A Retrieval Method for Chinese EMR Based on Semantic Knowledge Map

Hao Li¹ and Runtong Zhang¹*

¹School of economics and management, Beijing Jiaotong University, Beijing, 100044, China
*Corresponding author’s e-mail: rtzhang@bjtu.edu.cn

Abstract. Due to the diversity of natural language in Chinese electronic medical records, it is usually hard for traditional retrieval methods to provide ideal results. On this condition, this paper proposes a retrieval method for Chinese EMR based on semantic knowledge map. Through natural language processing and semantic analysis, we can build connections for medical knowledge, and organize all the entities into a visual knowledge map. After that, a novel retrieval method based on semantic knowledge map is proposed, which focuses on node connection of documents and terms. Through semantic extension and intention spread, the improved retrieval results are returned, and the results are reordered by correlation. Compared with general methods, this method can significantly improve the accuracy of Chinese EMR text retrieval and optimize the ranking strategy of retrieval results.

1. Introduction
Electronic medical record (EMR) is a kind of medical record which is constructed by medical staff using words, symbols, charts, graphs, data, images and other digital information in the process of medical activities[1]. Doctors and researchers all want to carry on effective data mining to massive medical information, in order to extract effective knowledge for themselves[2]. However, the traditional retrieval strategy is often difficult to meet this demand. There are three reasons:

1. At present, there is no widely used standard for the development of EMR in China.
2. The traditional retrieval strategy takes keywords-identifying as filtering means, while the value of EMR is often difficult to be expressed through specific keywords[3].
3. For the retrieval of valuable medical information, operators often do not have clear purpose, hoping that the system can give reasonable recommendation to solve their own problems[4].

Compared with general text retrieval, the researches on EMR retrieval pay more attention to the establishment of professional knowledge base and the relationship between information. Zhang et al.[5] proposed a text structuring method for Chinese EMR, reorganizing medical information in chronological order according to the temporal signs in the text. Vasanthakumar et al.[6] focus on extracting meaningful patterns from EMR, mapping EMR to SNOMED core subset to find important concepts. Meng et al.[7] also used semantic relationships in the SNOMED to expand the related terms of renal cancer, and proposed a text analyzing and early warning model based on EMR. By incorporating multiple medical semantic resources and query expansion techniques, Maree et al.[8] proposed a semantics-based retrieval system which improved the quality of the returned results. It can be seen that although previous studies have solved the problem of diversified representation of medical information by constructing domain ontology and realizing semantic query expansion, they...
have not considered the relationship between medical information. Their results are mainly based on the principle of words-matching.

This paper takes this as a breakthrough point, considers the connection mode of massive medical information, constructs a visual semantic knowledge map, expands the user's query intention through the connection between knowledge nodes, and makes a reasonable sorting according to the length of connection path. This model makes the retrieval results more accurate and comprehensive.

2. Methods
The new method is divided into three steps: 1) Construction of medical semantic knowledge map; 2) BILSTM+CRF entity recognition; 3) Relevance retrieval model. The specific process is shown in Figure 1.

2.1. Construction process of medical information semantic knowledge map
Medical information semantic knowledge map is the basic framework of this study. Its function is to extract relations from unstructured medical information, construct the connection between information, and attach searching tags to medical information. The topological structure of medical information semantic knowledge map model is shown in Figure 2.
Figure 2. Structure of semantic knowledge map

The structure of the model is a star topology without boundary.
1. As a key node in the model, document $D$ is nonredundant and interconnected;
2. $k_1$-$k_n$ is the keyword set of document $D$, which is marked as $K$;
3. $w_1(1)$-$w_t(m_t)$ denotes the set of related words belonging to the keyword $k_t$, where $m_t$ represents the number of related words that can be extracted for the keyword $k_t$. The set is denoted as $W$;
4. For each element in the set of related words, the additional related words can be extracted again to form a new set $S$. The elements in $S$ are uniformly marked as $w(s)$.

Medical information text has a high degree of freedom. In order to facilitate machine analysis, it is necessary to carry out natural language processing (NLP) to convert unstructured electronic medical records into structured descriptions[9].

Firstly, Jieba word segmentation tool is used to deal with Chinese EMR text, and stop words are removed to form word segmentation corpus. Then, the graph model is established by using the TextRank automatic summarization algorithm, and the keyword set of each document is extracted by calculating the score of each point in the graph[10]. Finally, this paper uses Word2Vec algorithm to build a word vector model and carry on the word correlation analysis[11]. The related word nodes are introduced into the model to strengthen the connection of medical knowledge.

2.2. BiLSTM+CRF named entity recognition

On the basis of the above work, in order to further enrich the knowledge dimension of semantic knowledge map, it is necessary to identify named entities of words in corpus. A BiLSTM+CRF method is used to establish the model[12]. BiLSTM+CRF method is a deep learning model integrating bidirectional LSTM (Long Short-Term Memory) and CRF (Conditional Random Field). It can effectively capture semantic dependency and is suitable for semantic recognition of EMR text.

Firstly, we train the model of EMR text. In order to process the EMR text presented in natural language, manual data annotation is necessary. For example, the original and annotated data format of an EMR text is as follows:

| EMR TEXT |
|----------|
| Male, 27 years old, in good health before, felt pain in his left upper limb when he was knocked down by a car three hours ago and was afraid to move, especially in his left elbow, which was obviously swollen. |

Figure 3. Chinese EMR text
Table 1. Annotation data format

| Entity          | The position of the first character | The position of the last character | Entity type |
|-----------------|-------------------------------------|-----------------------------------|-------------|
| Left upper limb | 17                                  | 19                                | Body        |
| Pain            | 14                                  | 14                                | Sign        |
| Left elbow      | 40                                  | 41                                | Body        |
| Swollen         | 46                                  | 46                                | Sign        |

In terms of entity type annotation, this study mainly refers to the classification of EMR named entities on the 2010 i2b2/VA challenge meeting[13], and expanded the classification method of "medical problems, examination and treatment". In order to carry out model training, this paper sets the following entity types: O (non-entity part), TREATMENT, BODY, SIGN, CHECK, DISEASE.

Combined with the forward and backward LSTM process, the stitching sequence is generated and substituted into CRF to train the named entity recognition model.

For a prediction sequence, $y = \{y_1, y_2, \ldots, y_n\}$, its probability can be expressed as follows:

$$K(X, y) = \sum_{i=0}^{n} A(y_i|y_{i+1}) + \sum_{i=1}^{n} P(y_i)$$  \hspace{1cm} (1)

In the training process, the likelihood function of the marker sequence is as follows:

$$\log(p(y|S)) = K(X, y) - \log \left( \sum_{y \in Y_x} e^{K(X, y)} \right)$$ \hspace{1cm} (2)

Among them, $Y_x$ represents all possible tag sets. In the prediction, (3) outputs a group of sequences with the largest overall probability.

$$y^* = \arg \max_{y \in Y_x} K(X, y)$$  \hspace{1cm} (3)

After getting the named entity recognition model, we can predict the key words, related words nodes in the semantic knowledge map to identify the corresponding entity type.

2.3. Relevance retrieval method of Chinese EMR

After the construction of medical information semantic knowledge map, we can optimize the retrieval scheme for Chinese EMR text based on it. The specific flow of the retrieval scheme is shown in Figure 4.

For the statements input by users, the text preprocessing is carried out firstly, including word segmentation, removal of stop words and keyword extraction. Then, we use the BiLSTM+CRF named entity recognition model to identify the entity type of each retrieval word, and different retrieval weights are given. Next, the query expansion based on Word2Vec model is carried out. Finally, two searching methods are realized: semantic extension search and intention spread search.
For semantic extension search, we search for documents with keyword correspondence in semantic knowledge map according to the extended search condition. We traverse the search terms and substitute them into Word2Vec model. If there is a related-words set, we traverse and add the qualified related words into the extended set; if there is no related-words set, we add the original words into the extended set. Finally, the extended search term set is formed by integrating and sorting.

For intention spread search, its essence is substituting the extended search terms into the semantic knowledge map, and spread the relationship of the search terms through the node connection in the knowledge map. The flow chart of the algorithm is shown as figure 5:

![Flow of the intention spread search](image)

**Figure 5. Flow of the intention spread search**

The process of the spread algorithm is as follows: At the beginning of the algorithm, the initial node $i$ obtained from the keyword-matching search is stored in the priority queue, and then the adjacent node $j$ of $i$ is searched (When there are multiple adjacent nodes, the one with greater similarity is preferred), and the input value of $j$ is calculated.

$$I(j) = \sum_i O(i)S_{ij}$$  \hspace{1cm} (4)

Where $O(i)$ is the output of node $i$, $O(i)=f(I(i))$. $S_{ij}$ represents the association weight between node $i$ and node $j$.

After getting the input value $I(j)$ of adjacent node $j$, if it is found that node $j$ is not a document node and is not in the priority queue, it will be added to the priority queue; if it is found that node $j$ is a document node and is not in the result queue, it will be added to the result queue. The above process is repeated until the termination condition is met, which is defined as: all the current nodes have no related nodes, or the distance of the nodes found by the algorithm is greater than the threshold.
Through the above process, several logical paths from the initial node to a specific document are found in the semantic knowledge map. The length of the path represents the intention relevance of the retrieval results. The specific calculation formula is as follows:

\[ R(D_r) = K(i) \cdot \sum_i \sum_j I(j) \]  \hspace{1cm} (5)

Where, \( i \) is the number of the paths ending at the result document; \( j \) is the number of nodes on a path; \( I(j) \) represents the input value of node \( j \); \( K(i) \) is the adjustment coefficient, generally expressed as the function:

\[ K(i) = i^{-\frac{1}{n}} \]  \hspace{1cm} (6)

\( n \) is a constant in the range of 1-5, \( K(i) \) forms a monotone decreasing function with a gradual gentle slope.

The larger \( R(D_r) \) is, the more paths to the result document \( T \) are, and the higher the path correlation is, which indicates that the result is more likely to meet the user’s expectations. Therefore, the model takes \( R(D_r) \) as the sorting basis, and the final result set of intention spread search is given.

3. Experiments

This paper selects a standard data set of Chinese EMR provided by CCKS2018 (China Conference on Knowledge Graph and Semantic Computing). A total of 1073 electronic medical records were retained. We construct a query set with ten query topics. Considering that the experimental text does not label the topic, we do artificial topic tagging.

In general retrieval evaluation tasks, the indicators include precision ratio and recall ratio.

\[ \text{precision} = \frac{T}{T + FP} \]  \hspace{1cm} (7)

\[ \text{recall} = \frac{T}{T + FN} \]  \hspace{1cm} (8)

Among them, \( T \) is the number of related documents in the query results, \( FP \) is the number of unrelated documents in the query results, and \( FN \) is the number of related documents not found.

Precision ratio and recall ratio respectively reflect the accuracy and comprehensiveness of the query. In order to integrate the two indicators, P-R curve can be introduced. The area formed by the curve and the coordinate axis can comprehensively reflect the effectiveness of the retrieval results.

In addition to the accuracy and comprehensiveness of the retrieval results, people also pay attention to the ranking of relevant contents in the retrieval results. Therefore, this paper uses MRR (Mean recursive rank) to detect the ranking effect of the results, which reflects the ranking status of the first correct result in the query.

\[ MRR = \frac{1}{|N|} \sum_{i=1}^{|N|} \frac{1}{\text{Rank}_i} \]  \hspace{1cm} (9)

Where, \( |N| \) is the number of query sets, \( \text{Rank}_i \) represents the ranking of the first correct document in the \( i \)-th query.

In order to verify the effectiveness of the proposed retrieval model, this paper compares the results of the following three algorithms: basic model based on keyword-matching (Kw-Base), semantic extension search model based on Word2Vec (W2V-Extend), and intention spread search model based on semantic knowledge map (SeMap-Spread).

The experiment simulates ten queries with different topics, and calculates the average precision of the three models for ten queries under 10 recall indexes. The experimental result is shown in Table 2 and figure 6:

| recall | precision |
|--------|-----------|
|        | Kw-Base   | W2V-Extend | SeMap-Spread |
| 10%    | 1         | 1          | 1            |
| 20%    | 0.88      | 0.92       | 0.97         |
We can find that under the same recall index, the results of precision are: SeMap-Spread > W2V-Extend > Kw-Base. This is because the Kw-base model only takes keywords-matching as the principle, and the W2V-Extend model introduces related-words set to expand the retrieval conditions, while SeMap-Spread model further utilizes the semantic relationship between EMR to expand the retrieval vision.

In order to further compare the effectiveness of different models, the ranking of the retrieval results was evaluated by calculating the MRR value. We evaluate the ranking of the first five relevant results of each retrieval, and compare the MRR-1, MRR-2, MRR-3, MRR-4, MRR-5 and MRR average indicators of the three models:

| MRR  | Kw-Base | W2V-Extend | SeMap-Spread |
|------|---------|------------|--------------|
| MRR-1| 1.00    | 1.00       | 1.00         |
| MRR-2| 0.49    | 0.48       | 0.50         |
| MRR-3| 0.30    | 0.31       | 0.33         |
| MRR-4| 0.20    | 0.23       | 0.24         |
| MRR-5| 0.13    | 0.15       | 0.18         |
| MRR-Average | 0.424 | 0.434 | 0.450 |

Figure 6. P-R curve

We can find that under the same recall index, the results of precision are: SeMap-Spread > W2V-Extend > Kw-Base. This is because the Kw-base model only takes keywords-matching as the principle, and the W2V-Extend model introduces related-words set to expand the retrieval conditions, while SeMap-Spread model further utilizes the semantic relationship between EMR to expand the retrieval vision.
Figure 7. MRR results graph

It can be seen from the table and figure that the MRR-1 and MRR-2 values of the three models are basically the same, but for MRR-3, MRR-4, MRR-5 and MRR average indicator, SeMap-Spread > W2V-Extend > Kw-base. This shows that W2V-Extend model is better than Kw-base model in ranking results, and SeMap-Spread model is better than W2V-Extend model. SeMap-Spread model’s ranking strategy for related documents is the most reasonable.

4. Conclusion

This paper focuses on the difficulties of Chinese EMR retrieval. In order to solve the problem of difficult keyword hit and limited retrieval coverage, a retrieval model for Chinese EMR based on semantic knowledge map is proposed. This model realizes retrieval by searching for semantic association of EMR texts, and further improves the ranking strategy of retrieval results to achieve efficient retrieval of Chinese EMR.

The new retrieval scheme proposed in this paper has the following advantages:

1. The semantic knowledge map of EMR is constructed in advance as the model background, which makes the domain pertinence much higher than that of general information retrieval.

2. The medical information knowledge map takes keywords and related words as the connecting nodes, and organizes all the synonyms together, so it is unnecessary to consider the standardization of medical terms when preprocessing retrieval statements.

3. Semantic knowledge map of medical information is a kind of knowledge organization mode with strong relevance. Compared with the retrieval strategy based on content matching, the scheme of this paper is based on the knowledge map for crawling query, which covers a wider range of retrieval results.

4. The model rearranges the retrieval results by calculating the number and length of target paths on the map, which makes the most relevant documents rendered at a higher position.

However, there are still some shortcomings in this method. In the follow-up research, we should further consider reducing model noise and establishing path connections more reasonably, so as to avoid query drift.

Acknowledgments

This work was partially supported by a major project of the National Social Science Foundation of China under grant 18ZDA086.
References

[1] Ministry of Health of the People’s Republic of China. (2013) The Basic Specifications of Electronic Medical Records (Trial). http://www.nhc.gov.cn/wjw/gfxwj/201304/a99a0bae95be4a27a8b7d883cd0bc3aa.shtml

[2] Chen, D., Zhang, R., Liu, K., Hou, L. (2018) Knowledge Discovery from Posts in Online Health Communities Using Unified Medical Language System. International Journal of Environmental Research and Public Health, 15: 6.

[3] Yang, S., Zheng, X., Yin, X., Mao, H., Zhao, D. (2020) An Algorithm of Query Expansion for Chinese EMR Retrieval by Improving Expansion Term Weights and Retrieval Scores. Ieee Access, 8: 200063-200072.

[4] Wang, H., Zhang, Q., Yuan, J. (2017) Semantically Enhanced Medical Information Retrieval System: A Tensor Factorization Based Approach. Ieee Access, 5: 7584-7593.

[5] Zhang, R., Chu, F., Chen, D., Shang, X. (2018) A Text Structuring Method for Chinese Medical Text Based on Temporal Information. International Journal of Environmental Research and Public Health, 15: 402-.

[6] Vasanthakumar, E. U., Bond, F. (2018) A Semantic Multi-Field Clinical Search for Patient Medical Records. Cybernetics and Information Technologies, 18: 171-182.

[7] Meng, J., Zhang, R., Chen, D. (2019) Utilizing Narrative Text from Electronic Health Records for Early Warning Model of Chronic Disease. In: 2018 IEEE International Conference on Systems, Man, and Cybernetics. Miyazaki. pp. 4204-4209.

[8] Maree, M., Noor, I., Rabayah, K. S., Belkhatir, M., Alhashmi, S. M. (2020) Head Concepts Selection for Verbose Medical Queries Expansion. Ieee Access, 8: 93987-93999.

[9] Patel, R., Yang, Y., Marshall, I., Nenkova, A., Wallace, B. C. (2018) Syntactic Patterns Improve Information Extraction for Medical Search. In: Association for Computational Linguistic. North American Chapter. pp. 371-377.

[10] Wen, Y., Yuan, H., Zhang, P. (2016) Research on Keyword Extraction Based on Word2Vec Weighted TextRank. In: 2nd Ieee International Conference on Computer and Communications. Chengdu. pp. 2109-2113.

[11] Shao, T., Chen, H., Chen, W. (2018) Query Auto-Completion Based on Word2vec Semantic Similarity. Journal of Physics Conference Series, 1004.

[12] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C. (2016) Neural Architectures for Named Entity Recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego.

[13] Uzuner, O., South, B. R., Shen, S., DuVall, S. L. (2011) 2010 I2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. Journal of the American Medical Informatics Association, 18: 552-556.