 doi:10.20944/preprints201708.0055.v1

Article

A Survey of Data Processing of EMR (Electronic Medical Record) Based on Data Mining

SUN Wencheng [1], LIU Fang [1] , CAI Zhiping [1], FANG Shengqun [1], and WANG Guoyan [2]

[1] College of Computer, National University of Defense Technology, Changsha 410073

[2] SysCan Biotechnology Company Limited, Suzhou215000

* Corresponding author: CAI Zhiping (zpcai@nudt.edu.cn)

Abstract: At present, medical institutes generally use EMR to record patient's condition, including diagnostic information, procedures performed and treatment results. EMR has been recognized as a valuable resource for large scale analysis. However, EMR has the characteristics of diversity, incompleteness, redundancy and privacy, which make it difficult to carry out data mining and analysis directly. Therefore, it is necessary to preprocess the source data in order to improve data quality and improve the data mining results. Different types of data require different processing technologies. Most structured data commonly needs classic preprocessing technologies, including data cleansing, data integration, data transformation and data reduction. For semi-structured or unstructured data, such as medical text, containing more health information, it requires more complex and challenging processing methods. The task of information extraction for medical texts mainly includes NER (Named Entity Recognition) and RE (Relation Extraction). In this paper, we introduce the process of EMR processing, including data collection, data preprocessing, data mining, evaluation and knowledge application, analyze the current status of the key technologies, such as data preprocessing and data mining, and provide an overview of the application domains and prospects of EMR mining technologies. Finally, we summarize the existing problems in the research of EMR mining, and review the development trends.

Keywords: EMR; data preprocessing; text mining; information extraction; medical decision support system

1. Introduction

With the development of information technology and HIS (Hospital Information System), EMR has also been popularized. Since 2010, China has introduced "Basic Norms of Electronic Medical Records (For Trial)", "Functional Norms of Electronic Medical Records System (For Trial)" and other policies, to guide hospitals at all levels to build their EMR systems.

EMR (Electronic Medical Record) or EHR (Electronic Health Record), which medical staff use to record texts, symbols, charts, graphics, data and other digital information generated by HIS, refers to medical records, which could be stored, managed, transmitted and reproduced efficiently [1]. With the tremendous growth of the adoption of EMR, various sources of clinical information (including demographics, diagnostic history, medications, laboratory test results, and vital signs, etc) are becoming available, which has established EMR as a treasure trove for large scale analysis of health data.

Data in EMR can be divided into three kinds, structured data, semi-structured data and unstructured data. Structured data, which is generally stored in fixed-mode databases, contains basic information (such as birth data, nationality,etc), drugs taken, allergies, vital signs (such as height, weight, blood pressure, blood type, etc), and so on. Semi-structured data usually has the flow chart

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 August 2017 doi:10.20944/preprints201708.0055.v1

format, similar to RDF (Resource Description Files), including name, value and time-stamp. Unstructured text is one kind of narrative data, including clinical notes, surgical records, discharge records, radiology reports, pathology reports, and so forth. Unstructured texts [2] store a lot of valuable medical information, but lack common structural frameworks, and there are many errors, such as improper grammatical use, spelling errors, local dialects, semantic ambiguities and so on, which increase complexity of data processing and analysis.

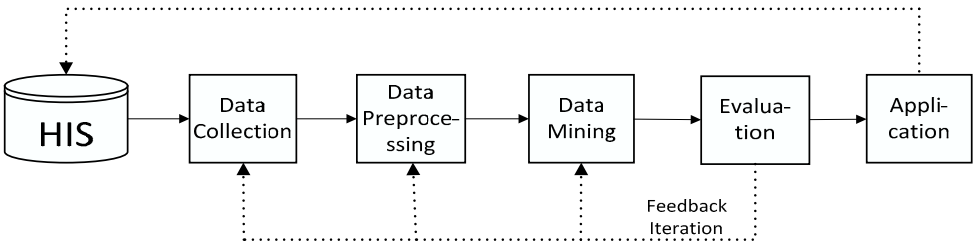

Figure 1. EMR data processing flow

2. General Procedure of EMR data processing

As shown in Figure 1, the general process of EMR data processing includes data collection, data preprocessing, data mining, evaluation and knowledge application.

Data collection is mainly carried out by the government and professional medical institutes. Knowledge application, which is not only the goal of data processing, but also the driving force, is more involved in medical management and treatment program disposal. There are many data mining technologies, such as classification, clustering, association rules, regression, etc. It is only after careful consideration of the data set that we can make a choice and establish a predictive model. Evaluation means that we need to arrange some tests for the model built, in order to grasp its performance. Besides, the patterns and knowledge excavated also need to be analyzed and optimized. Therefore, the data processing is a process of interactive iteration and requires continuous corrective feedback. Only in this way can we get a relatively better knowledge model.

However, it must be pointed out that the data complexity of EMR has made it difficult to analyze data directly, which needs to be effectively preprocessed. High-quality data is more likely to bring high-quality results. According to statistics, in the entire data processing process, the workload of the preprocessing stage is more than 60% [3]. This paper would sum up those diverse preprocessing technologies and analyze the future trends of Chinese EMR.

The characteristics of unstructured medical texts should be taken into account. The common method is to partition the unstructured data reasonably (also called word segmentation), and then store the segmentation results into a standard database. There are many kinds of word segmentation tools both for Chinese and English texts, but the effect of English word segmentation is generally better. This is because there is no spacer between Chinese words , but there is a space, used as the spacer, between English words.

For the moment, the popular Chinese word segmentation systems include ICTCLAS [4], Ansj (ICTCLAS's java implementation), HTTPCWS, SCWS, PhpanAlysis, MMSEG4J [5], PanGu Segment, IKAnalyzer, imdict-chinese-analyzer [6] and LTP -cloud. In the field of Chinese word segmentation, the Institute of Computer Research of the Chinese Academy of Sciences started earlier, and achieved more research findings.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 August 2017 doi:10.20944/preprints201708.0055.v1

3. Classical Data Preprocessing on EMR

Usually the EMR database is composed of a variety of heterogeneous data sources and the data retrieved from EMR database is of diversity, incompleteness and redundancy, which will affect the final mining result to a great extent. Therefore, the EMR data must be preprocessed to ensure that the EMR data is accurate, complete and consistent, and has protected privacy [7]. The process of data preprocessing includes data cleansing, data integration, data transformation, data reduction and privacy protection, as shown in Figure 2. It should be pointed out that the strategies adopted at each stage of the preprocessing are related. Accordingly, the preprocessing methods should be chosen reasonably, especially for medical data.

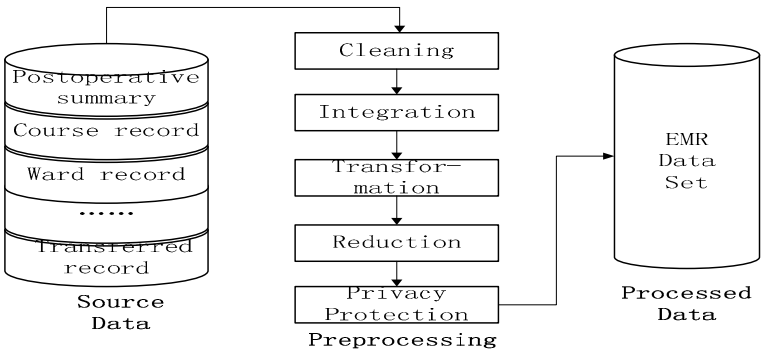

Figure 2. Process of Data Preprocessing on EMR mining

3.1. Data Cleansing

The EMR data, which is incomplete, noisy, and inconsistent, should be improved by filling defaults, smoothing noise and correcting data inconsistency.

1. Default processing

When gathering EMR data, some data attributes maybe lost due to manual errors and system failure [8]. For default data, there are several ways round this. We could ignore missing data, manually fill default values, use attribute averages, fill defaults with the most likely values, or retrieve other data sources.

When the missing value has great influence on the processing process, the missing data is usually ignored. For example, when extracting patient information, if the operation name is lost, the data should be ignored; but if the bed number information is lost, the data cannot be ignored. In the case where the data set is relatively small, the defaults could be manually filled. However, when dealing with larger sets with more defaults, it doesn't work. In addition, this method is time-consuming and costly, so it is generally not applied. In the case where the data distribution is uniform and the cost budget is not much, the defaults could be filled with the attribute averages. Besides, for default data, machine-learning methods can be utilized to determine the optimum value, including regression, Bayesian formal methods, decision tree induction, and so on. Although the prediction may show a relatively great deviation in extreme cases, these methods are still able to better deal with data defaults. Furthermore, when the missing data attribute exists in other data source, the data source should be retrieved.

2. Noise processing

Noise refers to an abnormal attribute value in a data source, also known as an illegal value [9]. For example, the patient has a temperature of 27.8 degrees centigrade, a pH of 3.26 (the normal range is 5.00-9.00), or a specific gravity of urine (SG) of 1.96 (the normal range is 1.01-1.03). The processing of noise data includes binning, regression, outlier analysis and retrieval of other data sources.

 doi:10.20944/preprints201708.0055.v1

The binning methods smooth the ordered data values by examining the values around the data. The key of binning methods is the size of sub-box. The regression method is to modify the noise value by setting up the function model that fits the data attribute value. Outlier analysis is to build clusters by clustering method. The attributes of data points within the same cluster are similar, but the attribute values of data points between different clusters have a large deviation.

3. Inconsistent data processing

There may be inconsistencies in different sources or homologous data, such as inconsistencies in measurement units and recorded values. The inconsistency of data can be corrected by analyzing the correlation between the data and retrieving the different data sources.

3.2. Data Integration

At the data integration stage, the data stored in different data sources needs to be consolidated, and the challenge is to deal with heterogeneous data and its redundancy. Through data integration, the accuracy and speed of data mining can be improved [10].

1. Heterogeneous data processing

EMR data may be collected from multiple EMR systems, and the different data sources will naturally lead to the heterogeneous problems. Heterogeneous problems are mainly represented by inconsistencies in data attributes, such as attribute names and measurement unit. For example, the expression of specific gravity of urine, which can be SG or Specific Gravity; the measurement unit of triglycerides can be mmole/L, but sometimes mg/dl.

2. Redundant data processing

In a nutshell, if an attribute can be derived from other attributes, then the attribute is redundant, which should be cleaned up. Redundancy is mainly reflected in the repeated records of data attributes or inconsistencies in the way of attribute expression. For example, when a patient needs to be transferred to other hospital for treatment, some inspections would be repeated in the latter hospital, which results in repeated medical records, that is redundancy.

Most redundant data can be detected by correlation analysis. When given two attributes, we can analyze how much one attribute has relevance to the other using the existing data. For nominal data, the commonly used analysis method is Chi-square test.

3.3. Data Reduction

On the premise of maintaining data integrity, data reduction can reduce the data set size, which can support data mining in terms of convenience and efficiency. In China, a large amount of EMR would be generated everyday. Given the circumstances, data reduction is quite necessary to perform. Data reduction methods include dimension reduction, quantity reduction and data compression. Among them, dimension reduction, which is easier to fulfill with a better effect, is a relatively popular method.

Dimension reduction method generally controls the size of the data set by reducing the number of random variables or attributes. Dimension reduction method includes wavelet transform and principal component analysis, which project the source data into a smaller data set. Attribute subset selection is also a method of dimension reduction, which reduces the size of a data set by detecting and deleting irrelevant, weak-correlated or redundant attributes or dimensions.

3.4. Data Transformation

Data transformation refers to the conversion of data set into a unified form suitable for data mining. Data transformation methods include smoothing noise, data aggregation and data

 doi:10.20944/preprints201708.0055.v1

normalization. According to the direction and target of data mining, data transformation method filters and summarizes EMR data. Data analysis can be more efficient by having a directional, purposeful data aggregation.

In order to avoid the dependency of the data attributes on the measurement units, data should be normalized to make the data fall into smaller common spaces, such as [0,10], which is more readable. There are three forms of normalization, including min-max normalization, zero-mean normalization, and fractional scale normalization. For neural network algorithms or classification algorithms based on distance measures (such as nearest neighbor classification), the normalization method works better.

3.5 Privacy Protection

Compared with paper medical records, the application of EMR has greatly promoted the development of medical care, but it has also brought a lot of security problems [11]. EMRs contain sensitive information about the patient's privacy, and can be very serious if they are obtained by lawbreakers. In 2011, the Chinese government issued the "Functional Norms of Electronic Medical Records System (For Trial)", and stressed that the EMR system should realize the function of protecting patient's privacy information [12].

There are two main ways to protect the privacy in EMR, including data protection protocols and access control methods [13]. The technical issues involved include data encryption, privacy anonymity processing [14] and access control. In addition, with the emphasis on privacy and sensitive information, the privacy protection system for EMR systems is also gradually established.

4. Information Extraction of EMR Based on Text Mining

Text mining, also known as text data mining, is designed to acquire implicit knowledge that is hidden in the unstructured text. A wealth of valuable information can be discovered from biomedical texts, such as identifying adverse drug reaction or making early judgments about the patient's symptoms.

As shown in Figure 3, the text mining process is usually composed of four stages: information retrieval, information extraction, knowledge discovery [15] and knowledge application. The process of text mining is similar to that of classical data processing. Information retrieval, intended to obtain the desired texts, is similar to data collection. Information extraction is used to extract predefined information, that is, preprocessing of the collected data. Knowledge discovery helps us to extract new knowledge from the text. Knowledge application is the ultimate goal of applying the unknown facts inferred from texts to practice. Medical text mining is mainly for the semi-structured and unstructured texts in the professional medical field, so the traditional preprocessing technology can not be applied directly. The main strategy is to convert semi-structured and unstructured texts into computer-readable structured data by means of information extraction and natural language processing (NLP) technologies. In this process, the key technologies involved include Named Entity Recognition (NER) and Relation Extraction (RE).

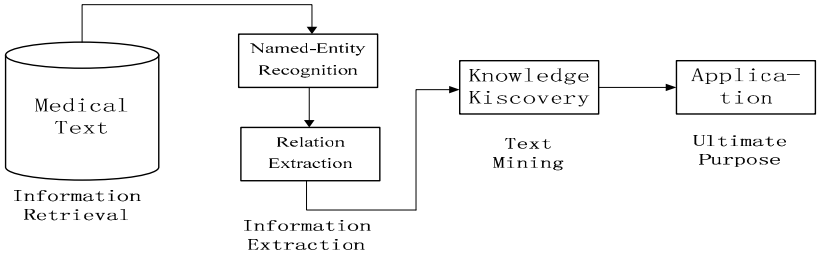

Figure 3. Progress of Text Mining

 doi:10.20944/preprints201708.0055.v1

4.1. Named Entity Recognition Technology

In 1995, the NER task, which refers to the process of identifying particular types of names or symbols in document collections, was introduced for the first time at the MUC-6 (Message Understanding Conference) conference [16]. In the field of EMR, the NER method is used to identify medical entities that have specific significance for the treatment, such as disease names, symptoms, drug names, and so on. Named entity recognition, the basic project of text mining, is an important part of information extraction. NER has two steps, entity boundary identification and entity class determination. In the medical field, NER encounters many obstacles, such as the doctor's writing styles (typos and grammar mistakes), different writing forms of medical terms (such as epilepsy and atrophy, which refer to the same disease) and ambiguity in term abbreviations (such as PC, which can refer to prostate cancer, phosphatidylcholine or personal computer). In addition, some medical terms are composed of phrases or compounds, or modified, which is particularly prominent in Chinese texts. All of these issues will reduce the effect of entity recognition.

Three evaluation indexes serve NER, that is, precision rate (P), recall rate (R) and F-score. P equals the total number of entities identified correctly / the total number of entities identified. R equals the total number of entities identified correctly / the total number of entities present in the test set. F-score equals P * R * 2 / (P + R). F-score, the harmonic average of the precision and recall rate, is a comprehensive evaluation of the test results. When the F-score is higher, the experimental results are better.

In the medical field, NER methods can be divided into three types: the rule-based approach, the dictionary-based approach and the machine-learning approach. The current research in the domain of NER is shown in Table 1.

1. The rule-based NER approach

Rule-based approaches need to identify the rules of the named entity from medical texts, and the identified rules are valid only in specific data sets, otherwise invalid [17]. In addition,the rule-based approach and dictionary-based approach require medical expert assistance to construct rule templates and dictionaries manually.

D Proux et al. [18] proposed a program for the identification of gene symbols and names inside sentences, and the program was made up of a series of sieves of different natures, lexical, morphological and semantic. K Fukuda et al. [19] proposed a new method of extracting material names, PROPER, using surface clue on character strings. It extracted material names in the sentence with 94.70% precision and 98.84% recall. D Hanisch et al. [20] constructed the ProMiner system by a rule-based approach and a pre-processed synonym dictionary, to identify potential name occurrences in the bio-medical text and associate protein and gene database identifiers with the detected matches. In blind predictions, the system achieved an F-measure of approximately 0.8 for the organisms mouse and fly and about 0.9 for the organism yeast.

2. The dictionary-based NER approach

The dictionary-based approach, which is well suited for accurate search, is widely utilized in large-scale biomedical literature annotation and indexing. However, due to the existence of many variants of medical terminology, it is difficult for a single dictionary to cover all of them. This results in that the entities, which are not defined in the dictionary, are missed easily. In view of this issue, the more popular methods are fuzzy dictionary matching method and post-processing method.

Z Yang et al. [21] presented a dictionary-based bio-entity name recognition approach, which expanded the bio-entity name dictionary via the Abbreviation Definitions identifying algorithm, improved the recall rate through the improved edit distance algorithm and adopted some post-processing methods. Y Tsuruoka et al. [22] presented a method using Bayes classifier and an

 doi:10.20944/preprints201708.0055.v1

expanded dictionary with a probabilistic variant generator, and experimental results using the GENIA corpus achieved an F-measure of 66.6%.

3. The machine-learning NER approach

The appropriate machine-learning algorithm is utilized to establish the entity recognition model using the statistical characteristics and parameters of the sample data. The machine-learning approach, which is data-driven and application-oriented, requires standard annotations training data set. Various machine-learning methods, such as Hidden Markov Models (HMM), Support Vector Machines (SVM), Conditional Random Field (CRF) and Maximum Entropy (ME), are available according to data characteristics. Among a variety of machine learning algorithms, CRF methods are more popular because they allow for the incorporation of various features that can be advantageous for the process of sequence labeling [23].

Zhou et al. [24] constructed a named entity recognition system in the biomedical domain by means of HMM and SVM algorithms to explore the widely used lexical-level features and the name alias phenomenon. Kazama et al. [25] used SVMs to identify proteins, DNA, cell types, cell lines and lipids, with a F-measure of 73.6%. Tsai et al [26] constructed a NER framework with CRF. On the GENIA 3.02 corpus, this system achieved an F -score of 78.4% for protein names. Lin et al. [27] proposed a hybrid method that used maximum entropy (ME) as the underlying machine learning method incorporated with dictionary-based and rule-based methods for post-processing, with an F-measure of 72%. Hsu et al. [28] trained bi-directional CRF models to extract tagging gene and gene product, with an F -score of 88.3%. Li et al. [29] constructed a CRF model to classify and identify the gene name entities with an F -measure of 89.1%.

Due to the professionalism and exclusivity of medical terms, the best F-score of biomedical NER systems is generally not particularly good [30]. By combining different approaches, the NER system would be improved, such as post-processing method or addition of the knowledge base.

Table 1. Summary of Named Entity Recognition Research

NER TYPE	REF.	Methods	Application	F-score	YEAR
Rule-based Approach	[18]	POS tagging	Gene name recognition	92.5%	1998
	[19]	Dependency rule	Target entity recognition	96.75	1998
	[20]	Regular expressions	Biomedical entity recognition	0.8-0.9	2005
Dictionary-based Approach	[21]	Post processing	Named entity recognition	53.7%	2008
	[22]	Naive Bayes	Named entity recognition	66.8%	2005
Machine-learning Approach	[24]	HMM	Bio-medical entity recognition		2004
	[25]	SVMs	Bio-medical entity recognition	73.6%	2004
	[26]	CRF	Protein name recognition	78.4%	2006
	[27]	ME	Biological entity recognition	72%	2004
	[28]	CRF	Gene marker	88.3%	2008
	[29]	CRF	Gene named entity recognition	89.1%	2009

 doi:10.20944/preprints201708.0055.v1

4.2 Research Progress of Relation Extraction (RE)

When the named entity is identified, the next task is to extract entity relation. According to the I2B2 2010 evaluation conference [31], the entity relations in EMR can be divided into three categories, including the relation between diseases, the relation between diseases and medical examinations, and the relation between diseases and treatment. In addition, the entity relation is limited to the relation between two named entity within a sentence [32].

In the medical field, three common methods are applied to extract entity relation, including co-occurrence-based [33], pattern-based [34], and machine-learning approaches [35]. When two entities appear in the same sentence, there is a co-relation between the two entities. The higher the frequency of co-occurrence, the stronger the relation. The most widely used method is the machine-learning approach. In addition, the hybrid system of two or more approaches is also developed gradually. For example, in order to handle more complex sentence structures and achieve better performance, a machine-learning system based on a knowledge base or feature dictionary would be proposed. Table 2 shows the current researches of RE.

Ben Abacha et al. [36] proposed a medical text annotation and extraction platform, MeTAE, which identified medical entities on the basis of MetaMap [37] and linguistic patterns. Chun et al. [38] applied a machine-learning method and a special dictionary, constructed from six public databases, to build a NER system that extracts disease-gene relations from Medline [39]. Shetty et al. [40] used the MeSH index terminology to establish a statistical document classifier based on MedLine to detect adverse drug reactions, which could complement current drug safety methods. Srinivasan et al. [41] constructed the mining function using the vector space model and applied this function towards exploring trends in disease research.

J Björne et al. [42] proposed a TEES event extraction system using the SVM classifier and the rule-based post-processing method, to extract the relations between genes and proteins from biomedical literature. In 2013, J Björne et al. [43] also presented the TEES version 2.1, which introduced an automated annotation scheme learning system. TEES 2.1 had good performance across the BioNLP 2013 task corpora, which was suitable to promote.

Li et al. [44] adopted a co-occurrence-based text mining approach through Bayesian and SVM methods to choose protein pairs with physical interactions from mouse protein-protein interaction database (MppDB). Taiai [45] proposed a text mining and visualization framework for finding details of that occur within a particular cellular location and the sequence of the amino acids at the interface of interaction. Lee et al. [46] proposed HiPub, a seamless Chrome browser plug-in that automatically recognized, annotated and translated biomedical entities from texts into networks for knowledge discovery.

Table 2. Summary of Relation Extraction Research

REF.	**System**	**Methods**	**Application**	Year
[36]	MeTAE	Semantic pattern	Treatment-disease relation extraction	2011
[38]		Machine learning	Gene-disease relation extraction	2006
[40]	Classifier	MeSH index	Adverse drug reaction	2011
[41]		Vector space model	Disease-disease relation extraction	2003
[42]	TEES	SVM	Gen- protein relation extraction	2009
[43]	TEES 2.1	Automatic analysis	Gen- protein relation extraction	2013

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 August 2017 doi:10.20944/preprints201708.0055.v1

[44]		Naive Bayes	Gen- protein relation extraction	2010
[45]		Visualization	Protein-protein relation extraction	2011
[46]	HiPub	Context feature	Medical entity relation extraction	2016

Table 3. Summary of Text Mining Research

REF.	System	Methods	Domain Oriented	Year
[47]	CRAB	Pipeline tool	Cancer risk assessment	2012
[48]	PKDE4J	Dictionary and rule based	Text mining system	2015
[49]	BeFree	Data priority strategy	Identification and extraction of drugs, genes and diseases	2015
[50]		Path statistics	Study on susceptibility of rectal cancer	2012
[51]	PWTEES	TEES；PathNER	Improving the understanding of molecular pathogenesis	2015
[52]		Subtype classification method	Identification of key genes and pathways	2014

4.3 Research Progress in Medical Text Mining

At present, the research direction of text mining is to integrate named entity recognition and relation extraction into a tool or system, which is convenient and efficient. However, the universal tools generally work worse in biomedical field for its complex and redundant specialized terminologies. Table 3 shows the current researches of text mining in medical field.

CRAB was a fully integrated text mining tool, developed by Korhonen et al. [47], to extract relevant data from the literature and to assess cancer risk through knowledge discovery techniques. Song et al. [48] proposed a comprehensive text mining system called PKDE4J, which integrated dictionary-based entity extraction and rule-based relation extraction in a highly flexible and extensible framework. Alex Bravo et al. [49] proposed the Befree system using data prioritization strategies, to automatically identify gene-disease, drug-disease and drug-target associations, which performed well in the MedLine database. Nam et al [50] utilized a text mining technique to integrate the current work, using a sub-pathway-based statistical model, to reveal the susceptibility of early colorectal cancer.

C Wu et al. [51] presented a system, PWTEES using an existing event extraction tool (TEES) and pathway named entity recognition (PathNER), to improve the understanding of the pathogenesis of thyroid disease. In addition, C Wu et al. [52] also developed a large-scale text mining system using a subtype classification method for the thyroid cancer literature, to generate a molecular profiling of thyroid cancer subtypes and develop targeted therapy.

5. Application of EMR Mining technology

Knowledge in EMR databases can be discovered using different data mining technologies [53]. In general, there are three main areas of application.

 doi:10.20944/preprints201708.0055.v1

1. Medical decision support and disease risk prediction

It takes a lot of effort to keep doctors in possession of everything about the patient's treatment. Although medical experts will do their best to provide diagnosis and treatment, the suggestion is still based on the subjective judgment of clinical experience, and misdiagnosis and missed diagnosis are therefore likely to appear. Medical decision support [54] systems allow medical experts to gain advice on the treatment plan of the symptoms, which is based on factual data [55]. If the mechanism can be further developed and applied, it will play an important role in medical experts' diagnosis of diseases, especially for doctors who have less clinical experience.

Medical decision support system (MDSS) is rapidly applied, and some systems, such as the Archimedes IndiGo system [56], the Auminence system, the Micromedex system developed by Thomson Reuters, the Zynx medical system developed by Scott Weingarten, and the DXplain system [57] developed by the Massachusetts General Hospital (Boston), have been put into use. Take the DXplain system as an example. DXplain has been widely used since its inception in 1986, and the system is still widely used in US hospitals until now [58]. DXplain uses a pseudo probability algorithm to generate a sequence of diseases by inputting patient signs, disease symptoms, laboratory results, and clinical treatment. In the experiment for the accuracy of DXplain diagnosis [59], DXplain's performance was affirmed. DXplain's knowledge system has expanded from the initial approximately 500 diseases to 2400 now, and there are 5000 new clinical findings and 230000 data points.

In china, the research on medical decision support systems also has preliminary development [60], [61], which have been applied in a small range. For example, HanDan Central Hospital has deployed medical decision support systems in more than 70 departments in both the East and West districts [62]. However, most medical institutes are still improving their hospital information system, and the promotion of medical decision support system is still in the theoretical and experimental research stage.

Besides, risk prediction models [63] can also be constructed that assist doctors to judge the possibility of disease deterioration or improvement and provide better health-care for patients with limited medical resources. Furthermore, patients can also reasonably purchase medical insurances to reduce medical costs.

2. Mobile health, network medical treatment and personalized health-care

Relying on the support of the government and the commercial operation, the mobile health system [64] takes EMR system as the core, based on medical facts rather than experience. Mobile health [65] and network medical treatment [66] can greatly simplify the work of hospital staff and make it more convenient to seek medical advice and more accurate to grasp physical quality.

In addition, for the concern about health, people are increasingly interested to participate in their own medical decision-making [67]. Personalized health-care will take into account the views of patients, and formulate treatment plans and nursing methods, more in line with the actual situation of patients, such as personalized nutrition catering.

3. Disease evolution prediction and drug reactions detection

Traditional disease and drug knowledge discovery costs the huge space-time price. However, the medical data mining technology can quickly find out the medical trajectory of the disease over time and study its natural history, with the auxiliary role for disease diagnosis and treatment. For example, in some areas with high incidence of epidemic diseases, the risk factors can be accurately identified by medical data mining technologies. In addition, after the development of new drugs, considerable funds and energy would be invested to study their effects [68], but the medical data mining technology can detect adverse drug events (ADEs) in a cost-effective way [69].

 doi:10.20944/preprints201708.0055.v1

6. Conclusions

On the development of the Chinese EMR data processing, we believe that the following aspects need attention.

1. Public annotated corpus

Under the government regulations, the quantity and quality of EMR are gradually improved. However, the lake of sufficient public annotated corpus that results in the lack of the Chinese EMR processing tools and clear research tasks, is the biggest obstacle to the Chinese EMR study. Therefore, the establishment of a set of hierarchical and complete Chinese public corpus is imperative. The English corpus study is more mature and systematic, so we can learn their technical implementation methods.

2. Professional dictionary and knowledge base

The truth is that the study of medical dictionary and knowledge base is far behind other professional fields. Many institutes have published their own medical dictionaries, but the little useful content can not meet the application requirements. In addition, the dictionary quality needs more appraisal and certification of specialized agencies. So, the standardization of dictionary in medical field is worthy for attention.

3. Privacy protection

With the deepening of EMR research, the communication between hospitals and research institutes will increase in the future and data transmission of EMR is bound to be more frequent, so more emphasis should be attached to the protection of personal privacy in EMR. However, the current simple methods, such as anonymization or security protocols, can't meet the market demand, which needs a more manageable data protection system.

4. Reasonable selection of processing tools

Processing tools should be selected according to the characteristics of EMR data and the principles of data sets design. The designed method which is of great performance in general contexts may appear performance variation in biomedical field. We used the similarity modeling algorithm Word2vec and the word segmentation tool Ansj, to deal with pneumonia data. This method eventually achieved an accuracy rate of 25%, which presented a very poor result.

In the future, the larger scale and more complex structure of EMR will make it harder to process data in EMR, but the social and economic benefits it brings will be more remarkable and the EMR research will play a greater role in the medical field.

References

1. Ministry of Health of the People0s Republic of China. The basic specifications of electronic medical records (trial)[OL]. [2017.4.6] http://www.gov.cn/zwgk/2010-03/04/content_1547432.htm（in Chinese）
2. Feldman K, Hazekamp N, Chawla N V. Mining the Clinical Narrative: All Text are Not Equal[C]// 2016 IEEE International Conference on Healthcare Informatics (ICHI). IEEE Computer Society, 2016:271-280.
3. Han J, Kamber M. Data Mining: Concepts and Techniques, Morgan Kaufmann[J]. Machine Press, 2001 (in Chinese), 2006, 5(4):394-395
4. Zhang H P, Yu H K, Xiong D Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS[C]// Sighan Workshop on Chinese Language Processing. Association for Computational Linguistics, 2003:págs. 758-759.
5. Huang R. rmmseg4j: R interface to the Java Chinese word segmentation system of mmseg4j[J]. International Journal of Radiation Oncologybiologyphysics, 2012, 66(1):83-90.
6. Huang Y B. A Comparative Study of Chinese Word Fingerprints for Lucene Interface[J]. Science & Technology Information, 2012(12):246-247.(in Chinese)
7. Kop R, Hoogendoorn M, Teije A T, et al. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records[J]. Computers in Biology & Medicine, 2016, 76:30-38.
8. Hu Z, Melton G B, Simon G J. Strategies for Handling Missing Data in Detecting Postoperative Surgical Site Infections[C]// International Conference on Healthcare Informatics. IEEE, 2015:499-499.
9. Lai K H, Maxim T, Goss F R, et al. Automated misspelling detection and correction in clinical free-text records.[J]. Journal of Biomedical Informatics, 2015, 55(C):188-195.
10. Monroe M, Lan R, Lee H, et al. Temporal event sequence simplification.[J]. IEEE Transactions on Visualization & Computer Graphics, 2013, 19(12):2227-36.
11. Xu Y H, Zhou T S, Tian Y, et al. Application of Chinese medical document anonymization in EMR system[C]// IEEE International Conference on Signal Processing, Communications and Computing. IEEE, 2015:1-4.
12. Ministry of Health of the People0s Republic of China. The functional specifications of electronic medical records system(trial)[OL]. [2017.4.8] http://www.gov.cn/gzdt/2011-01/04/content_1778059.htm(in Chinese)
13. Yarmand M H, Sartipi K, Down D G. Behavior-Based Access Control for Distributed Healthcare Environment[C]// IEEE International Symposium on Computer-Based Medical Systems. IEEE, 2008:126-131.
14. Alhaqbani B, Fidge C. Privacy-preserving electronic health record linkage using pseudonym identifiers[C]// International Conference on E-Health Networking, Applications and Services, 2008. Healthcom. IEEE Xplore, 2008:108-117.

15. Zhu F, Patumcharoenpol P, Zhang C, et al. Biomedical text mining and its applications in cancer research.[J]. Journal of Biomedical Informatics, 2013, 46(2):200-211.

16. Grishman R, Sundheim B. Message Understanding Conference-6: a brief history[C]// Conference on Computational Linguistics. Association for Computational Linguistics, 1996:466-471.

17. Rebholzschuhmann D, Jimeno Y A, Li C, et al. Assessment of NER solutions against the first and second CALBC Silver Standard Corpus.[J]. Journal of Biomedical Semantics, 2011, 2(5):1-12.

18. Proux D, Rechenmann F, Julliard L, et al. Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction.[C]// CiteSeer, 1998:248-255.

19. Fukuda K, Tamura A, Tsunoda T, et al. Toward information extraction: identifying protein names from biological papers.[C]// Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. Pac Symp Biocomput, 1998:707-718.

20. Hanisch D, Fundel K, Mevissen H T, et al. ProMiner: rule-based protein and gene entity recognition[J]. Bmc Bioinformatics, 2005, 6 Suppl 1(Suppl 1):S14.

21. Yang Z, Lin H, Li Y. Brief Communication: Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature[J]. Computational Biology & Chemistry, 2008, 32(4):287.

22. Tsuruoka Y, Tsujii J. Improving the performance of dictionary-based approaches in protein name recognition[J]. Journal of Biomedical Informatics, 2005, 37(6):461-470.

23. Munkhdalai T, Li M, Kim T, et al. Bio Named Entity Recognition Based on Co-training Algorithm[C]// International Conference on Advanced Information NETWORKING and Applications Workshops. IEEE, 2012:857-862.

24. Guodong Z, Jian S. Exploring deep knowledge resources in biomedical name recognition[C]// International Joint Workshop on Natural Language Processing in Biomedicine and ITS Applications. Association for Computational Linguistics, 2004:96-99.

25. Kazama J, Makino T, Ohta Y, et al. Tuning Support Vector Machines for Biomedical Named Entity Recognition[C]// In Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain. 2004:1--8.

26. Tsai T H, Chou W C, Wu S H, et al. Integrating linguistic knowledge into a conditional random fieldframework to identify biomedical named entities[J]. Expert Systems with Applications, 2006, 30(1):117-128.

27. Lin Y F, Tsai T H, Chou W C, et al. A Maximum Entropy Approach to Biomedical Named Entity Recognition.[C]// ACM SIGKDD Workshop on Data Mining in Bioinformatics. DBLP, 2004:56-61.

28. Hsu C N, Chang Y M, Kuo C J, et al. Integrating high dimensional bi-directional parsing models for gene mention tagging.[J]. Bioinformatics, 2008, 24(13):i286-94.

29. Li Y, Lin H, Yang Z. Incorporating rich background knowledge for gene named entity classification and recognition[J]. Bmc Bioinformatics, 2009, 10(1):1-15.

30. Hong-Jie, Chang, Richard, et al. New Challenges for Biological Text-Mining in the Next Decade[J]. Journal of Computer Science and Technology, 2010, 25(1):169-179.

31. Özlem Uzuner, South B R, Shen S, et al. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text[J]. Journal of the American Medical Informatics Association Jamia, 2011, 18(5):552.

32. Yang JF, Yu QB, Guan Y, et al. An Overview of Research on Electronic Medical Record Oriented Named Entity Recognition and Entity Relation Extraction[J]. Acta Automatica Sinica, 2014, 40(8):1537-1562.(in Chinese)

33. Jelier R, Jenster G, Dorssers L C, et al. Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes.[J]. Bioinformatics, 2005, 21(9):2049-2058.

34. Auger A. Pattern-based approaches to semantic relation extraction: a state-of-the-art[J]. Terminology, 2008, 14(1):1-19.

35. Song M, Yu H, Han W S. Combining active learning and semi-supervised learning techniques to extract protein interaction sentences[J]. 2011, 12 Suppl 12(12):S4.

36. Abacha A B. Automatic extraction of semantic relations between medical entities: a rule based approach[J]. Journal of Biomedical Semantics, 2011, 2(5):S4.

37. Aronson A R, Lang F. An overview of MetaMap: historical perspective and recent advances[J]. Journal of the American Medical Informatics Association, 2010, 17(3):229-236.

38. Chun H W, Tsuruoka Y, Kim J D, et al. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning.[J]. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 2006, 11:4.

39. Craig T B. MEDLINE.[J]. Journal of Family Practice, 1991, 33(2):128.

40. Shetty K D, Dalal S R. Using information mining of the medical literature to improve drug safety[J]. Journal of the American Medical Informatics Association, 2011, 18(5):668-674.

41. Srinivasan P, Wedemeyer M. Mining Concept Profiles with the Vector Model or Where on Earth are Diseases being Studied?[J]. 2003.

42. Björne J, Heimonen J, Ginter F, et al. Extracting Complex Biological Events with Rich Graph-Based Feature Sets[C]// The Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. OAI, 2009:10-18.

43. Björne J, Salakoski T. TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task[C]// Bionlp Shared Task 2013 Workshop. 2013:16-25.

44. Li X, Cai H, Xu J, et al. A mouse protein interactome through combined literature mining with multiple sources of interaction evidence.[J]. Amino Acids, 2010, 38(4):1237-1252.

45. Tsai F S. Text mining and visualisation of Protein-Protein Interactions.[J]. International Journal of Computational Biology & Drug Design, 2011, 4(3):239-244.

46. Lee K, Shin W, Kim B, et al. HiPub: translating PubMed and PMC texts to networks for knowledge discovery[J]. Bioinformatics, 2016, 32(18):btw511.

47. Korhonen A, Séaghdha D O, Silins I, et al. Text mining for literature review and knowledge discovery in cancer risk assessment and research.[J]. Plos One, 2012, 7(4):e33427.

48. Song M, Kim W C, Lee D, et al. PKDE4J: Entity and relation extraction for public knowledge discovery.[J]. Journal of Biomedical Informatics, 2015, 57:320-332.

49. Àlex Bravo, Piñero J, Queraltrosinach N, et al. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research[J]. BMC Bioinformatics, 2015, 16(1):55.

50. Nam S, Park T. Pathway-based evaluation in early onset colorectal cancer suggests focal adhesion and immunosuppression along with epithelial-mesenchymal transition.[J]. Plos One, 2012, 7(4):e31685.

51. Wu C, Schwartz J M, Brabant G, et al. Constructing a molecular interaction network for thyroid cancer via large-scale text mining of gene and pathway events[J]. BMC Systems Biology, 2015, 9(6):1-10.

52. Wu C, Schwartz J M, Brabant G, et al. Molecular profiling of thyroid cancer subtypes using large-scale text mining[J]. BMC Medical Genomics, 2014, 7(3):1-11.

53. Tekieh M H, Raahemi B. Importance of Data Mining in Healthcare: A Survey[C]// Ieee/acm International Conference. 2015:1057-1062.

54. Li P, Yates S N, Lovely J K, et al. Patient-like-mine: A real time, visual analytics tool for clinical decision support[C]// IEEE International Conference on Big Data. 2015:2865-2867.

55. Potters L, Raince J, Chou H, et al. Development, Implementation, and Compliance of Treatment Pathways in Radiation Medicine[J]. Frontiers in Oncology, 2012, 3:105.

56. Bellows J, Patel S, Young S S. Use of IndiGO individualized clinical guidelines in primary care[J]. Journal of the American Medical Informatics Association, 2014, 21(3):432-437.

57. Barnett G O, Cimino J J, Hupp J A, et al. DXplain. An evolving diagnostic decision-support system.[J]. Jama the Journal of the American Medical Association, 1987, 258(1):67-74.

58. London S. DXplain: a Web-based diagnostic decision support system for medical students.[J]. Medical Reference Services Quarterly, 1998, 17(2):17-28.

59. Feldman M J, Barnett G O. An approach to evaluating the accuracy of DXplain.[J]. Computer Methods & Programs in Biomedicine, 1991, 35(4):261-6.

60. Liu Y, Wei L, Yao Z, et al. The Practice and Experience of Emergency Information System Construction[J]. China Digital Medicine, 2016, 11(5):53-55.(in Chinese)

61. Zhang Y, Yan X, Gao X, et al. Demand Analysis of Decision Support System of Grass-roots Health[J]. Chinese General Practice, 2016, 19(22):2636-2639.(in Chinese)

62. Shao W, Wang Y, Yan GT, et al. Research on Construction of a Clinical Decision Making Support System[J]. China Medical Devices, 2016, 31(8):87-88.(in Chinese)

63. Choi E, Du N, Chen R, et al. Constructing Disease Network and Temporal Progression Model via Context-Sensitive Hawkes Process[J]. 2015:721-726.

64. Lomotey R K, Deters R. Efficient mobile services consumption in mHealth[C]// Ieee/acm International Conference on Advances in Social Networks Analysis and Mining. ACM, 2013:982-989.

65. Kai E, Rebeiro-Hargrave A, Inoue S, et al. Empowering the Healthcare Worker Using the Portable Health Clinic[J]. 2014:759-764.

66. Carchiolo V, Longheu A, Malgeri M. Personal Health Record feeding via Medical Forums[C]// IEEE, International Conference on Computer Supported Cooperative Work in Design. IEEE, 2015.

67. Zengtreitler Q, Gibson B, Hill B, et al. The effect of simulated narratives that leverage EMR data on shared decision-making: a pilot study[J]. BMC Research Notes, 2016, 9(1):359.

68. An L, Ravindran P P, Renukunta S, et al. Co-medication of pravastatin and paroxetine-a categorical study.[J]. Journal of Clinical Pharmacology, 2013, 53(11):1212–1219.

69. Karlsson I, Henrik Bostrom. Predicting Adverse Drug Events Using Heterogeneous Event Sequences[C]// 2016 IEEE International Conference on Healthcare Informatics (ICHI). IEEE Computer Society, 2016:356-362.