基于 HL7 的电子病历关键信息抽取技术研究

徐永东，权光日，王亚东

（哈尔滨工业大学（威海）计算机科学与技术学院，264209 山东 威海，ydxu@insun.hit.edu.cn）

摘　要：结合 HL7（Health Level Seven）标准的数据存储特点对目前电子病历的内容和结构进行了深入分析，提出了医疗信息五元组模式，以及更为细化的二元组和语义类描述，并在此基础上提出了模式泛化、模式获取、医疗信息自动抽取等一系列算法. 通过实际 312 份住院病历数据下的实验表明，系统在查准率与查全率方面，获得了较好的结果，而且由于有自动学习的特性，随着训练语料的增加，系统的整体性能表现将更加优异.

关键词：电子病历；信息抽取；HL7；模式自动抽取

中图分类号：TP391.12　　**文献标志码**：A　　**文章编号**：0367－6234(2011)11－0089－06

Research of electronic medical record key information extraction based on HL7

XU Yong-dong, QUAN Guang-ri, WANG Ya-dong

(School of Computer Science and Technology, Harbin Institute of Technology at WeiHai, 264209 WeiHai, China, ydxu@insun.hit.edu.cn)

Abstract: We analyzed the contents and structure of current electronics medical records, and proposed a definition of Five-Tuples pattern and another more fine-grained definition of two-turples pattern and semantic classes. On this foundation, we proposed a series of algorithms including patterns generalization, patterns automatic extraction and medical information extraction. The experiments with 312 actual medical records show that the system performs well both in the precision and recall. And because of the functionality of self-learning, the system will be more outstanding with an increase in the training corpus.

Key words: electronic medical record; information extraction; HL7; automatic pattern discovery

在医疗信息化的大背景下，随着计算机及互联网技术的发展，以电子病历为载体的数字化医疗信息呈现海量增长的趋势，为快速准确获得所需信息，并将其以规范化的形式存储，使之方便地实现医疗数据的存储、传输、共享以及数据挖掘已成为人们一种迫切的需求.

电子病历（EMR）在美国、日本和英国等[1-2]国家已投入巨资进行了深入的研究和应用. 在国内医学领域[3]，电子病历及相关的医疗信息存储、传输、共享、挖掘等研究仍属刚刚起步阶段. 而目前的电子病历大多以非结构化或者半结构化的文本形式为主，越来越难以满足现代医学研究的需求. 因此，将病历信息进行数字化与可计算化转换是具有重大意义的研究课题.

目前普遍采用基于结构化病历采集模板的方式采集病历信息，获得大规模结构化病历库. 由于人工创建模式的方法不仅耗时费力，而且由于信息抽取模式的分布不均匀，获取数目庞大的低频模式往往变得非常困难. 因此目前常用模式自动生成技术来获取信息抽取模式，包括基于人工语料标注的模式学习方法、基于人工语料分类的模式学习方法、基于种子的模式学习方法等[4-9]. 国内医疗模式获取方面主要进行了对中医药学的局部信息抽取的研究. 文献[10]提出了 Bubble-bootstrapping 方法，对中医药学中的复方名称和疾病名称进行了自动抽取. 文献[11]构建了医学本体来处理医学知识. 本文通过对大量实际病历的分析；根据其描述独特的句法、语法特点以及极强的领域特殊性. 根据人工给出的种子对大量训练

收稿日期：2010－11－29.

基金项目：国家自然科学基金资助项目(60803092)；山东省优秀中青年科学家奖励基金资助项目(2010BSA10014)；山东省科技攻关资助项目(2009GG10002053).

作者简介：徐永东(1974—)，男，博士，副教授；
王亚东(1964—)，男，教授，博士生导师.

病历进行学习,最终得到每类诊察、诊断信息的抽取模式,然后通过匹配信息抽取模式的方法完成对病历诊察信息的抽取.

1 医疗信息模式自动获取

病历信息的抽取通常基于模式匹配的方法,即模式以空槽的形式给出应从文本中获取的各项内容. 因此模式的自动获取是关键信息抽取算法的核心.

1.1 基于 HL7 的病历信息描述特征分析

本文采用满足 HL7 标准的数据库作为数据存储方案. 因此必须首先结合 HL7 参考信息模型与病历自身的特点来分析,进而达到对病历信息有效覆盖的同时方便抽取的数据进行结构化存储.

HL7(Health Level Seven)是由美国国家标准局(ANSI)授权的标准开发机构 Health Level Seven INC.(HL7 组织)研究开发的一个专门用于医疗卫生机构及医用仪器、设备数据信息传输、存储以及处理的标准. 目前已成为全球各国医疗卫生信息化的重要基础和关键技术. 在 HL7 标准中特别定义了用于数据的存储与交换的信息表示模型——参考信息模型(RIM),它是结构化的信息规范. 这个模型包括属性(Attributes)、关系(Relationships)、约束(Constraints)和状态(States). 它简明、完整的定义一套结构和词汇,用于实现广泛的临床场景的信息表达需求.

本文利用自动获取的模式来抽取的信息是诊察信息,该类信息在 HL7 设置的数据模型中,属于观察类表的信息. 在该模型中很多类信息都属于观察类的信息,包括诊察信息、诊断信息、实验结果、过敏反应以及生理现象. 如表 1 所示的是一个观察类的实例——病述表.

表 1 病述表

属性名	属性描述
ID	表的主码
发作时间	疾病发作的时间
症状号	疾病的编号
位置号	疾病位置的编号
持续时间	疾病持续的时间
症状值	疾病的严重程度代码

在病述表中,除了用来识别表内项的主码“ID”外,用来描述诊察症状的项分为:症状开始时间、症状部位、症状描述、程度/频度和持续时间 5 个部分.

为了能用机器学习的方法自动完成病历信息的抽取,还需要对病历本身的结构和特征有充足的认识. 表 2 所示的是一份有代表性的烫伤病历的诊察信息.

表 2 病历诊察信息

信息类型	信息内容
诊察	发热,左手背、左腕、左前臂见烫伤创面,左手臂局部轻度红肿,手背散布大小不等水疱,部分腐皮脱落,基底红白相间,渗出少. 创面污染不重,总面积约 2%.

诊察信息除了在表达方式上具有显著的医学领域特征外,与 HL7 的参考信息模型相比有以下特点:

1)与参考信息模型的数据相比,普通诊察信息没有症状开始时间与持续时间 2 个部分. 上述信息在病历中以病史及入院经过信息出现. 病历中出现的诊察信息全部以静态信息出现,时间默认为病历生成时间.

2)普通诊察信息的表述具有自然语言的特点,信息的表述与参考信息模型的代码表述相比较复杂.

因此,模式的表达方式经过消减与细分后可分为:对象的修饰、对象、程度/频度、性质和描述 5 个部分. 根据上述信息特点本文设计了病历信息的五元组表示形式:< 对象修饰,对象,程度,性质,对象描述 >. 其中对象与描述部位必须出现. 而病历信息抽取模式就是五元组的子集.

另外,由于语言使用的灵活性,五元组中的主体修饰与程度部分的位置也具有一定的灵活性. 因此模式格式的定义中各元素的相对位置并不是唯一的.

根据对大量的病历中诊察信息的特征结合 HL7 参考信息模型的分析,发现诊察信息基本上是由主体和描述 2 部分组成,这部分诊察信息本文称为第 1 类诊察信息(文中没有明确说明是哪一类的都是指第 1 类诊察信息),通过诊察抽取模式对它进行获取,如“头痛、面色无华、口干、舌质暗红、心烦”等.

其中信息的主体是由对象和描述 2 部分组成,如“头痛、面色无华、口干”等,本文可以采用二元组来表示:< 对象,描述 >.

例如,对于语句“右臂前部剧烈肿痛”,经过词法分析处理后为“右 /*f*, 臂 /*n*, 前部 /*f*, 剧烈 /*a*, 肿 /*v*, 痛 /*a*”,采用 < 对象, 描述 > 二元关系模式可以得到特例模式:< 臂 /*n*,痛 /*a* >. 采用五元组对它进行匹配可得到一个六元组的特例模式:< 右 /*f*, 臂 /*n*,前部 /*f*,剧烈 /*a*,肿 /*v*,痛 /*a* >.

五元组中除二元关系外其他 3 个部分（即主体修饰词、程度和性质）称为模式抽取语义类. 语义类虽然不是诊察信息必须的组成部分，但它们对抽取模式的表示具有非常重要的作用.

1）解决同义词表述问题，简化二元关系模板.

2）去掉修饰语中的否定性词汇，使模板得到统一和简化，解决语句歧义问题.

3）可以更加精确的对信息进行量化，例如含有修饰词“无”的症状可以量化成 0；含有修饰词“轻微的可以量化成 0.5；含有修饰词“极”、“非常”可以量化成 2 等等；方便后续数据的转换、量化分析、数据挖掘.

4）去掉导致模板抽取的错误的无效信息.

表 3 是对语义类的一些应用举例. 本文利用《同义词词林》并针对它们在医学领域的语义特点，设计了病历信息语义类同义词林，并用一个对象来替换模板中出现的多个同类词. 这个对象可能为词组，如“部位”，也可能为各类复合词组如表示数值的“程度 1.5”或者直接删除.

表 3　语义类举例

语义类	同类词举例	替换对象
对象修饰	左、右、上…	<部位>
程度	稍微、严重…	数值<程度 1.5>
性质	肿、胀、浮肿…	代码<性质 001>
否定性词组	无、没有、未…	数值<0>
无效信息	患者、感到…	空

1.2　模式泛化

二元关系特例模式只能概括生成它的文本，要想使其能概括别的文本片段，则需要对特例模式泛化，生成泛化模式. 一般对特例模式进行泛化有 2 种方法：基于语法的泛化和基于语义的泛化. 基于语法的泛化是把特例模式中具有相同语法的对象进行抽象，得到泛化模式；基于语义的泛化是指把多个特例模式中具有相同语义类型的单元进行抽象，得到泛化模式. 本文采用语法与语义相结合的泛化方法，进行二元关系泛化和语义类泛化 2 个阶段的处理.

1.2.1　模式的二元关系泛化算法

二元关系泛化的主要目的为：去除否定性词汇；生成可发现新二元关系的模式. 二元关系泛化算法如图 1 所示.

例如文本片段“右膝无肿痛” 与二元关系“膝痛”进行匹配，找到一个对象类命名实体和一个描述类命名实体. 可以生成一个特例模式为：<｛右 /*adj*｝，｛膝 /*n*｝，｛无 /*v*｝，｛肿 /*v*｝，｛痛 /*v*｝>. 该特例模式经过二元关系泛化后可得到二元关系泛化模式：<｛右 /*adj*｝，对象，｛肿 /*v*｝，描述｝>.

输入：诊察信息文本片段 *S*

输出：二元关系泛化模式 *SP*

1. 对文本片段进行预处理，获得一个包含某个二元关系的词集合，从该集合中查找含有否定语义类的词，并删除.

2. 从二元关系库中找出该集合含有的二元关系. 若有，转入 3，否则转入 4.

3. 确定二元关系在该集合中的对象和描述的位置. 并将相应位置进行泛化，生成二元关系泛化模式，存入模式库.

4. 二元关系模式泛化算法结束.

图 1　二元关系泛化算法描述图

1.2.2　模式的语义类泛化算法

由于本文处理的信息具有医学领域的独特特性，特别是句法结构比较简单. 易于建立简单的语义理解规则. 因此提出了一个基于规则与语义类表的语义类泛化算法. 该算法的主要思想为：利用二元关系泛化模式的对象与描述部位的槽，将模式分割为 3 部分，分别为：对象前部、对象与描述中间部位和描述后部. 将对象前部的全部词作为对象描述信息，与对象修饰语义类进行比较替换. 中间部位的信息的词集合一定属于对象修饰类、程度类与性质类构成的总集合的子集，且具有顺序性，因此这部分信息按照该顺序规则进行匹配替换. 描述后部的词作为程度类信息进行匹配替换. 替换后的泛化模式存入模式库.

语义类泛化算法具体描述如图 2 所示.

输入：二元关系泛化模式集合 *U*

输出：语义类泛化模式 *P*

1. 从 *U* 中找出一个二元关系泛化模式 *SP*. 若找到，转到 2，否则转到 7.

2. 将模式 *SP* 依照二元关系的 2 个位置进行分割. 获得 3 段词组集合，分别为 *SU*1、*SU*2、*SU*3.

3. 对 *SU*1 进行处理，逐个将 *SU*1 中的词汇与对象修饰语义类进行匹配，若找到，则将其替换为相应词汇.

4. 利用对象修饰语义类、程度语义类、性质语义类对 *SU*2 进行拆分，并满足对象修饰语义类——程度语义类——性质语义类的顺序进行排列. 若出现顺序错误，将出现错误的位置处还原.

5. 逐个将 *SU*3 中的词汇与程度语义类进行匹配，若找到，则将其替换为相应词汇.

6. 删除原有模式，将替换后的词组按照 *SU*1-*SU*3 的顺序重新组合为模式. 存入模式库中，转 1.

7. 算法结束.

图 2　语义类泛化算法描述图

例如将二元关系泛化模式 < {右/adj},对象,{关节/n},{肿/v},描述,{严重/adj} > 按照对象实体和描述实体进行分割,形成3个片段 < {右/adj} >、< {关节/n}{肿/v} >、< {严重/adj} >. 分别对其进行语义类泛化并重新组合,生成语义类泛化模式{ < 部位 >,对象, < 部位 >, < 性质001 >,描述, < 程度2 >}.

1.3 抽取模式自动获取算法设计

在抽取模式和模式泛化的定义的基础上,针对诊察信息的特点,本文设计了一个基于 Bootstrapping 算法的模式自动获取算法,用于获取病历信息的二元关系,从而得到该类信息的抽取模式. Bootstrapping 算法是一种从自由文本中进行信息抽取实现结构化数据存储的新的信息抽取模式获取方法. 它不需要预先标注的手工训练集,只需要以少数数据(种子)和大量的未标注语料为基础,通过由种子词产生模式,再由模式产生种子词的不断循环迭代,最终产生所需的模式库. 其整体流程如图3所示.

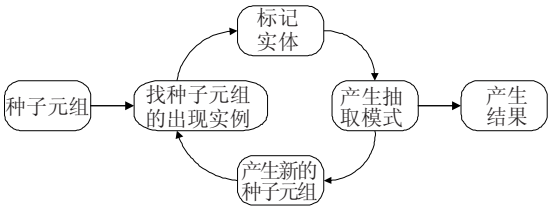

图3　Bootstrapping 算法流程图

基于 Bootstrapping 的模式获取方法由3个部分组成:

l) 初始条件. 给定一个医院病历文档集合 $D = \{d_1, d_2, \cdots, d_n\}$、医学术语库、语义类库、词法分析模块和人工给初的少量种子二元关系. 定义二元关系类别 $SR = < n, v >$,其中:n 为文本串中名词性短语 NP 的中心词;v 为文本串中修饰 n 的动词或形容词短语的中心词.

2) Bootstrapping 迭代. 目标为从 D 中学习 SR 的二元关系模式集合 P 和属于 SR 的二元关系集合 R. 首先利用种子关系检索病历文本集合,产生可信度高的新二元关系实例;然后利用新模式去匹配语料集,抽取新的种子关系. 由此循环迭代,至达到规定的迭代次数或种子关系不再增加为止.

3) 输出结果. 目标词和用于抽取目标词的模式.

算法的具体描述如图4所示.

输入:病历文档集合 D、种子二元关系集合 S
输出:二元关系模式集合 P、全部二元关系集合 SP
1. 对病历信息进行预处理操作,包括分类,内容的分词、词性标注、单句切分等操作.
2. 读出二元关系库中未处理的二元关系.
3. 根据种子词学习特例模式集合. 主要包括:
① 根据给出的种子二元关系,遍历整个训练集合的每个单句,找出所有包含种子二元关系的单句.
② 以 n 和 v 为中心,将每个单句分割为3部分.
③ 根据抽取模式的表示方式,将每个单句转化成一个特例模式,形成种子二元关系的特例模式集合.
4. 利用泛化算法生成二元关系的泛化模式集合.
5. 使用同义词林对得到的二元关系进行重复性验证,非重复的种子保存到二元关系库中去,否则抛弃.
6. 二元关系库中有新的种子,则转到2,否则转到7.
7. 使用语义类模式泛化算法,对模式库中的二元关系泛化模式进行泛化. 获得最终的泛化模式,并存入模式库.
8. 算法结束.

图4　Bootstrapping 算法描述图

2　基于自动获取模式的信息抽取算法

利用已有的模式库对病历中的诊察信息进行抽取. 问题在于:模式库中的诊察信息抽取模式经过模式的2阶段泛化,已经成为抽象的字符串组合. 这种组合已经脱离了自然语言的范畴,无法直接用来与诊察信息中的语句进行匹配,并抽取相应信息. 因此,需要将抽象化的模式与诊察信息语句建立对应关系.

算法思想为:二元关系库中的二元关系既是用来进行模式生成的种子,也是模式抽取的对象. 因此,若一条语句能够被诊察信息抽取模式所匹配,则在二元关系库中至少有一个二元关系能与该语句进行匹配. 找出与所要抽取的语句能够匹配的所有二元关系,从这些二元关系中依次取出二元关系并对待抽取语句调用一次模式泛化算法,获得的模式与模式库中的模式进行匹配. 算法的流程如图5所示.

输入:待抽取语句 S
输出:由 S 生成的特例模式 SP
1. 从二元关系库中找出所有可以与 S 进行匹配的二元关系,并放入二元关系集合 U.
2. 若 U 不为空,则中取出一个二元关系,并使用该二元关系,从 S 中获取特例模式 SPT. 否则转到5.
3. 对 SPT 调用模式泛化算法,获得泛化模式 SSP.
4. 将 SSP 与模式库中的模式进行匹配,若成功,则转到5,否则从 U 中删除该二元关系并转到2.
5. 返回特例模式 SPT,或者失败标志,算法结束.

图5　自动获取模式的信息抽取算法流程图

3　结果与分析

衡量信息抽取系统性能主要根据 3 个常用的评价指标：召回率 R、准确率 P 和 F 度量值. 对于需要抽取的 n 个状态，则评价指标为

$$R = ce/(ce + te), \quad (1)$$

$$P = ce/(ce + fe), \quad (2)$$

$$F = (\beta^2 + 1)P \times R/(\beta^2 P + R). \quad (3)$$

式中：ce 为对于第 i 个状态抽取出的正确信息个数；te 为没有抽取出的信息个数；fe 为抽取出的错误信息个数.

实验数据采用由长春市市医院烧伤科提供的 312 份 2005 年～2008 年间烧伤科住院病历. 本文从中取出 30 份病历作为抽取测试数据，其余病历作为训练语料.

系统主要包括模式库建立、信息抽取 2 个处理部分，总体流程如图 6 所示. 首先进行语料预处理，根据各个部位相应的关键词提示，如“病人特点”、“查体”、“入院经过”等，将相关信息分类. 然后对含有并列词语句拆分成单句，并对单句进行分词与词性标注，本文采用的资源是《新编全医药学大辞典》公布的医学术语库. 最后进行模式自动抽取和结构化转换处理.

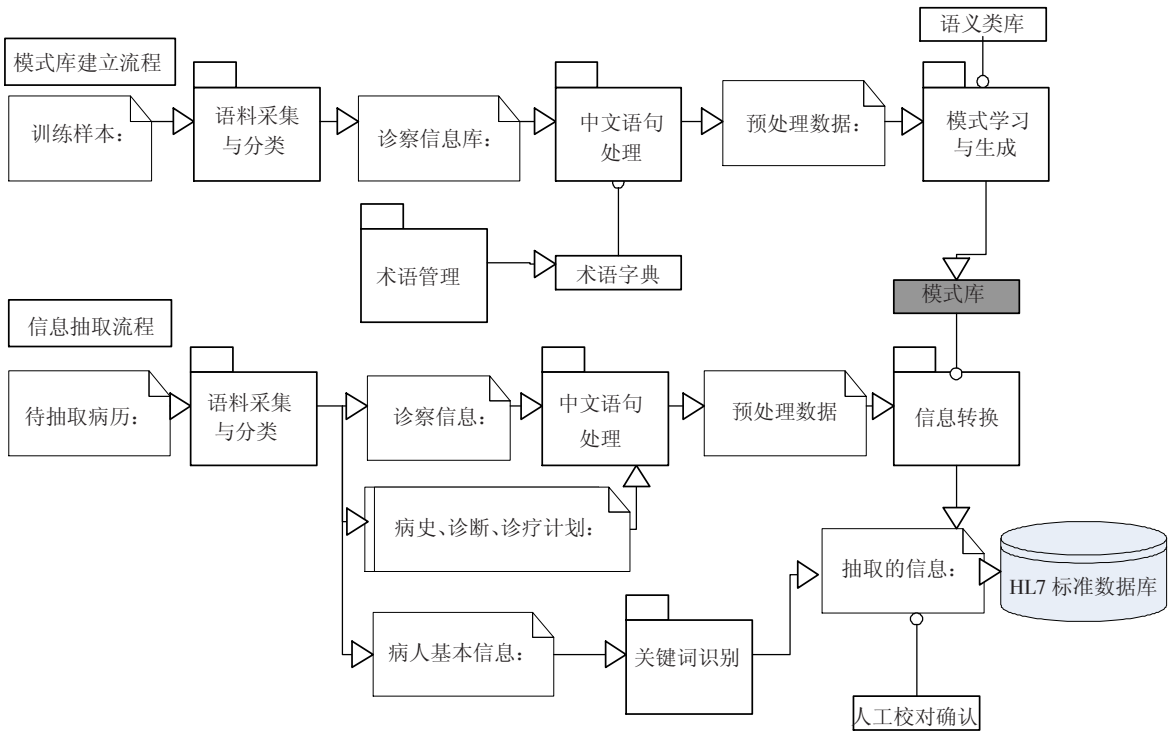

图 6　系统总体流程图

本文选取的对象词包括：皮肤、伤口、肩、膝、腺、腹、头、创面；选取描述词：肿胀、痛、色、活动和伤口等. 组合后生成 29 个种子二元关系. 根据这 29 个种子二元关系匹配病历库，共得到 865 个特例模式. 经过泛化归并后得到 110 个泛化模式. 然后在这些泛化模式的指导下，对 30 份病历进行了测试. 实验数据如表 4～表 6 所示.

表 4　种子二元关系举例

主体	皮肤	皮肤	皮肤	伤口	伤口	肩	肩
描述	肿胀	痛	色	痛	肿胀	肿胀	痛

从上述实验数据可以看出，二元关系达到了较高的准确率与召回率，主要是因为病历信息相对来说句子结构简单，没有长句子，且语法简单，而且对二元关系的抽取没有经过语义类泛化，模式比较准确. 对性质语义类的抽取得到了较高的召回率. 这主要是因为性质的种类较少，易于抽取. 但准确率较低，这是由于部分的描述词前面直接出现程度类词，而没有出现性质类词，导致识别错误.

表 5　泛化模式举例

所属二元关系	获取的模式
<皮肤，肿胀>	{<部位>，BODY，<部位>，<性质 001>，DESCRIBE，<程度 2>}
	{<局部>，BODY，DESCRIBE}
<肩，黑>	{BODY，<皮肤>，DESCRIBE}
<手，活动>	{BODY，DESCRIBE，<受限>}
<背，脱落>	{BODY，<部分>，<腐皮>，DESCRIBE}

表6 模式抽取实验结果

实验结果	二元关系	主体修饰类	程度类	性质类
抽取的正确项	151	101	67	47
抽取的全部项	165	104	72	77
实际正确项	187	132	93	47
召回率/%	80.75	76.50	72.04	100
准确率/%	91.52	97.10	93.05	61.03
F值/%	85.80	85.58	81.21	75.80

对于主体修饰和程度语义类来说,准确率很高,但召回率很低,主要是因为这2类词汇出现的位置变化比较大,而且出现的种类也比较多,特别是语义类的引入易导致出现识别错误.

4 结 论

1)通过对电子病历文本信息的描述特征的分析,提出了符合HL7标准的五元组抽取模式,并进一步细分为二元关系和语义类2部分,为进一步的模式自动获取提供了抽取对象和标准的形式化描述.

2)利用基于Bootstrapping算法的模式自动学习技术,实现了诊察信息抽取模式的自动学习,并提出了基于模式的电子病历关键信息抽取算法.

3)通过实际住院病历作为测试数据的实验表明本文提出的方法可以有效地抽取二元关系模式,并对各种语义类也有较好的识别.

参考文献:

[1] FLETCHER M. President promotes switching to electronic medical records: Bush says paperless system would cut costs and improve care[N]. Washington Post, 2005-01-28(A07).

[2] HALAMKA J. Electronic health record and revolution of clinical information technology—current status and future impact[J]. Zhonghua Yixue Za Zhi, 2005, 85(22): 1513-1515.

[3] 金伟. 电子病历系统的存储与交换的研究[D]. 上海: 上海交通大学, 2006.

[4] AGICHTEIN E, GRAVANO L. Snowball: Extracting relations from large plain-text collections[C]//Proceedings of the 5th ACM International Conference on Digital Libraries. New York, NY: ACM, 2000: 213-219.

[5] BRIN S. Extracting patterns and relations from the World Wide Web[C]//Proceedings WebDB 98 Selected papers from the International Workshop on The World Wide Web and Databases. London, UK: Springer-Verlag, 1999: 172-183.

[6] AGICHTEIN Y. Extracting relations from large text collections[C]//Doctoral Dissertation Extracting Relations from Large Text Collections. New York, NY: Columbia University, 2005: 157.

[7] MUSLEA I. Extraction patterns for information extraction task: A survey[C]//Procedings of AAAI-99 Workshop on Machine Learning for Information Extraction. Orlando, Florida: AAAI Press, 1999: 1-6.

[8] RILOFF E. Automatically generating extraction patterns from untagged text[C]//Proceedings of the Thirteenth National Conference on Artificial Intelligence. Portland, Oregon: AAAI Press, 1996: 1044-1049.

[9] HUFFMAN S B. Learning information extraction patterns from examples[C]//Proceedings of Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing. London, UK: Springer-Verlag, 1996: 246-260.

[10] 周肖彬,曹存根. 基于本体的医学知识获取[J]. 计算机科学. 2003, 30(10): 35-39.

[11] 郭玉峰,刘保延, 周雪忠. 面向中医临床科研需求的术语分类框架研究[J]. 环球中医药, 2008 (2): 9-12.

(编辑 张 红)