第 35 卷 第 6 期 计 算 机 工 程 2009 年 3 月
Vol.35 No.6 Computer Engineering March 2009
·开发研究与设计技术· 文章编号：1000—3428(2009)06—0251—03 文献标识码：A 中图分类号：TP311.52

基于本体的医疗信息搜索技术

赵修文，刘伍颖，王 挺
(国防科技大学计算机学院，长沙 410073)

摘 要：针对医疗信息联合搜索中存在的问题，提出一种基于医疗领域本体的多信息融合搜索方法。该方法采用信息抽取技术自动构建本体实例，运用医疗领域本体对用户查询请求进行语义处理，同时实现了基于该方法的原型系统。实验结果表明，该原型系统能有效返回多种相关信息，从而说明了本体在多信息融合搜索方面的重要性。

关键词：医疗本体；信息抽取；匹配；搜索引擎

Medical Information Search Technology Based on Ontology

ZHAO Xiu-wen, LIU Wu-ying, WANG Ting
(School of Computer, National University of Defense Technology, Changsha 410073)

【Abstract】Aiming at the problems existed in the process of medical information combination search, a search method for multi-information fusion based on medical ontology is proposed, which sets up ontology entity by using information extraction technique, and deals with the users's query by using medical ontology. A prototype system based on this method is also implemented. Experimental results show this system can return multiple relevant information effectively, and show the importance of ontology in the field of multi-information fusion search.

【Key words】medical ontology; information extraction; matching; search engine

1 概述

随着互联网的普及和网上医疗信息的不断丰富，越来越多的普通用户和专业人员倾向于使用互联网查询、获取各种医疗信息。目前，用户获取互联网上医疗信息主要有 2 种手段：(1)通用搜索引擎，如百度、Google 等；(2)使用医学专业搜索引擎，如 Medical Matrix, Medscape 等。使用通用搜索引擎搜索医疗信息，因其没有对医疗信息进行专门处理，不能胜任专业化的医疗信息检索，因此，出现了大量医学专业搜索引擎。而现有的医学专业搜索引擎大多采用传统信息检索方法：(1)基于关键词的检索；(2)基于分类目录的检索。通用搜索引擎的优点是具有很高的召回率，缺点是检索结果过于庞大，准确率不高，无法处理语义关系，用户难以快速准确地找到自己所需要的信息；医学专业搜索引擎虽然对医学知识进行了专门处理，提高了查准率，但要用户具备一定医学分类专业知识，且目前国内尚没有专业的中文医疗搜索引擎。

针对这些现状，本文在 Nutch 开源搜索引擎的基础上，提出一种基于本体的医疗搜索引擎系统原型。该系统使用信息抽取技术，有效地将疾病、药品、医生和医院信息通过本体进行融合，使医护人员和患者能快速准确地获取当前的医疗信息。该系统使用本体对用户的查询请求进行语义分析，能够提高搜索医疗信息的准确性。

2 本体在综合信息搜索中的作用

随着 Web 信息的急剧膨胀，如何有效地进行知识表示和信息组织，以帮助用户快速准确地获取信息成为亟待解决的难题。为解决这些问题，人们提出将本体引入信息处理。本体能够更好地进行知识表示，且具有语义处理能力。目前对本体公认的定义是：本体是对共享概念模型的形式化的明确的描述。下面将详细阐述构建医疗本体的原因及其作用。

2.1 构建本体的意义

当前的主流搜索引擎都是对文档进行词法分析，然后对整篇文档的所有词语构建全文索引。由于对信息的处理粒度过大，使得索引词(或索引词的组合)无法表达文档的真实意义，丢失了文档的语义信息；另外，中文文字之间没有分隔符，当前的分词技术只能在一定程度上解决歧义问题和新词的识别，而医疗领域新词问题尤为突出，使得通过索引词表达的文档含义更加不准确。

搜索引擎在执行用户查询的时候，只是简单地提取用户查询请求中的关键词，同样丢失了用户查询的语义信息。这些因素直接导致检索结果的不准确，因此，在医疗搜索引擎中引入本体技术，使用信息抽取技术对文档中的信息进行更小粒度的分析，保留文档中原有的语义信息。本体提供了计算机可理解的语义信息，使得计算机和人对知识的理解达到一致，且语义信息使计算机具有一定推理能力，使计算机能够自动挖掘潜在的知识；通过本体对用户的查询请求进行语义分析，使计算机能够理解用户查询请求的实际意义，返回用户真正需要的结果。

2.2 医疗本体的作用

在医疗领域中构建医疗本体是为了有效地融合和关联疾病信息、药品信息、医生信息和医院信息，旨在帮助用户获取准确的医疗信息。医疗本体的作用主要有以下 6 个方面：

基金项目：国家自然科学基金资助项目(60403050)；新世纪优秀人才基金资助项目(NCET-06-0926)；国家“863”计划基金资助项目(2006AA02A312)

作者简介：赵修文(1978－)，男，硕士研究生，主研方向：信息检索；刘伍颖，博士研究生；王 挺，教授、博士生导师

收稿日期：2008-09-10 **E-mail：**xwzhao@nudt.edu.cn

(1)对医疗领域的实体以及实体之间的关系进行严格的语义描述，利于该领域的数据集成；

(2)利于分析该领域的知识；

(3)澄清医疗领域知识；

(4)能更好地进行知识表示；

(5)指导知识获取，并且使获取的知识更加准确；

(6)可实现中文分词词典的动态更新。

在利用医疗本体进行检索的时候，首先提取用户查询语句中的概念(或实体)，通过医疗本体进行语义分析，返回本体中用户所需要的知识；其次通过对用户查询的语义分析进行关键词扩展，挖掘潜在概念，实现更深层次的检索。

3 医疗搜索引擎设计

系统实现对网络上的各种医疗信息，包括药品信息、医生信息、疾病信息和医院信息的搜索。用户可以设定搜索引擎的抓取列表，启动新的抓取任务，可以修改索引和合并数据。普通用户通过浏览器搜索网络上的各种医疗信息。

系统采用面向对象方法，基于 B/S 体系结构进行设计。在不影响理解和沟通的前提下，尽量减少模型图的绘制和文档的创建。本文使用 IBM Rational Rose 工具创建类图和部分模块的顺序图，设计过程中使用的设计模式主要有简单工厂模式和单例模式[1]。

3.1 总体架构

医疗搜索引擎是在 Nutch 基础上进行的二次开发，主要功能通过 Nutch 的插件(plug-in)扩展方式实现，系统体系结构如图 1 所示。

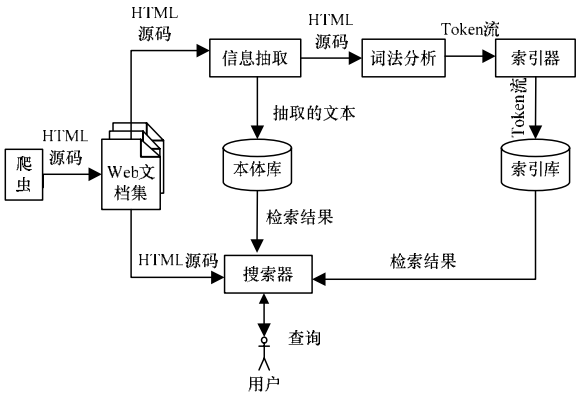

图 1 医疗搜索引擎体系结构

从图 1 可以看出，先通过 Nutch 爬虫抓取网页，然后通过信息抽取构建本体实例，最后通过索引器构建倒排索引。此外，Nutch 的爬虫和搜索器通过配置可以运行在独立的硬件平台上，具有更强的适应性。

爬虫首先生成抓取列表，然后根据此列表下载 Web 页面内容并存储，通过文档中的超链更新抓取列表；索引器对 Web 文档集进行分析，构建倒排索引，将索引存入索引库中；搜索器分析用户查询，并进行查询匹配，向用户返回搜索结果；信息抽取将网页中我们需要的医疗信息进行抽取，转化为结构化数据，并将抽取结果存入本体库中；词法分析在索引和查询时对文本进行中文分词或词干分析。

3.2 医疗本体

在描述医疗本体时使用 OWL[2] 语言，同时采用 Protege3.1 建立医疗本体模型，其类结构如图 2 所示。其中，方框表示类；弧线表示对象属性(Object Property)；数据类型属性(Datatype Property)在图中未标记。

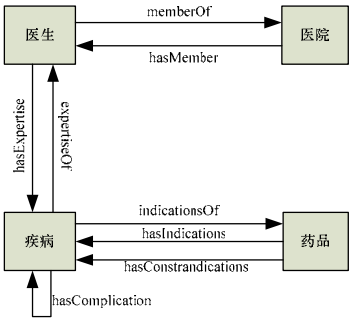

图 2 医疗本体模型

图 3 为医疗本体中部分疾病类和实例树，其中，“恶性疟疾”是实例，其余每个节点均是类。类之间的关系定义为 ISA 关系。

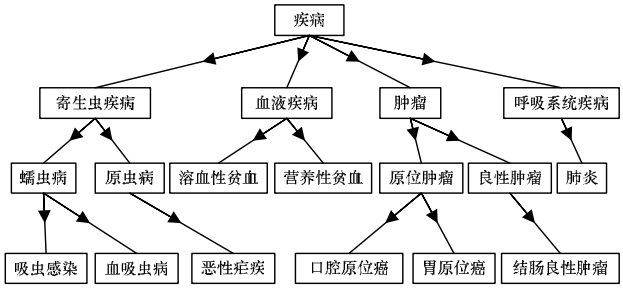

图 3 疾病类和实例树

医疗本体的详细定义如表 1 所示，表中分别列出了类的对象属性和数据类型属性的属性名和类型，其含义为字面值。

表 1 医疗本体类定义

类名	对象属性	数据类型属性
疾病	并发症(hasComplication)::类型 疾病实例	英文名称::类型 字符串
	治疗药物(indicationsOf) ::类型 药品实例	病因::类型 字符串
	医生(expertiseOf) ::类型 医生实例	症状::类型 字符串
药品	适应症(hasIndications) ::类型 疾病实例	英文名称::类型 字符串
	禁忌症(hasContraindications)::类型 疾病实例	别名::类型 字符串
		曾用名::类型 字符串
医生	所属医院(memberOf) ::类型 医院实例	医生简介::类型 字符串
	特长(hasExpertise) ::类型 疾病实例	
医院	医生(hasMember)::类型 医生实例	

表 1 中的对象属性 hasXXX 和 XXXOf 互为逆属性。如 hasMember 和 memberOf 对象属性。

4 搜索关键技术研究

医疗本体构建是项非常复杂和繁琐的工作。为了能自动构建医疗本体，笔者根据上述内容对 Web 信息抽取技术以及采用医疗本体后的用户查询方法进行重点研究。

4.1 信息抽取

Web 所表示的信息缺乏语义，机器无法理解。当前很多信息抽取系统能够识别命名实体而不能识别实体之间的关系，这导致信息的意义非常小。本文使用基于正则表达式的方法进行信息抽取，将获得的知识封装为本体，同时完成实体间关系的自动创建。

对于要进行信息抽取的 Web 站点都有较好的结构，每个站点的同类信息都采用统一的模板生成，这使得使用正则表达式进行信息抽取就可以达到很好的效果。目前共创建了 20 组规则集分别对 20 个站点进行信息抽取。每组规则由 URL、医疗本体某一个类的所有属性和类别组成，其中，URL 为所要抽取站点的 URL，属性为医疗本体中某一类具体属性所对应的正则表达式，类别用于标识医疗本体中的类名。规

则示例如下(中国疾病知识总库疾病信息的抽取规则)：

URL=http://cdd1.juhe.com.cn/cdd/Disease/

disease.name =(.*?)((.*?))

disease.pathogeny=<DIV>病因：(.*?)</DIV>

disease.symptom=<DIV>临床表现：(.*?)</DIV>

disease.complication=<DIV>并发症：(.*?)</DIV>

disease.treat=<DIV>治疗：(.*?)</DIV>

type=disease

每条规则用于提取 HTML 文档中与之匹配的子串。正则表达式中每个(*?)称为组，每个组有一个编号。如 disease.symptom 规则匹配所有起始于“ <DIV>临床表现：”，结束于“ </DIV> ”的子串，在程序中通过 group(1)即可取得相应的子串。

匹配抽取算法如下：

取得当前网页的 URL，并加载该 URL 的抽取规则

对网页进行预处理 //去除 W3C 规定的空字符，如 0x09

for(每一条规则){

if(匹配结果不为空){

去除抽取结果中的 html 标记

if(属性为数据类型属性)将抽取结果加入本体中相应的类

if(属性为对象属性)将实例加入本体，并设置实例之间的关系(包括逆属性)

}

}

图 4 是中国疾病知识总库中对“ 更年期综合征 ”进行抽取，并存入本体的结果。

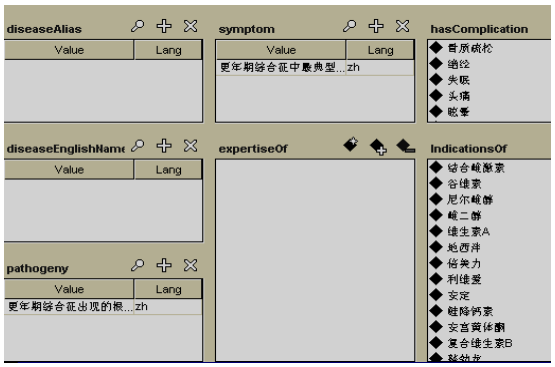

图 4　信息抽取结果

4.2　查询分析

查询过程主要包括用户查询请求的规范化处理和查询匹配。用户查询规范化处理是指将用户的查询请求进行词法分析，转化为系统可理解的形式，应用于本体的查询。检索由 2 个部分构成：(1)基于本体的关键词检索；(2)本体实例的检索，其中，本体实例的检索使用 ARQ[3]查询语句。关键词检索流程如下：(1)对用户的查询请求进行中文分词；(2)将用户的查询请求表示为二元组的形式；(3)使用 MeSH(医学主题词表)进行查询扩展；(4)使用本体进行查询扩展，同时返回本体实例检索结果；(5)返回关键词检索结果。

这里所描述的二元组形式为：(实体，属性)。通过本体查询分别确定实体和属性。笔者对医疗本体中每一类的每个属性定义了一系列的属性关键词，如药品类的适应症属性的关键词集合为{功能，适应症，作用，功用，主治功能，…}。检索流程中第(2)步的算法如下：

由第(1)步产生的关键词集合 A

for(A){

检索本体取得关键词(实体)的类名

加载该类属性的关键词集合 B

若 A 中有一项包含在 B 中，取得该类对应的标准属性(本体中类的属性)

返回二元组(实体，属性)

}

例如，查询为“ 治疗感冒的药物 ”，则该查询的二元组表示形式为(感冒，药物)。在本体进行关键词扩展的时候，使用实体的属性值进行关键词扩展，使查询具有一定的语义关系，提高关键词检索的准确率。如不使用本体进行语义扩展，则上述查询的关键词为“ 治疗 感冒 药物 ”，若使用本体进行查询扩展，则查询关键词被扩展为“ 感冒通 白加黑 感康 感冒 药物 治疗 ”(假设本体查询结果为“ 感冒通 白加黑 感康 ”)。

5　系统实现

在医疗搜索引擎系统中使用 Eclipse3.2+Jena2.4+Nutch 0.8.1+Lucene1.9.0 进行开发，采用 Nutch 插件扩展方式进行系统集成。Nutch 的插件系统是基于 Eclipse 中对插件的使用。信息抽取模块通过 Nutch 的 org.apache.nutch.parse.HtmlParseFilter 扩展点进行扩展；词法分析模块通过 Nutch 的 org.apache.nutch.analysis. NutchAnalyzer 扩展点进行扩展；查询模块使用 Lucene 进行重写。查询接口使用 JSP 和 Servlet 编程实现，用户在查询页面中输入关键词，系统完成查询匹配并返回索引检索和本体实例检索结果。

图 5 展示了用户检索“ 感冒 ”的示例，左侧为基于本体的关键词检索结果；右侧为本体实例的检索结果，给出了疾病的详细信息和治疗药物(无治疗专家，由于感冒属于常见病)，点击超链可获得实例的所有属性；底部相关搜索条目是系统对用户的查询请求进行分析后，在 MeSH 中查找相应的概念，并通过此概念的子节点、父节点和兄弟节点进行扩展的结果。

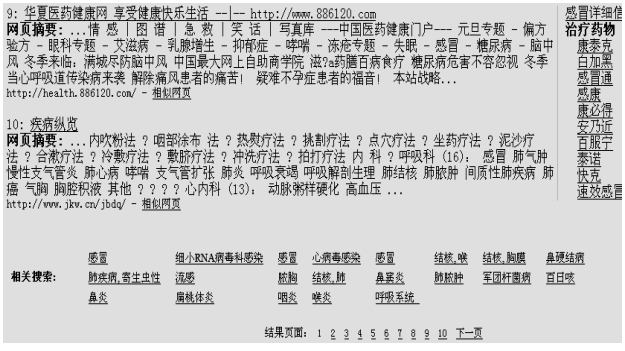

图 5　检索示例

6　结束语

本文分析了通用搜索引擎和医学专业搜索引擎在医疗领域搜索中存在的问题，将本体引入医疗信息搜索，并描述系统框架和实现的关键技术，阐述了本体以及信息抽取模块和检索模块的设计。该原形系统的实现为医学专业搜索引擎的设计和开发提供了良好的理论依据。下一步工作将引入机器学习算法进行信息抽取，并对本体进行语义相似度度量。

(下转第 256 页)