artificially can be judged. When the input medical records are also stored in the medical record database, the similarity between the two is 100%; medical records with similarity between 80% and 100% are usually of reference significance; the larger the medical record database, the greater the probability of retrieving medical records with high similarity. The search results also reflect many problems to be solved: first, because the description of symptoms is not standardized, it is necessary to collect approximate dictionaries of symptoms, such as "fatigue", "weakness", etc. Second, because the symptom dictionary is not rich enough, the common words of some departments or diseases are not considered as important features. After the common words dictionary is added, the similarity results will be more accurate. Third, the parts of symptom modification, such as "both lower limbs" and "left lower limbs" have not been established. When such a relationship is added to the ontology database, the retrieval results will be more accurate. The processing of medical record description language involves complex natural language processing technology. If more feature points are considered, the system shall be optimized and improved for a long time. This system is based on natural language processing and ontology technology, and makes a preliminary study on similar medical record retrieval.

4. Conclusion

A similar medical record retrieval system is described in this paper based on the medical big data platform. It does feature extraction and similarity calculation for the unstructured data in the files stored on the platform, i.e., the natural language description part, and displays the retrieval results to the user.

Users can further view relevant examination and image data information after similar medical records are retrieved. With this system, users can refer to similar medical records for auxiliary diagnosis, and can also analyze a special type of medical records according to their own scientific study needs and mine new knowledge from them.

References

[1] Song Bin, Chen Hai-dong, Lei Yong, et al. Application of Data Warehouse in Digital Hospital [J]Southeast Defense Medicine, 2010,12 (6): 519-522
[2] Zhao Bo-cheng, Zhou Bin, Lyu Yao-xin, et al. Actual Effect and Experience of Monitoring the Quality of Electronic Medical Records in Our Hospital [J] Southeast National Defense Medicine, 2010,12 (3): 276-277
[3] Zhang Zhi-chang, Lou Yan, 2013-2015 Cluster Analysis of Subject Words in SCI Papers Based on Electronic Medical Records [J] China Digital Medicine, 2016,11 (3): 26-27
[4] Meng Yan, Li Shan-shan, Song Hai-qing, et al. Deep Application and Experience of Electronic Medical Record [J]. China Digital Medicine, 2016,11 (7): 111-113
[5] An Zhi-ping, Gao Zhi-jun, Zhang Yun-hong, et al. Construction and Application of Remote Medical Record Information Inquiry System [J] Journal of Medical Graduate Students, 2016, 29 (12): 1325-1327
[6] Zou Bei-ji, Big Data Analysis and Its Application in Huang Therapy [J] Computer Education, 2014, 7:24-29
[7] Xue Yi-feng, Gu Guang-li, Zhao Bo-cheng, et al. Study and Implementation of Medical Big Data Platform Based on Metadata File Storage [J] China Digital Medicine, 2015, 10 (10): 73-75
[8] Zhou Bin, Yang Hui, Xue Song, et al. Application of Solr in Medical Big Data Retrieval. China Digital Medicine, 2016, 11 (9): 21-23
[9] Huo Qing, Liu Pei-zhi, Using Sok to Build a Search Engine for Large Databases [J] Software, 2011, 32 (6): 11-14
[10] Wang Huan. Study on Semantic Retrieval System Based on Domain Ontology and Lucene [J] Computer Applications, 2010, 30 (6): 1656-1660