IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Research on the Aided Diagnosis Method of Diseases Based on Domain Semantic Knowledge Bases

DEYAN CHEN[1,2,3,4], HONG ZHAO[1,2] and XIA ZHANG[1,2,3]

[1]School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
[2]National Engineering Research Center for Computer Software (Northeastern University), Shenyang 110179, China
[3]Neusoft Institute of Intelligent Healthcare Technology, Co. Ltd., Shenyang 110179, China
[4]Healthcare IT Division, Neusoft Corporation, Shenyang 110179, China

Corresponding author: D. Chen (e-mail: chendeyan@neusoft.com).

This work was supported in part by the Key projects of National Natural Science Foundation of China under Grant 61232015, the National High-Tech Research and Development Plan of China under Grant 2015AA020103 and the National Key Research and Development Program of China under Grant 2016YFC1303000.

ABSTRACT The health care domain is a knowledge-intensive domain. The quality of clinical diagnosis relies mainly on the medical knowledge and experience held by doctors. However, the ability of a single doctor is very limited, so the quality of clinical diagnosis is currently not high. In this paper, an aided diagnosis method based on domain semantic knowledge bases is proposed. Firstly, a domain semantic knowledge base is established by extracting and refining the knowledge of the medicine subject matter domain from the Freebase RDF dumps. Then, based on the semantic knowledge base, the algorithms for calculating the weights of the symptoms in the knowledge base, the relative weights of the diseases related to the input symptom set from a patient, and the related symptom set related to the input symptom set from the patient are proposed. Finally, the clinical medical record data of several common diseases are selected to make an evaluation on the proposed method. For each medical record, the symptom information is extracted from the chief complaint as the patient's input symptom set. Based on the input symptom set, the method of this paper is used to obtain the list of related diseases and the ranking of disease relative weights. From the disease relevance rankings, the Top 1 (first diagnosis) and Top-3 (first 3 diagnoses) are compared with the doctor's diagnoses in the medical records. Among them, *ovarian cyst* has the highest Top-1 and Top-3 hit rates of 67.3% and 89.1%, respectively. Followed by *acute upper respiratory tract infection*, Top-1 and Top-3 hit rates are 56.6% and 85.2%, respectively. The average Top-1 and Top-3 hit rates are 47.9% and 79.7%, respectively. Compared with the relevant methods, the method of this paper is better. The evaluation results show that based on the domain semantic knowledge base and the aided diagnosis method of diseases constructed in this paper, it is possible to provide aided diagnosis services of a large number of common diseases for general practitioners (especially inexperienced doctors) at the grass-roots level as well as self-diagnosis services of diseases for patients.

INDEX TERMS Ontology, domain semantic knowledge bases, aided diagnosis of diseases, symptom weights, disease relative weights, related symptoms

I. INTRODUCTION

Diagnosis of diseases is one of the most important aspects of medical activities. It provides a solid foundation for the treatment and prognosis of patients [1]. The quality of disease diagnosis depends mainly on the medical knowledge and medical experience that doctors have mastered. However, the individual doctor's medical knowledge and medical experience are still limited. How to improve the level of clinical diagnosis and treatment of doctors (especially inexperienced doctors) and reduce the workload of doctors is a problem that needs to be solved urgently. In the early days, research in this area mainly focused on expert systems. The idea of the expert systems is to formalize experts' experience and knowledge and to use them to diagnose. Extracting

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2875722, IEEE Access

empirical knowledge from experts is a labor-intensive task. Because experts' diagnosis process is often intuitive, many experts are unable to provide this kind of empirical knowledge with direct causation.

With the development of medical science and the improvement of hospital informatization, a large amount of clinical knowledge and electronic medical record data have been accumulated in the clinic. Doctors' experience in diagnosis and treatment is also hidden in these medical records. Correspondingly, the emergence and development of Cloud Computing, Big Data, AI (Artificial Intelligence), and other technologies provide favorable support for the mining and utilization of these data. Under these favorable conditions, computer-aided disease diagnosis and prediction research based on data mining and machine learning [2]-[5] algorithms has mushroomed. However, most of these studies are aimed at a single disease or specialist disease. The resulting aided diagnosis model cannot provide a large number of basic general practitioners with the aided diagnosis services for common diseases, nor can it provide patients with self-diagnosis services for common diseases. The current intelligent guidance service robots can replace the guidance nurses in the hospital. To provide more efficient and accurate triage services to patients, they also rely on the aided diagnostic services for a large number of common diseases.

An ontology is an explicit and formal specification of a shared conceptualization [6]. Ontologies provide a formalized method for structurally representing domain knowledge and provide reasoning capabilities. Constructing ontology can achieve some degree of knowledge sharing and reuse. Because ontologies have powerful knowledge representation and reasoning capabilities, they have been widely used in many domains, such as Semantic Web [7], Knowledge Engineering, Natural Language Processing, Information Acquisition, Information Integration, Biomedicine, and other domains. In the domain of biomedicine, there have been a large number of domain semantic knowledge bases built on ontologies [8], such as gene ontology [9], human phenotype ontology [10], and disease ontology [11]. Correspondingly, in the domain of health care, a large number of aided diagnosis research [1],[12]-[14] and other applied research [15]-[17] have also appeared based on domain semantic knowledge bases. These aided diagnosis methods of diseases based on the domain semantic knowledge bases can quickly support the diagnosis of a large number of common diseases.

Diagnosis of diseases is an iterative and complex process, including prospective diagnosis and retrospective diagnosis [1]. During a prospective diagnosis, doctors continually collect detailed information about patients, such as symptoms, examination results, and medical history, to narrow the range of possible diseases. At some point during this process, doctors may have accumulated enough information to give a final diagnosis. The final diagnosis may include one or several of the most likely diseases. The prospective diagnosis is a forward reasoning process based on collected patient information. After the final diagnosis is made, doctors must also verify the final diagnosis by a retrospective diagnosis. On the one hand, it is verified whether the signs, symptoms, abnormal indicators, etc., associated with diseases in the final diagnosis are consistent with the information collected by doctors; on the other hand, diseases in the final diagnosis may also show some other information not collected by doctors, and they need to collect and confirm them. This process is a backward reasoning process.

In the semantic knowledge bases of health care domain, with diseases as the center, the static relationships between diseases and signs, symptoms, examinations, etiology, drugs, surgery, etc., are established respectively. A disease may manifest multiple symptoms. Different diseases may show one or more of the same symptoms. When doctors screen for diseases based on collected patient information, how do they measure the likelihood of the screened diseases and rank them based on this? The conventional idea is: assuming that the five symptoms of the patient are collected, if the disease d_1 in the domain semantic knowledge base matches four of the symptoms, and the other disease d_2 matches three of the symptoms, then the probability that the patient has a disease d_1 is considered to be larger than one of the disease d_2. For this reason, d_2 is ranked in front of d_2 in the ranking of diagnosis results. The assumption above is that all the symptoms are of the same importance in the diagnosis of diseases, but the actual situation is not the case. Although different patients with the same disease may show different symptoms due to individual differences, usually a certain disease will show some of the same typical symptoms in most people. For example, the typical symptoms of a *cold* are *cough*, *runny nose*, *running tears*, *fever*, and *loss of appetite*. Typical symptoms of *diabetes* include *polydipsia*, *polyphagia*, *polyuria*, *weight loss*, etc. Therefore, although different diseases may show some of the same symptoms, the typical symptoms that these diseases show may not be the same. In other words, the weight of the same symptoms in different diseases may not be the same.

Aiming at the above problems, the key point for diagnosing diseases based on symptoms is to give each symptom its importance in the diagnosis of diseases. In this regard, the literature [1] studied and suggested that this importance will be given based on the number of diseases associated with each symptom in the domain semantic knowledge base. For example, *muscle weakness* is a symptom that occurs in many diseases, so its contribution to the diagnosis of diseases is small. Another symptom *bradycardia* is a symptom specific to another small cluster of diseases. If the symptoms provided by a patient include this symptom, the patient's diagnosis is likely to fall in this small cluster of diseases. Based on this reasoning, the literature [1] proposed an algorithm for calculating the weight w_s of the symptom s in the domain semantic knowledge base, and based on w_s, the relative weight w_i of the disease d_i related to the patient's input symptom set S was calculated. Then, the screened diseases

were sorted based on w_i. However, the algorithm for calculating w_i in [1] has the following three obvious problems:

1) The value of w_i is greater than 1 and the minimum value is 1. When the disease d_i is associated with all the symptoms in the patient's input symptom set *S*, the value of w_i is then 1; otherwise, the value of w_i is greater than 1. Since the relative weight between the disease d_i and the patient's input symptom set S is a probability, from the standpoint of probability, the value of w_i cannot be greater than 1. If the value of w_i is 1, it indicates an inevitable event, that is, the disease d_i must be related to the patient. Therefore, the algorithm for calculating w_i in [1] has obvious errors.

2) According to the algorithm for calculating w_i, the value of w_i is larger when the $\sum w_s$ of the symptoms in the patient's input symptom set *S* associated with the disease d_i is smaller. This is obviously contrary to the inference of the importance of the symptoms in the above literature [1].

3) In the algorithm for calculating w_i, the value of the numerator is always $\sum_{s \in S} w_s$. Symptoms in the symptom set *S* that are not associated with the disease d_i have no effect on the calculation of the relative weight w_i of the disease d_i. However, all the symptoms associated with disease d_i in the domain semantic knowledge base contribute to the diagnosis of disease d_i, which is related to the assumptions underlying the static domain semantic knowledge base. See below for an analysis of the lack of literature [1].

In addition, the algorithm for calculating w_i in [1] has the following two shortcomings:

1) The effects of other symptoms associated with the disease d_i in the domain semantic knowledge base (not in the patient's input symptom set *S*) are not considered. Each disease in the domain semantic knowledge base is associated with a certain number of symptoms. The underlying assumption of this static association is that all the symptoms associated with the disease act together on the disease. That is, if a patient's input symptom set *S* completely covers all the symptoms associated with a certain disease, there are no superfluous symptoms, and there are no missing symptoms, then the patient can be considered as having acquired the disease. Under this assumption, if both diseases contain all the symptoms or the same symptoms that the patient has entered, it cannot be assumed that each of these two diseases has the same relative weight with the patient. It is also necessary to consider the effect of other symptoms associated with these two diseases.

2) Retrospective validation of screened diseases was not performed. That is, it is not confirmed whether a patient also has other symptoms other than the symptoms in the input symptom set *S* according to the screened diseases. Based on the initially collected symptom set *S* and re-confirmed symptoms, the screened diseases are adjusted and the relative weight w_i of diseases is recalculated.

In view of the above problems and deficiencies in the aided diagnosis algorithm in the literature [1], this paper has improved it. This paper proposes an aided diagnosis method based on domain semantic knowledge bases, including prospective diagnosis and retrospective diagnosis. In the prospective diagnosis, this paper proposes two algorithms: the algorithm for calculating the weight w_s of the symptom *s* in the domain semantic knowledge base, and the algorithm for the relative weight w_i of the disease d_i associated with one or more symptoms in the collected patient symptom set *S*. In the retrospective diagnosis, the algorithm for symptom set S_{rel} that is most relevant to the symptoms in the symptom set *S* is proposed.

The rest of this paper is organized as followed. Section II introduces the related work. Section III gives the definition of domain semantic knowledge bases. Section IV introduces the construction method of domain semantic knowledge bases. The three algorithms proposed in this paper are given in Section V. Section VI evaluates the aided diagnosis method presented in this paper. The last section summarizes this paper and points out further work.

II. RELATED WORK

The quality of disease diagnosis depends mainly on the medical knowledge and medical experience of medical experts. Early expert systems attempted to meet or exceed the expert's abilities to solve problems by modeling the problem solving abilities of human experts, using knowledge representation and knowledge reasoning techniques in AI to simulate complex problems that are usually solved by experts [18]. The knowledge in the expert systems is usually the heuristic empirical knowledge possessed by the experts described by the rules, and the Rule-Based Reasoning (RBR) method is used to provide the domain problem solving service. The process of acquiring knowledge from experts is a time-intensive process and relies on the opinions of experts, which are sometimes subjective. Literature [19]-[20] combined ontology knowledge base and Semantic Web Rule Language (SWRL) [21] to realize a diagnosis model of hypertension, and provided diagnosis and reasoning of hypertension based on RBR.

In the domain of health care, the expert's experience knowledge is often contained in the medical records of the patients who have been treated. If you can directly use the empirical knowledge of the experts contained in these medical records data, you will avoid the bottleneck of obtaining empirical knowledge directly from experts, because knowledge acquisition is nothing more than collecting cases that have occurred in the past. The study of Case-Based Reasoning (CBR) classification method [22] draws on this idea. The underlying idea of CBR is based on the assumption that similar problems have similar solutions. For example, in the domain of health care, the medical history and treatment plan of patients diagnosed by medical experts are collected and stored as a source case library, and the target case is solved based on the source case library, which is used to help diagnose and treat new patients. Due to the knowledge acquisition, memory, maintenance and other issues of

traditional AI technology, the literature [23] discusses the CBR methodology, research issues and technical aspects of implementing intelligent medical diagnosis systems. The Medical Informatics Research Group at Ain Shams University has developed a system for cancer and heart disease diagnosis based on CBR technology, which is also discussed in [23]. Literature [24] discusses the suitability of CBR in the medical care domain, pointing out the problems, limitations and the possibility of partially overcoming these problems and limitations. In the domain of health care, expert knowledge includes theoretical knowledge and empirical knowledge. For typical and complex cases, experts will make comprehensive diagnosis and recommendation based on theoretical knowledge, empirical knowledge, specific space, time and individual patient conditions. Although the historical medical treatment case may contain some experts' theoretical knowledge and experience knowledge, it is still out of the support of a large number of expert knowledge, so the rationality of the new and old case adaptation is the main problem facing CBR.

The emergence of health care big data also offers the possibility to use data mining and machine learning techniques to obtain knowledge directly from a large amount of historical medical record data. Literature [2] proposed a new method for the diagnosis of heart diseases based on decision tree and naive Bayes algorithm, which can reduce the number of attributes that need to be input for diagnosis, thereby reducing the number of tests that need to be performed on patients during the diagnosis. This method can improve the efficiency of diagnosis. Literature [3] developed a fuzzy inference system using a subtractive clustering algorithm and used this system to classify patients' MRI images to identify Mild Cognitive Impairment, Alzheimer's Disease and Normal Controls. The literature [4] uses BP (Backpropagation) learning algorithm to train a Multilayer Perceptron for diagnosing and predicting neonatal diseases. Literature [5] evaluated the feasibility of using supervised machine learning algorithms in the clinical diagnosis of Parkinson's disease and Progressive Supranuclear Palsy. Disease diagnosis models constructed using data mining and machine learning techniques can provide higher disease recognition rates and detection efficiencies than manual methods. However, for different diseases, separate disease diagnosis models need to be constructed separately, so it is impossible to provide aided diagnosis services for a large number of common diseases in a short time.

In the domain of health care, a large amount of structured knowledge has been accumulated, such as domain semantic knowledge bases built on the ontology model. A method based on the domain semantic knowledge bases can quickly provide aided diagnostic services for a large number of common diseases. The literature [12] discusses some of the technical problems of ontology-based medical systems for cancer diseases and also proposes an ontology-based methodology for the diagnosis of cancer diseases. The methodology can be applied to help patients, students and doctors to determine the type of cancer, the stage of the cancer, and how to treat it. The literature [13] proposed a new mathematical model for the differential diagnosis of genetic diseases, rather than the traditional method of gene mutation analysis. It describes the "genotype-phenotype" association via ontologies. New gene mutations in patients are mapped to a standardized vocabulary in the Human Phenotype Ontology. These terms are then used for differential diagnosis. Combining information theory with fuzzy relation theory, the differential diagnosis can be achieved by measuring the semantic similarity based on ontology. The system can diagnose the occurrence of 5 complex diseases, namely Lymphedema-Distichiasis Syndrome, Cornelia de Lange syndrome, Cohen Syndrome and Smith-Lemli-Opitz syndrome.

Human gene sequences and new biometric data generation technologies provide an opportunity to uncover mechanisms in human diseases. Using "gene-disease" data, recent studies have increasingly shown that many seemingly different diseases have similar or identical molecular mechanisms. Understanding the similarities between diseases helps in the early diagnosis of diseases and the development of new drugs. Informatics methods can be used with ontology to discover the similarities between diseases and to gain insights into the causes of these diseases. It helps to discover the fundamental methods of treating the disease, not just symptomatic treatment. Since the combination of different genes may be related to similar diseases, especially complex diseases, the assessment of disease-likeness based on shared genes alone may be misleading. Searching for identical or similar biological processes, not just explicit genetic matching between diseases, can help overcome this deficiency. In addition to identifying new biological processes related to disease, the use of semantic similarities between biological processes to assess disease similarity can enhance the identification and characterization of disease similarities. In addition, if the disease has a similar molecular mechanism, the drugs currently in use may be used to treat diseases beyond their original indications. This is of great benefit to patients who do not have adequate treatment, especially those with rare diseases. This will also greatly reduce medical costs, because developing new drugs is much more expensive than using existing drugs. In [14], based on co-occurrence and information content, a method of measuring the similarity of terms in ontologies and using terms in ontology to annotate the semantic similarities between entities is proposed. New methods of similarity measurement have been shown to be better than existing methods using biological pathways. The similarity measure uses disease-related biological processes to assess the similarity between diseases and evaluates the method using a manually-planned data set of known disease similarities. In addition, ontologies are used to code diseases, drugs, and biological processes, and demonstrate a method that uses network-based algorithms to combine biological data about diseases with drug information to find new uses for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2875722, IEEE Access

TABLE 1 COMPARISON OF THE AIDED DIAGNOSIS METHODS OF DISEASES

Method	Description	Advantages	Disadvantages
RBR based methods [19]-[20]	Based on the rules to describe the knowledge of disease diagnosis or treatment, or the heuristic experience knowledge possessed by experts, and based on RBR to provide domain problem solving services.	High accuracy and high efficiency.	It is difficult to obtain and maintain rule knowledge; when there are few rules, the problem solution will not be provided; when the number of rules is relatively large, the rule reasoning efficiency is low.
CBR based methods [23]-[24]	Based on the idea that "similar problems have similar solutions", the historical medical records diagnosed and treated by experts are directly stored as source case bases, and based on the similarity comparison between target cases and source cases, the problem solutions or corrected solutions are directly provided for the target cases.	There is no need for a clear domain model to avoid the bottleneck of knowledge acquisition; the solution to the problem can be quickly generated; the problem solution is easy to understand and has direct case evidence; even with a small number of cases, CBR can run.	Because the case involves patient privacy, the case is difficult to obtain; the case is out of the support of some theoretical knowledge and empirical knowledge, the rationality of case adaptation is a major problem; the CBR reasoning process is not reusable.
Methods based on statistical analysis [2]-[5]	Use data mining and machine learning techniques to obtain model knowledge of disease diagnosis and treatment from health care big data, and provide disease diagnosis and treatment services.	It can provide higher disease recognition rate and detection efficiency than manual means.	For different diseases, separate disease diagnosis models need to be constructed separately, and the aided diagnosis services for a large number of common diseases cannot be provided in a short time.
Methods based on domain semantic knowledge bases [1],[12]-[14]	The structured domain semantic knowledge base is used to directly establish relevant knowledge for disease diagnosis and treatment, and provide disease diagnosis and treatment services based on knowledge inquiry and knowledge reasoning.	Aided diagnostic services for a large number of common diseases can be quickly provided.	Low accuracy; knowledge building and maintenance is a time-intensive task.

existing drugs. The effectiveness of the method was verified by comparison with existing drug-related clinical trials.

The research work in [1] is part of the European project K4CARE [62]. The goal of the project is to combine health care with information and communication technology (ICT) experiences in Western and Eastern European countries to establish, implement and validate a knowledge-based healthcare model to provide professional assistance to elderly patients in the home. The project focuses on 9 chronic diseases, 2 syndromes, and 5 social problems. It uses CPO (Case Profile Ontology) ontology to describe knowledge related to these diseases, and uses the SDA (State-Decision-Action) diagram to describe related intervention plans related to these diseases. The literature [1] shows the methods and tools for disease diagnosis and ontology personalization developed in this project. There are some obvious problems and deficiencies in the diagnosis method of diseases, which are described in detail in the introduction part of this paper.

Table 1 summarizes the existing research on disease-aided diagnostic methods and analyzes their respective advantages and disadvantages. These methods use different data in addition to their technical principles. For example, the methods in [1],[12],[19],[20] use the patient's symptoms and signs data, the method in [13] uses the patient's genetic data, and the method in [3] uses the patient's MRI image data. Some data (for example, symptoms and signs) are better to obtain, while some data (for example, genetic data) are difficult to obtain.

In order to provide aided diagnostic services of a large number of common diseases to primary general practitioners and to provide self-diagnosis services for patients, this paper adopts the aided diagnosis method based on the domain semantic knowledge bases. In view of the shortcomings of the existing research, this paper has carried out correction and improvement.

III. RELATED DEFINITIONS

Definition 1 (Domain Ontology Schema): The domain ontology schema describes domain knowledge by capturing concepts, concept attributes, semantic relations between concepts, and related constraints that are commonly accepted in the domain. The domain ontology schema is O_{domain}, which is defined as follows:

$$O_{domain} = \langle C, A, R, X, I \rangle.$$

Where *C* stands for classes, which describe concepts in a domain. A class represents a set of instances that have certain similar characteristics. For example, people with different characteristics belong to the class *people*. *A* stands for a set of attributes (also called data type properties), for instance, a person's name, sex, date of birth, height, and weight. *R* stands for a set of semantic relations (also called object properties). There are two types of semantic relations, i.e., taxonomic relations (e.g., the class *people* can be further divided into the class *man* and the class *woman*) and non-taxonomic relations (such as, a good friend relationship, a father-son relationship, and a sibling relationship). *X* stands for a set of axioms, and axioms are used to define the constraints on *C*, *A* and *R*. For example, a person has only one date of birth, but some people can have the same date of birth; a person's biological parents are unique; the domain and range values of the property *date_of_birth* are *people* and *date* respectively. *I* represents a set of instances, which describes the commonly accepted knowledge, such as, "*diabetes*" and "*hyperthyroidism*" are instances of "*endocrine and metabolic diseases*". The domain

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2875722, IEEE Access

ontology schema usually doesn't contain instances, other than domain common sense. Nevertheless, an RDF (Resource Description Framework) [25] description fragment which only contains instances is not the domain ontology schema [26]. The Ontology standard description languages recommended by the W3C include RDF, RDFS (RDF Schema) [27] and OWL (Web Ontology Language) [28].

Definition 2 (Domain Instance Data): The domain instance data is about the knowledge of the individuals described by the classes in the domain ontology base, for example, a person's basic information and health status. It is noted as I_{domain}:

$$I_{domain} = \{(s, p, o) | s \in I, p \in A \cup R, o \in I \cup V\}.$$

Where (s, p, o) represents a statement or a triple. s stands for an instance, and I stands for a set of instances. p stands for an attribute of an instance or a semantic relationship between instances. A represents a set of data type properties and R represents a set of object properties. o represents a property value, which is either an instance or a literal value [25]. V represents a set of all literal values.

Definition 3 (Domain Semantic Rule Set): Semantic rules are used to supplement the description capabilities of ontology description languages and are often used to describe empirical knowledge from experts. Semantic rules are typical conditional statements: *if-then* clauses, which permit the adding of knowledge in *then* portion when *if* portion is true. The domain semantic rule set is noted as F_{domain}:

$$F_{domain} = \{r_1, r_2, \dots, r_{\mathrm{i}}, \dots, r_{\mathrm{n}}\}, \mathrm{n} \geq 0.$$

Where r_{i} stands for the i-th semantic rule. The Semantic Web layer cake [29] provides a variety of knowledge representation, ranging from RDF to the latest version of the OWL and other formats, expanding expressivity at each level and allowing users to use a given representation based on the amount of semantics needed for a particular application. However, there are drawbacks such as lack of descriptive vocabulary and flexibility in expression. This situation is constantly improved by adding an additional level of expressivity based on user-defined rules. The W3C recommended semantic rule description language is SWRL.

Definition 4 (Domain Semantic Knowledge Base): O_{domain}, I_{domain} and F_{domain} together form a domain semantic knowledge base. The domain semantic knowledge base is SKB_{domain}, which is defined by:

$$SKB_{domain} = \langle O_{domain}, I_{domain}, F_{domain} \rangle.$$

In reality, there is a fine line where the ontology ends and the knowledge base begins [30].

SPARQL query language is recommended by W3C to be specifically used for the semantic layer (RDF layer) query language over the domain semantic knowledge base. SPARQL can be used to express queries across diverse RDF query tools, for it supports RDF syntax, RDF models, and RDF vocabulary [29].

IV. DOMAIN SEMANTIC KNOWLEDGE BASE CONSTRUCTION

Due to the professionalism of domain knowledge, it is widely accepted that the participation of experts is inseparable for the construction of domain semantic knowledge bases. Then it is the knowledge engineers who model and formalize the domain knowledge provided by the experts for the purpose of building domain knowledge bases that could be shared and processed by computers. Due to the complexity of domain knowledge body, it is impossible to build an ontology artificially, not to mention the total time it may consume. Therefore, ontology engineering [29] and almost all ontology modeling methods [31]-[36] emphasize the consideration of integrating and reusing already existing domain ontology knowledge bases before constructing domain semantic knowledge bases. For example, Ontology Integration [37] and Ontology Mapping [38]-[39] methods are used to quickly build the required domain ontology knowledge base. Alternatively, domain knowledge is automatically or semi-automatically acquired from domain data sources using Ontology Learning [40] technology and described on the basis of ontologies. Domain data sources include structured, semi-structured and unstructured data in the domain, as well as other open knowledge bases, such as Freebase [41]-[43], DBpedia [44]-[45], YAGO [46]-[48].

For the construction of domain semantic knowledge bases, Freebase is a good and reusable data source, which can be used as a starting point to build domain semantic knowledge bases. Freebase is a practical, scalable, graph-shaped database of structured general human knowledge, where users collaboratively create, structure and maintain content over an open platform. Freebase data is expressed in triples (also known as facts or statements) format and can be visually represented as a directed graph. Freebase data comes from a large number of high-quality open data sources, such as Wikipedia [49], MusicBrainz [50], WordNet [51], and others. Freebase is also an important data source of LOD (Linked Open Data) [52] project. On a weekly basis, Freebase releases its data as an N-Triples [54] RDF dump file under the CC-BY [53] license. This file is a single text file that is compressed using *gzip*. For example, the size of the *gzip* archive downloaded in August 2014 was 22GB, and the size of the unzipped file was 250GB, which contained a total of about 1.9 billion triples. The Freebase RDF dump includes 11 Freebase Implementation Domains, 5 OWL Domains and 89 Subject Matter Domains [55]. For example, Freebase's *medicine* subject matter domain describes domain ontology schema and domain instance data (i.e., domain common sense knowledge) in the domain of health care. The medicine ontology schema describes related concepts such as *disease*, *symptom*, *cause*, *risk_factor*, and *drug*. Based on these concepts, the domain common sense knowledge is described, for example, the symptoms of diseases, the causes and risk factors of diseases, and the therapeutic drugs for diseases.

For this reason, the study of this paper chooses to extract the knowledge of *medicine* domain from the Freebase RDF dumps. However, extracting domain knowledge from the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2875722, IEEE Access

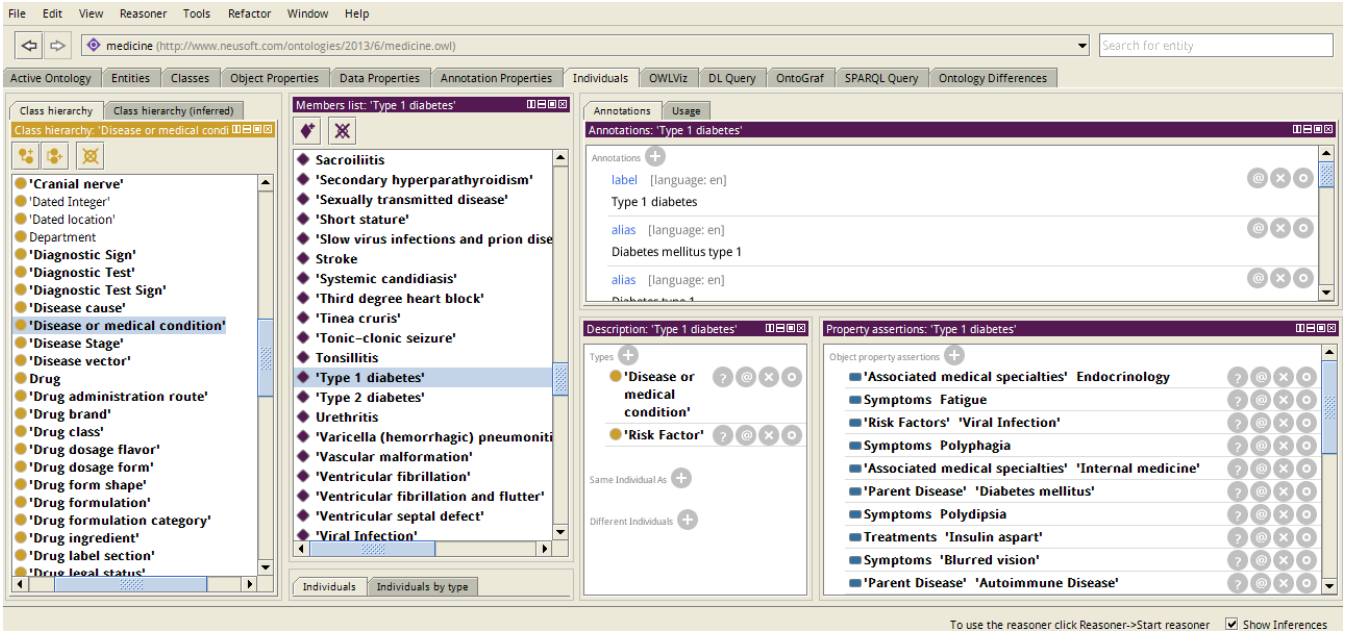


FIGURE 1. The semantic knowledge base of the medicine domain.

Freebase RDF dumps on a particular domain, such as the *medicine* domain, will face many difficulties. With the complete understanding of Freebase-related concepts, Freebase's knowledge representation model, and the structural features of the Freebase RDF dumps, a method for extracting domain knowledge from the Freebase RDF dumps is proposed and implemented [56]. The method ensures a fast, precious, complete knowledge extraction over one or more domains from the Freebase RDF dumps. And the extracted domain semantic knowledge base is converted into a form being described by standard ontology description languages. The detailed extraction and processing method is not the subject of this paper.

The extracted *medicine* domain semantic knowledge base is represented as Turtle [57] RDF format, and the file size is 1.6 GB. The TDB [58] store provided by Apache Jena [59] is used here as a triple store. After the file is loaded into the TDB store, the file system space occupied by the TDB storage is 1.3GB (i.e., it's smaller than the original file). Jena Fuseki [60] is used as a SPARQL server to publish the TDB store. SPARQL queries are performed using the SPARQL Query Endpoint provided by the Jena Fuseki.

As the method of the aided diagnosis of diseases discussed in this paper only relies on diseases, symptoms and their semantic relations in *medicine* domain, medical experts finally proofread and perfect this part of knowledge. The final scale of the *medicine* domain semantic knowledge base $SKB_{medicine}$ is as follows:

1) The *medicine* domain semantic knowledge base contains 70 concepts, 63 data type properties, 156 object properties, 886,272 instances, and 7,073,580 triples.

2) There are 7,367 disease instances, but do not include synonymous disease instances of these disease instances. Of these, 3,590 disease instances include a total of 3,802 synonymous disease instances. Through the *owl:sameAs* semantic construct, these synonymous disease instances are normalized to their corresponding standard disease instances. The disease instances mentioned below all refer to standard disease instances.

3) There are 1,444 symptom instances, but do not include synonymous symptom instances of these symptom instances. Of these, 1,112 symptom instances include a total of 1,352 synonymous symptom instances. Through the *owl:sameAs* semantic construct, these synonymous symptom instances are normalized to their corresponding standard symptom instances. The symptom instances mentioned below all refer to standard symptom instances.

4) There are 6,028 semantic relationships from standard disease instances to standard symptom instances.

Fig. 1 shows the *medicine* domain semantic knowledge base $SKB_{medicine}$ using Protégé4.3 [61].

V. THE AIDED DIAGNOSIS METHOD

A. THE CALCULATION OF SYMPTOM WEIGHTS

The calculation of the weight w_s of the symptom s in the knowledge base is based on the assumption that the current $SKB_{medicine}$ contains all disease instances, symptom instances, and their semantic relationships. Once the disease instances, the symptom instances, or their semantic relationships in $SKB_{medicine}$ are updated, the w_s will be recalculated.

Assume that the total number of diseases contained in $SKB_{medicine}$ is N. For each symptom s, we define N_s as the number of diseases that have a semantic relationship with the symptom s. w_s is the weight of symptom s in the diagnosis of diseases. Then w_s is calculated as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2875722, IEEE Access

序号	Symptom code 症状代码	Symptom name 症状名称	Symptom introduction 症状简介	Symptom weight 症状权重	gender 性别	operating 操作
1001	-	High heat 高热	由于多种不同原因致人体产热大于散热，使体温超过正常范围称为发热(fever)，临床上按热度高低将发热分为低热、中等度热、高热及超高热。高热指体温超过39.1℃。	0.94	全部	修改 删除
1002	-	High fever 高热不退	发热是多种疾病的常见症状。高热(High Fever)在临床上属于危重症范畴。正常体温常以肛温36.5～37.5℃，腋温36～37℃衡量。通常情况下，腋温比口温(舌下)低0.2～0.5℃，肛温比腋温约高0.5℃左右。肛温虽比腋温准确，但因种种原因常以腋温为准。若患者所测腋温的值长时间高达39.1～40℃称为高热不退。	1	全部	修改 删除
1003	-	High heat chill 高热寒战	寒战大多发生在急性发热性疾病之前。感染性疾病的致病原，作用于机体引起发热时，病人全身发冷、起鸡皮疙瘩和颤抖，即肌肉不自主活动，此称为恶寒战栗，简称寒战。 寒战是高热的先声，寒战期间，体温已有升高，在发热不太高的前期，有时病人仅有全身发冷感，而无战栗，称为发冷。	0.99	全部	修改 删除

FIGURE 3. **The weight of each symptom in the knowledge base.**

$$w_s = \left(\frac{N - N_s}{N - 1}\right)^2 \quad (1)$$

Among them, $N_s \geq 1$, $w_s \leq 1$. From formula (1), it can be seen that the larger the number N_s of diseases that are semantically related to the symptom *s*, the smaller the weight w_s of the symptom *s* in the diagnosis of diseases. The purpose of squaring the equation is to emphasize the difference between symptom weights as the number of symptom-associated diseases increases. The denominator takes *N*-1 because when the number N_s of the diseases that are semantically related to the symptom *s* is 1, it is ensured that w_s is 1, that is, a disease can be uniquely determined based on the symptom *s*. The relationship between w_s and N_s is shown in Fig. 2.

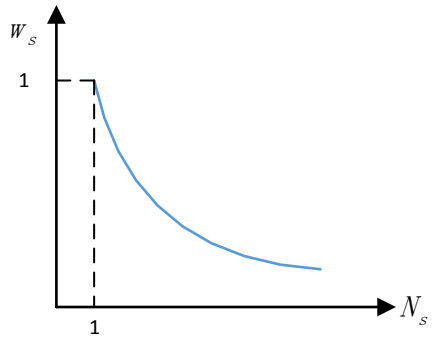


FIGURE 2. **The relationship between w_s and N_s.**

Based on formula (1), a w_s can be calculated for each symptom s in $SKB_{medicine}$, as shown in Fig. 3. Since the language of system implementation pages is Chinese, an English description of some keywords is marked on the pages. The same below.

B. THE CALCULATION OF DISEASE RELATIVE WEIGHTS

Based on the w_s of each symptom s in $SKB_{medicine}$, the relative weight w_i between the disease d_i in $SKB_{medicine}$ and the patient's input symptom set *S* can be calculated. Based on the relative weight w_i of the disease d_i, the doctor or the patient may be recommended for possible diseases and disease relevance rankings.

Assume that the number of symptoms is *M* and the number of diseases is *N* in $SKB_{medicine}$. The patient's input symptom set is $S = \{s_1, s_2, \ldots, s_j\}$, $1 \leq j \leq M$. The set of diseases in $SKB_{medicine}$ that are semantically related to one or more symptoms in the symptom set *S* is $D = \{d_1, d_2, \ldots, d_i\}$, $1 \leq i \leq N$. The set of symptoms associated with the disease d_i in $SKB_{medicine}$ is S_i, $S_i' = S_i \cap S$. The relative weight between the disease d_i and the patient's input symptom set *S* is w_i, then w_i is calculated as follows:

$$w_i = \frac{\sum_{s \in S_i'} w_s}{\sum_{s \in S_i} w_s} \quad (2)$$

The numerator of formula (2) only considers the symptoms associated with the disease d_i in the symptom set *S* because the other symptoms have no effect on the relative weight calculation of the disease d_i. The denominator takes into account all the symptoms associated with the disease d_i in $SKB_{medicine}$ because they work together to diagnose the disease d_i.

Assume that a patient enters the symptom "*pharyngeal foreign body sensation*". Based on formula (2), the relative weights of diseases associated with the symptom in $SKB_{medicine}$ can be calculated, as shown in Fig. 4(a).

C. THE RECOMMENDATION OF RELATED SYMPTOMS

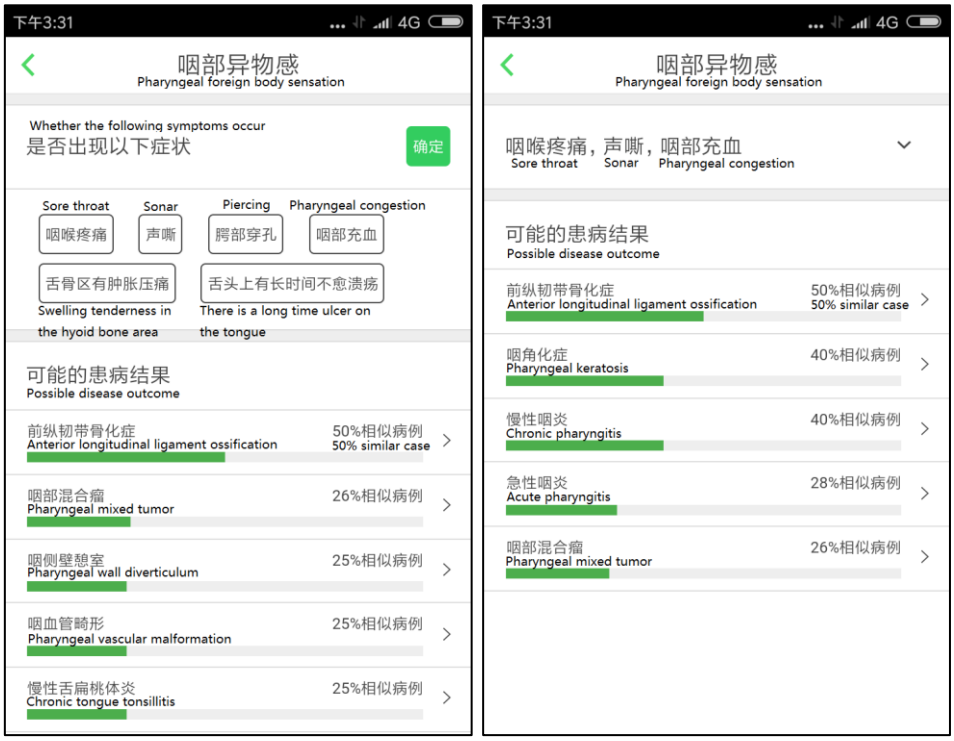
After screening the disease set *D* associated with the patient's input symptom set *S*, the diseases in the set *D* needs to be retrospectively verified. That is, according to the screened set of diseases *D*, other symptoms that are most relevant to the symptoms in the patient's input symptom set *S* are evaluated and recommended for confirmation by the doctor or the patient. Then, based on the initially collected patient symptom set *S* and the patient's reconfirmed symptoms, the disease screening result set is adjusted, and the relative weights of diseases in the adjusted disease set *D* are recalculated.

The recommended algorithm for the symptom set S_{rel} that most closely relates to the symptoms in the patient's input symptom set *S* is described as follows:

Input: The initial symptom set *S* entered by a patient.

Output: The symptom set S_{rel} consisting of Top-6 symptoms that most closely correlate with the symptoms in the patient's input symptom set *S*. Here, only 6 symptoms remain in S_{rel} and can be adjusted as needed.

S-1: Firstly, a recommendation is made from the recorded history input symptom combinations. The system automatically records the combinations of historical symptoms entered



(a) (b)

FIGURE 4. The calculation of disease relative weights and the recommendation of related symptoms.

and selected by different patients. The recording method is $\{s_1, s_2, \ldots, s_i\}_f$. The symptoms in the symptom combinations are in no particular order. *f* indicates the frequency at which a combination of symptoms occurs. If the symptom set *S* entered by the patient falls within one or more historical symptom combinations, the Top-6 symptoms other than the ones in the set *S* are selected from the one or more historical symptom combinations by the *f* value from high to low. Take the Top-6 symptoms as S_{rel} and then go to step S-6. If there are not enough 6 symptoms, the actual number of symptoms can be selected as S_{rel}, and then go to step S-6. If there are no more symptoms or *S* does not fall into any of the historical symptom combinations, go to step S-2. Note that when selecting symptoms in this step, the weights of the symptoms are not considered, but in the order in which the symptoms appear.

S-2: The set of diseases *D* that are semantically related to one or more symptoms in the symptom set *S* is queried from $SKB_{medicine}$. Here, $D = \{d_1, d_2, \ldots, d_i\}$, $1 \le i \le N$, N represents the number of diseases in $SKB_{medicine}$.

S-3: Let the symptom set associated with the disease d_i in $SKB_{medicine}$ be S_i, and get $S' = S_1 \cup S_2 \cup \ldots \cup S_i$.

S-4: The symptoms in the set S' are sorted in descending order according to the value of w_s to get S^*.

S-5: Select the Top-6 symptoms from S^* as the most relevant symptom set S_{rel} with the patient's input symptom set *S*. If these Top-6 symptoms contain symptoms from the patient's input symptom set *S*, these symptoms are skipped and then selected one after the other.

S-6: Output S_{rel}.

As shown in Fig. 4(a), the six most relevant symptoms are recommended based on the symptom "*pharyngeal foreign body sensation*" entered by the patient. After further selecting three symptoms "*throat pain*", "*sonar*", and "*pharyngeal hyperemia*" from the six recommended symptoms, the relative weights of diseases will be recalculated based on the four symptoms that the patient has entered twice, and the ranking will be adjusted, as shown in Fig. 4 (b). At the same time, the most relevant symptom set S_{rel} will be re-recommended based on these 4 symptoms.

The entire aided diagnosis process is a cyclical iterative process with the participation of doctors and patients. For general practitioners at the primary level, detailed information about the disease can be viewed from the list of recommended diseases, including disease introduction, treatment departments, high-incidence groups, contagiousness, symptoms, tests, differential diagnosis, treatment, dietary taboo, prevention, etc. Based on the detailed information, they can determine if further inspections are needed and what inspections to do. Since the diseases in the list of recommended diseases have some similarities in the symptoms, it is possible to make further judgments through the differential diagnostic information presented, as shown in Fig. 5(b). In the case of self-diagnosis for patients,, a patient can obtain medical resource recommendations for the region where he is located by selecting a disease from the list of recommended diseases, including hospitals, departments, and doctors in the region. The recommendation process can also consider the objective evaluation of medical institutions in the region, for example, the level of diagnosis and treatment, the

(a) (b)

FIGURE 5. Medical resource recommendations and disease details.

cost of diagnosis and treatment, and the rate of diagnosis and treatment. The history of the patient's visit reflects the patient's preferences, so the recommendation process also needs to consider the patient's preferences, such as distance, cost, as shown in Fig. 5(a). Patients can also view detailed information about diseases and compare them with their own performance.

VI. METHOD EVALUATION

A. EVALUATION METHOD

In order to evaluate the diagnostic methods proposed in this paper, the clinical outpatient medical records of six common diseases are selected from the top three hospitals in a city in China. Each medical record includes information such as a patient's gender, age, chief complaint, medical history, allergy history, physical examination, treatment advice, and a doctor's diagnosis. The method of this paper only uses patients' chief complaint and doctors' diagnosis information. However, the quality of outpatient medical record data is not high. For example, the value of the main complaint fields of most medical records are empty, or their value is "unfilled". Here, when selecting the medical record data, the data of these two cases are filtered out. However, there are still other data quality problems. For example, the chief complaint content is "*review of acute pharyngitis*", "*consultation*", and "*requiring color ultrasound*". Therefore, medical experts are invited to screen the selected medical record data and filter out some invalid medical record data. The distribution of the medical records collected is shown in Table 2.

The evaluation method is as follows:

1) For each medical record of each disease, the symptom information is extracted from the chief complaint as the patient's input symptom set *S*.

2) Based on the symptom set *S*, the method of this paper and the method in the literature [1] are used to obtain the list of related diseases and the ranking of disease relative weights.

TABLE 2 THE DISTRIBUTION STATISTICS OF SELECTED MEDICAL RECORDS

Diagnosis	Total number of medical records	Number of invalid medical records
Acute upper respiratory tract infection	500	35
Acute bronchitis	500	98
Acute pharyngitis	500	76
Chronic pharyngitis	500	52
Chronic gastritis	500	112
Ovarian cyst	500	23

3) From the disease relevance rankings, the Top 1 (first diagnosis) and Top-3 (first 3 diagnoses) are compared with the doctor's diagnoses in the medical records. If the Top-1 diagnosis is consistent with the diagnosis given by the doctor, this indicates a Top-1 hit. Otherwise, if a diagnosis in Top-3 is consistent with the diagnosis given by the doctor, it indicates a Top-3 hit.

4) For the medical record data of each disease, the hit rates of Top-1 and Top-3 are respectively counted.

B. EVALUATION RESULTS

Table 3 is the diagnostic hit rate statistics obtained based on the method of this paper, and Table 4 is the diagnostic hit rate statistics obtained based on the method of the literature [1]. The following is the analysis of the evaluation results:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2875722, IEEE Access

TABLE 3 DIAGNOSTIC HIT RATE STATISTICS BASED ON THE METHOD OF THIS PAPER

Diagnosis	Number of medical records participating in the assessment	Top-1 hit medical record number	Top-1 hit rate	Top-3 hit medical record number	Top-3 hit rate
Acute upper respiratory tract infection	465	263	56.6%	396	85.2%
Acute bronchitis	402	99	24.6%	316	78.6%
Acute pharyngitis	424	179	42.2%	313	73.8%
Chronic pharyngitis	448	207	46.2%	344	76.8%
Chronic gastritis	388	178	45.9%	281	72.4%
Ovarian cyst	477	321	67.3%	425	89.1%
Overall evaluation	2604	1247	47.9%	2075	79.7%

TABLE 4 DIAGNOSTIC HIT RATE STATISTICS BASED ON THE METHOD OF [1]

Diagnosis	Number of medical records participating in the assessment	Top-1 hit medical record number	Top-1 hit rate	Top-3 hit medical record number	Top-3 hit rate
Acute upper respiratory tract infection	465	0	0	0	0
Acute bronchitis	402	0	0	0	0
Acute pharyngitis	424	0	0	97	22.9%
Chronic pharyngitis	448	0	0	105	23.4%
Chronic gastritis	388	0	0	0	0
Ovarian cyst	477	80	16.8%	198	41.5%
Overall evaluation	2604	80	3.1%	400	15.4%

1) *Ovarian cyst* is one of the most common diseases in gynecological diseases. Compared with several other diseases, the quality of the chief complaint in the medical records is the best. Most of the symptoms in the chief complaint are only related to gynecological diseases, so the Top-1 hit rate and Top-3 hit rate in Table 3 are highest compared to other diseases.

2) According to the statistics of outpatient medical records of a certain city in China, *acute upper respiratory tract infection* is the disease with the highest incidence rate in outpatient diagnosis. The description of the symptoms in the chief complaint is more typical, and the similarity between different medical records is higher. Therefore, the diagnostic hit rate of this disease is also relatively high in Table 3.

3) Due to some problems with the method itself in the literature [1], the Top-1 hit rate and the Top-3 hit rate in Table 4 are very low. Because *ovarian cyst* is a typical gynecological disease, its symptoms are only related to gynecological diseases, so although there are problems in the method in [1], Top-1 and Top-3 have respectively hit some medical records. Similarly, some symptoms of *acute pharyngitis* and *chronic pharyngitis* are related to the pharynx, so Top-3 also hit some medical records. The symptoms of the remaining diseases are related to most diseases, so no medical records are hit.

4) Overall, using the method proposed in this paper, the average hit rate of Top-1 is 47.9%, and the average hit rate of Top-3 is 79.7%. This result is already relatively good. It can be used to provide basic diagnostic services for common diseases for general practitioners, and it can also be used to provide patients with self-diagnosis services. However, using the method in [1], the average hit rate is very low and cannot be used to provide aided diagnosis services. Moreover, the method in [1] does not provide a retrospective diagnosis.

C. THE SHORTCOMINGS OF THE METHOD IN THIS PAPER

The overall quality of outpatient medical records is poor. The quality of inpatient medical records is better. The quality of medical record data in the top three hospitals is better than that of other grade hospitals. Therefore, the quality of outpatient medical record data has a great influence on the evaluation results of the method. But the method of this paper still has the following shortcomings:

1) The domain semantic knowledge base $SKB_{medicine}$ needs to continue to improve. The quality and scale of domain semantic knowledge bases also have a great influence on the accuracy of the diagnosis results. For example, some synonymous symptoms may be overlooked because they cannot be matched with the symptoms in the knowledge base.

2) The method of this paper only uses the symptom information in the patient's chief complaint, but the actual outpatient diagnosis also needs to refer to the patient's gender, age, history, examination and other information.

VII. CONCLUSI0N

In this paper, the knowledge of the *medicine* subject matter domain is extracted from the Freebase RDF dumps, and the domain semantic knowledge base $SKB_{medicine}$ is constructed. Finally, the medical experts correct and improve $SKB_{medicine}$, including diseases, symptoms and their semantic relationships. On this basis, an aided diagnosis method based on $SKB_{medicine}$ is proposed. The entire aided diagnosis process of diseases is a cyclical iterative process with the participation of doctors and patients, including prospective diagnosis and retrospective diagnosis. Finally, based on the real medical record data screened by medical experts, the *medicine* domain semantic knowledge base constructed in this paper and the aided diagnostic method of diseases proposed in this paper are

evaluated. Overall, the average hit rate for Top-1 is 47.9%, and the average hit rate for Top-3 is 79.7%. This result is already relatively good. It can be used to provide basic diagnostic services for common diseases for general practitioners, and it can also be used to provide patients with self-diagnosis services.

Further research work is as follows:

1) In order to verify the scientificity in clinic of the method proposed in this paper, further clinical trials are needed.

2) Introduce more information, such as age, gender, history, and tests, into the disease relevance calculation method.

3) Directly establish the aided diagnosis rules for some common diseases and conduct rule-based reasoning for disease diagnosis first. In the absence of any matching rules, the disease relevance calculation method is applied again.

4) Disease diagnosis should be personalized. Some diseases have different performances among people of different ages, sexes, and regions. The same check-up indicator may differ in the normal range of different populations and even different individuals. Therefore, the bolder idea is to turn the knowledge base around disease types into a knowledge base around the crowd or "standardized patients", and to provide personalized aided diagnosis services for common diseases based on this knowledge base.

REFERENCES

[1] K C. Romero-Tris, D. Riaño, and F. Real, "Ontology-Based Retrospective and Prospective Diagnosis and Medical Knowledge Personalization," in *KR4HC 2010*, 2010, pp 1-15.

[2] N. Bhatla and K. Jyoti, "A Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic," *International Journal of Computer Applications*, vol. 54, no. 17, pp. 16-21, Sep. 2012.

[3] I. Krashenyi, A. Popov, J. Ramirez, and et al, "Application of fuzzy logic for Alzheimer's disease diagnosis," in *Signal Processing Symposium*, 2015, pp 85-88.

[4] D. R. Chowdhury, "An Artificial Neural Network Model for Neonatal Disease Diagnosis," *International Journal of Artificial Intelligence & Expert Systems*, vol. 2, no. 3, pp. 96-106, Jan. 2011.

[5] C. Salvatore, A. Cerasa, I. Castiglioni, et al, "Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy," *Journal of Neuroscience Methods*, vol. 222, pp. 230-237, Jan. 2014.

[6] R. Studer R, V. R. Benjamins, and D. Fensel, "Knowledge engineering: Principles and methods," *Data and Knowledge Engineering*, vol. 25, no. 1/2, pp. 161-197, Mar. 1998.

[7] T. BERNERS-LEE, J. HENDLER, and O. LASSILA, "The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities," *Scientific American*, vol. 284, no. 5, pp. 34-43, May 2001.

[8] B. Smith, M. Ashburner, C. Rosse, et al, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature Biotechnology*, vol. 25, no. 11, pp. 1251-1255, Nov. 2007.

[9] G. Licata, "Employing fuzzy logic in the diagnosis of a clinical case," *Health*, vol. 2, no. 3, pp. 211-224, Jan. 2010.

[10] S. Köhler, M. H. Schulz, P. Krawitz, et al, "Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies," *American Journal of Human Genetics*, vol. 85, no. 4, pp. 457-64, Oct. 2009.

[11] O. Bodenreider, "Disease Ontology," *Encyclopedia of Systems Biology*, pp. 578-581, 2013.

[12] M. Alfonse, M. M. Aref, and A. B. M. Salem, "An Ontology-Based Cancer Diseases Diagnostic Methodology," *Recent Advances in Information Science*, pp. 95-99, 2013.

[13] L. Jayaratne, "Ontology Based Approach for Diagnosis in Personalized Medicine," in *International Conference on Computer Games, Multimedia and Allied Technology*, 2015.

[14] S. Mathur, "Ontology-based methods for disease similarity estimation and drug repositioning," Ph.D. dissertation, Computer Science and Mathematics, University of Missouri-Kansas City, 2012.

[15] S. Izumi, K. Dai, G. Itabashi, et al, "An ontology-based advice system for health and exercise," in *Tenth Iasted International Conference on Internet and Multimedia Systems and Applications*, 2006, pp 95-100.

[16] J. Cantais, D. Dominguez, V. Gigante, et al, "An example of food ontology for diabetes control," in *International Semantic Web Conference Workshop on Ontology Patterns for the Semantic Web*, 2005.

[17] K. Kostopoulos, I. Chouvarda, V. Koutkias, et al, "An ontology-based framework aiming to support personalized exercise prescription: application in cardiac rehabilitation," in *Conf Proc IEEE Eng Med Biol Soc*, 2011(4), pp 1567-1570.

[18] *Expert system*. [Online]. Available: http://www.intsci.ac.cn/ai/es.html, accessed on: Aug 5, 2018.

[19] M. Gong, "Research on ontology based knowledge base of hypertension electronic medical records," M.S. thesis, Xi'an Electronic and Science University, Xi'an, China, 2010.

[20] M. Gong and Y. Wen, "Ontology based knowledge base for diagnosis of hypertension," *Journal of Intelligence*, vol. 29, no. b06, pp. 169-172, 2010.

[21] *SWRL: A Semantic Web Rule Language Combining OWL and RuleML*. [Online]. Available: https://www.w3.org/Submission/SWRL/, accessed on: May 19, 2018.

[22] M. Kamber, J. Han, and J. Pei, *Data Mining: Concepts and Techniques*. 3rd Edition. Elsevier, 2011.

[23] A. B. M. Salem, "Case Based Reasoning Technology for Medical Diagnosis,", in *Proceedings of World Academy of Science Engineering & Technolog*, 2007.

[24] R. Schmidt and L. Gierl, "Case-based reasoning for medical knowledge-based systems," *Stud Health Technol Inform*, vol. 77, no. 2-3, pp. 720-725, 2000.

[25] *RDF 1.1 Primer*. [Online]. Available: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/, accessed on: May 19, 2018.

[26] J. Hebeler, M. Fisher, R. Blace, and A. Perez-Lopez, *Semantic Web Programming*. Indianapolis: Wiley Publishing, 2009.

[27] *RDF Schema 1.1*. [Online]. Available: https://www.w3.org/TR/2014/REC-rdf-schema-20140225/, accessed on: May 19, 2018.

[28] *OWL 2 Web Ontology Language Primer (Second Edition)*. [Online]. Available: http://www.w3.org/TR/2012/REC-owl2-primer-20121211/, accessed on: May 19, 2018.

[29] G. Antoniou and F. V. Harmelen, *A Semantic Web Primer*. 2nd ed. Cambridge: MIT Press, 2008.

[30] N. F. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," Stanford Knowledge Systems Laboratory, Technical Report: KSL-01-05, Jan. 2001.

[31] I. W. Kim and K. H. Lee, "A model-driven approach for describing semantic Web services: From UML to OWL-S," *IEEE Transaction on Systems, Man, and Cybernet-ics Part C: Applications and Reviews*, vol. 39, no. 6, pp. 637-646, Nov. 2009.

[32] L. Iribarne, N. Padilla, J. A. Asensio, J. Criado, R. Ayala, J. Almendros, and M. Menenti, "Open-environmental ontology modeling," *IEEE Transaction on Systems, man and humans*, vol. 41, no. 4, pp. 730-745, Jul. 2011.

[33] C. S. Lee, Y. F. Kao, Y. H. Kuo, and M. H. Wang, "Automated ontology construction for unstructured text documents," *Data & Knowledge Engineering*, vol. 60, no. 3, pp. 155-176, Mar. 2007.

[34] B. Raufi, F. Ismaili, and X. Zenuni, "Modeling a complete ontology for adaptive Web based systems using a top-down five layer framework," in *Proceedings of the ITI 2009 31st Int. Conf. on Information Technology Interfaces*, 2009, pp. 511-518.

[35] J. Li and L. S. Meng, "Comparison of seven approaches in constructing ontology," *New Technology of Library and Information Service*, vol. 7, pp. 17-22, Jan. 2004.

[36] R. Subhashini R and J. Akilandeswari, "A survey on ontology construction methodologies," *International Journal of Enterprise Computing and Business Systems*, vol. 1, no. 1, Jan. 2011.
[37] S. B. Jadhav and S. N. Pardeshi, "Ontology intergration with semantic similar entity classes amongst different ontologies for enhanced information retrieval," *International Journal of Recent Trends in Engineering*, vol. 2, no. 3, pp. 132-134, Nov. 2009.
[38] K. Zaiß, T. Schlüter, and S. Conrad, "Instance-based ontology matching using different kinds of formalisms," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 3, no. 7, pp. 1716-1724, Jul. 2009.
[39] P. Lambrix and T. He, "Ontology alignment and merging," *Computational Biology*, vol. 6, pp. 133-150, Jan. 2008.
[40] X. Y. Du, M. Li, and S. Wang S, "A survey on ontology learning research," *Journal of Software*, vol. 17, no. 9, pp. 1837-1847, Sep. 2006.
[41] Freebase. [Online]. Available: https://en.wikipedia.org/wiki/Freebase, accessed on: May 19, 2018.
[42] K. Bollacker, R. Cook, and P. Tufts, "Freebase: A Shared Database of Structured General Human Knowledge," in *AAAI Conference on Artificial Intelligence*. Vancouver, British Columbia, Canada: DBLP, 2007, pp 1962-1963.
[43] K. Bollacker, C. Evans, P. Paritosh, and et al, "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge," in *SIGMOD'08*, 2008, pp 1247-1250.
[44] *DBpedia*. [Online]. Available: http://wiki.dbpedia.org/, accessed on: May 19, 2018.
[45] S. Auer, C. Bizer, G. Kobilarov, and et al, "DBpedia: a Nucleus for a Web of Open Data," in *ISWC'07/ASWC'07*, 2007, pp 722-735.
[46] F. M. Suchanek, G. Kasneci, and G. Weikum, "YAGO - A Core of Semantic Knowledge," in *WWW'07*, 2007, pp 697-706.
[47] J. Hoffarta, F. M. Suchanekb, K. Berberich, and et al, "YAGO2: Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages," in *WWW'11*, 2011, pp 229-232.
[48] J. Hoffarta, F. M. Suchanekb, K. Berbericha, and et al, "YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia," *Artificial Intelligence*, vol. 194, pp. 28-61, Jan. 2013.
[49] *Wikipedia*. [Online]. Available: https://www.wikipedia.org/, accessed on: May 19, 2018.
[50] *MusicBrainz*. [Online]. Available: http://musicbrainz.org, accessed on: May 19, 2018.
[51] *WordNet*. [Online]. Available: http://wordnet.princeton.edu/, accessed on: May 19, 2018.
[52] *Linked Data*. [Online]. Available: http://linkeddata.org/, accessed on: May 19, 2018.
[53] *Attribution 2.5 Generic (CC BY 2.5)*. [Online]. Available: https://creativecommons.org/licenses/by/2.5/, accessed on: May 19, 2018.
[54] *RDF 1.1 N-Triples*. [Online]. Available: https://www.w3.org/TR/n-triples/, accessed on: May 19, 2018.
[55] C. Niel, "freebase-triples: A Methodology for Processing the Freebase Data Dumps," 2017. [Online]. Available: https://arxiv.org/abs/1712.08707v1, accessed on: May 19, 2018.
[56] D. Chen and H. Zhao, "Research on the Method of Extracting Domain Knowledge from the Freebase RDF Dumps," *IEEE Access*, to be published. DOI: 10.1109/ACCESS.2018.2868516.
[57] *Turtle-Terse RDF Triple Language*. [Online]. Available: https://www.w3.org/TeamSubmission/turtle/, accessed on: May 19, 2018.
[58] *Apache Jena - TDB*. [Online]. Available: http://jena.apache.org/documentation/tdb/index.html, accessed on: May 19, 2018.
[59] *Apache Jena*. [Online]. Available: http://jena.apache.org/, accessed on: May 19, 2018.
[60] *Apache Jena Fuseki*. [Online]. Available: http://jena.apache.org/documentation/fuseki2/index.html, accessed on: May 19, 2018.
[61] *Protege ontology editor*. [Online]. Available: http://protege.stanford.edu/, accessed on: May 19, 2018.
[62] *The European Project K4CARE*. [Online]. Available: http://www.k4care.net/, accessed on: May 19, 2018.

DEYAN CHEN received the M.S. degree in Computer Application Technology from Northeastern University, Shenyang, China, in 2003. He is currently working towards the Ph.D. degree in the School of Computer Science and Engineering, Northeastern University, Shenyang, China. From 2003 to 2005, he worked as a Senior Software Engineer at ZTE Corporation, Shenzhen, China. From 2005 to now, he has been working as a Senior Researcher at Neusoft Corporation, Shenyang, China. His research interests include natural language processing, semantic web, knowledge engineering, data mining, machine learning, network and information security.

HONG ZHAO received the Ph.D. degree in Computer Science and Technology from Northeastern University, Shenyang, China, in 1990. He is a professor in the School of Computer Science and Engineering, Northeastern University, Shenyang, China. He has been supervising Ph.D. students since 1996. His research interests include the next generation network, network management, network and information security.

XIA ZHANG received the Ph.D. degree in Computer Application Technology from Northeastern University, Shenyang, China, in 1994. She is a professor in the School of Computer Science and Engineering, Northeastern University, Shenyang, China. Her research interests include cloud computing application support and management platform, remote health management platform based on Internet of Things, data parallel processing.