Chen, Q. & Li, B. (2018). Retrieval method of electronic medical records based on rules and knowledge graph. In *Proceedings of The 18th International Conference on Electronic Business* (pp. 570-577). ICEB, Guilin, China, December 2-6.

Retrieval Method of Electronic Medical Records Based on Rules and Knowledge Graph

(Full paper)

Qiangui Chen, University of International Business and Economics, China. 13261785787@163.com
Bing Li*, University of International Business and Economics, China. lb0501@126.com

ABSTRACT

Every day, the hospital has accumulated a large number of electronic medical records, which record the doctor's description of various patients' conditions, including a large number of medical knowledge related to the patient's health status. So intelligent retrieval research of EMR is an important extension of the medical field. Its accurate retrieval plays an important guiding role in research in the medical field. In this paper, we come up with the combination of rules and knowledge graph, and experimental results demonstrate that the electronic medical record retrieval method based on rules and knowledge map technology can significantly improve the retrieval effect.

Keywords: EMR, information retrieval, knowledge graph.

*Corresponding author

INTRODUCTION

With the informatization of the medical system, the hospital has accumulated a large number of electronic medical records. These electronic medical records are good learning materials for doctors when they conduct medical research or encounter incurable diseases (Yang *et al.,* 2014). Therefore, accurate retrieval of electronic medical records has important guiding significance for inexperienced doctors. At the same time, it is also of great significance to the development of the entire medical research. However, when doctors describe the condition on an electronic medical record, they are not accustomed to this way of filling in the blanks, but describe them according to their own writing habits. Although this method can accurately record the patient's condition, but different doctors write differently, the terms are different, so it brings a lot of trouble to the medical record retrieval. The current electronic medical record management system has certain defects in retrieval, which are embodied in the following aspects: (1) Focus on patient needs and ignore the needs of doctors. Patients or their families have a clear purpose in querying medical records. Usually, they only need one person's medical record and do not care about other people's medical records. Of course, they do not have the right to inquire. For medical or scientific purposes, doctors need to check the medical records of multiple patients. At present, most electronic medical record management systems only support the direct search of a patient's medical record by name and other information, ignoring the needs of doctors. (2) Ignore the connection between keywords. Only through the most basic keyword matching, although it is possible to search for these subjective items with strong description of the condition, if the contact of the keyword in the sentence is neglected, it will find that the precision has certain limitations. Medical descriptions are relatively small, but the combination is complex, and different combinations may represent different meanings.the complexity of electronic medical record content poses great difficulties for accurate retrieval. Therefore, how to improve the retrieval effect of electronic medical records has become a hot issue in the field of information retrieval.

The retrieval task of the electronic medical record is to filter out the electronic medical record set that meets the requirements according to the query statement. The keywords in these queries usually include descriptions of the organs and symptoms of the disease, such as: left lung lobe, nodules, thyroid, and swelling. These keyword groups represent query intents and there is a clear logical relationship between these keywords. Therefore, this paper proposes a query method based on rules and knowledge graphs. Under the premise that doctors use punctuation correctly, we use the punctuation to semantically segment these texts, and use dependency parsing analysis to judge the logical relationship between keyword groups. Experiments show that this method can significantly improve the retrieval effect of electronic medical records. Finally, using knowledge graph to achieve query reconstruction, the results of the eight groups of experiments show that the knowledge graph can further improve the retrieval effect.

PREVIOUS WORK

In this section, we briefly review the related work. The retrieval effect of text depends on the semantic analysis of the text and the validity of the representation method. One idea of text representation technology is to build semantic models based on knowledge of statistics and probability. In the late 1980s, Deerwester and others (Deerwester *et al.*, 2010) proposed Latent Semantic Analysis (LSA), which extracts the meaning of vocabulary in context from the corpus through statistical methods. Its core idea is statistics. The frequency of occurrence of words, thereby forming a representation matrix of the corpus, and then using the corresponding algorithm to convert the matrix into the weight of the words for the document itself. Later, Hoffman (Hoffman *et al.*, 1999)

proposed Probability Latent Semantic Indexing (pLSI)*et al.*, which solved the problem of synonym and polysemy to some extent. Compared to the LSA, it considers the document to be a mixture of topics based on certain weights, and the words are generated according to probability under each topic. Blei *et al.* (2003) proposed a potential Dirichlet distribution in 2003 for the large computational complexity of pLSI and the uncertainty of the probability calculation of new documents. The basic idea is still based on the corpus-based probability generation model. Potential topics are described, and documents can be represented as a random mix of potential topics. On the basis of LDA, another person conducted a research question and answer in the medical health questionnaire (Seelig *et al.*, 2015), and also used it to test the similarity degree of disease types (Frick *et al.*, 2015), and later appeared many variants, such as Dynamic topic Model (Blei *et al.*, 2006), Approximate distributed-LDA (Ihler *et al.*, 2012) and some studies of simplified probability models (Griffiths *et al.*, 2004; Wallach., 2006) and topic models combined with metadata (Chang *et al.,* 2009; Boydgraber *et al.,* 2010; Wang *et al.,* 2009; Newman *et al.*, 2006).In order to improve the accuracy of the search, some people try to use structured electronic medical records. Although this brings convenience to retrieval, the structured electronic medical records lack flexibility, it is difficult to adapt to complex and variable diseases, and the structured electronic medical records rely on human experiences. It is difficult to fully consider the various expressions of the disease record. In order to avoid this situation, and to improve the search results, there are two method, namely expanding query keywords (Weekamp *et al.,* 2012; Xu *et al.,* 1996; Gao *et al.,* 2013) and keyword weight adjustment (Chang *et al.,* 2006).The keywords are extracted from the medical dictionary and added to the original query to form a new query (Zhu *et al.,* 2011), which can improve the effect of the electronic medical records retrieval. However, such a query can improve the retrieval effect to some extent, but it is easy to bring about the problem of query drift, and the retrieval effect cannot be further improved. Therefore, some people propose a query reconstruction method based on word weight adjustment (Dinh *et al.,* 2011). They adds medical information to the weight adjustment algorithm, and take into account the weight adjustment of medical terms in the query statement (Wang *et al.,* 2016).Experimental results demonstrate that there are improvements in the three indicators of map, bpref and p10.In addition to query reconstruction, some people try to use the machine learning algorithm to calculate the similarity between texts through text presentation. However, most of text presentation methods are based on the statistical characteristics of words such as whether they appear (Rudin *et al.,* 2012; Rudin *et al.,* 2010) or the number of appearance (Qiu *et al.,* 2016; Xie *et al.,* 2016), and do not consider the relationship between words and words, so there are certain defects. Therefore, Word2vec is proposed. The word vector originated from the paper proposed by Hinton in 1986 (Hinton *et al.,* 1989). Later, Bengio summarized a set of neural network language model (NNLM), and formally proposed Word vector (Bengio, 2008a, 2012b; Bengio *et al.,* 2009a, 2011b). Later, more and more researchers have invested in the study of neural network learning. They hope to replace the word frequency statistics and complex feature extraction work with artificial neural network models, and get a digital representation corresponding to each word. Previously, both Bengio's paper and "deep Neural Networks with Multitask Learning" (Collobert & Weston, 2008) published in ICML in 2008, word vectors were a by-product of training, and later in Google's open source Word2vec (Mikolov *et al.,* 2013a, 2013b). There is an efficient way to train word vectors. In 2014, Stanford's Richard published "GloVe: Global Vectors for Word Representation" on EMNLP (Richard *et al.,* 2015). It combines the concept of LSA and the training method of Word2Vec, and proposes a global vocabulary information to make greater use. The contextual semantic information of text data has achieved better results in some practical applications. Finally, the emergence of knowledge graph provides us with a new way of thinking. The ontology-based electronic medical records retrieval method (Zhao *et al.,* 2010) can realize the semantic expansion of the concept, and has improve the precision and recall. However, ontology research is a complex project. Therefore, this paper attempts to use the dependency syntax to analyze the logical relationship between words and words based on the grammatical features of texts; and building a small-scale knowledge graph to further improve the search results.

RETRIEVAL FRAMENWORK BASED ON RULES AND KNOWLEDAGE GRAPH

Retrieval Framework

According to the characteristics of electronic medical records from Shengjing Hospital , this paper proposes an electronic medical record retrieval framework combining rules and knowledge graph. The rules are mainly based on the grammatical features of EMRs, such as punctuation and dependency syntax analysis. The entire search framework is shown in the following figure.

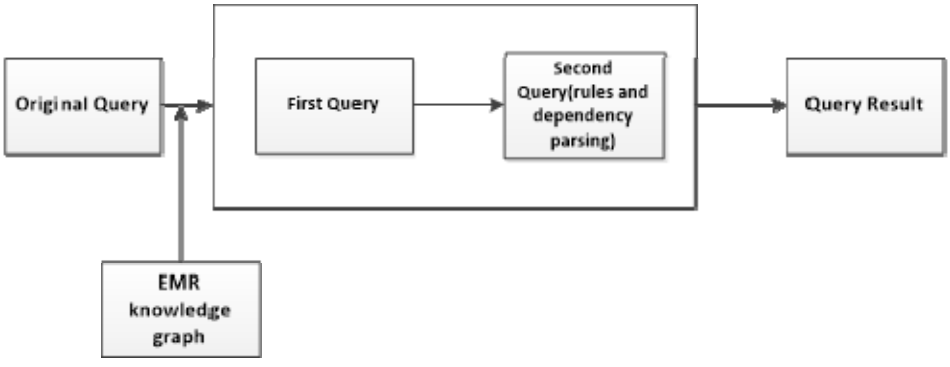

Figure 1. Framework of EMRs retrieval

Step 1: Reconstructing the original query using knowledge graph.
Step 2: Perform the first search in the original data set and get the result1.
Step 3: Perform a second search on result 1, using rules and dependency parsing.
Step 4: Get the query result.

Dependency Parsing Analysis
There are relatively few keywords for a single disease, but the combination is complicated. Sometimes even if we can retrieve the field in a record, it may not be what we want actually. For example, the high-density blurred patch of the upper lobe of the left lung and the high-density nodules of the basal segment of the right lower lobe. When we use keywords to retrieval such as "left lung nodule", we can find a record like that, but that isn't what we want. Keywords mismatch apparently happened here. So the key is to identify the collocation of the keywords "left lung" and "nodule" in the record.

This paper mainly identifies the keyword collocation problem by dependency parsing analysis, and uses punctuation to divide complex sentences into multiple independent simple sentences, thus improving the effect of dependency parsing. The basic steps are as follows:
Step 1: Use the punctuation to divide the long sentence into a sequence of clauses.
Step 2: Determine whether the query keyword is in the same clause, if not, output no; if yes, proceed to step 3.
Step 3: Perform a dependency syntax analysis on the clause to obtain a dependency tree with the highest probability.
Step 4: Determine whether the keyword group has a collocation relationship according to the dependency parsing result. If yes, the output is; otherwise, the output is no.

For example, there are blurred patch on the upper part of the left lower lobe. The dependency parsing result is shown below:

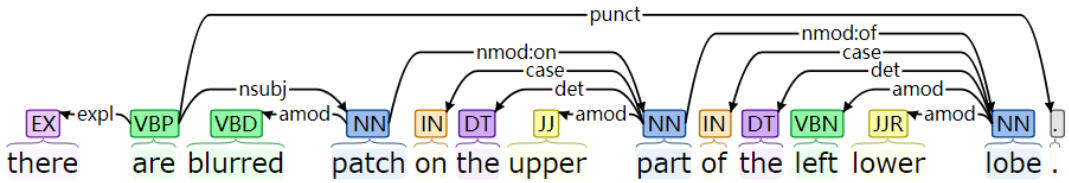

Figure 2. Result of dependency parsing

It can be seen from the figure that there is a dependency between the left lower lobe and the patch, so this record is what we want.

Building EMR Knowledge Graph
Knowledge graph are generally divided into open domain knowledge graph and closed domain knowledge graph. The former involves a wide range and complex relationships, so the current application depth is limited. The latter refers to the knowledge graph applied to a specific domain, and its scope is relatively narrow. It is relatively simple and the application is deeper. In this paper, the knowledge graph constructed is applied to the electronic medical record text.

When describing a patient's condition, it is usually the form of the diseased part and symptoms, such as left lung and nodules, etc. According to the characteristics of its text, the steps of building EMR knowledge graph are as follows:

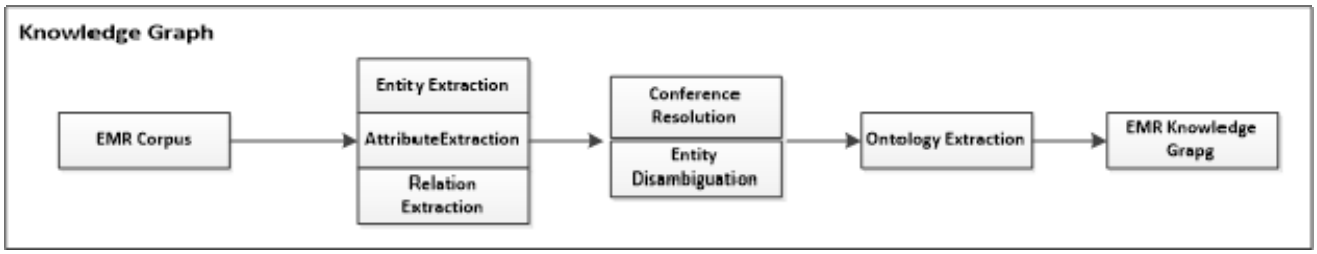

Figure 3.Steps of building EMRs knowledge graph

Entity Extraction
Entity extraction, also called named entity identification, mainly extracts words representing entities in the electronic medical record description. The main steps are word segmentation and part-of-speech tagging.

1. Word segmentation.Firstly, we segment the text of the patients' condition description of the electronic medical records by jieba. At the same time, enhancing the effect of word segmentation through a custom dictionary. For some high-frequency phrases like "left lobe", this article classifies it as a word phrase. In order to identify more such high frequency phrases, it is mainly achieved by the following formula:

$$p(A,B) = \frac{count(A,B) - \theta}{count(A) + count(B) - count(A,B)} \quad (1)$$

For a example, A=left, B=lobe.if left and lobe are often used together, the value of p (A, B) is relatively big, so AB can be seen as a high-frequency phrase.And these two words can be seen one word when segmentation, which could improve the effect of word segmentation. The θ is mainly to prevent the problem that count (A, B) is too small and p (A, B) is too large. For an AB phrase that appears only once, if there is no θ, p (A, B)=1, but this is not a high frequency phrase.

2. Part-of-speech tagging. This paper mainly uses the LTP cloud platform of Harbin Institute of Technology to mark the words in the electronic medical record text. The words are mainly divided into verbs, nouns, conjunctions, adverbs, and punctuation. For example, "high-density patchy shadows in the medial segment of the right lung are essentially absorbed" can be marked as:

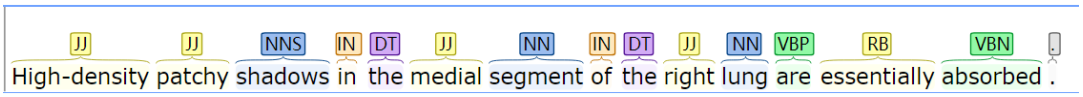

Figure 4. Result of POS tagging

According to the medical dictionary, the words are mainly divided into three categories. 1. words describing the parts of body, such as the left lung, thyroid and so on. 2. words describing the disease, such as nodules, patchy shadows and so on. 3. Adverbs describing the extent of the condition. For an instance, that record can be divided as follows:

Table 1.Example of categories

Categories	Words
Describing the parts of body	The medial segment of right lung
Describing the disease	Patchy shadows
Describing the extent of the condition	High-density, essentially

Relation Extraction

After extracting entities from the text corpus, we obtain the entities. The main purpose of relation extraction is to find the relation between the entities, and connect these individual entities to make it a knowledge network. There are three main relations between entity A and entity B: A contains B, B contains A or has no relationship. Because the research object of this paper is the electronic medical record of the CT scan results of the lungs, the number of entities involved is small, and the relationship between entities mainly depends on manual judgment. For example, "double lungs" include "left lung", "left lung" contains "left lung upper lobe", "left lung" has no relationship with "right lung", etc.

Attribute Extraction

Attributes are inseparable from entities and are a kind of information supplement to entities. Therefore, attribute extraction can be regarded as the relationship extraction between entities and attributes. Here, the attributes of the entity mainly refer to the symptoms, such as "nodule", "plaque shadow" and so on.

Conference Resolution

The main task of conference resolution is to find synonyms that represent entities or attributes. This paper mainly uses word2vec to train the corpus to get the word vector, then put the words together according to the part of speech, and finally select the similar words according to the cosine similarity. In order to ensure the accuracy of the conference resolution, the finally selected synonyms must be manually reviewed.

Entity Disambiguation

The corpus used in this paper is the electronic medical record related to the lungs, and the medical terms are relatively standardized. Therefore, the problem of entity ambiguity does not exist, and the steps of entity disambiguation are omitted.

EXPERIMENTS

In order to test the experimental results of the retrieval method based on rules and knowledge graph, this paper takes about 30000 EMRs related to lungs in Shengjing Hospital as an example. At the same time, the traditional string matching result is used as the benchmark result, and the result of the retrieval is recorded as QF; then the second retrieval result of using the rule and the

dependency syntax analysis is recorded as QS; finally, adding the knowledge graph. the retrieval result is recorded as QT. And eight different sets of queries are set in this paper to compare the retrieval effects of these three methods.

Evaluation Standard

The evaluation indicators used to measure the retrieval results mainly include mean average precision (MAP), binary preference (bpref), and precision. This paper uses precision as the main evaluation indicator.

$$\text{precision} = \frac{\text{the number of related documents in the result}}{\text{the number of documents in the result}} \quad (2)$$

It should be noted that for each retrieval result, the first 100 samples are taken as statistical samples, that is, the precision here is equivalent to the Top-100 precision

Experimental Results and Analysis

In this paper, eight different queries are selected for experiment, and getting QF, QS, and QT in turn. Among them, QF is regarded as baseline. The results of the experiment are shown in the following figure:

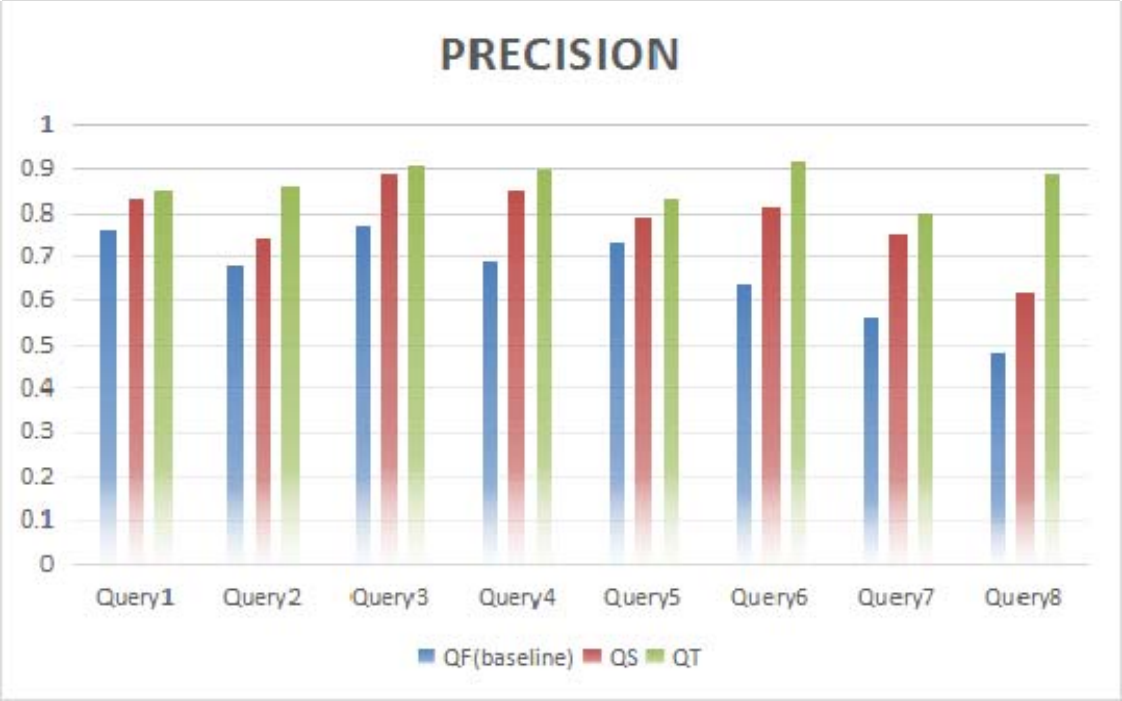

Figure 5. Precision of eight queries result

As can be seen from the above figure, in the eight examples of this experiment, the retrieval effect QS>QF , which indicates that the doctor mostly follows the Chinese grammar rules when describing the condition, so that the method can improve the retrieval effect; QT>QS explains that query reconstruction through knowledge graph can effectively improve the retrieval effect. Specifically, in each of the examples, the retrieval effect is not same. Maybe after the query reconstruction, it may not optimize the query, but add noise, so that the precision improvement is not obvious. See the table below for details.

Table 2. Comparison of eight queries

	QF (baseline)	QS	QT
Query1	0.76	0.83 (+9.21%)	0.85 (+2.41%)
Query2	0.68	0.74 (+8.82)	0.86 (+16.22%)
Query3	0.77	0.89 (+15.58%)	0.85 (-4.49%)
Query4	0.69	0.85 (+23.19%)	0.90 (+5.88%)
Query5	0.73	0.79 (+8.22%)	0.83 (+5.06%)
Query6	0.64	0.81 (26.56%)	0.92 (+13.58%)
Query7	0.56	0.75 (+33.93)	0.80 (+6.67%)
Query8	0.48	0.62 (29.17%)	0.89 (+43.55%)

For example, in Query 3, the retrieval effect QT>QS. This may be because after the query reconstruction, it may not function as a query optimization, but instead add noise, so that the precision improvement is not obvious. Overall, QS's performance improvement is higher than QT, but query8 is the exception. The reason is that query8 is relatively complicated, which indicates that the more complicated the query conditions are, the more important the role of the knowledge graph is. At the same time, the more complex the query, the worse the effect of QF, indicating that the traditional method is not suitable for complex queries. In order to compare the effects of QF, QS and QT intuitively, this paper introduces the concept of APFQ.

$$APFQ = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{3}$$

Pi represents the precision of each retrieval

Table 3.Compare of APFQ

method	APFQ
QT(overall)	0.87
QS	0.78
QF(baseline)	0.66

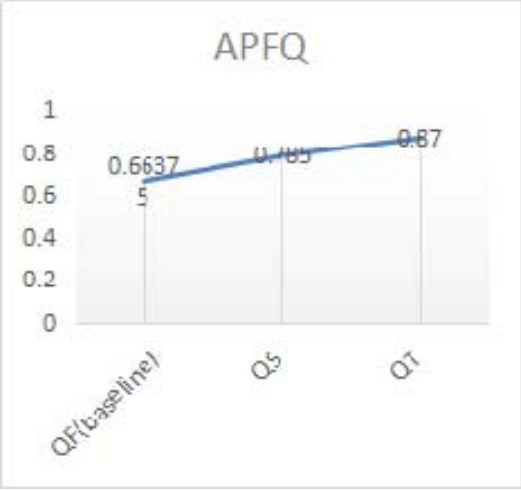

Figure 6.compare of APFQ

As can be seen that the retrieval effect of the combination of knowledge graph and rules is better than the effect based on syntax and rules, better than baseline. All in all, retrieval framework based on rules and knowledge graph can effectively improve the retrieval effect of electronic medical records

CONCLUSION

In the electronic medical record retrieval, this paper proposes a retrieval method combining rules and knowledge graph, which partially solves the problem of synonym and the matching between keyword groups, which makes the results of the query more accurate. It is also proved by experiments that the retrieval effect of this method is better than that of the rule retrieval alone, which makes the electronic medical record play an important role in medical assistant decision-making, providing more learning and guidance for medical staff with poor experience.

LIMITATIONS AND FUTURE WORK

However, the experimental data used in this experiment is relatively simple, and whether it is valid on large-scale data remains to be verified. Moreover, when performing query reconstruction, it is possible to produce noise due to the expansion of semantics, thereby decreasing the accuracy of the query. How to reduce noise when expanding semantics is also the direction that needs to be

continued in the future. Finally, the small-scale knowledge graph constructed in this paper is not suitable for other kinds of diseases, but the idea of this construction can be used for reference. Therefore, building a large-scale knowledge graph that can be applied to more diseases is also the focus and direction of the next step.

ACKNOWLEDGEMENT

This work is supported by National Social Science Fund Project, China (No. 16BTQ065).

REFERENCES

[1] Bengio, Y. (2008). Neural net language models. Scholarpedia, 3 (1), 3881.
[2] Bengio, Y. (2012). Deep Learning of Representations for Unsupervised and Transfer Learning. international conference on machine learning.
[3] Bengio, Y., & Delalleau, O. (2009). Justifying and generalizing contrastive divergence. Neural Computation, 21 (6), 1601-1621.
[4] Bengio, Y., & Delalleau, O. (2011). On the expressive power of deep architectures. algorithmic learning theory.
[5] Blei, M.D. and Ng, Y.A. and Jordan, I.M. (2003) Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.
[6] Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. international conference on machine learning.
[7] Boydgraber, J. L., & Blei, D. M. (2008). Syntactic Topic Models. neural information processing systems, , 185-192.
[8] Chang, J., & Blei, D. M. (2010). Hierarchical relational models for document networks. The Annals of Applied Statistics, 4 (1), 124-150.
[9] Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. international conference on machine learning.
[10] Chang, Y., & Chen, S. (2006). A new query reweighting method for document retrieval based on genetic algorithms. IEEE Transactions on Evolutionary Computation, 10 (5), 617-622.
[11] Dinh, D., & Tamine, L. (2011). IRIT at TREC 2011: Evaluation of query expansion techniques for medical records retrieval.Proceeding of the 20th Text Retrieval Conference Proceeding TREC.
[12] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. . (2010). Indexing by latent semantic analysis. Journal of the Association for Information Science & Technology, 41 (6), 391-407.
[13] Gao, J., Xu, G., & Xu, J. (2013). Query expansion using path-constrained random walks. International ACM SIGIR Conference on Research And Development in Information Retrieval.
[14] Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2004). Integrating Topics and Syntax. neural information processing systems.
[15] Hinton, G. E. . (1989). Learning distributed representations of concepts. Eighth Conference of the Cognitive Science Society.
[16] Hofmann, T. . (1998). Unsupervised learning from dyadic data. Advances in Neural Information Processing Systems, 11.
[17] Ihler, A. T., & Newman, D. (2012). Understanding Errors in Approximate Distributed Latent Dirichlet Allocation. IEEE Transactions on Knowledge and Data Engineering, 24 (5), 952-960.
[18] Jian, Q., Huifang, W., Gaoliang, Y., Bo, Z., Guoping, Z., & Benteng, H. E. . (2016). Text mining technique and application of lifecycle condition assessment for circuit breaker. Automation of Electric Power Systems.
[19] Jin-Feng, Y., Qiu-Bin, Y. U., Yi, G., Zhi-Peng, J., Laboratory, W. I., & Center, L. T. . (2014). An overview of research on electronic medical record oriented named entity recognition and entity relation extraction. Acta Automatica Sinica, 40 (8), 1537-1562.
[20] Frick, J. M., Guha, R., Peryea, T., & Southall, N. (2015). Evaluating disease similarity using latent Dirichlet allocation. bioRxiv, .
[21] Mikolov T, Chen K, Corrado G, *et al.,* Efficient estimation of word representations in vector space.arXiv preprint arXiv:1301.3781, 2013.
[22] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. . (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26, 3111-3119.
[23] Newman, D., Chemudugunta, C., & Smyth, P. (2006). Statistical entity-topic models. knowledge discovery and data mining.
[24] Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global Vectors for Word Representation. empirical methods in natural language processing.
[25] Rudin, C., Waltz, D. L., Anderson, R. N., Boulanger, A., Sallebaouissi, A., Chow, M., ... & Wu, L. (2012). Machine Learning for the New York City Power Grid. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34 (2), 328-345.
[26] Rudin, C., Passonneau, R. J., Radeva, A., Dutta, H., Ierome, S., & Isaac, D. (2010). A process for predicting manhole events in Manhattan. Machine Learning, 80 (1), 1-31.
[27] Seelig, H. P., & Seelig, R. . (2015). Use of a latent topic model for characteristic extraction from health checkup questionnaire data. Methods of Information in Medicine, 54 (06), 515-521.
[28] Weerkamp, W., Balog, K., & de Rijke, M. (2012). Exploiting external collections for query expansion. *ACM Transactions on the Web (TWEB)*, *6* (4), 18.
[29] Wang, W. Gu, J. & Zhou, Z. (2016). Query reconstruction based on word weight adjustment in electronic medical record

retrieval. Journal of Computer Applications and Software, 33 (4), 80-83.
[30] Wang, C., Thiesson, B., Meek, C., & Blei, D. M. (2009). Markov Topic Models. International conference on artificial intelligence and statistics.
[31] Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. International conference on machine learning.
[32] Xu, J., & Croft, W. B. (2017, August). Query expansion using local and global document analysis. In *ACM SIGIR Forum* (Vol. 51, No. 2, pp. 168-175). ACM.
[33] Xie, C., Zou, G., Wang, H., & Jin, Y. . (2016). A new condition assessment method for distribution transformers based on operation data and record text mining technique. China International Conference on Electricity Distribution. IEEE.
[34] Zhu, D., & Carterette, B. . (2012). Improving health records search using multiple query expansion collections. IEEE International Conference on Bioinformatics & Biomedicine. IEEE Computer Society.
[35] Zhao Y, Li W, Bai J. (2010).Electronic Medical Record Retrieval System Based on Ontology. Computer Technology and Development, 20 (3), 211-213.