International Journal of Innovative Computing, Information and Control
Volume **8**, Number **5(B)**, May **2012**

ICIC International ©2012 ISSN 1349-4198
pp. 3601–3616

ON AUTOMATIC CONSTRUCTION OF MEDICAL ONTOLOGY CONCEPT'S DESCRIPTION ARCHITECTURE

Yao Liu[1], Zhifang Sui[2], Qingliang Zhao[2], Yongwei Hu[2] and Ruijia Wang[1]

[1]Institute of Scientific and Technical Information of China
No. 15, Fuxing Road, Beijing 100038, P. R. China
liuy@istic.ac.cn

[2]Institute of Computational Linguistics
Peking University
No. 5, Yiheyuan Road, Haidian District, Beijing 100871, P. R. China

Received March 2011; revised July 2011

Abstract. *Using the techniques and theories of natural language processing (NLP), this paper puts forward research ideas and methods to automatically build domain ontologies based on text content. And for the clinical medical domain of modern medicine, this paper also constructs a concept system with multidimensional model by restructuring and utilizing the generally acknowledged domain knowledge, realizes the automatic construction and acquisition of modern medical knowledge descriptive system. Furthermore, in the ontology evolution, this paper comes up with an automatic method of retrieving attribute values based on Internet, which makes up for the defects of limited information and slow update in the automatic ontology construction based on single text. Our methods have achieved good results, and further provide theoretical foundation and technical support to automatic construction of professional domain ontologies.*

Keywords: Domain ontology, Knowledge engineering, Natural language processing, Ontology construction, Medical ontology

1. Introduction.

As a huge knowledge infrastructure project, the construction of Ontology is complex and should be carried out by stages. We choose modern medicine as the experimental subject in the beginning of our work, and establish the scientific methods and basic principles in the building of domain ontology's knowledge system in the medical field. After that, we establish the construct platform and service demonstration platform in this field, in the hope of providing preparations for building domain ontology in other fields in the future [1,2].

2. Design Philosophy and Construction Flow.

2.1. Design philosophy.

In the sight of the overall designing of ontology, the construction of ontology should form a concept model in the base of objective phenomena's concept abstraction, and the knowledge and conceptual set in the model should be widely recognized so that the model could be shared. What is more, the concept types, constraint conditions and utilization should have distinct description, and could be disposed automatically by machine to illustrate its explicit formalization. In engineering practice, the design and development of domain ontology usually depend on the specific description language and ontology utilities [3,4]. In other words, the ontology contains two models: one is knowledge representation model, and the other is concept model.

The knowledge representation model contains two parts.

(1) Building the Formalization Model

At present, domain knowledge bases are usually presented by ontology. Ontology is the explicit and formalized explanation of the concept model, and there are several kinds of formalized models which could demonstrate ontology in computer science, including first-order predicate logic, semantic web and frame system, and description logic system developed from the latter two.

(2) Determing the Description Language

There are various ontology description languages in the research, and two of the most famous description languages used in the present study are AI-based language and Web-based language.

The concept model: a model obtained by abstracting related concepts of objective phenomena [5,6].

In the remainder of this paper, we will introduce the "Construction of the Multidimensional Medical Ontology (construction of concept model)" and "Development of the Chinese Domain Ontology Construction Platform (knowledge representation)" respectively.

2.2. Construction flow (Figure 1).

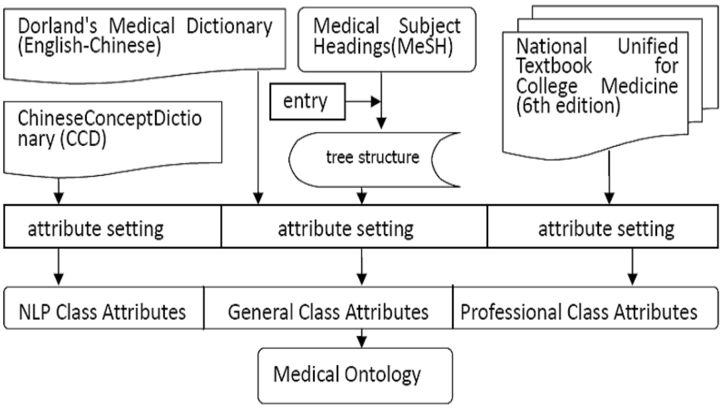

Dorland's Medical Dictionary (English-Chinese)
Medical Subject Headings(MeSH)
National Unified Textbook for College Medicine (6th edition)
entry
ChineseConceptDictionary (CCD)
tree structure
attribute setting
attribute setting
attribute setting
NLP Class Attributes
General Class Attributes
Professional Class Attributes
Medical Ontology

Figure 1. Construction flow of the medical ontology

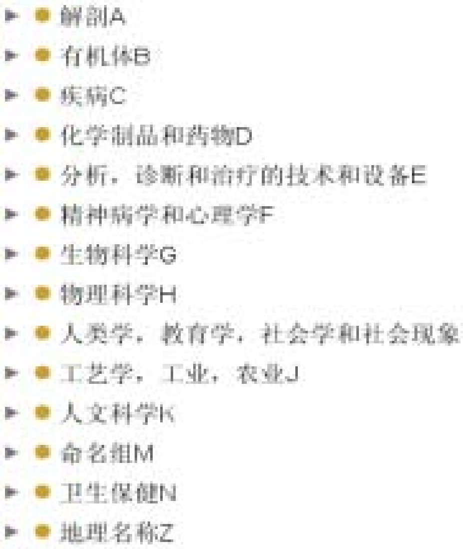
3. Construction of the Multidimensional Medical Ontology.

No matter from the specific perspective of the objective world or the abstract perspective of logic, the design of the concept in domain ontology should be closed to the relation rules between objective objects in the professional domain the researchers are studying [7,8]. Therefore, according to knowledge characteristics in medical domain, we come up with a medical domain ontology construction scheme based on a multidimensional model. By multi-dimension we mean to construct the ontology concept models of medical domain in different levels and categories on the basis of current medical knowledge structure. For instance, attributes such as Symptoms and Physical Signs, Laboratory and Auxiliary Examinations, Locations of Diseases, Treatment and Nursing Care are used to describe the disease class, while the attributes of Chemicals and Drugs are described by Effections and Purposes, Dosage Forms and Specification, Characteristics, Instruction and Dosage, Untoward Effects, Cautions, Storage, etc. Now we will illustrate in detail.

3.1. Prototype building.

After investigate the current medical knowledge systems both domestic and overseas, we choose the Medical Thesaurus (MeSH) compiled by American National Library of Medicine (NLM) as the base of our knowledge description system, because MeSH contains relatively complete subject terms and their hyponymies. We consider the subject terms in MeSH as the knowledge elements of modern medical ontology and their hyponymies as the tree structure of the knowledge elements. By this means we build the prototype of the medical knowledge description system.

We have 14 main classes in the medical knowledge description system, and each of them has their own subclasses. The main classes in our system are shown in the following figure (Figure 2).



► ● 解剖A
► ● 有机体B
► ● 疾病C
► ● 化学制品和药物D
► ● 分析，诊断和治疗的技术和设备E
► ● 精神病学和心理学F
► ● 生物科学G
► ● 物理科学H
► ● 人类学，教育学，社会学和社会现象
► ● 工艺学，工业，农业J
► ● 人文科学K
► ● 命名组M
► ● 卫生保健N
► ● 地理名称Z

FIGURE 2. Schematic diagram of the medical tree structure

However, most thesauri are mainly edited by librarians and information personnel for the purpose to describe library and information science, so they failed to deeply reflect the internal relations within the professional domain. To describe the knowledge more deeply and comprehensive, we must restructure the knowledge during the building of knowledge description system of domain ontology based on the thesaurus. In this paper, we restructure the knowledge in three levels according to the practical application requirements.

3.2. **Knowledge restructuring of class I: transformation from tree structure to multilayer-nested network structure.** In thesauri, subject words always distributed parallelly in several tree structures for the convenience of document indexing. For instance, MeSH distributes the key subject terms in medical domain into 14 main classes parallelly with no status distinction between the classes. But when we describe the knowledge of a certain domain, we always only choose the concepts in one class as the knowledge element, and take the concepts in other classes as the attribute of the concepts in the chosen class from different point of view. Since our application is used to offer retrieving service for disease counseling, we choose the subject term “Disease Class” as the basic unit for knowledge description to build a hierarchical structural system, in order to establish the longitudinal association of knowledge. Other classes (A dissection, B organism, D chemicals and drugs, E diagnosis and treatment techniques and equipment, N health care) are treated as attribute descriptions for the knowledge element of “illness classes” so as to establish the lateral association of disease class knowledge.

The lateral association is shown in the following figure (Figure 3).

We describe the attributes of attribute knowledge elements deeply to build the knowledge description frame of other 13 classes (Figure 4).

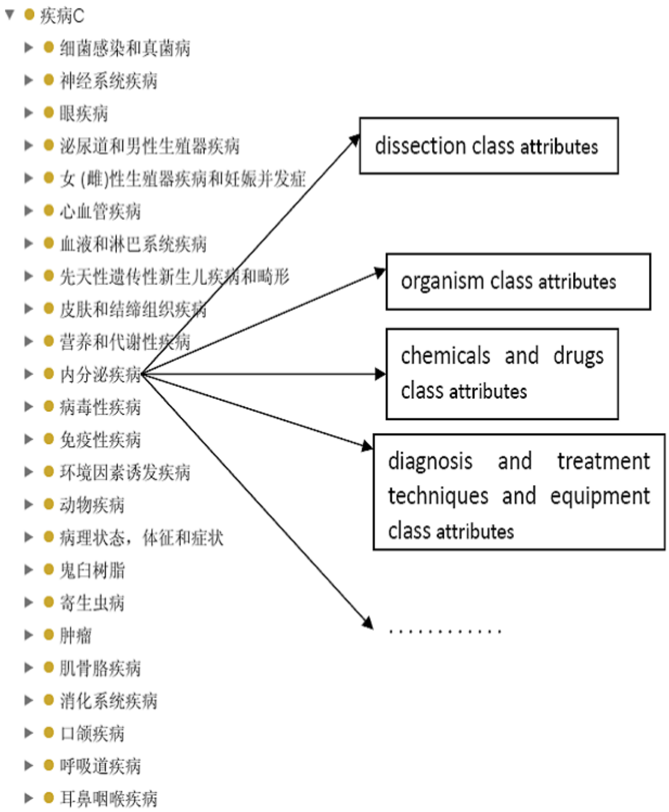


FIGURE 3. Schematic diagram of the lateral association of disease class knowledge

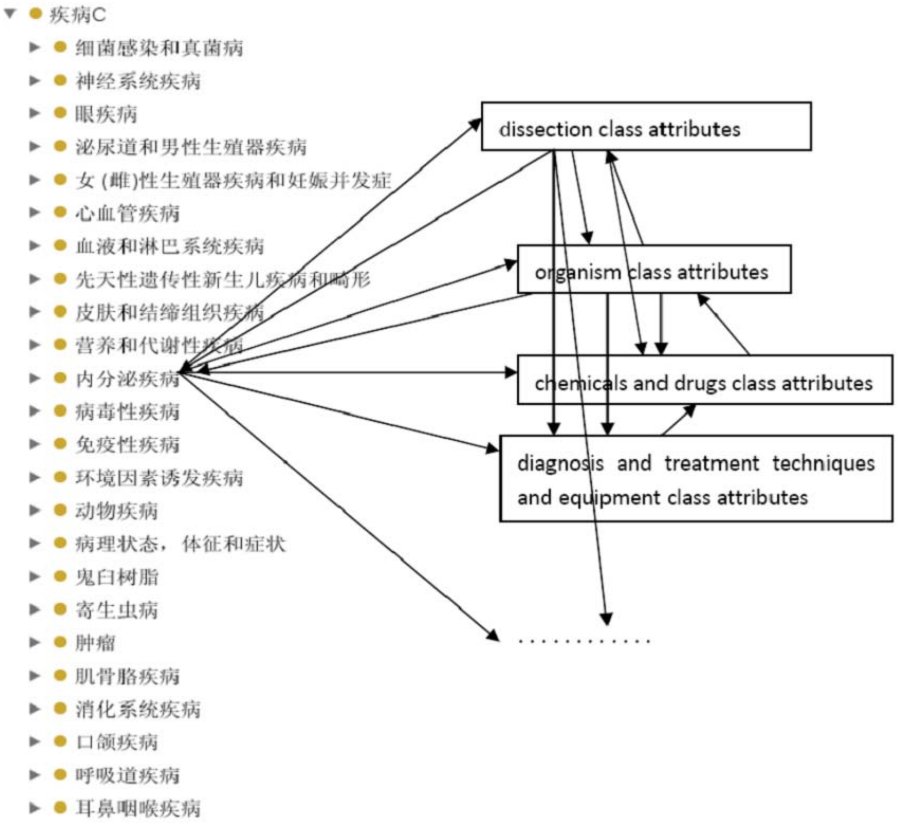


FIGURE 4. Schematic diagram of the knowledge description frame

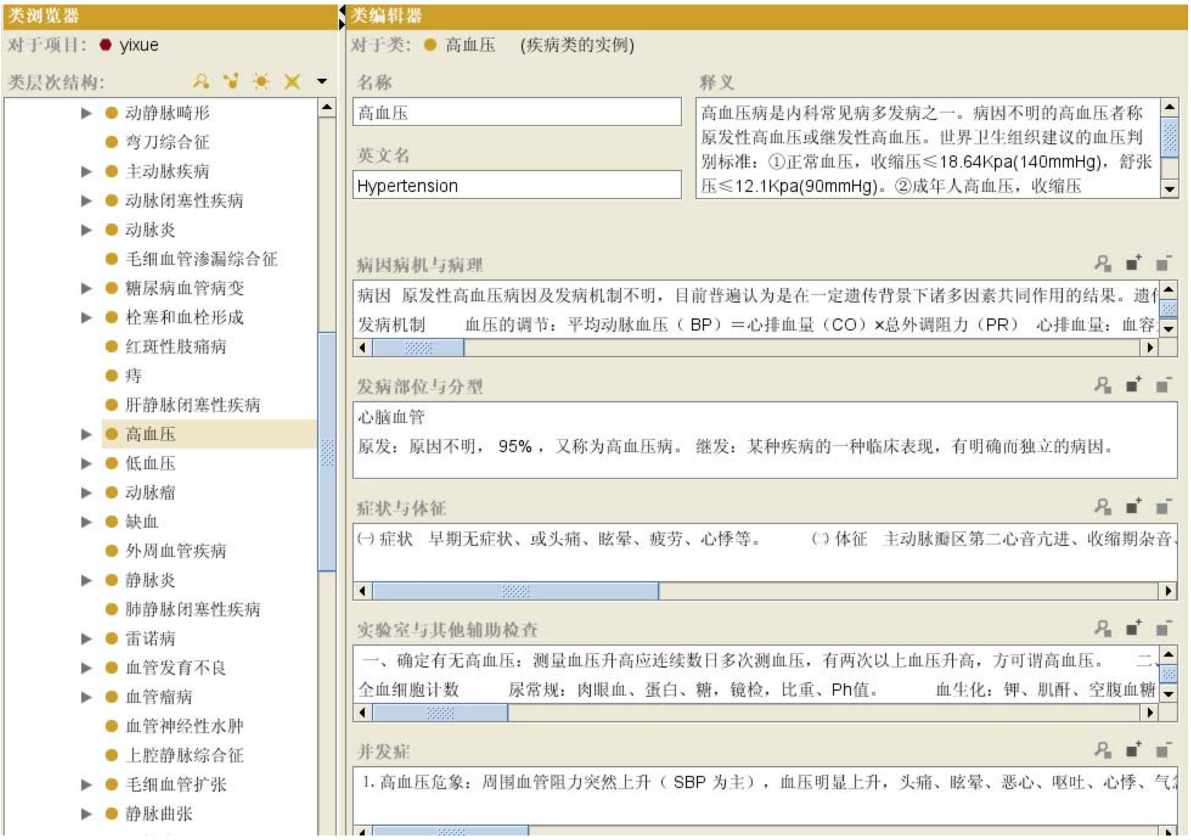
3.3. Knowledge restructuring of class II: transformation from document retrieval to expert system, describing knowledge with the core of application. The Knowledge restructuring of class I can change the description structure of concept system, but it cannot change the essential description of knowledge, so it is necessary to restructure the knowledge in the second level according to the professional knowledge of certain domain. We organize the medical domain knowledge from the perspective of clinic application, merge and split the rest classes in the knowledge description frame of disease class, get the clinic attributes descriptions of knowledge elements in this class, including Symptoms and Physical Signs, Laboratory and Auxiliary Examinations, Locations of Diseases, Treatment and Nursing Care, etc. Similarly, the attributes of the Chemicals and Drugs are described by the Effections and Purposes, Dosage Forms and Specification, Characteristics, Instruction and Dosage, Untoward Effects, Cautions, Storage, etc.

These restructurings can make the knowledge system also serve the clinical diagnosis and treatment as well as the document retrieval and labeling.

The clinical attributes description frame of the knowledge elements in disease class (Figure 5).

The description frame of the clinical attributes in dissection class (Figure 6).

The description frame of the clinic attributes in chemicals and drugs class (Figure 7).



▶ ● 动静脉畸形
● 弯刀综合征
▶ ● 主动脉疾病
▶ ● 动脉闭塞性疾病
▶ ● 动脉炎
● 毛细血管渗漏综合征
▶ ● 糖尿病血管病变
▶ ● 栓塞和血栓形成
● 红斑性肢痛病
● 痔
● 肝静脉闭塞性疾病
▶ ● 高血压
▶ ● 低血压
▶ ● 动脉瘤
▶ ● 缺血
● 外周血管疾病
▶ ● 静脉炎
● 肺静脉闭塞性疾病
▶ ● 雷诺病
▶ ● 血管发育不良
▶ ● 血管瘤病
● 血管神经性水肿
● 上腔静脉综合征
▶ ● 毛细血管扩张
▶ ● 静脉曲张

对于类: ● 高血压 (疾病类的实例)

名称: 高血压

英文名: Hypertension

释义: 高血压病是内科常见病多发病之一。病因不明的高血压者称原发性高血压或继发性高血压。世界卫生组织建议的血压判别标准：①正常血压，收缩压≤18.64Kpa(140mmHg)，舒张压≤12.1Kpa(90mmHg)。②成年人高血压，收缩压

病因病机与病理
病因 原发性高血压病因及发病机制不明，目前普遍认为是在一定遗传背景下诸多因素共同作用的结果。遗
发病机制 血压的调节：平均动脉血压（BP）＝心排血量（CO）×总外调阻力（PR） 心排血量：血容

发病部位与分型
心脑血管
原发：原因不明， 95%，又称为高血压病。继发：某种疾病的一种临床表现，有明确而独立的病因。

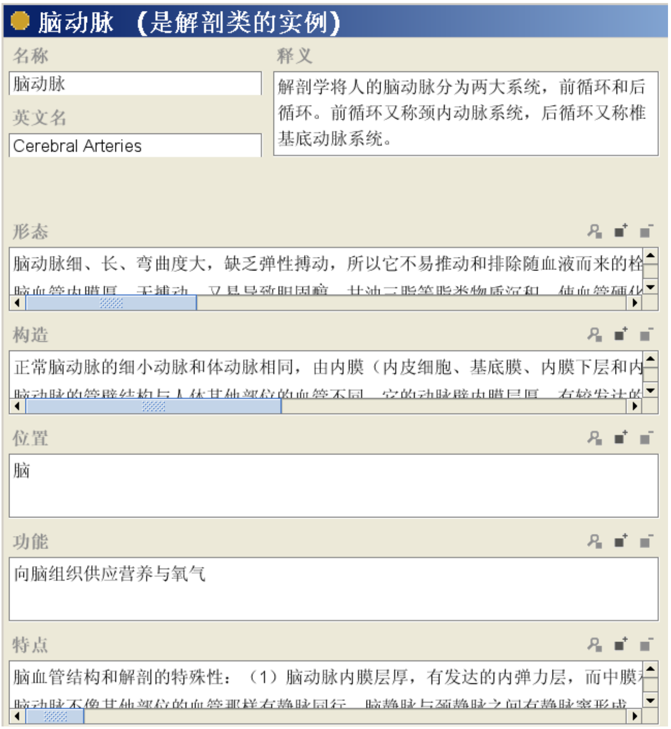
症状与体征
(一) 症状 早期无症状、或头痛、眩晕、疲劳、心悸等。 (二) 体征 主动脉瓣区第二心音亢进、收缩期杂音

实验室与其他辅助检查
一、确定有无高血压：测量血压升高应连续数日多次测血压，有两次以上血压升高，方可谓高血压。 二、
全血细胞计数 尿常规：肉眼血、蛋白、糖，镜检，比重、Ph值。 血生化：钾、肌酐、空腹血糖

并发症
1. 高血压危象：周围血管阻力突然上升（SBP 为主），血压明显上升，头痛、眩晕、恶心、呕吐、心悸、气

Figure 5. The clinical attributes description frame of knowledge elements in disease class

3.4. Knowledge restructuring of class III: transformation from natural language description to subject term description. Ontology should contain the connections between various concepts, but the current knowledge description methods describe the attributes of concepts by natural language, which only connects the concepts (subject terms of disease class) and the natural language descriptions instead of the connections



脑动脉 （是解剖类的实例）

名称

脑动脉

英文名

Cerebral Arteries

释义

解剖学将人的脑动脉分为两大系统，前循环和后循环。前循环又称颈内动脉系统，后循环又称椎基底动脉系统。

形态

脑动脉细、长、弯曲度大，缺乏弹性搏动，所以它不易推动和排除随血液而来的栓

脑血管内膜厚，无搏动，又易导致胆固醇、甘油三脂等脂类物质沉积，使血管硬化

构造

正常脑动脉的细小动脉和体动脉相同，由内膜（内皮细胞、基底膜、内膜下层和内

脑动脉的管壁结构与人体其他部位的血管不同，它的动脉壁内膜层厚，有较发达的

位置

脑

功能

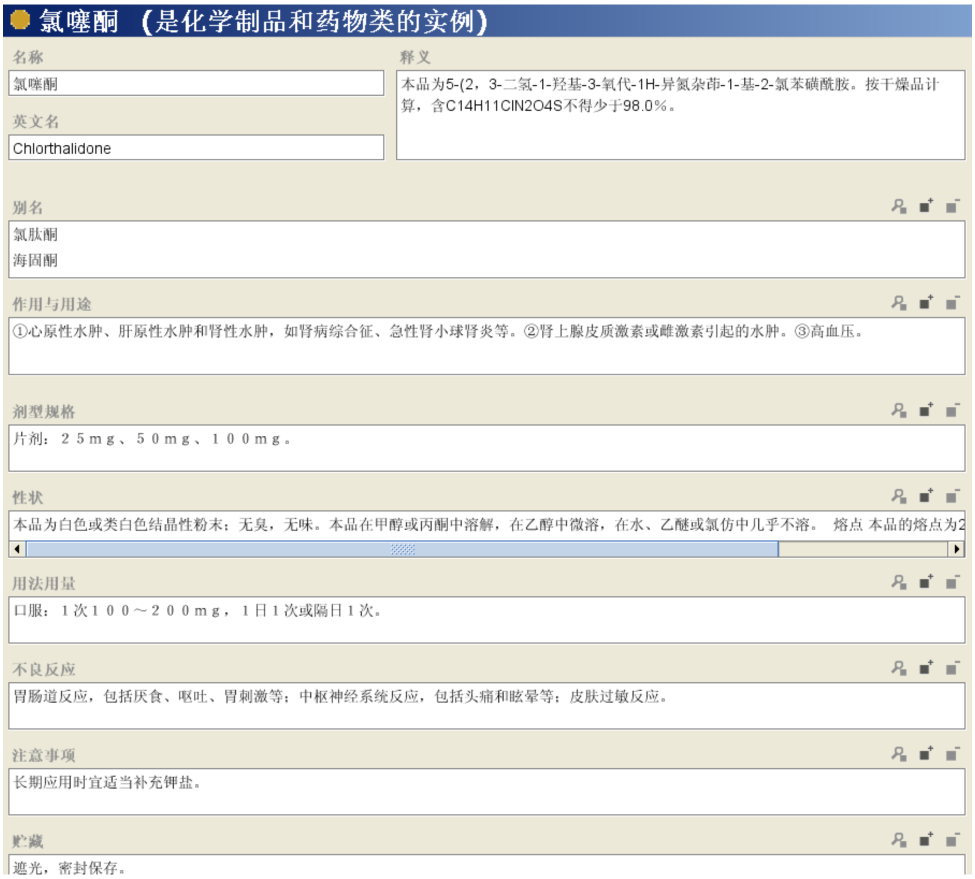
向脑组织供应营养与氧气

特点

脑血管结构和解剖的特殊性：（1）脑动脉内膜层厚，有发达的内弹力层，而中膜

脑动脉不像其他部位的血管那样有静脉同行，脑静脉与颈静脉之间有静脉窦形成

FIGURE 6. The description frame of the clinical attributes in dissection class



氯噻酮 （是化学制品和药物类的实例）

名称

氯噻酮

英文名

Chlorthalidone

释义

本品为5-(2，3-二氢-1-羟基-3-氧代-1H-异氮杂茚-1-基-2-氯苯磺酰胺。按干燥品计算，含C14H11ClN2O4S不得少于98.0%。

别名

氯肽酮
海固酮

作用与用途

①心原性水肿、肝原性水肿和肾性水肿，如肾病综合征、急性肾小球肾炎等。②肾上腺皮质激素或雌激素引起的水肿。③高血压。

剂型规格

片剂：２５ｍｇ、５０ｍｇ、１００ｍｇ。

性状

本品为白色或类白色结晶性粉末；无臭，无味。本品在甲醇或丙酮中溶解，在乙醇中微溶，在水、乙醚或氯仿中几乎不溶。 熔点 本品的熔点为2

用法用量

口服：１次１００～２００ｍｇ，１日１次或隔日１次。

不良反应

胃肠道反应，包括厌食、呕吐、胃刺激等；中枢神经系统反应，包括头痛和眩晕等；皮肤过敏反应。

注意事项

长期应用时宜适当补充钾盐。

贮藏

遮光，密封保存。

FIGURE 7. The description frame of the clinic attributes in chemicals and drugs class

between concepts. So the current language descriptions are not highly formalized. Therefore, we proceed the Knowledge Restructuring of Class III: transformation from natural language description to subject term description. Using the semantic analysis and the subject indexing techniques in natural language, we index the subjects indexing of the concept attributes described by natural language, add the knowledge element – subject term description for concept attributes, as shown in the parts within the red circle in the following figure.

Furthermore, we build the corresponding NLP semantic class descriptions for all the knowledge elements – subject terms of "disease class" (C), including free words, synonyms, related words, CCD conception dictionary words, etc. After that, we establish the corresponding relations between the natural language and the subject terms, and build the bridge between the user's natural language queries and formalized/structured knowledge bases [9].

3.5. **Knowledge description system of medical domain ontology.**

We use subject terms as the knowledge representation units (include more than 20,000 bilingual (Chinese-English) subject terms), the subject terms are divided into 15 classes, such as A-dissection, B-organism, D-chemicals and drugs, E-techniques and equipments of diagnosis and treatment and N health care.

Description of all kinds of diseases includes (Figure 8):

Concept class description: including name, English name, paraphrase, code and constraints.

Clinical class description: including symptoms and signs, laboratory and auxiliary examinations, locations of diseases, cause of disease, pathogenesis and pathology, etc.

Every kind of attributes is described in two forms: natural language text description and knowledge element description. NLP semantic description includes the freed words, synonyms, relative words and so on. The description of the knowledge elements in other class is deeply and widely connected with each other.

Our knowledge description system has the following features: based on the international standards, clinical practice oriented, and multilayer-nested network knowledge description.

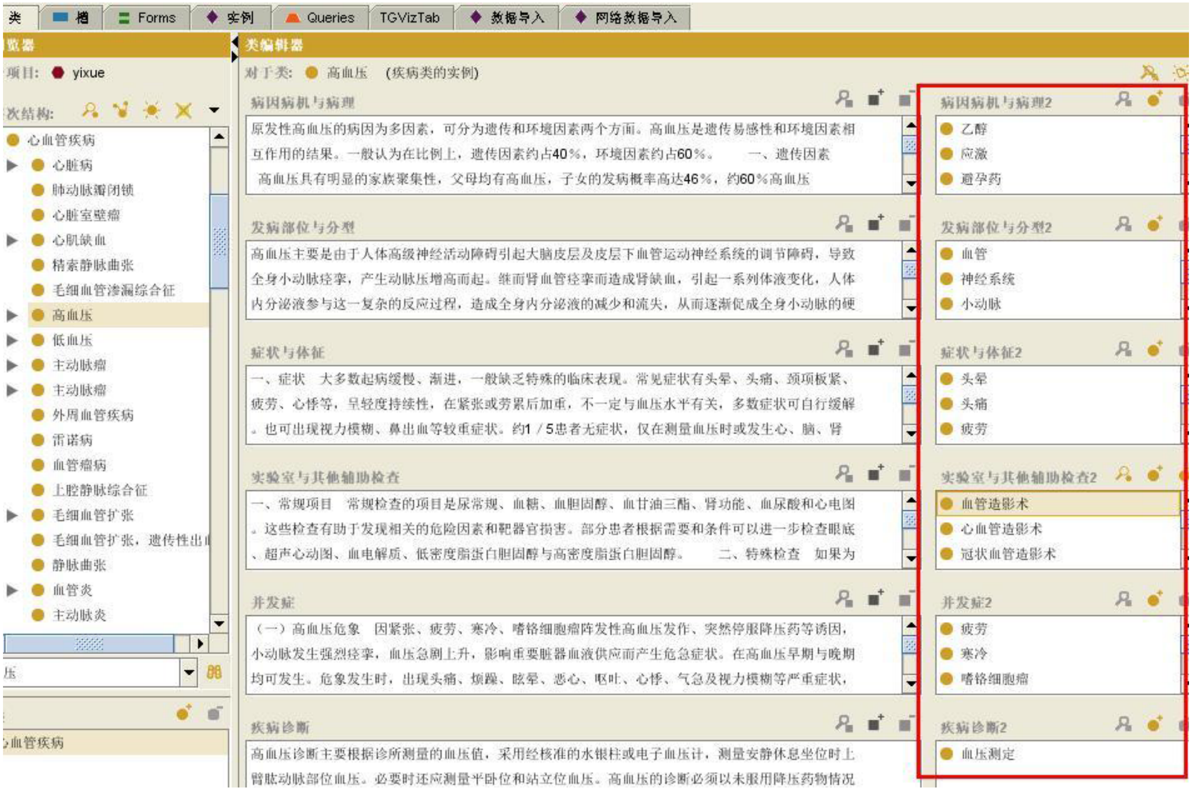
4. **Development of the Chinese Domain Ontology Construction Platform.**

The Chinese domain ontology construction platform is developed on the basis of Protégé, an open source tool. In this chapter we will introduce main functions of this platform, the process of ontology construction using this platform, key technologies in the platform's development and conclusion of the whole work [10].

4.1. **Differences between the knowledge representation model used by Protégé and OKBC.**

We need to know about some basic concepts of ontology to develop the ontology construction platform, that is, the knowledge representation model of ontology. So we will introduce the knowledge representation model of ontology defined by the agreement of Open Knowledge-Based Connectivity (OKBC).

Protégé is based on the OKBC knowledge representation model. The OKBC knowledge representation model is mainly designed to maximize the generalization and the synergy among knowledge representation systems. Therefore, the OKBC knowledge representation model makes the least restrictions of knowledge representation and is commonly used. OKBC integrates the concepts of all systems based on frame, some of which are conflicted and not very compatible. But OKBC is aimed at pursuing the maximum flexibility, so it always permits their coexistence when two concepts are conflicted. Although this design



对于类: 高血压 (疾病类的实例)

病因病机与病理

原发性高血压的病因为多因素，可分为遗传和环境因素两个方面。高血压是遗传易感性和环境因素相互作用的结果。一般认为在比例上，遗传因素约占40％，环境因素约占60％。　　一、遗传因素　高血压具有明显的家族聚集性，父母均有高血压，子女的发病概率高达46％，约60％高血压

发病部位与分型

高血压主要是由于人体高级神经活动障碍引起大脑皮层及皮层下血管运动神经系统的调节障碍，导致全身小动脉痉挛，产生动脉压增高而起。继而肾血管痉挛而造成肾缺血，引起一系列体液变化，人体内分泌液参与这一复杂的反应过程，造成全身内分泌液的减少和流失，从而逐渐促成全身小动脉的硬

症状与体征

一、症状　大多数起病缓慢、渐进，一般缺乏特殊的临床表现。常见症状有头晕、头痛、颈项板紧、疲劳、心悸等，呈轻度持续性，在紧张或劳累后加重，不一定与血压水平有关，多数症状可自行缓解。也可出现视力模糊、鼻出血等较重症状。约1／5患者无症状，仅在测量血压时或发生心、脑、肾

实验室与其他辅助检查

一、常规项目　常规检查的项目是尿常规、血糖、血胆固醇、血甘油三酯、肾功能、血尿酸和心电图。这些检查有助于发现相关的危险因素和靶器官损害。部分患者根据需要和条件可以进一步检查眼底、超声心动图、血电解质、低密度脂蛋白胆固醇与高密度脂蛋白胆固醇。　　二、特殊检查　如果为

并发症

（一）高血压危象　因紧张、疲劳、寒冷、嗜铬细胞瘤阵发性高血压发作、突然停服降压药等诱因，小动脉发生强烈痉挛，血压急剧上升，影响重要脏器血液供应而产生危急症状。在高血压早期与晚期均可发生。危象发生时，出现头痛、烦躁、眩晕、恶心、呕吐、心悸、气急及视力模糊等严重症状，

疾病诊断

高血压诊断主要根据诊所测量的血压值，采用经核准的水银柱或电子血压计，测量安静休息坐位时上臂肱动脉部位血压。必要时还应测量平卧位和站立位血压。高血压的诊断必须以未服用降压药物情况

病因病机与病理2
- 乙醇
- 应激
- 避孕药

发病部位与分型2
- 血管
- 神经系统
- 小动脉

症状与体征2
- 头晕
- 头痛
- 疲劳

实验室与其他辅助检查2
- 血管造影术
- 心血管造影术
- 冠状血管造影术

并发症2
- 疲劳
- 寒冷
- 嗜铬细胞瘤

疾病诊断2
- 血压测定

FIGURE 8. Schematic diagram of concept attributes

can assure the maximum commonality, the designer has to make some rules to restrict its generalization.

However, the Protégé knowledge representation model [11] is fully compatible with OKBC: Protégé realized every concept required by OKBC knowledge model. Every concept in Protégé keeps correspondence with OKBC in logic. For some concepts with flexibility in OKBC, Protégé keeps its flexibility as much as possible, only makes some necessary adjustments according to the needs of present user interface. The knowledge representation model used by Protégé increases its practicability at the cost of some flexibility. The following Table 1 lays out the main differences between Protégé-2000 and OKBC.

TABLE 1. Differences between knowledge representation models of Protégé and OKBC

OKBC	Protégé-2000
A frame can be the instance of multiple classes	A frame can be the instance of only one class
A frame doesn't have to be the instance of any class	A frame must be the instance of one class
A private slot can be attached directly to any frame	All private slot attached to any frame must be derived from the corresponding template slots
Classes, slots, descriptions and entities may not be frames	Any classes, slots, descriptions and entities may be frames
A frame can be a class, a slot and a description at the same time	A frame can only be a class, a slot or a description at one time

4.2. **The overall structure of the ontology construction platform.** Figure 9 shows the overall structure of the construction platform.

In addition to the basic functions of the Chinese domain ontology construction platform of creating and editing the ontology, there are other functions in this ontology platform: supporting the data importing in various forms, saving current ontology as all kinds of forms and ontology data retrieval. Figure 10 is the main interface of the ontology construction platform, which shows the medical domain ontology constructed on the platform.

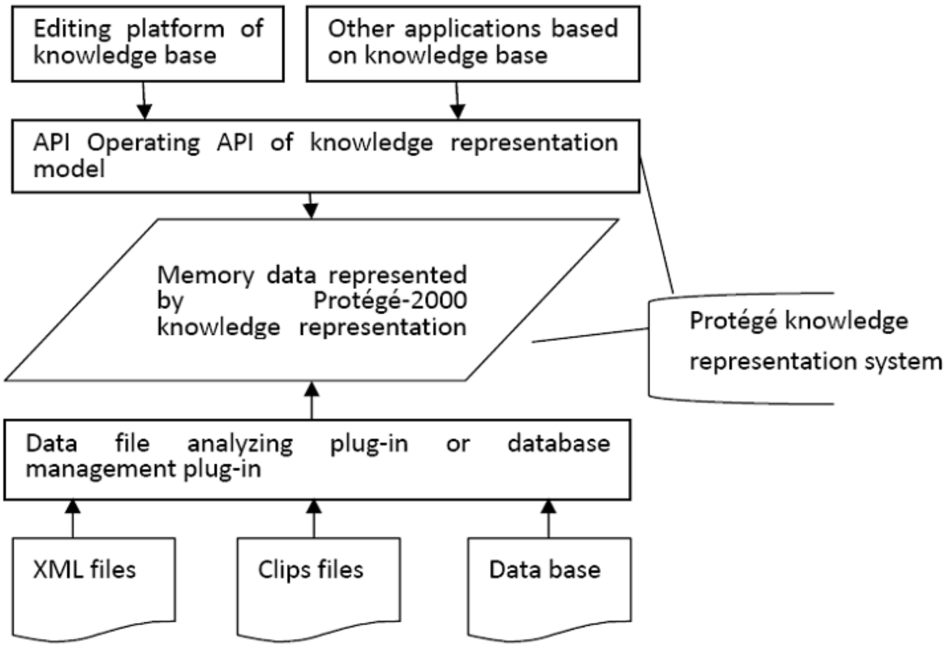
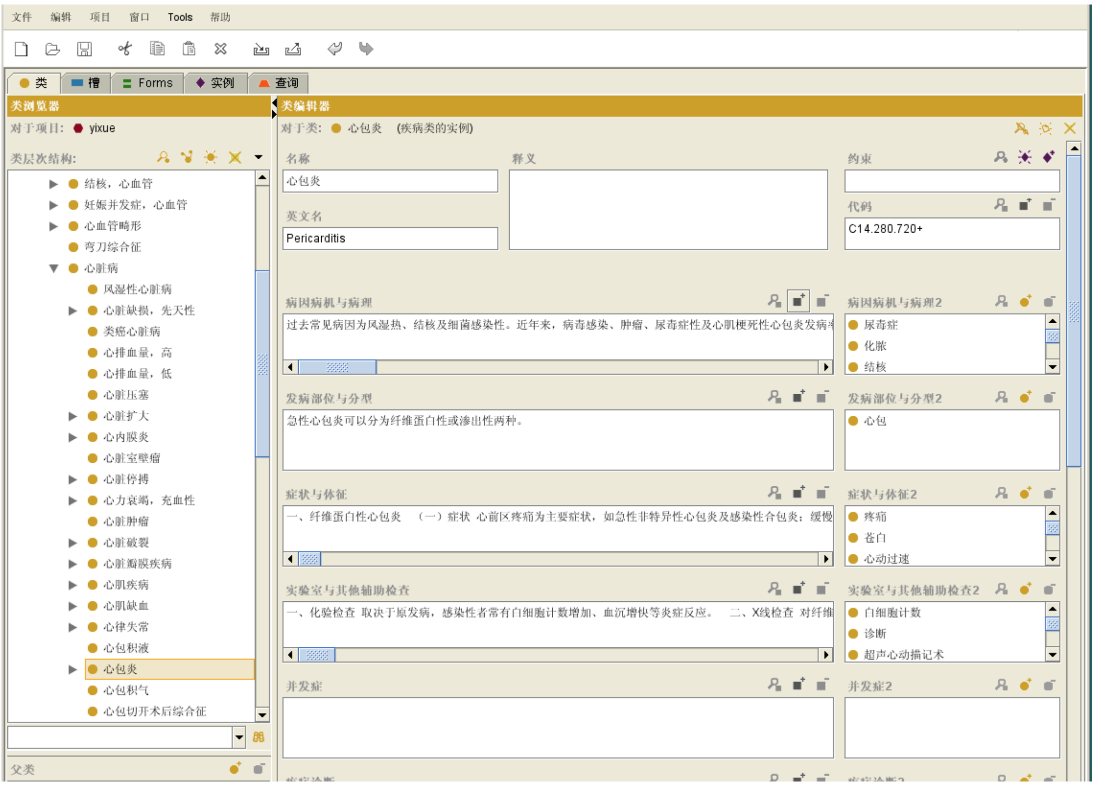

Editing platform of knowledge base
Other applications based on knowledge base
API Operating API of knowledge representation model
Memory data represented by Protégé-2000 knowledge representation
Protégé knowledge representation system
Data file analyzing plug-in or database management plug-in
XML files
Clips files
Data base

Figure 9. The overall structure of the ontology construction platform

4.3. **Creating an ontology.** The simplest way to create ontology in the editing platform is to create a blank ontology, then add and modify ontology elements like class and attributes. There are also other ways to create ontology, such as importing current existing ontology data into the platform. The ontology construction platform has some limitations to the storage formats of ontology. So we need do some extra work in either of the following two ways if the current domain ontology is incompatible to the formats supported: one is to transfer the current ontology data to a supported format before importing it; the other is to developing some plug-ins. It is also a feature of the platform to support user-defined plug-ins. The editing platform can support any format currently used in ontologies through developing different plug-ins.

4.4. **Editing and searching the ontology.** The edit of the ontology includes the addition, deletion and amendment of the basic concepts defined by the ontology knowledge representation model. Here we mainly address the adjustment of the hierarchical structure between different classes. On the interface, the hierarchical structure between different classes is shown as a tree structure. When adjusting the structure, we can randomly change the position of the class in the tree structure by a drag and drop way. Searching essential elements in the ontology is another function of the platform, the editing platform offers a range of searching functions including single knowledge element searching, simple exact matching and some advanced functions like approximate matching. For instance, in this platform we can search for the class with a particular noun or a particular attribute,



对于项目: yixue

类层次结构:

- 结核，心血管
- 妊娠并发症，心血管
- 心血管畸形
- 弯刀综合征
- 心脏病
  - 风湿性心脏病
  - 心脏缺损，先天性
  - 类癌心脏病
  - 心排血量，高
  - 心排血量，低
  - 心脏压塞
  - 心脏扩大
  - 心内膜炎
  - 心脏室壁瘤
  - 心脏停搏
  - 心力衰竭，充血性
  - 心脏肿瘤
  - 心脏破裂
  - 心脏瓣膜疾病
  - 心肌疾病
  - 心肌缺血
  - 心律失常
  - 心包积液
  - 心包炎
  - 心包积气
  - 心包切开术后综合征

对于类: 心包炎 (疾病类的实例)

名称: 心包炎

释义:

约束:

英文名: Pericarditis

代码: C14.280.720+

病因病机与病理: 过去常见病因为风湿热、结核及细菌感染性。近年来，病毒感染、肿瘤、尿毒症性及心肌梗死性心包炎发病

病因病机与病理2: 尿毒症
化脓
结核

发病部位与分型: 急性心包炎可以分为纤维蛋白性或渗出性两种。

发病部位与分型2: 心包

症状与体征: 一、纤维蛋白性心包炎 （一）症状 心前区疼痛为主要症状，如急性非特异性心包炎及感染性合包炎；缓慢

症状与体征2: 疼痛
苍白
心动过速

实验室与其他辅助检查: 一、化验检查 取决于原发病，感染性者常有白细胞计数增加、血沉增快等炎症反应。 二、X线检查 对纤维

实验室与其他辅助检查2: 白细胞计数
诊断
超声心动描记术

并发症:

并发症2:

FIGURE 10. The main interface of editing platform

we can also search for the classes connected with a certain class in some ways, and the specific classes with certain constraint.

4.5. **Transformation of knowledge storage formats.** Another primary feature of the Chinese domain ontology construction platform is the transformation between knowledge storage formats. The platform has "import" and "export" functions. At present the formats supported for mutual transformation are limited, but we can solve the problem of unsupported format transformation by developing plug-ins. Taking creating the computer domain ontology as an example: the input of the ontology construction platform – the ontology information of computer domain is saved in a half structured text, this storage format is different to those provided by the construction platform. In this paper we solve this problem by developing a plug-in for the platform. The plug-in is designed to import the ontology data in current task into the Chinese domain ontology construction platform so that we can edit the domain ontology through the platform (Figure 11).

4.6. **Developing plug-ins.** In this section we will mainly introduce the plug-in developed in this paper. In order to realize the data import from designated texts, the type of the developed plug-in is "Create Project". This plug-in is used for creating new Protégé project in the second way we have talked about above, and its input is the knowledge base that saved in other formats. The plug-in can analyze the contained knowledge elements, the relation between the knowledge elements and the attributes of the knowledge elements, and save them in the way that can be recognized by Protégé.

Protégé has already offered certain interfaces to this type of plug-ins as long as the class can be realized. The class that needs to be inherited is "AbstractCreateProject Plugin" class. As for this class, we need to implement the function of Project createProject(), whose function is to create a project and return it. In the class of ComputerPlugin, the implement process of the createProject function is: firstly create a new Project

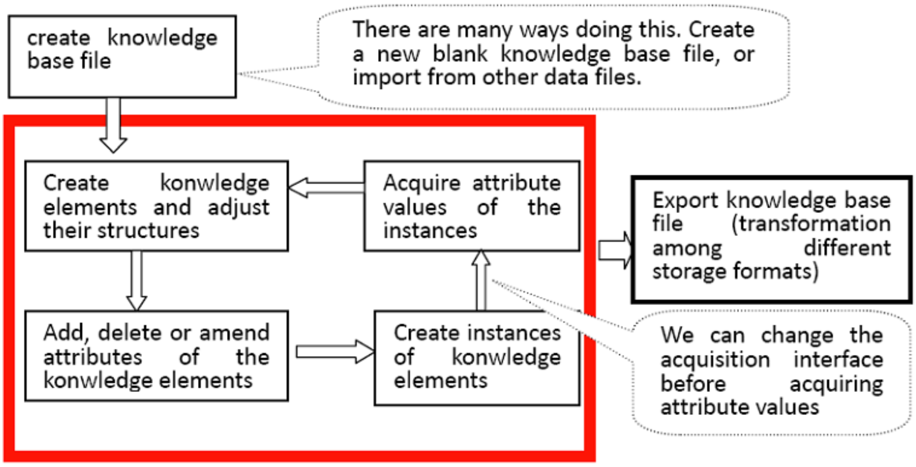

create knowledge base file
There are many ways doing this. Create a new blank knowledge base file, or import from other data files.
Create konwledge elements and adjust their structures
Acquire attribute values of the instances
Export knowledge base file (transformation among different storage formats)
Add, delete or amend attributes of the konwledge elements
Create instances of konwledge elements
We can change the acquisition interface before acquiring attribute values

Figure 11. Creating an ontology

_medicineProject, then add knowledge elements and attributes into the "Project", lastly add hierarchical structure of the knowledge elements. And the process of adding knowledge elements to Protégé is to create metaclasses which will determine the attributes contained in the knowledge elements first, then abstract the knowledge elements and the values of their attributes from text and add them to the knowledge elements.

4.7. **Process of creating ontology with the platform.** Figure 11 shows the main process of creating Chinese domain ontology with the "Chinese domain ontology construction platform". We can build domain ontology with the following process: firstly use the platform to create a new blank knowledge base file, which is saved as the default form of the platform. Then add the classes in the domain, and lastly add slots to different classes. It should be noted that, we need to add the slots and their attributes in the slot editing window before adding slots to classes, because the slot is also a type of frame in the platform. We also need to establish the hierarchical relations between the concepts when creating concepts. If you want to save the ontology created by this platform in other formats, you can export the current editing knowledge base file into other formats such as XML format and plain text format. Figure 11 is a brief introduction to the functions of the platform. The detailed process is shown in Figure 12.

The main features of this ontology construction platform are as follows:

Diversified import and export formats (RTF/XML/OWL): offering convenient to the knowledge communication, knowledge sharing and reusing of related international ontologies;

Powerful edit functions: including hierarchical structure adjustment, attribute relation adjustment, the adding, deleting and amending of the attribute values;

Powerful searching functions: supporting both exact and fuzzy searching for knowledge elements or attributes.

5. **Ontology Evolution.** Ontology is an open and integrated system, whose underlying knowledge bases and concept sets should be amended and updated with the update and development of disciplines and domains [9]. Therefore, we put forward a method of automatically abstracting ontology attribute values based on Internet. We first put forward an interactive way of selecting sentences which contain certain attribute values and extracting certain attribute values based on small scale seed-sets of attribute values. Then, utilizing the redundancy of the Internet information, we automatically abstract

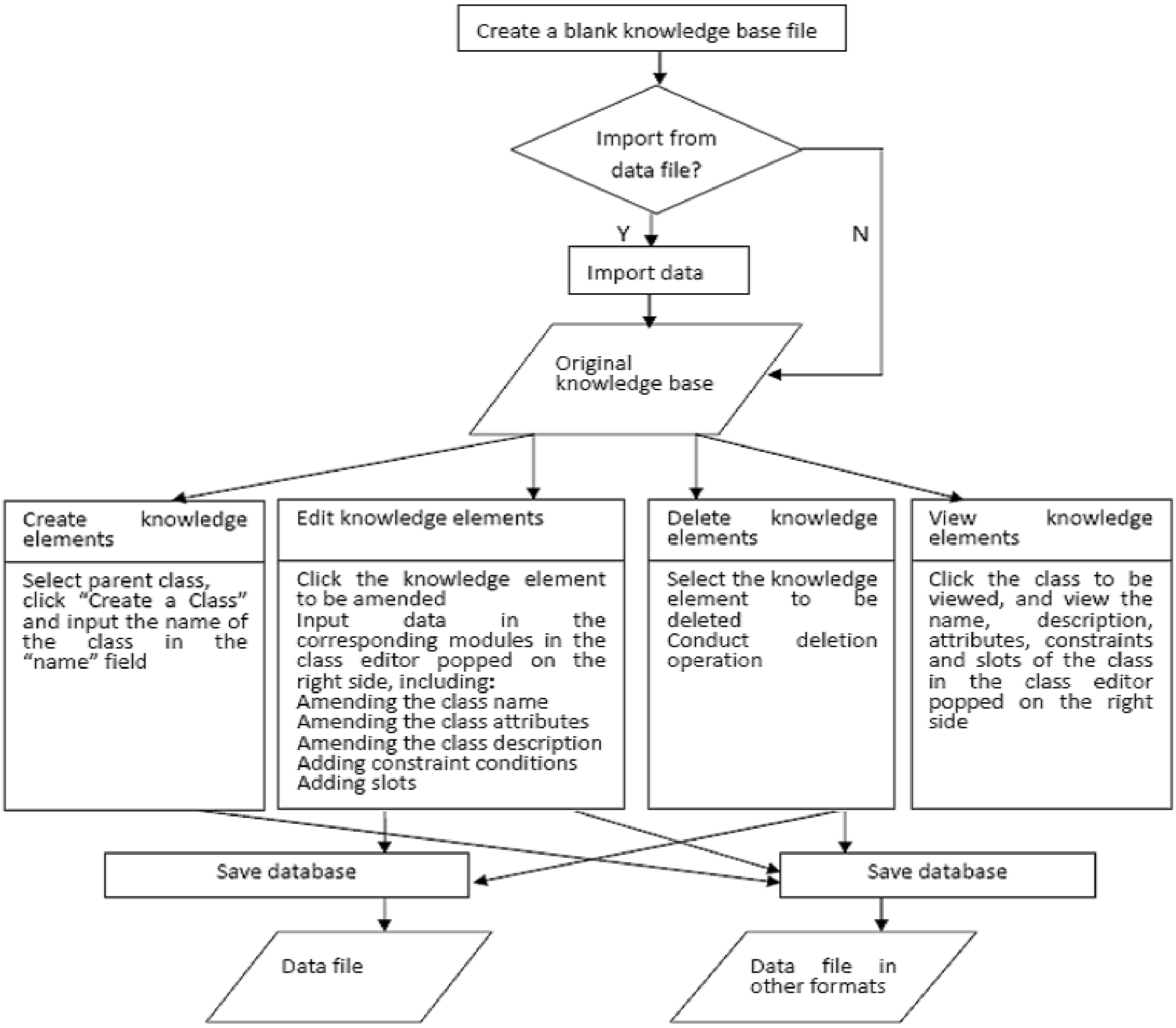

Create a blank knowledge base file
Import from data file?
Y
N
Import data
Original knowledge base
Create knowledge elements
Select parent class, click "Create a Class" and input the name of the class in the "name" field
Edit knowledge elements
Click the knowledge element to be amended
Input data in the corresponding modules in the class editor popped on the right side, including:
Amending the class name
Amending the class attributes
Amending the class description
Adding constraint conditions
Adding slots
Delete knowledge elements
Select the knowledge element to be deleted
Conduct deletion operation
View knowledge elements
Click the class to be viewed, and view the name, description, attributes, constraints and slots of the class in the class editor popped on the right side
Save database
Save database
Data file
Data file in other formats

FIGURE 12. Detailed process of creating an ontology knowledge base

and expand the target attribute value sets. Furthermore, to avoid artificially structuring any attribute value seed-set, we put forward a way of automatically generating seed-sets.

5.1. **Fundamental ideas.** As is well known, the enormous quantity of webpage information gained from Internet is always redundant. The information redundancy will affect the efficiency of information acquisition for people; however, it is helpful for computers to automatically acquire knowledge. To be specific, the webpage information redundancy will help the automatically ontology attribute values abstracting in two ways:

Firstly, it assists computers to judge the reliability and the authority of webpage information. For instance, we type the key words "感冒 (cold) 症状 (symptom)" and search in Google. In the 10 most relevant web pages, 5 of them describe the same cold symptoms, accounting for 50%. So we can add up the candidate attribute value phrases in relevant webpages and the frequently occurred phrases are likely the target attribute values.

Secondly, we only need to use simple grammatical structures to ensure the completeness of information. We choose the most relevant 100 web pages in the searching above, we find that although the cold symptoms are identical in those web pages, they are distributed in various grammatical structures. So it would not affect much on the completeness of information even if we abandoned some of the structures. For instance, in this paper we only consider the parallel structure which is easy to handle and also ensure the effects, the experimental results showed that it has very little influence on the completeness of the results.

The most difficult challenge of NLP techniques in handling enormous quantity of information is that its precision and processing speed failed to meet the actual needs when handling complicated language structures. This paper makes use of the redundancy in webpage information and avoids the irregularities of the languages in the web pages. Consequently, the NLP techniques can be better applied to handle with the tremendous information in web pages.

5.2. **Experiment setting and results analysis.**

5.2.1. *Experiment setting.* We use the modern medical domain ontology in the experiment (see Subsection 4.3 for detail). Our experiment employs the Chinese word segmentation and POS tagging software developed by the Computational Linguistics Institute, Peking University for word segmentation and POS tagging, we will manually judge the automatically filled results and further measure them by F value:

$$F\text{-}value = \frac{2 * precision * recall}{precision + recall} \tag{1}$$

Precision means the accuracy rate of the filled results and recall is the recall rate of the filled results.

5.2.2. *The preprocessing of web pages.* We use Google API to get the original web pages. We extract information from the top 100 related web pages among the search results. Before using them, we need to denoise the original web pages, i.e., to extract the body text of the web pages. The denoising work is based on two hypotheses: A. Most web pages on the Internet are blocked by Table or Div; B. The links in the body takes a small proportion of the overall body of the web pages while the links of guidance, advertisement and index take a very large part. Based on the two hypotheses, we first construct the Dom tree of web pages, then analyze the proportion of links under the leaf tags of Table and Div. If the proportion of the links exceeds 50%, this part will be considered as noise and removed.

5.2.3. *Experimental results and analysis.*

(1) Extracting attribute values of mutual supervision under the weak guidance

We set the seed-set {"咳嗽 (cough)"} to fill the attribute value of Class "感冒 (cold)"and sort according to the weight, choose the top 30 candidate attribute values as follows:

We can see from the results that on the condition of inputting Class "感冒 (cold)" and one of its "symptoms" "咳嗽 (cough)", we can obtain other symptoms tightly related to "感冒 (cold)". Particularly, only with the basic operation of word segmentation and POS tagging, we can extract terms related to cold, such as "鼻塞声重 (nasal tampon and low voice speaking)", "恶风寒 (badly chill cold)" and "全身酸楚 (systemic distressed)".

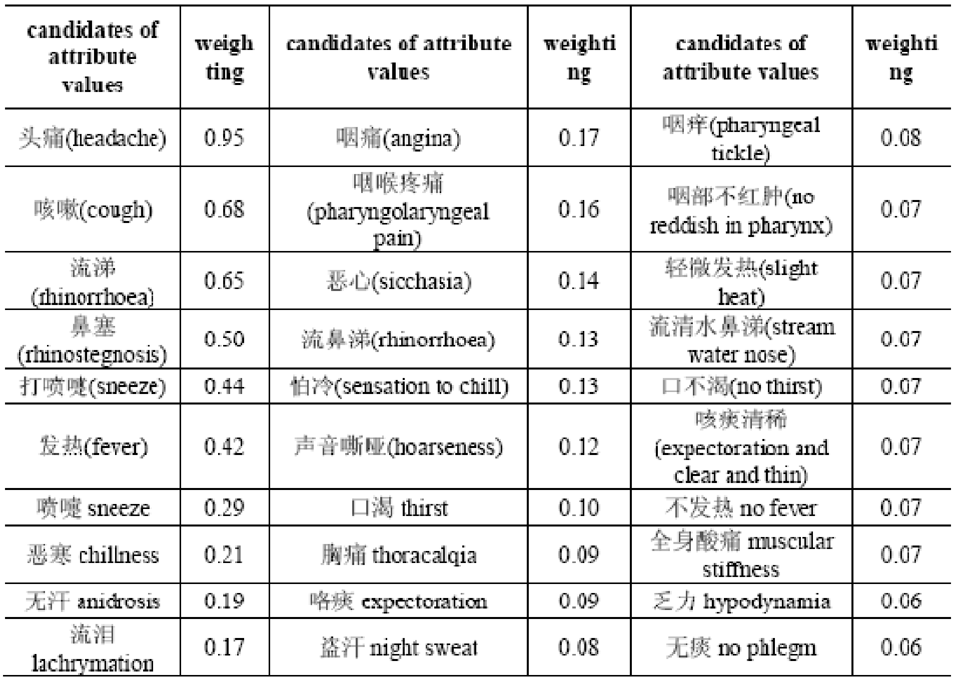
The mean of F value: 0.802. From the results we can see the automatically extracted results are satisfactory and the recall rate is higher than precision.

(2) Extracting attribute values with no guidance

We set a seed-set "咳嗽 (cough)" for the attribute "symptom" of Class "感冒 (cold)" through the process of Extracting attribute values with no guidance. The table following (Table 4) lists the results of attribute value extraction of attribute "symptom" of Class "慢性咽炎 (chronic pharyngitis)" (top 30 according to their weight).

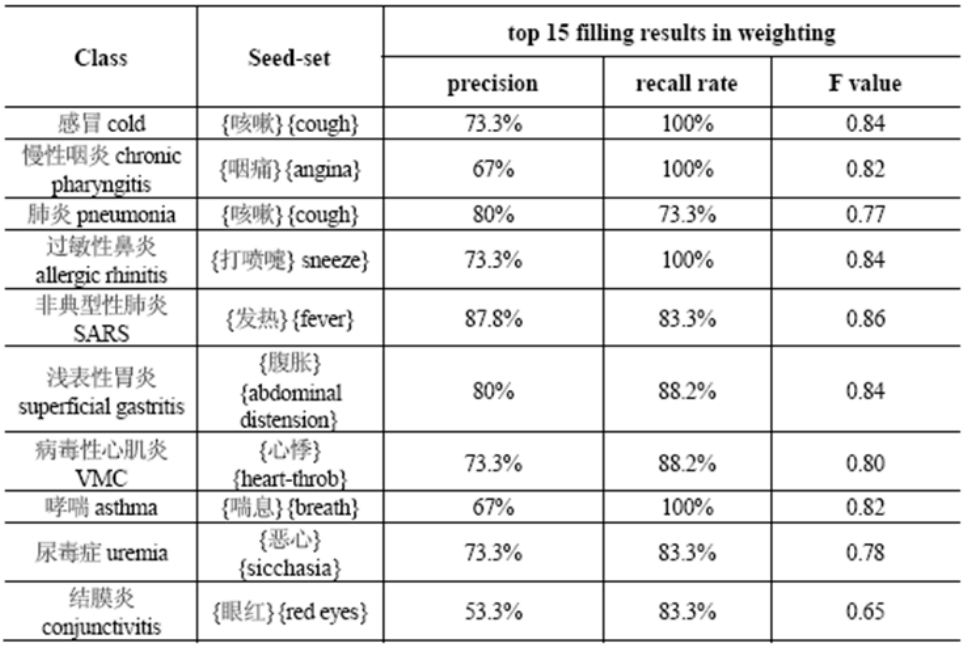
From this example we can see that with no attribute value of target class, this method can also achieve good results. Certainly this is a visualized and concrete example, and a greater scale of test is needed for more objective appraisal.

Table 2. The extraction results of the single class attribute value with seed-sets



candidates of attribute values	weighting	candidates of attribute values	weighting	candidates of attribute values	weighting
头痛(headache)	0.95	咽痛(angina)	0.17	咽痒(pharyngeal tickle)	0.08
咳嗽(cough)	0.68	咽喉疼痛 (pharyngolaryngeal pain)	0.16	咽部不红肿(no reddish in pharynx)	0.07
流涕 (rhinorrhoea)	0.65	恶心(sicchasia)	0.14	轻微发热(slight heat)	0.07
鼻塞 (rhinostegnosis)	0.50	流鼻涕(rhinorrhoea)	0.13	流清水鼻涕(stream water nose)	0.07
打喷嚏(sneeze)	0.44	怕冷(sensation to chill)	0.13	口不渴(no thirst)	0.07
发热(fever)	0.42	声音嘶哑(hoarseness)	0.12	咳痰清稀 (expectoration and clear and thin)	0.07
喷嚏 sneeze	0.29	口渴 thirst	0.10	不发热 no fever	0.07
恶寒 chillness	0.21	胸痛 thoracalqia	0.09	全身酸痛 muscular stiffness	0.07
无汗 anidrosis	0.19	咯痰 expectoration	0.09	乏力 hypodynamia	0.06
流泪 lachrymation	0.17	盗汗 night sweat	0.08	无痰 no phlegm	0.06

Table 3. The extraction results of multiple class attribute value



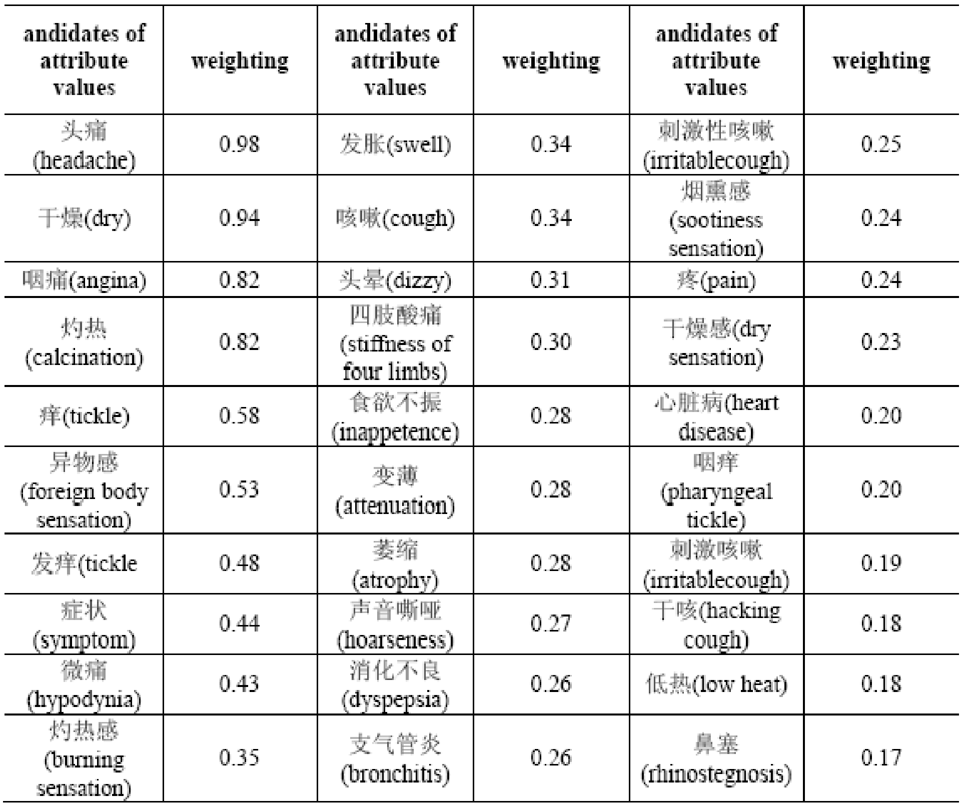
Class	Seed-set	top 15 filling results in weighting		
		precision	recall rate	F value
感冒 cold	{咳嗽}{cough}	73.3%	100%	0.84
慢性咽炎 chronic pharyngitis	{咽痛}{angina}	67%	100%	0.82
肺炎 pneumonia	{咳嗽}{cough}	80%	73.3%	0.77
过敏性鼻炎 allergic rhinitis	{打喷嚏} sneeze}	73.3%	100%	0.84
非典型性肺炎 SARS	{发热}{fever}	87.8%	83.3%	0.86
浅表性胃炎 superficial gastritis	{腹胀} {abdominal distension}	80%	88.2%	0.84
病毒性心肌炎 VMC	{心悸} {heart-throb}	73.3%	88.2%	0.80
哮喘 asthma	{喘息}{breath}	67%	100%	0.82
尿毒症 uremia	{恶心} {sicchasia}	73.3%	83.3%	0.78
结膜炎 conjunctivitis	{眼红}{red eyes}	53.3%	83.3%	0.65

The mean of F value: 0.793. Since there is no seed-set, the final filling results descend to some extent, but it does not get strong impact. So we conclude that the appraisal procedure of the candidate attribute values is robust at a certain degree.

6. Conclusions.

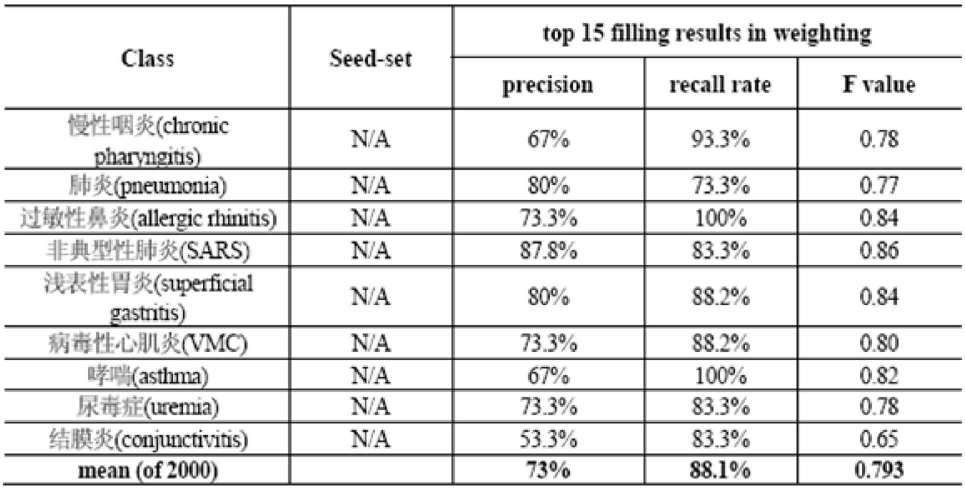
Using the techniques and theories of natural language processing(NLP), this paper puts forward research ideas and methods to automatically build domain ontologies based on text content. And for the clinical medical domain of modern medicine,

Table 4. The extraction results of single class attribute value with no seed-set



andidates of attribute values	weighting	andidates of attribute values	weighting	andidates of attribute values	weighting
头痛 (headache)	0.98	发胀(swell)	0.34	刺激性咳嗽 (irritablecough)	0.25
干燥(dry)	0.94	咳嗽(cough)	0.34	烟熏感 (sootiness sensation)	0.24
咽痛(angina)	0.82	头晕(dizzy)	0.31	疼(pain)	0.24
灼热 (calcination)	0.82	四肢酸痛 (stiffness of four limbs)	0.30	干燥感(dry sensation)	0.23
痒(tickle)	0.58	食欲不振 (inappetence)	0.28	心脏病(heart disease)	0.20
异物感 (foreign body sensation)	0.53	变薄 (attenuation)	0.28	咽痒 (pharyngeal tickle)	0.20
发痒(tickle	0.48	萎缩 (atrophy)	0.28	刺激咳嗽 (irritablecough)	0.19
症状 (symptom)	0.44	声音嘶哑 (hoarseness)	0.27	干咳(hacking cough)	0.18
微痛 (hypodynia)	0.43	消化不良 (dyspepsia)	0.26	低热(low heat)	0.18
灼热感 (burning sensation)	0.35	支气管炎 (bronchitis)	0.26	鼻塞 (rhinostegnosis)	0.17

Table 5. Automatic extraction results of multiple-class attribute value with no guidance



Class	Seed-set	top 15 filling results in weighting		
		precision	recall rate	F value
慢性咽炎(chronic pharyngitis)	N/A	67%	93.3%	0.78
肺炎(pneumonia)	N/A	80%	73.3%	0.77
过敏性鼻炎(allergic rhinitis)	N/A	73.3%	100%	0.84
非典型性肺炎(SARS)	N/A	87.8%	83.3%	0.86
浅表性胃炎(superficial gastritis)	N/A	80%	88.2%	0.84
病毒性心肌炎(VMC)	N/A	73.3%	88.2%	0.80
哮喘(asthma)	N/A	67%	100%	0.82
尿毒症(uremia)	N/A	73.3%	83.3%	0.78
结膜炎(conjunctivitis)	N/A	53.3%	83.3%	0.65
mean (of 2000)		**73%**	**88.1%**	**0.793**

this paper also constructs a concept system with multidimensional model by restructuring and utilizing the generally acknowledged domain knowledge, realizes the automatic construction and acquisition of modern medical knowledge descriptive system. In the "Development of the Chinese Domain Ontology Construction Platform" and "Construction of the Multidimensional Medical Ontology" sections, this paper has addressed the key techniques of automatic construction of domain ontology in detail, such as "knowledge representation techniques and methods" and "construction of domain knowledge concept models". Furthermore, in the ontology evolution, this paper comes up with an automatic method of retrieving attribute values based on Internet, which makes up for the defects of limited information and slow update in the automatic ontology construction based on single text. Our methods achieved good results, and further provide theoretical foundation and technical support to automatic construction of professional domain ontologies.

The Chinese ontology construction platform developed in this paper is adjusted and customized on the basis of the open source software Protégé. Thus, the future work of the platform developing will focus on optimizing current ideas and develop our own ontology construction software platform.

Acknowledgment. This work is partially supported by National Key Project of Scientific and Technical Supporting Programs No. 2011BAH10B04, Pre-research found of Institute of Scientific and Technical Information of China (YY-201125). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Y. Liu, H. Duan and Z. Sui, Non-interactive literature-based knowledge discovery using NLP technique, *Journal of Computational Information Systems*, vol.3, no.3, pp.885-894, 2007.

[2] Y. Liu, Y. Zhou, Z. Sui and Z. Wang, Research on the key technology of based-NLP about Chinese medicine pulse presentations' mathematical quantifying, *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology – Workshops*, pp.307-310, 2008.

[3] Y. Liu, X. Chen, Z. Sui, H. Wang and Y. Zhou, On automatic construction of based-NLP Chinese medicine ontology concept's description architacture, *Proc. of International Conference on Audio, Language and Image Processing*, pp.50-55, 2008.

[4] Y. Liu, X. Chen and Z. Sui, Study on evolution of domain ontology, *The 2nd International Conference on Innovative Computing, Information and Control*, Kumamoto, Japan, pp.139-142, 2007.

[5] Y. Liu, Y. Zhou, Z. Sui and Z. Wang, Research on non-interactive literature-based knowledge discovery, *International Conference on Computer Science and Software Engineering*, pp.747-752, 2008.

[6] Y. Liu and Y. Zhao, Research on ancient literature corpus creation and development of Chinese traditional medicine, *ICIC Express Letters*, vol.3, no.4(B), pp.1227-1232, 2009.

[7] Y. Liu, X. Chen, Y. Zhou and Z. Wang, Development and usage of Chinese medicine supporting system based on post-controlled machinery, *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp.263-266, 2009.

[8] Y. Hu, *The Development of Ontology Construction System and Research on Related Techniques*, Ph.D. Thesis, Peking University, 2008.

[9] Y. Liu, Z. Sui and X. Chen, On method and automatic construction theory of domain ontology based on depended text, *The 1st International Conference on Innovative Computing, Information and Control*, Beijing, China, pp.63-66, 2006.

[10] Y. Liu, Z. Sui, Y. Hu and T. Ji, Automatic construction on of domain ontology, *Journal of Beijing University of Posts and Telecommunications*, vol.29, pp.65-69, 2006.

[11] N. F. Noy, R. W. Fergerson and M. A. Musen, The knowledge model of Protégé-2000: Combining interoperability and flexibility, *Lecture Notes in Computer Science*, pp.69-82, 2000.