Research on Similar Medical Records Recommend Model based on Natural Language Processing

Yong Shuai¹*, Chunjiang Zhou² Qiaoyue Pang¹ Can Yuan¹ and Zheng Xie³

¹Chongqing Ceprei Industrial Technology Research Institute, Chongqing, 401332, China
²The First Affiliated Hospital of Chongqing Medical, Chongqing, 400042, China
³Chongqing Public Health Medical Center, Chongqing, 400030, China

*Corresponding author’s e-mail: alexshuai@sina.com

Abstract. Aiming at the problem of low efficiency and poor credibility in medical record text data mining, this paper proposes a similar medical record recommend model based on natural language processing. Firstly, extract the samples by structured data in the medical record data and narrow down the extraction of similar medical records, then use the text keyword extraction algorithm to obtain the keyword text data about chief complaint, medical history and physical examination in the medical record text, use these data as input data, get the similarity between the input data and the patient data by text similarity algorithm, and use collaborative filtering algorithm to calculate the most similar medical records and obtain the most similar medical records diagnosis conclusions, take the chief complaint, medical history, physical examination data and diagnosis conclusions as input data to calculate the text similarity and collaborative filtering values and obtain the final recommend treatment plan. The case analysis proves that the model proposed in this paper has high credibility and practicality.

1. Introduction

In recent years, the improvement of deep learning frameworks and artificial intelligence algorithms performance have provided new ideas for text data mining, and the accuracy and efficiency of the results obtained by using natural language processing techniques for natural language processing also surpass the traditional methods. For text data such as medical record data, the use of natural language processing technology has wide application value, especially for some specific diseases, using similar medical record data can help junior doctors and doctors in underdeveloped medical resources region to improve the accuracy of diagnosis and rationality of treatment plans, optimize auxiliary diagnosis and treatment methods, and promote the development of smart medical technology.

The current methods for processing medical record data include:

1. Use machine learning algorithms to analyze text statistical data, such as statistical analysis of drug properties [1].
2. Use statistical methods to analyze the frequency of drug use and obtain the most frequently used words [2-4].
3. Text clustering analysis based on machine learning algorithms, clustering medical record data and discovering the rules of medication [5].

The above methods mainly use the traditional statistical learning algorithms and machine learning algorithms to process the structured data and text data in the medical records, resulting in lower
accuracy of calculations and poor credibility of the results. Therefore, this paper proposes a similar medical record recommend model based on natural language processing. First, extract the sample from the structured data in the medical record data to narrow down the extraction of similar medical records, then use the chief complaint, medical history, and physical examination data in the medical record as input data to recommend the diagnostic result, and use the selected diagnostic data as input data to get the optimal diagnostic program, achieve auxiliary diagnosis and treatment based on natural language processing.

2. Modeling Process
This paper constructs a similar medical record recommend model, which mainly includes SQL-based text data acquisition from medical record database, input text keyword extraction, text data similarity calculation, recommending the optimal diagnosis conclusion by similarity data collaborative filtering, and recommending for final processing scheme based on collaborative filtering. The overall modeling process is shown in Figure 1.

### 2.1. Text data acquisition of medical record database based on SQL
Currently, all medical records are issued through the Hospital Information System (HIS). Therefore, the medical record data is stored in the database of the HIS system.

After the doctor enters the basic information of the patient, we can use SQL language to obtain data similar to the basic information of the patient based on the basic situation of the current patient. The pseudo code is ‘Select medicine_cases from case_table where Gender = ‘male’ and Visiting_department = ’Visiting_department’ and ages in (Minimum_age, Maximum_age)’.

The range of similar data acquisition is shown in Table 1.

| Content            | Range          |
|--------------------|----------------|
| Gender             | Equal          |
| Visiting Department| Equal          |
| Age                | Adult: +360 days\Child: +180 days |

The obtained text data is divided into two parts, where the data of chief complaint, medical history (including current medical history and past history), physical examination are set as input data for diagnostic result data, the diagnostic result data is set as output data for the first three types of data, and the data of chief complaint, medical history, physical examination and diagnostic result are set as input data for treatment program data.
2.2. Keyword extraction
To the three types of input text data, due to the large amount of data and the importance of containing information, it is impossible to make full use of these data. Therefore, we consider using keywords to replace these text data, which can speed up the calculation speed, ensure the important content to be not lost, and guarantee the accuracy of the algorithm. The keyword extraction algorithm used in this paper contains four algorithms: TF-IDF, TextRank, LSI, and LDA, and the extraction results of each algorithm are combined by using intersection method to avoid the loss of important keywords due to different algorithms and prevent over-production large amount of computation.

Assuming that the keyword set extracted by each model of TF-IDF, TextRank, LSI, and LDA is \( S_{TF} \), \( S_{Te} \), \( S_{LAS} \), and \( S_{LDA} \), then the total text data set is
\[
S_{in} = S_{TF} \cup S_{Te} \cup S_{LAS} \cup S_{LDA} \quad (1)
\]

2.3. Text data similarity calculation
For the extracted keywords, a similarity calculation with the keywords of the current patient is required, so as to obtain a sample matrix collaborative filtering algorithms needed. The text similarity calculation steps are as follows:

1. Calculate the frequency of words
2. Filter out low-frequency words
3. Build a dictionary through a corpus
4. Load the documents to be compared
5. Convert the compared documents to be sparse vectors through doc2bow
6. Process the sparse vectors to obtain new corpus
7. Process the new corpus through TF-IDF model to get TF-IDF
8. Get the feature number by token2id
9. Calculate sparse matrix similarity and build indexes
10. Get the final text data similarity result

Assume that the obtained medical record text is \( n \), then the similarity matrix of the \( i \)-th text is
\[
X_{i} = [X_{1i}, X_{2i}, X_{3i}] \quad (2)
\]
where \( X_{1i} \) represents the text similarity of the chief complaint, \( X_{2i} \) represents the text similarity of the medical history, and \( X_{3i} \) represents the text similarity of the physical examination.

Medical history text similarity \( X_{2i} \) is calculated as:
\[
X_{2i} = 0.6 \times X_{2i1} + 0.4 \times X_{2i2} \quad (3)
\]
Where \( X_{2i1} \) indicates the text similarity of the current medical history, and \( X_{2i2} \) indicates the similarity of the text the previous medical history.

2.4. Recommending diagnosis result and processing program based on collaborative filtering
Collaborative filtering is a recommendation technology. It can filter information that is difficult for the machine to automatically analyze based on some complex and difficult to expression, and ensure the feasibility of recommendations, so it is very suitable for text recommendation. Therefore, this paper adopts collaborative filtering to recommend the optimal diagnosis conclusion and final treatment plan.

Set the similarity matrix of each text as \( X_{i} = [X_{1i}, X_{2i}, X_{3i}, X_{4i}] \), define the current patient input data matrix as \( X_{0} = [1, 1, 1, 1] \). Use the sklearn.metrics.pairwise.cosine_similarity function to calculate the collaborative filtering value between the text similarity matrix and the current patient input data matrix, takes the diagnosis result corresponding to the maximum value as the recommended diagnostic conclusions. Using this conclusion as input data, recalculate the similarities of the four types of texts in the medical record, including the main complaint, medical history, physical examination and diagnostic results, and record them as matrices \( X_{i} = [X_{1i}, X_{2i}, X_{3i}, X_{4i}] \). Set the current patient input data matrix \( X_{0} = [1, 1, 1, 1] \), calculate the collaborative filtering value of each text similarity matrix and
the new matrix of the current patient input data, and take the processing program corresponding to the maximum value as the recommended program.

3. Case analysis
The case data is derived from the medical records of pediatric visits within a 5-day period in a hospital. The method proposed in this paper is used to process these data.

(1) Use SQL statements to get text from HIS and get a total of 10 hits. Categorize the text in the data, including the main complaint, medical history, physical examination, diagnostic conclusions, and treatment program.

(2) Extract text keywords, among which 1000 pieces of medical records of the department were collected for the keyword learning samples.

For example, for a medical examination text: ‘pharyngeal congestion, 1 degree of double tonsil, no secretions, thick lung breathing, no wet and dry rales, no abnormalities in the abdomen, red tongue, less moss, flat pulse’, the keywords obtained by using TF-IDF / TextRank / LSI / LDA algorithms are shown in Table 2.

| Algorithm | Keywords |
|-----------|----------|
| TF-IDF    | Tongue quality / Pulse flatness / Unheard / No rales / Tonsils / Heart belly / Pharynx / Secretions / |
| TextRank  | Respiration / Belly / Unheard / No rales / Tonsils / Pharynx / Congestion / Tongue / Secretion / |
| LSI       | Secretions / pharynx / tonsils / |
| LDA       | Pharynx / secretions / tonsils / |

(3) Based on the hypothesis of this article, the key word matrix of the modified medical history text is [Tongue quality, flat pulse, unheard, no rales, tonsils, confidants, pharynx, secretions, breathing, congestion].

(4) Use collaborative filtering to find the optimal recommendation medical record. The final recommendation matrix for collaborative filtering of is [0.8115, 0.8350, 0.7515, 0.8710, 0.8564, 0.8115, 0.9790, 0.7924, 0.9893, 0.9330]. It can be seen that the medical record No. 9 is most similar to the case of the current patient. Taking the diagnosis result of No. 9 as the diagnosis program of the current patient, calculate the text similarity of each medical record and the collaborative filtering value of each medical record text data. Finally, the 9th program has the highest collaborative filtering value, so the 9th medical record processing program is pushed.

The collaborative filtering values based on the similarity of 3 text data and 4 text data are shown in Figure 2.
Figure 2. Collaborative filtering values based on the similarity of 3 text data and 4 text data.

(5) Algorithm credibility analysis. Because there is no similar algorithm to compare, in order to verify the effectiveness of the algorithm, this paper analyzes 16 existing medical records. Assume that the medical record data is diagnostic data, and the diagnosis conclusions and processing schemes are pushed according to the method in this paper. The textual similarity data between the pushed medical record data and the original medical record data are shown in Table 3. From Table 3 we could find that the similarity of the medical records pushed by the algorithm in this paper is mostly above 0.9, which illustrates the credibility and practicability of the algorithm.

| Medical Record Number | Diagnosis Result | Treatment Program | Medical Record Number | Diagnosis Result | Treatment Program |
|-----------------------|------------------|-------------------|-----------------------|------------------|-------------------|
| 1                     | 0.8607           | 0.8047            | 9                     | 0.9585           | 0.8266            |
| 2                     | 0.9053           | 0.8277            | 10                    | 0.9261           | 0.8631            |
| 3                     | 0.8524           | 0.8236            | 11                    | 0.8283           | 0.8735            |
| 4                     | 0.9392           | 0.8553            | 12                    | 0.8925           | 0.8686            |
| 5                     | 0.996            | 0.8563            | 13                    | 0.8503           | 0.8371            |
| 6                     | 0.9142           | 0.9481            | 14                    | 0.9034           | 0.9335            |
| 7                     | 0.8456           | 0.8713            | 15                    | 0.996            | 0.8873            |
| 8                     | 0.9424           | 0.9766            | 16                    | 0.9918           | 0.9513            |

Table 3. Similarity of Medical record text.

4. Conclusion
In recent years, natural language processing has become an important branch of artificial intelligence, and it will also be widely used in the field of smart medicine in the future. This paper uses natural language processing technology including keyword extraction, text similarity model and collaborative filtering algorithm to achieve the pushing of similar medical records, which has good practicability and credibility. Meanwhile, this paper only analyzes the medical record data of the first diagnosis.
Because the re-examination data is more relevant, the recommending of the re-examination medical records needs to consider more complicated models.

Acknowledgments
This research was financially supported by Intelligent Manufacturing Innovation and Development Project of the Ministry of Industry and Information Technology (MGY1804080), Chongqing Science and Technology Bureau Project (cstc2019jscx-fxyd0298, cstc2020jscx-fyyx0065, cstc2020jscx-fyyx0098).

References
[1] Zhang, G.L., Liu, J., Xin, L., Huang, D. (2019) Study on the Drug Use Rule of Traditional Chinese Medicine in Treating Ankylosing Spondylitis Based on 895 Hospitalized Cases Data Mining. JOURNAL OF JIANGXI UNIVERSITY OF TCM, 31: 30-34.
[2] Li, Y., Zhang, J.X., Wang, X.L., Wang, X.L. (2017) Screening of basic prescriptions for flu-like cases based on data mining of famous medical records. Hebei J TCM, 39: 1334-1336.
[3] Li, Q., Zhang, N.G., Gao, T. (2018) Analysis of Chinese Medicine Treatment of Type 2 Diabetes Mellitus Basedzone Data Mining. Aisa-Pacific Traditional Medicine, 14: 103-106.
[4] Song, Y., He, J., Shu, H.P., Zheng, J.L., Lang, F.R., Ren, Y.L., Wen, L.Y. (2014) The Research of Disease Symptoms Weighting Mining based on Text Mining Method MAPTF-IDF. Journal of CHengdu University of Information Technology, 29: 52-58.
[5] Mo, T.R., Li, S.B., Yun, J. (2017) The Research of Data Mining on the Medication Rule of Chronic Glomerulonephritis Treating by Professor Song Liqun’S Clinic Cases. CJITWN, 18: 786-788.
[6] Tu, M., Liu, X., Liu, C.S. (2018) Key word extraction method. In: Zhang, X.P.(Eds.), Natural Language Processing Core Technology and Algorithm with Python. China Machine Press, Peking. pp. 85-105.