书籍搜索领域 Deep Web 数据集成系统

钟 昕,伏玉琛
(苏州大学 计算机科学与技术学院,江苏 苏州 215006)

摘 要:随着在线数据库的迅速增长,可以访问的数据库资源大大增多,但它们的信息传统搜索引擎无法获得,它隐藏在网站背后,成为人们快速有效获取信息的障碍。为了获得 Deep Web 中大量有价值的隐藏信息,需要整合各在线异构数据源,以便在同一领域内比较某一事物的大量相关信息。目前,越来越多的人采取网上买书的消费方式,针对这个消费热点问题,设计了一个书籍搜索领域的 Deep Web 数据集成系统,提供一个集成的查询接口,使得用户可以方便地进行查找和比对。

关键词:Deep Web;Web 数据集成;书籍搜索
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2008)09-0050-03

A Deep Web Data Integration System for Book Searching Domain

ZHONG Xin, FU Yu-chen
(School of Computer Sci. & Tech., Soochow University, Suzhou 215006, China)

Abstract: With the rapid development of Web databases, there are more and more Web databases available for users to access. But their information which is hidden behind the web sites, is not available for traditional search engines. So, to get so much resource quickly and effectively is very difficult. To obtain mass valuable information in deep Web, and compare related information for one issue in the same field, need to integrate all heterogeneous databases. Now, more and more people buy books online, aiming at this popular problem, design an integrated book search system over deep Web data, provide an integrated search interface, which will make users search and compare books easily and effectively.

Key words: Deep Web; Web data integration; book search

0 引 言

近年来,在线数据库迅速增长,它们的信息不能通过传统搜索引擎获得,也不能通过静态 URL 链接得到,是隐藏在网站背后的数据库信息。2004 年 4 月的最新调查[1]表明大约有 450000 个在线数据库。为了获得 Deep Web 中大量有价值的隐藏信息,在同一领域内比较某一事物的大量相关信息,需要把在线异构数据源进行整合集成。目前,越来越多的人采取网上买书的消费方式,通过大量的书籍提供网站搜索自己想要的书。用户需要通过在他们各自的查询接口上提交查询来获得想要的信息。面对日益增多的网站,手工一个一个的查找显然很不实际,而且这种方式下得到的信息不仅不全面,而且是独立的。比如,想要知道某本书到底有哪些网站提供,哪个网站的信誉度最好,售价最便宜,运费最低,送货最及时等信息,需要手工去查找书籍提供网站,输入查询条件,记录结果,这显然是费时烦琐的。考虑到这种问题,设计了一个 Deep Web 数据集成系统,提供给用户一个统一的查询接口,在该统一接口的查询界面上输入查询条件后,由接口自动向同一领域各在线数据库提交查询,进行查找,最终反馈给用户最优的一系列查询结果。Deep Web 数据集成是当前的热门研究领域,有大量的研究成果,本系统的实现是 Deep Web 数据集成的一个典型应用,在借鉴大量已有成果的基础上,进行了适当改进与创新。

收稿日期:2007-12-14
基金项目:国家自然科学基金(60673092);江苏省高校自然科学基金(07KJD520187)
作者简介:钟 昕(1984-),女,硕士研究生,研究方向为 Deep Web、Web 数据挖掘;伏玉琛,副教授,研究方向为数据挖掘、地理信息系统。

1 系统框架

本系统的目的是实现一个书籍的元搜索引擎,架构流程如图 1 所示。主要包括以下几个步骤:第一,书籍网站的发现,用一个特定领域的爬虫程序找到书籍领域的在线数据库资源,并找出可以提交查询的查询

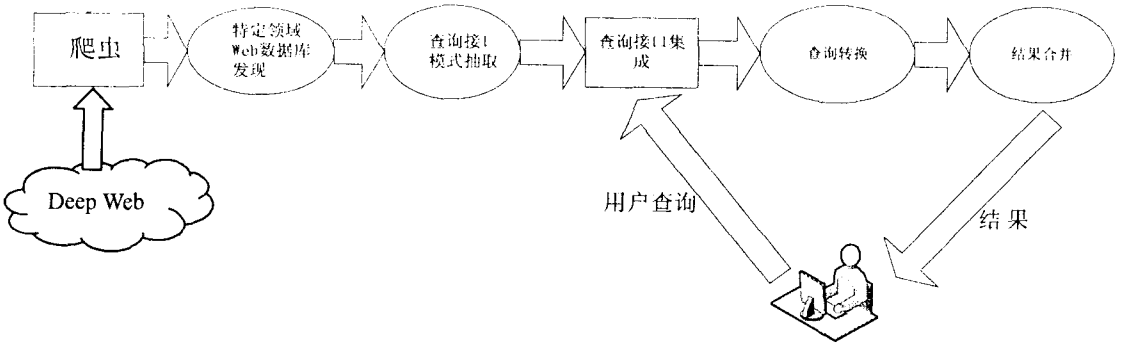

图 1　系统流程图

接口;第二,查询接口模式抽取,将查询接口的模式信息抽取出来,以很好的结构存储;第三,查询接口集成,利用抽取好的模式信息,进行模式匹配,得到一个统一的查询接口;第四,查询转换,将在统一接口上提交的查询转换到各查询接口上处理;第五,结果合并,将各个查询接口得到的结果抽取,进行语义注释、实体识别、结果合并,最终反馈给用户。

2　主要实现技术

本集成系统实现过程中的每个步骤都是一个当前的研究热点,下面介绍每个步骤的主要实现技术。

2.1　书籍网站的发现

Web 信息急剧膨胀使搜索引擎专用化成为发展趋势,定向采集信息成为搜索引擎一个重要研究方向,聚焦爬虫[2]应运而生。聚焦爬虫是一个自动下载网页的程序,它根据既定的抓取目标有选择地访问万维网上的网页与相关链接,获取所需信息。笔者基于朴素贝叶斯分类算法设计一个聚焦书籍领域的爬虫。根据观察到的 94% 的查询接口深度不超过 3,可以大大缩小搜索范围。

找到书籍领域的网页后,要从中发现可以提交查询的接口,采用一种简单的判断规则:首先页面中要有 Form 标签;其次 Form 标签中必须有 Text 输入控件;最后,至少出现一组关键词中的一个,如"查询"、"搜索"等。实验证明这种简单方法的准确性比利用 C4.5 决策树方法的更好。

2.2　查询接口模式抽取

查询接口模式抽取主要是查询接口属性和控件约束元素的获取与分析,把它们按照逻辑关系重组成一个个属性添加到接口模式集合中。为了让用户更容易理解和使用查询接口,设计者通常会融入多种类型的视觉特征,主要包括位置、布局和外貌等特征。基于此,本系统采用基于视觉的网页分割算法[3,4]来进行模式抽取,充分利用 Web 页面的视觉特征,结合 DOM 树进行页面语义块的分析。首先采用 VIPS 算法把查询接口所在页面转换成 Visual Block Tree,将包含 < form > 标签的最小 Block 作为查询接口区域。在查询接口中存在多个文本 block,需要根据位置、布局、外貌等特征的相似度对其进行聚类来确定标签 block。最后将得到的每个标签属性进行结构化存储。

2.3　查询接口集成

查询接口的集成是为了给用户提供一个对属于同一领域的 Web 数据库统一的访问途径。通过属性分析是查询接口集成最主要的途径,这种方式主要发掘给定查询接口的模式信息和语义信息,利用这些语义信息来识别不同查询接口上属性之间的匹配关系,在这些具体的查询接口之上获得一个集成的查询接口。模式匹配是实现接口集成中的关键技术,也是整个 Deep Web 数据集成中的关键。目前,已经有了很多技术[5~9]。文中提出一种基于数据挖掘技术的算法(Correlated - clustering)来实现。首先在输入的属性组 I 中找出使用频率最高的若干属性形成词表,再根据两项间的积极相关度量标准[8] m_p 和给定的阈值[8] T_p 进行组发现,将组属性加入到词汇表后,得到更新的词汇表。在此基础上把组属性和其他频繁使用的属性分别对应一个概念,通过计算概念相似度,进行概念聚类,最后把侯选概念集进行排序,再用贪婪算法进行最优选择,得到最终匹配结果。其中概念相似度通过公式(1)[7]来计算:

$$\text{conceptSim}(C_1, C_2) = \lambda_{ls} * \text{lingSim}(C_1, C_2) + \lambda_{ds} * \text{domSim}(C_1, C_2) \quad (1)$$

其中 λ_{ls} 和 λ_{ds} 是反应组成相似度的各部分的权值系数,$\text{lingSim}(C_1, C_2)$ 表示语言学上的相似度,$\text{domSim}(C_1, C_2)$ 表示数据域的相似度。

得到一个较为完整的匹配模型后,按照属性无冗余、属性全面、界面友好等原则,构造一个良好的统一查询接口。

2.4　查询转换

当用户在集成查询接口上填写并提交查询时,是为了同时得到从多个 Web 数据库中获得符合该查询

的结果,并把这些异构的数据以统一的模式存储或展现。首先要为用户选择合适的Web数据库,把查询近似等价地转化到在这些具体Web数据库查询接口上的查询。一个领域中存在大量可访问的Web数据库,都访问需要花很大代价,要选出合适的Web数据库进行查询转换,使得花尽可能少的访问代价,获得冗余度足够小,且满足特定查询的结果。采用文献[10]提出的基于直方图的Top-N的方法实现,先判断数据库与特定查询之间的相关性,再确定最适合提交查询的数据库和从返回的结果中选择最合适的记录。

查询转换涉及到集成接口与各Web数据库接口之间属性的匹配、约束的映射以及查询重写方面的问题。属性匹配部分上面已经作过叙述,约束的映射部分本系统采用Z.Zhang等人提出的基于数据类型的谓词匹配方法[11],这种方法在观察谓词模板中发现,150个Web数据库中共有37个模板,有两个使用最多:[attr;default;$val]和[attr;default;$val∈{D}]。当谓词模板出现在不同数据源中,而且表达的是同一概念时,才是对应的。谓词的对应关系经常存在于数据类型相同的谓词模板之间,这一特征与具体的Web数据库及具体的域无关。由于模式、约束和查询能力的不同,一般只能近似进行转换。查询重写主要是基于查询能力的查询重写,即尽可能填写接口的每一个谓词,以最大限度使用各数据库的查询能力。本系统中主要考虑三方面因素:属性约束、谓词映射、查询能力。

2.5 结果合并

多个Web数据库返回的结果,需要进行合并和去重,最后反馈给用户。对于结果的抽取仍采用基于视觉特征的抽取方法[3,4],但是结果模式和接口模式有不同,结果页面上都是结构化的文本内容,而查询接口每个文本标签总会对应一个表单控件,因此,对结果页面的抽取比接口模式难度更大。从方位上来说,若块A后面紧挨":",或在其后面很小的距离之内(一般是两到三个空格的距离)存在另一个块,认为A是字段块。从数量上来看,若存在个数大大小于n的类C_i,将它去除,认为这是由某些记录中含有相同属性值内容而形成的。经过两层判断,最终得到的就是字段块。

语义注释方面的工作目前还在起步阶段。本系统通过机器学习的方式预先在一组样本页面上训练形成一个添加语义的程序,学习出数据与对应语义之间的关系,再对各个Web数据库的模式之间建立匹配关系,将这两种关系以互补的方式达到对数据语义的添加。但是由于页面的结构化程度很差,实验结果并不理想。

目前的实体识别工作也很不成熟,只是在关系模式和半结构化的XML模式上开展。其中的关键问题是:建立实体之间属性的映射关系和属性之间值的比较。实体间属性的映射即模式匹配。属性间值的比较则首先选取能够代表实体的属性,然后在这些代表性的属性上值的比较。本系统借鉴基于实例的方法[10]来完成最终的结果合并,这种方法考虑了数据类型、与属性相关的高频词,实验证明效果不错。

3 相关工作

本系统不仅是Deep Web数据集成的应用,而且是对数据集成领域已有工作的集成。国内Deep Web的研究才刚刚起步,成型的系统更是稀少。本系统中的各部分并不紧密联系,每个部分都可进行独立研究,可作为一个很好的实验平台。当然,也存在很多问题。在查询接口发现过程中,不能把代表Web数据库的查询接口与搜索引擎区的查询接口区分开。模式匹配、查询重写、实体识别方面准确度还达不到很好的要求,有待进一步改善。

目前,代表性的Deep Web数据集成系统有:Wise-Integrator[12]和Metaquerier[13]。Wise-Integrator是对e-commerce进行数据集成的一个系统,它是一个综合的解决方案,首先对每个查询接口进行分析,获取其中的属性信息,在语义分析的过程中用到了Wordnet,然后进行属性匹配,在完成对所有查询接口的属性匹配后,要为匹配的属性在集成的查询接口上确定它的全局名称和它的类型和取值范围,这样就得到集成的查询接口。但是也存在不足:首先把查询接口看作是一个"平"的结构,实际上查询接口具有很丰富的结构信息;其次是只考虑了查询接口属性间1:1的匹配关系,无法进行复杂匹配。Metaquerier可以进行动态发现和灵活的查询转换,也可以进行属性间$m:n$的复杂匹配,但它着重关注查询接口的处理过程,查询结果的处理没有详细阐述。

4 结束语

随着在线数据库的迅速增长,所设计的书籍元搜索引擎为用户提供了统一的访问接口,使得用户可以方便地进行书的挑选和比对。将本系统进行适当的参数调整,也可以应用于其他领域。Deep Web数据集成还有巨大的研究空间,仍需不断进行探索。

参考文献:

[1] Chang K C-C,He B,Li C,et al. Structured databases on the web:Observations and implications[J]. SIGMOD Record,

(下转第56页)

由 TCL 脚本向 MplsModule 类发起 CR(Constraint - based Routing)请求,Simulator 类通过调用显示路由计算模块将一个显示路由 ER(Explicit Route)返回,然后 TCL 脚本再次发起 CRLSP 建立请求,最后由 LDPAgent 根据 CRLSP 请求中指定的显示路由,使用标记映射消息建立一个显示标记交换路径 ER - LSP。

2.4 包转发机制的设计和实现

MPLS 节点的数据包转发由 MPLS 分类器完成。数据包到达 MPLS 节点后,MPLS 分类器负责执行的操作过程如图 5 所示。

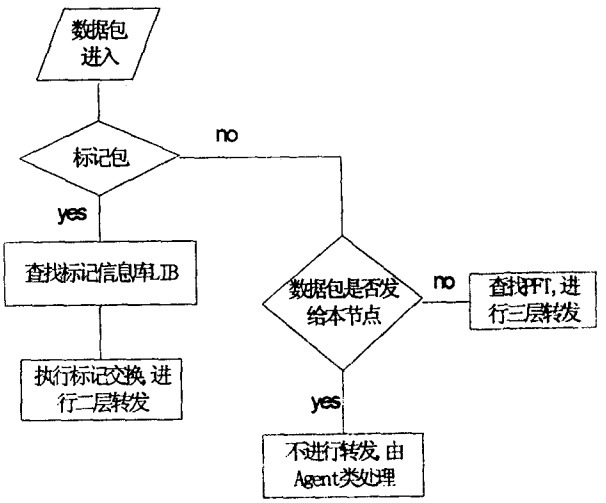

图 5 MPLS 分类器的操作过程

3 结束语

利用 MPLS 实施流量工程可以在现有网络体系结构基本不变的前提下,合理利用网络资源,优化网络性能。利用显示路由和故障恢复机制,MPLS 提供了基本的流量工程能力,这使得大型网络中利用 MPLS 实施性能优化变得十分便利。文中使用 NS2 模拟环境对 MPLS 流量工程进行性能评估和测试。通过对 NS2 的扩展,实现了一个 MPLS 流量工程仿真器,可以支持标记分发,标记交换路径 LSP,显示路由以及区分服务等。

参考文献:

[1] Rosen E, Viswanathan A, Callon R. Multiprotocol Label Switching Architecture[S]. IETF RFC 3031, 2001.

[2] 石晶林,丁 炜. MPLS 宽带网络互联技术[M]. 北京:人民邮电出版社,2001.

[3] 谢金星,刑文训. 网络优化[M]. 北京:清华大学出版社,2000.

[4] Awduche D. Requirements for Traffic Engineering Over MPLS[S]. IETF RFC 2702, 1999.

[5] 黄 河,李伟琴. MPLS 流量工程体系结构优化研究[J]. 北京航空航天大学学报,2003,29(3):221 - 224.

[6] Kodialam M, Lakshman T V. Minimum Interference Routing with Applications to MPLS Traffic Engineering[C]//IEEE INFOCOM. 2000. [s. l.]: IEEE Communications Society Press, 2000:884 - 893.

(上接第 52 页)

2004,33(3):61 - 70.

[2] 周立柱,林 玲. 聚焦爬虫技术研究综述[J]. 计算机应用,2005,25(9):1965 - 1969.

[3] Cai D, Yu S, Wen J, at al. Extracting Content Structure for Web Pages Based on Visual Representation[C]//In APWeb, 2003. Xi'an:[s. n.],2003:406 - 417.

[4] Cai D, Yu S, Wen J, et al. VIPS: a Vision - based Page Segmentation Algorithm[R]. Microsoft Research Technical Report, MSR - TR - 2003 - 79, 2003.

[5] Rahm E, Bernstein P A. A survey of approaches to automatic schema matching[J]. VLDB Journal, 2001, 10(4):334 - 350.

[6] WANG J, WEN J - R, Lochovsky F, et al. Instance - based schema matching for web databases by domain - specific query probing[C]//In VLDB 2004 Conference. Toronto, Canada: [s. n.],2004.

[7] WU W, YU C T, Doan A, et al. An interactive clustering - based approach to integrating source query interfaces on the deep web[C]//In SIGMOD 2004 Conference. Paris, France: [s. n.],2004.

[8] He B, Chang K C - C, Han J. Automatic complex schema matching across web query interfaces: A correlation mining approach[R]. Technical Report UIUCDCS - R - 2003 - 2388, Dept. of Computer Science, UIUC, 2003.

[9] Wu W, Doan A, Yu C T. WebIQ: Learning from the Web to Match Deep - Web Query Interfaces[C]//In ICDE 2006. Atlanta, GA, USA:[s. n.],2006.

[10] Yu C T, Philip G, Meng W. Distributed top - N query processing with possibly uncooperative local systems[C]//In: Proceedings of the 29th International Conference on Very Large Data Bases. Berlin:[s. n.],2003:117 - 128.

[11] Zhang Z, He B, Chang K C - C. Light - weight domain - based form assistant: Querying Web Databases On the Fly[C]//In VLDB Conference. Trondheim, Norway:[s. n.],2005:97 - 108.

[12] He H, Meng W, Yu C, et al. Wise - integrator: An automatic integrator of web search interfaces for e - commerce[C]//In VLDB 2003 Conference. Berlin:[s. n.],2003.

[13] Chang K C - C, He B, Zhang Z. Toward large scale integration: Building a metaquerier over databases on the web[C]//In CIDR 2005 Conference. Asilomar, CA, USA:[s. n.],2005.