ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.2, February 2008, pp.237−245
DOI: 10.3724/SP.J.1001.2008.00237

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于本体的Deep Web数据标注*

袁 柳+, 李战怀, 陈世亮

(西北工业大学 计算机学院,陕西 西安 710072)

Ontology-Based Annotation for Deep Web Data

YUAN Liu+, LI Zhan-Huai, CHEN Shi-Liang

(School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China)

+ Corresponding author: Phn: +86-29-88495821 ext 112, E-mail: yuanl@mail.nwpu.edu.cn, http://www.nwpu.edu.cn

Yuan L, Li ZH, Chen SL. Ontology-Based annotation for deep Web data. *Journal of Software*, 2008,19(2): 237−245. http://www.jos.org.cn/1000-9825/19/237.htm

Abstract: A semantic annotation method for Web database query result is proposed in this paper by adopting the deep annotation procedure in semantic Web. As a global schema Web database should be followed, domain ontology is introduced to the annotation procedure for a completed and consistent annotation result. The query interface and the query result features are analyzed in detail, the strategy of query condition reconfigured is adopted, and then the semantic markups of query result are determined. By collecting Web database from different domains, the experiments indicate that the approach proposed can annotate the Web database query result properly under the support of domain ontology.

Key words: deep Web; ontology; semantic annotation; deep annotation; schema mapping

摘 要: 借鉴语义 Web 领域中深度标注的思想,提出了一种对 Web 数据库查询结果进行语义标注的方法.为了获得完整且一致的标注结果,将领域本体作为 Web 数据库遵循的全局模式引入到查询结果语义标注过程中.对查询接口及查询结果特征进行详细分析,并采用查询条件重置的策略,从而确定查询结果数据的语义标记.通过对多个不同领域 Web 数据库的测试,在具有领域本体支持的条件下,该方法能够对 Web 数据库查询结果添加正确的语义标记,从而验证了该方法的有效性.

关键词: deep Web;本体;语义标注;深度标注;模式匹配
中图法分类号: TP311 **文献标识码**: A

随着越来越多的数据库资源可通过 Web 页面中的查询接口进行访问,Deep Web[1]中的信息在迅速增长.Deep Web 一般是指 Web 中可访问的在线数据库,简称为 Web 数据库(Web database,简称 WDB),其内容存储在真正的数据库中.由于 Deep Web 数据的异构性和动态性,如何有效地利用 Deep Web 中的信息是一项具有挑战性的工作.用户通过查询接口向后台数据库提交查询请求,查询结果动态地呈现在结果页面中.目前的查询结果一般仅供人工浏览,为了使获得的数据具有更高的使用价值,这些数据应该是机器可处理的,因此必须为其添加

* Supported by the National Natural Science Foundation of China under Grant No.60573096 (国家自然科学基金); the NSFC-JST Major International (Regional) Joint Research Project under Grant No.60720106001 (NSFC-JST 重大国际(地区)合作项目)

Received 2007-08-31; Accepted 2007-10-19

语义标注.简单地说,就是要为抽取到的数据分配一个有意义的标记来表示数据的语义.

当 WDB 中的信息以 HTML 页面的方式展现时,数据库相关模式结构信息会部分或完全丢失,从页面抽取信息的一个主要目的就是重现数据的模式结构信息.抽取到的数据根据其存储信息的结构化程度可划分为结构化信息、文档信息和非文本文件等形式.由于对结构化信息的处理更有实用意义并且可采用的技术也更多样,本文将主要讨论结构化查询结果的标注.

1 Deep Web语义标注研究现状

有关 Web 信息抽取的研究已经较为成熟,已提出了许多自动/半自动的生成 Wrapper 的方法[2].在 WDB 研究领域也是如此,有多种方法和技术从理论及实践上可以解决数据抽取的问题,但均很少涉及对抽取到的结果添加语义标注.总体上看,目前对 WDB 数据进行标注的研究工作还处于起步阶段,大多以启发式规则的方式对抽取到的数据进行语义注释[3,4],不能对抽取到的全部数据添加语义标注,而且准确性较低.对于 WDB 数据的语义标注,如果仅仅针对一个特定的 Web 数据库,则可以利用机器学习算法预先在一组样本页面上进行训练,挖掘出数据与对应语义之间的关系,推导出一系列的标注规则,然后将其用于新页面的标注.但这种方法与特定的 WDB 相关,即使在同一领域内,不同 WDB 的标注方法也不能相互适用.因此,单纯地使用机器学习的方法不适用于 Web 环境下对大量 WDB 数据的自动标注.

获得 WDB 的模式信息有助于对抽取结果添加语义标注,然而,如果没有 WDB 提供者的合作与支持,则很难直接获取 WDB 的模式信息.通常情况下,由于接口模式和结果模式都直接呈现给用户,因此对查询接口和查询结果的分析是获取 WDB 模式信息的主要途径[5].接口模式反映了 WDB 中可用于查询的属性,结果模式说明了对用户来讲可见的属性.从一些 WDB 的接口模式或结果模式中能够得到某个属性的语义,而有的 WDB 对相应的属性则没有语义注释.如果能够在各个 WDB 模式之间建立匹配关系,则利用这种模式匹配关系可以互补的方式添加数据的语义,但保证语义正确性的前提是要保证这种模式匹配关系的正确性.由于页面结构化程度较差,目前还很难保证页面中模式匹配具有较高的正确性.已有的研究利用层次聚类、模式集成等多种技术,将多种基本的标注方法进行组合,实现对结构化 Web 数据库数据的标注.由于需要通过训练集获得标注过程中使用的重要参数,因此该方法的通用性仍有待提高[6].

通过对大量 WDB 查询接口及查询结果的观察与分析可知,同一领域的 WDB 资源存在着一个隐含的模式,可对该领域的 WDB 资源进行较为准确的描述.该隐含模式可以看作特定领域的 WDB 都遵循的一个全局模式,这个全局模式中包含的属性的语义是确定的.本体作为特定领域共享概念的规范说明,可用于表示一个领域内的 WDB 的特征.建立一个概念的层次树结构,最底层节点是属于父节点概念的实例集合,这样,通过实例查询可以估计每层的每个分类在一个 Web 数据库中所拥有的信息比例,从而能够更好地刻画 Web 数据库在这个属性上的特征.已有一些方法将本体用于 WDB 数据的抽取过程,本体在处理语义相关的问题上具有其特有的优势.本文正是利用了本体具有强语义表达能力的特点,将其用于查询接口和查询结果的分析,从而解决标注 WDB 数据的问题.

2 WDB数据标注的定义及评价准则

本文所关注的语义标注是指利用一组语义明确的词汇,标注 WDB 查询结果中的每个数据,使查询结果不但人容易理解,而且是机器可处理的.这是一个对 WDB 查询结果添加机器可处理的语义标记的过程.设一个词汇集合 $L=\{l_1,l_2,\ldots,l_n\}$,一个待标注的查询结果的属性值集合 $V=\{v_1,v_2,\ldots,v_m\}$,语义标注就是要对每个 $v_i\in V$ 找出一个合适的 l_j,l_j,可以较为准确地描述 v_i 的语义,即建立集合$\{(v_i,l_j)|v_i\in V,l_j\in L,l_j$ 是 v_i 的说明$\}$.

虽然目前仍缺少一个统一的标准来评价 WDB 数据标注的质量,但可以确定的是,在考虑效率的同时,标注方法至少应该满足以下要求:

- 完整性:对查询结果页面中所呈现的 WDB 模式中的属性值都能够正确地标注,即每个 v_i 都能找到一个 l_j 与其匹配.

- 有效性:不受页面中其他附加性和装饰性内容的影响,仅对查询结果页面中的查询结果部分进行标注,以减少不必要的页面分析处理工作,从而提高标注方法的效率,即仅对集合 V 中的元素进行标注,不关心对集合之外的数值进行标注.
- 一致性:若同一领域中的不同 WDB 对同一实体的同一特征使用不同的属性名称进行描述,则对这种属性值的标注应该是一致的;若同一属性值在不同的查询结果页面中具有不同的表现形式,则对其标注也应该是一致的.设两个查询结果集合 $V_1=\{v_1,v_2,\ldots,v_m\}$,$V_2=\{u_1,u_2,\ldots,u_k\}$,若 $v_i=u_j$,则对 v_i 与 u_j 的标注应该是一致的.
- 确定性:在特定领域内,对 WDB 查询结果的标注应该是确定的,标注不会导致产生歧义.若 L 中的多个词汇都可以用于标注 v_i,则这些词汇一定是语义上有关联的.

3 Deep Web查询结果语义标注

3.1 接口模式与结果模式分析

对于一个 WDB,呈现在 HTML 页面上的查询接口通常包含一些 WDB 的数据属性.对于这样一个查询接口,可定义其接口模式:$S_I=\{a_1,a_2,\ldots,a_k\}$,其中,$a_i$ 为属性名称;a_i 与 WDB 模式中的某个属性名相对应,是用户容易理解且具有明确语义的名称.用户可以指定其中部分或全部属性的值,然后提交给后台数据库系统进行查询.查询接口通常以 label 与 text/select/radio 等 Form 表单的基本组件相结合的形式表示,如图 1(b)所示,label 的取值通常可看作为 a_i,label 一般用来标记其右方或下方的表单组件.用户指定或输入的属性值将作为查询条件,在 WDB 中查找满足这些约束的记录.接口模式相对比较容易获得,一般只要能够找到表单,获得 label 的取值,就可推导出接口模式.

查询结果也会包含 WDB 中的若干数据属性,可定义结果模式:$S_R=\{r_1,r_2,\ldots,r_m\}$,$r_i$ 为属性名称.与 S_I 不同,S_R 中的属性名称一般不会直接显示在页面上,需要根据查询结果进行判断.用户在查询接口中指定的属性值一般都会出现在查询结果中,因此,结果页面中该属性值的语义根据接口中的 label 就容易确定.接口模式与结果模式之间通常含有若干共同属性,如果能找到接口模式属性和结果模式属性之间的对应关系,就可以对查询结果数据进行标注.经分析,WDB 模式中的许多重要属性都会出现在查询接口或查询结果中.接口模式与结果模式具有较高的相似性,即两者在包含较多数量的相同属性的情况下,使用模式匹配的方法找到两种模式间的对应关系,利用查询接口中 label 提供的语义信息,确定查询结果的语义信息是比较有效的方法.

3.1.1 查询接口页面特征

对于详细的查询接口,含有较多的属性值,允许用户指定更为详细的查询条件,其接口模式与结果模式具有较高相似性的概率较大.但还有大量的查询接口非常简单,甚至仅提供一个类似于搜索引擎的文本框输入属性值,如图 1(a)所示.为了向用户提供满意的查询结果,许多网站在提供基本查询接口的同时,还支持 Advanced Search 的方式,如图 1(b)所示,细化查询条件,从 Advanced Search 界面中获取的接口模式显然更加完整.因此,本文假定总可以从查询接口页面中获得一个具有语义价值的接口模式.例如,从图 1(b)可获得接口模式如下:

$$S_I=\{\text{Keywords,Author,Title,ISBN,Publisher,Section}\}.$$

3.1.2 查询结果页面特征分析

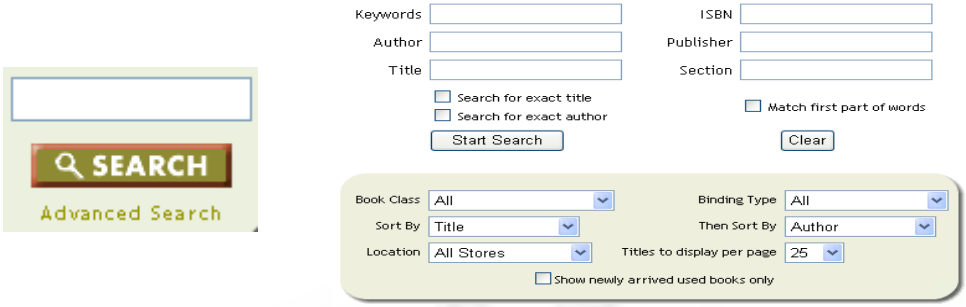
WDB 中满足用户查询要求的数据在结果页面中的组织方式和表现形式是有规律可循的.WDB 中每一条符合查询要求的记录 R 在结果页面中视觉上以数据块 B_d 的方式呈现,如图 2 所示的查询结果中包含了相对独立的两个数据块.一个查询结果页面通常包含若干数据块.设 P_R 为结果页面,则 P_R 可表示为集合:

$$P_R=\{B_{d_1},B_{d_2},...,B_{d_n}\}.$$

B_d 为数据块,$B_{d_i}=(v_{i_1},v_{i_2},...,v_{i_k})$,$i$=1,2,…,$n$,其中,$v_{i_j}$ (j=1,2,…,k)为数据块包含的 WDB 中某个属性的值.

查询结果页面一般都是基于模板自动生成的.为了便于用户浏览并展现网站的特色,一个网站的所有页面往往包含相同的导航栏、目录、广告等装饰性信息,页面的核心内容呈现在特定的区域.针对页面的这一特征,

有研究者提出了抽取 Web 页面中“data-rich”区域的方法[7].本文在对 WDB 查询结果进行分析时,借鉴了该方法对 WDB 查询结果,也就是结果页面的核心内容进行标注,并排除其他装饰性信息的干扰,对核心内容进行分析,保证标注的完整性和有效性.



(a) Simple query interface (b) Advanced query interface

(a) 简单的查询接口 (b) 高级查询接口

Fig.1 Query interface

图 1 查询接口

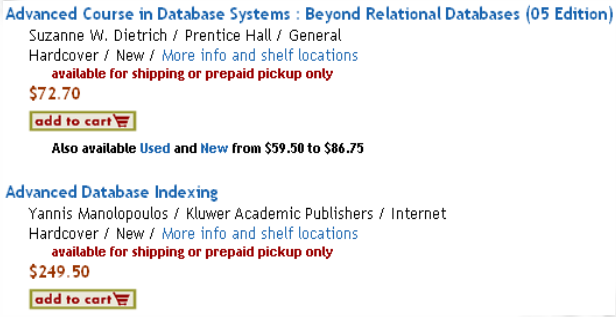


Fig.2 Query result display

图 2 查询结果显示

从数据内容上看,每个块具有相同个数的属性值,即每个 B_d 具有相同的分量个数,且一定包含用户在查询接口中指定的关键词,即存在 v_{i_j} 正是用户从接口输入的查询关键词.

从表现形式上看,不同块之间对应的属性值在 font face,font size,font color,font weight,text decoration(如 underline,strike,italic)等方面具有相同的形式.也就是说,对指定的 j 值,每个 v_{i_j} (i=1,2,…,n)在表现形式上是相同的.

从数据类型上看,每个块所包含的属性值 v_{i_j} 都有其特定的数据类型,即对某个确定的 j 值,每个 v_{i_j} (i=1,2,…,n)的类型都是相同的.

从数据块的页面布局上看,数据块 B_{d_i} 之间是并列的,即将结果 HTML 页面解析,不同块对应属性值 v_{i_j} 的标签路径是相似的.

这些特征对抽取属性值很有帮助,查询结果数据块是同构的,呈现方式也相同.因此,每个相同属性的取值都有同样的表现形式,这样就容易区分每个数据块中的不同属性值,不同数据块中相同属性的取值也容易确定,有助于保证对同一个查询请求所产生的结果页面的标注的一致性.

获得结果模式的一种途径就是根据数据块的组织形式,找到对数据块中每个属性的注解文字,类似于查询接口中的 label 标签的作用,将这些注解文字的集合看作结果模式.数据块最常见的组织形式是表格,一次的查询结果用一张表格呈现,表格中的每一行代表一个数据块,每一列表示一个属性的取值.表格形式清晰,不但易于

人工浏览,也使提取单个数据块或属性值的工作变得容易.在这种情况下,表头标题的向量表示就可以看作结果模式.但更多的时候,使用表格是为了达到更好的显示效果,利用⟨table⟩⟨tr⟩⟨td⟩等对页面进行布局和组织,每个数据块显示为一个独立的块.此时,就需要利用上述的页面特征对数据块进行分析,发现标注性说明文字.在数据块中以相同的文字外观及位置重复出现是说明性文字的主要特征,通过解析 HTML 页面,分析节点的路径,可以容易地找到结果页面中的标注性文字.图 3 是一个股票信息的查询结果,以“说明文字+属性值”的方式呈现.

Aug. 27, 2007 Market Closed Common Stock		Market : NASDAQ-GM	
Last Sale:	$ 37.04	Net Change:	3.42 ▲ 10.17%
Share Volume:	8,059,775	Previous Close:	$ 33.62
Today's High:	$ 38.33	Today's Low:	$ 34.84
Best Bid:	N/A	Best Ask:	N/A
52 Week High:	$ 35.16	52 Week Low:	$ 10.38
P/E Ratio:	NE	Shares Outstanding:	54,586,000
Earnings Per Share (EPS):	$ -1.97	Market Value:	$ 2,021,865,440
NASDAQ Official Open Price:	N/A	Date of Open Price:	N/A
NASDAQ Official Close Price:	N/A	Date of Close Price:	N/A

Fig.3 Query result with text annotation

图 3 具有标注文本的查询结果

获得了结果模式,也就获得了对查询结果数据的一种解释.但仅使用结果模式作为查询结果的语义标注是不够的.首先,结果模式形式多变,不具有稳定性,即使针对同一个 WDB,抽取方法的不同和页面呈现方式的差异也都会导致结果模式的变化.因此,仅使用结果模式作为对查询结果的语义标注,对同一领域内的 WDB 难以建立一个统一的标注,就不能实现标注结果机器可处理的目标.其次,并非所有的数据块中的属性都有标注,如图 2 所示的查询结果,在结果页面中直接显示属性值,没有说明性文字,此时,要获得结果模式信息的主要途径就是将查询接口中用户指定的查询条件字符串与数据块中的属性值进行匹配,如果匹配成功,则可使用查询接口中的标签文字对该属性进行解释;但如果匹配不成功,或查询接口过于简单,就不能利用接口模式信息推导结果模式.为了解决上述问题,本文将领域本体引入 WDB 数据查询结果的标注过程中,利用其共享性、规范性以及一致性的特征,协同结果模式,完成对结果数据的标注.

3.2 利用本体标注结构化信息

对 WDB 数据进行标注的主要目的是机器可处理,这恰好与语义 Web 的初衷相符.对现有 Web 信息添加语义标注是实现语义 Web 必须解决的关键问题.Deep Annotation[8]是语义 Web 领域深入研究的一个问题,主要目的是对动态 Web 页面进行标注.目前提出的 Deep Annotation 过程主要分两部分:WDB 所有者根据数据库的信息结构对 Web 页面进行服务器方的标记;标注者根据服务器方的标记结果和客户本体(client ontology)进行客户端的页面标记,根据标记结果可建立数据库和本体之间的映射规则.Deep Annotation 中结构化的标注结果使得 WDB 的信息是机器可处理的.

目前的 Deep Annotation 方法的一个重要前提是,假设所有 Web 数据库的所有者都参与到标注过程中,在服务器端对 WDB 已进行过标记,即 WDB 的模式是已知的.这只是一种理想化的情况,现有大多数 WDB 的模式对用户来讲都是透明的.但提供一个用于标注的客户本体是可能的,本文正是借鉴了 Deep Annotation 过程中利用客户本体的思想来解决 WDB 查询结果的语义标注问题.

虽然语义 Web 发展还在起步阶段,但是本体在知识表示及推理上的能力已经得到了广泛的认可,使用也越来越广泛,很多行业领域都提出了概括其主要信息的本体模型.同一领域内不同应用所采用的信息尽管可能有差异,但一般都会遵循该领域本体所规定的约束.从这个角度分析,可以将领域本体看作不同应用都遵循的一种“全局模式”.在对 WDB 进行语义标注的过程中,如果有这样一个“全局模式”的支持,则不但可以解决接口模式与结果模式相差太多而给标注带来的问题,而且可以保证标注的一致性并提高标注效率.

一个本体 O 可以表示为$\langle C,P,R,A\rangle$,其中:$C=\{c_1,c_2,\ldots,c_n\}$为概念的集合; $P=\bigcup_{i=1}^{n}P_i$, $P_i=\{p_1,p_2,\ldots,p_m\}$为某个概念 c_i 的属性的集合; $R=\bigcup_{i=1}^{n}R_i$, $R_i=\{r_1,r_2,\ldots,r_k\}$为某个属性 p_j 上的约束集合,对 p_j 的取值进行限制; $A=\bigcup_{i=1}^{n}A_i$, $A_i=\{a_1,a_2,\ldots,a_l\}$为公理的集合,说明了某个属性的性质.一般本体词汇的命名不会使用无意义的字符串,名称可直观反映其所描述的内容,单个英文单词、单词缩写、单词组合等是常用的命名形式.在此不对本体 O 的性质作更深入的分析,仅考虑本文WDB数据标注过程中主要利用的性质.本体采用树型结构对概念进行组织,IS-A 是其中一种重要的关系,也是最常用的一种关系.

设 $n(c)$表示概念 c 的名称,$Instance(c)$表示所有满足概念 c 的实例集合,$c(i)$表示个体 i 为概念 c 的实例,则如下规则是容易理解的:

R1. 有概念 c_i,c_j,若 $n(c_i)$与 $n(c_j)$的编辑距离小于给定阈值,则认为概念 c_i 与概念 c_j 有较高的相似性;

R2. 设 i 为概念 c_s 的实例,若概念 c_p 为 c_s 的父类,则 i 也为 c_p 的实例;

R3. 若 $Instance(c_i)$与 $Instance(c_j)$包含大量相同的个体,即 $\frac{|Instance(c_i)\cap Instance(c_j)|}{|Instance(c_i)\cup Instance(c_j)|}$ 大于给定阈值,则可认为概念 c_i 与概念 c_j 具有较高的相似性.

使用领域本体的目的是提供一个所有 WDB 都遵循的统一规范,但是与被标注词汇直接对应的是接口模式或结果模式中的属性,因此,需要建立领域本体与接口/结果模式间的映射关系,这样才能得到统一的标注结果.以上规则正是建立不同模式间的映射关系、确定 v_{i_j} 语义标注的主要依据.标注过程中使用领域本体有以下优势:

(1) 领域本体的参与保证了 WDB 查询结果中的数值总可以找到一个相对较合适的词汇来标注;

(2) 由于本体适用于特定领域内的每一个应用,因此,当不同应用对现实中同一实体的描述有所差异时,本体作为中介可以协调这种差异,使得对同一实体的标注满足一致性;

(3) 目前,在本体描述语言、本体管理以及本体推理等方面已有一定数量可实用的研究成果,可将其用于 WDB 的研究,以解决 WDB 语义相关的问题.

3.2.1 基于概念实例统计的标注

由于接口模式与结果模式也是对查询结果的最直接的说明,因此需要在接口模式、结果模式与本体概念之间建立映射关系,确保语义上的统一.这种映射关系的建立类似于传统的模式匹配过程,所不同的是,本文强调概念的层次关系及概念-实例关系对标注结果的影响.这种映射关系的建立分两种情况:

1. 如果可以同时获得 WDB 的接口模式 S_I 和结果模式 S_R,那么:

首先,在两者所包含的属性之间建立映射关系,遵循如下原则:

(1) 字符匹配:使用编辑距离对模式属性的名称进行比较,名称相似度越高,语义上越相近;

(2) 值匹配:将用户从接口输入的查询条件字符串与查询结果中的字符串进行匹配,根据相同字符串出现的位置判断具有对应关系的属性.

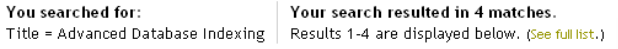
其次,建立结果模式与本体之间的映射,先前的工作已对 Ontology Alignment 方法进行了较为深入的研究[9],综合实体名称、概念实例、本体性质等多方面因素,提出了一种新的本体对齐方法.本文使用该方法实现结果模式与本体之间的映射.

2. 如果结果模式信息不完整,则查询结果数据块中存在从页面中找不到合适语义说明的属性值,此时,选择一个最恰当的本体词汇对该数值进行标注.对此,本文提出了一种“查询条件重置”的策略:

首先,对 B_{d_i} 中的每个分量 v_{i_j}, 逐一将其作为接口模式中 $a_l(l=1,2,\ldots,k)$的值,生成一个新的查询条件 $Cond(v_{i_j},a_l)$,用 $T(v_{i_j},a_l)$ 表示使用查询条件 $Cond(v_{i_j},a_l)$ 从WDB中得到的结果所包含的数据块个数.对于v_{i_j},找到取值最大的 $T(v_{i_j},a_l)$ 所对应的 a_l,则 a_l 为解释 v_{i_j} 的最合适的接口属性.

其次,从本体词汇中选择与 a_l 匹配的词汇,如果存在多个本体词汇与 a_l 相匹配,则选择最特化的词汇,即在概念树中最靠近叶节点的词汇.

使用这种策略的依据是:合理的查询条件可以获得更多的查询结果.例如,如图 4 所示,在“Advanced Database Indexing”的标注未知的情况下,将其作为“Title”的值比作为“author”的值可以得到更多的查询结果,因此判定“Advanced Database Indexing”作为 book title 的可能性更高.



(a) Query result when “Advanced Database Indexing” annotated with “Title”
(a) “Advanced Database Indexing”标注为“Title”的查询结果

(b) Query result when “Advanced Database Indexing” annotated with “Author”
(b) “Advanced Database Indexing”标注为“Author”的查询结果

Fig.4 Query result
图 4 查询结果

3.2.2 具有领域本体支持的WDB数据标注过程

综上所述,在有领域本体支持的条件下,对特定领域的一个 WDB 数据进行语义标注的步骤可描述如下:

输入:WDB 查询接口页面,查询结果页面,WDB 所在领域的本体.

Step 1. 解析查询接口页面,抽取接口模式 S_I.

输出:接口模式 S_I.

Step 2. 输入查询条件,获得查询结果,获取查询结果数据块.

输出:结果模式 S_R;需要被标注的数据对象集合 V.

Step 3. 如果从结果页面中可获取结果模式,则:(1) 在接口模式与领域本体之间建立映射;(2) 在领域本体与结果模式之间建立映射.

输出:(1) 接口模式与领域本体间的映射规则;(2) 结果模式与领域本体间的映射规则.

Step 4. 对于没有文字解释的查询结果,使用“查询重置”策略找到合适的标注词汇.

输出:所有查询结果及其标注,即生成集合$\{(v_i,l_j)|v_i\in V,l_j\in L,l_j$ 是 v_i 的说明$\}$.

Step 5. 根据查询结果及其标注,生成标注文件.

输出:以结构化形式(如采用 OWL 语言)表示的 WDB 标注结果.

4 实验及结果分析

为了对标注方法的有效性进行检验,本文选择了Book,Movie,Music,Auto,Stock领域中共计 80 个WDB进行统计分析.由于目前还缺少通用的测试语义标注方法性能的标准数据集,因此,实验用的WDB资源均通过Google搜索引擎得到,从每个领域的WDB搜索结果中人工选择若干结构相对清晰、包含信息相对丰富的资源作为测试数据.同时,对每个领域选定一个相对设计完整、表述规范的本体作为实验使用[**].正确率(precision)和召回率(recall)是检验标注结果的常用标准.标注的正确率为本方法正确标注的数据块个数与查询得到的全部数据块个数之比;召回率为本方法正确标注的数据块个数与手工方式应该正确标注的数据块个数之比.为了计

** Book: eBiquity Publication Ontology, http://ebiquity.umbc.edu/ontology/publication.owl;
Movie: IMDB Mapping Movie Ontology, http://www.schemaweb.info/schema/SchemaInfo.aspx?id=284;
Music: Music Ontology Specification, http://musicontology.com;
Auto: http://www.cs.umd.edu/~sengcy/classes/828y/scy_auto-ont.daml;
Stock: NASDAQ stock ontology, http://www.daml.org/2002/08/nasdaq/nasdaq-ont.

算正确率与召回率,本文对查询结果页面进行人工分析,提取数据块,为属性值添加语义,以人工处理结果作为比较基准,对方法的有效性进行评价.

在接口模式、结果模式、本体间建立映射关系是本文的标注方法的关键.表 1 说明了本文所使用的模式映射方法的有效性,同样采用正确率和召回率作为评价准则.表 2 分别说明了对 5 个领域内 80 个 WDB 标注的正确率和召回率以及平均正确率和召回率.图 5 展示了"查询条件重置"对方法性能的影响,使用 *fMeasure* 作为评价准则,可以看出,"查询条件重置"策略可以改善标注的效率.

$$fMeasure = \frac{2\times(precision\times recall)}{precision + recall}.$$

Table 1 Performance of the used schema mapping

表 1 模式映射性能

Domain	Precision (%)	Recall (%)
Book	96.8	97.2
Movie	97.6	98.3
Music	96.2	96.6
Auto	99.3	99.3
Stock	99.3	99.3
Average	97.8	98.1

Table 2 Performance of the proposed annotation method

表 2 标注方法性能

Domain	Precision (%)	Recall (%)
Book	97.7	97.0
Movie	94.5	93.5
Music	90.2	90.2
Auto	99.6	94.5
Stock	100	96.0
Average	96.4	94.2

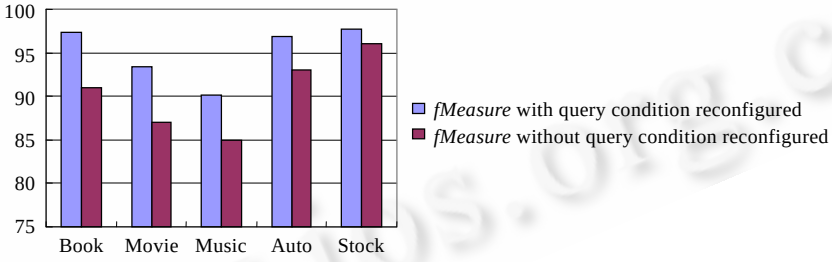


Fig.5 The effect of "Query Condition Reconfigured" on *fMeasure*

图 5 "查询条件重置"对 *fMeasure* 的影响

以上结果表明,使用本文方法所标注的 WDB 查询数据具有较高的正确率和召回率.方法的效率主要受以下几方面因素的影响:一是如何从结果页面中抽取查询结果数据块;二是抽取到的接口模式和结果模式的质量直接影响标注词汇的选择,由于关于信息抽取的研究已比较成熟,获得准确抽取结果的前提条件可以得到满足;三是接口模式、结果模式以及本体间的对应关系的建立.因此,WDB 查询结果数据语义标注的添加是一个将多种任务有机结合在一起的复杂过程,为了提高标注的效率,需要考虑到各方面因素的影响.本文正是在充分借鉴 Web 数据研究领域成果的基础上,为了解决数据的语义标注问题而提出了一种新的标注方法.

在本文方法的实现过程中,编辑距离的计算是模式匹配的重要步骤,其计算复杂度是指数级的,这在实际应用中是不可接受的,但可采用动态规划算法将编辑距离的计算复杂度降低到多项式级别.因此,模式匹配过程中采用编辑距离是可行的.此外,查询重置策略将查询结果值与每个标注词汇进行组合,并进行再次查询.这些都是费时的操作,但由于查询结果 B_d 中属性值的排列是相对固定的,同一位置显示的查询结果值一般情况下属于同一个属性,因此,可参考对 B_{d_j} 的标注来标注 B_{d_j}, $i<j$.

5 结 论

对 WDB 查询结果添加语义标注仍然是一个新的研究问题.为了获得完整、一致的标注结果,本文提出的方法将领域本体看作“全局模式”参与到标注过程中,并充分考虑到了 WDB 查询接口和查询结果的特征,所提出的“查询条件重置”的策略有效提高了标注结果的正确率.今后拟在以下两个方面对现有工作进行改进:一方面,增强对属性的标注能力.本文主要考虑利用本体概念-实例关系,使用概念对抽取到的信息进行标注,忽略了对概念属性的标注.而表达实体属性是本体的主要特征之一,从结构化的 WDB 查询结果中提取实体间的关系并进行标注是实现机器可理解 Web 内容的重要任务.完整的语义标注应该考虑对实体属性的标注.另一方面,为了实现机器对标注结果的可理解、可处理,标注结果的表现形式和存储方式需要深入分析,如标注结果的自动生成以及标注文件的管理.

References:

[1] Bergman MK. The deep Web: Surfacing hidden value. White Paper on the Deep Web. 2001. http://www.brightplanet.com/pdf/deepweb whitepaper.pdf

[2] Liu W, Meng XF, Meng WY. Deep Web data integration. Technical Report, WAMDM-TR-2006-3, WAMDM, 2006 (in Chinese with English abstract). http://idke.ruc.edu.cn/reports/report2006/seminar%20summary/Deep%20Web.pdf

[3] Arlotta L, Crescenzi V, Mecca G, Merialdo P. Automatic annotation of data extracted from large Web sites. In: Christophides V, Freire J, eds. Proc. of the 6th Int'l Workshop on Web and Databases. San Diego: ACM Press, 2003. 7−12.

[4] Wang JY, Lochovsky FH. Data extraction and label assignment for Web databases. In: Proc. of the 12th Int'l World Wide Web Conf. Budapest: ACM Press, 2003. 187−196.

[5] He H, Meng WY, Lu YY, Yu C, Wu ZH. Towards deeper understanding of the search interfaces of the deep Web. World Wide Web, 2007,10(2):133−155.

[6] Lu YY, He H, Zhao HK, Meng WY, Yu C. Annotating structured data of the deep Web. In: Proc. of the IEEE 23rd Int'l Conf. on Data Engineering. Istanbul: IEEE Computer Society Press, 2007. 376−385.

[7] Wang JY, Lochovsky FH. Data-Rich section extraction from HTML pages. In: Keong W, Ling TW, eds. Proc. of the 3rd Int'l Conf. on Web Information Systems Engineering. Singapore: IEEE Computer Society Press, 2002. 313−322.

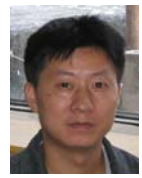
[8] Handschuh S, Staab S, Volz R. On deep annotation. In: Proc. of the 12th Int'l World Wide Web Conf. San Diego: ACM Press, 2003. 431−438.

[9] Yuan L, Li ZH, Chen SL. Inference rules guided ontology alignment. Journal of Computational Information Systems, 2006,2(3): 1085−1090.

附中文参考文献:

[2] 刘伟,孟小峰,孟卫一.Deep Web 数据集成问题研究.科技报告,WAMDM-TR-2006-3,WAMDM,2006. http://idke.ruc.edu.cn/reports/report2006/seminar%20summary/Deep%20Web.pdf

袁柳(1979－),女,陕西西安人,博士生,主要研究领域为语义 Web,信息检索.



陈世亮(1968－),男,博士生,主要研究领域为多媒体信息管理.

李战怀(1961－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据管理技术.