

Deep Web 信息集成系统中查询转换

王 兵¹, 刘彩虹²

(1. 上海理工大学 信息办, 上海 200093;

2. 上海理工大学 管理学院, 上海 200093)

摘 要:随着 Internet 信息的迅速增长, 许多 Web 信息已经被各种各样的可搜索在线数据库所深化, 并被隐藏在 Web 查询接口下面。传统的搜索引擎由于技术原因不能索引这些信息——Deep Web 信息。由于 Deep Web 惟一“入口点”是查询接口, 为使查询接口自动产生有意义有查询, 给出了 Deep Web 信息集成系统框架, 提出了基于数据类型的搜索驱动的用户查询转换方法, 基于此设计并实现了一个针对中文 Deep Web 信息集成原型系统。通过在实际 Deep Web 站点上的实验证明了此方法是非常有效的。

关键词:Deep Web; 信息集成; 表单抽取; 查询转换

中图分类号:TP393

文献标识码:A

文章编号:1673-629X(2008)07-0176-05

Query Translation on Deep Web Information Integration System

WANG Bing¹, LIU Cai-hong²

(1. Information Office, University of Shanghai for Sci. and Techn., Shanghai 200093, China;

2. Business School, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: As the amount of information on the Web increases rapidly, much Web information has been deepened by myriad searchable on-line databases, been hidden behind query interfaces. Traditional search engine does not index these information. Since the only “entry point” to a hidden web sit is a query interface. Provides the frame of Deep Web information integration system, and the method of mapping a user query to a set of Deep Web source query interfaces is proposed to automatically generate meaningful queries. And based on it, a Deep Web information integration system frame is designed and implemented. The Deep Web site experiment shows that this method is very effective in reality.

Key words: Deep Web; information integration; form extracting; query translation

0 引 言

为了帮助人们在 Internet 中找到自己需要的信息, 出现了搜索引擎。但是, 目前主流搜索引擎大多只能搜索互联网表面可索引的信息, 而更加丰富的、有价值的信息被隐藏在网络深处, 这些信息受当前搜索引擎技术限制不能被检索到, 被称为 Deep Web 信息^[1]。Deep Web 信息量比静态页面信息量多, Deep Web 页面内容存储在可搜索的数据库中, 这些页面是用户特定查询被提交后, 由后台数据库动态创建产生的。由于人工对每个 Web 数据库查询接口提交查询来获取感兴趣信息是费时费力的, 因此对构建 Deep Web 信息

集成系统来说, 用户查询转换技术相当关键。

1 Deep Web 信息检索框架

文中设计的 Deep Web 信息检索系统框架见图 1。其中查询转换是 Deep Web 信息检索系统框架中重要组成部分, 它负责将用户查询请求转发到各个主题领域内所选择的多个目标查询接口上。如果每个主题领域内数据源是固定不变的, 那么可以通过分析目标查询接口模式将用户查询请求转发给所有的目标数据源。然而, 随着 Deep Web 数据源的迅速发展, 每个主题领域内的数据源都在不断更新或增加。因此要把用户查询请求转发给所有动态变化的未知模式结构的目标数据源存在很大困难。

2 用户查询转换

将用户查询转换(Query Translation)到多个动态

收稿日期: 2007-10-14

基金项目: 上海市重点学科资助项目(T0502)

作者简介: 王 兵(1973-), 男, 硕士研究生, 工程师, 主要研究领域为智能信息处理; 刘彩虹, 博士研究生, 主要研究领域为信息系统与信息系统管理, 供需网。

选择的目标数据源的查询接口上,这就需要一个查询转换器来完成这项任务。为提高用户 Deep Web 信息检索的质量,当用户选择了某个主题领域后,针对每个主题领域构造了一个统一的查询 Deep Web 信息的接口界面。同时使用相应于特定应用领域对象的谓词集来描述某领域统一查询接口。如构造的汽车领域的统一查询接口可以使用品牌、车型、里程、价格、引擎、颜色、日期等谓词来描述。系统存储用户输入谓词的结构[标签名;属性名;修饰语;值域]以用于模型化查询接口。

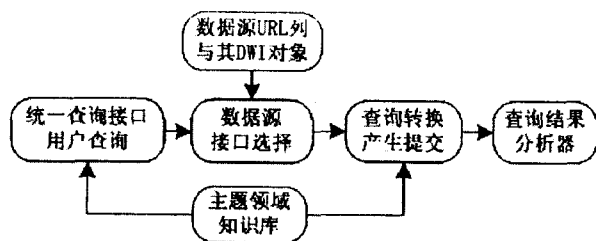


图1 Deep Web 信息检索框架

2.1 查询转换形式化定义

查询转化是把统一接口 S 上的用户查询“翻译”到多个目标表单 T 上以在目标表单上产生查询,它是 Deep Web 信息集成系统的一个关键部件。文中设计的查询转换器是一个可扩展的中间件。

假设要建立关于汽车主题的查询转换器,它可以将汽车领域内的用户查询从源表单翻译到同领域内任意目标表单上。例如将源表单(见图2)上的查询翻译到目标表单(见图3)上。假设 Q_s 是源表单查询在目标表单谓词上映射的所有可能查询。源表单上的谓词有 $P_{s1} = [\text{中国售价}; \text{CHNPrice}; \text{在...之间}; \{E_{11}, E_{12}\}]$, $E_{11} = [\text{万元到}, \text{first_price}, 10]$, $E_{12} = [\text{万元}, \text{second_price}, 20]$; $P_{s2} = [\text{品牌}; \text{chx_pinpai}; \text{含有}; \text{本田}]$; $P_{s3} = [\text{产地}; \text{chx_local}; \text{含有}; \{E_{31}, E_{32}\}]$, $E_{31} = [\text{国产}, \text{local}, \Phi]$, $E_{32} = [\text{进口}, \text{foreign}, \Phi]$; $P_{s4} = [\text{品牌关键词}; \text{csle_model}; \text{含有}; \text{雅阁2005}]$ 。而目标表单 T 上只有四个谓词: $P_{T1} = [\text{品牌}; \text{brand}; \text{含有}; \Phi]$; $P_{T2} = [\text{系列}; \text{series}; \text{含有}; \Phi]$; $P_{T3} = [\text{价位}; \text{cost}; \text{在...之间}; \Phi]$; $P_{T4} = [\Phi; \text{chandi}; \text{其中之一}; \{E_{41}, E_{42}\}]$, $E_{41} = [\text{国产}, \text{是}, \Phi]$, $E_{42} = [\text{进口}, \text{是}, \Phi]$ 。

将源表单 S 上的源查询转换到目标表单 T 上,需要协调好不同数据源接口上的三个问题:

(1)谓词识别:源查询接口和目标查询接口上出现的谓词形式可能不同但表示相同的概念。例如源表单上出现的谓词为“中国售价”,而在目标表单上出现的是“价位”。为了准确地转换用户查询必须协调好谓词标签名之间的匹配关系。

(2)谓词映射:两个数据源可能使用不同的谓词结

构来表示相同的概念,如源表单上使用含有两个元素的谓词“中国售价”来表示价格而目标表单上使用含有一个元素的谓词“价位”来表示价格,这就需要转换器尽可能地做好谓词映射。

(3)查询产生:两个数据源中的谓词值域可能不同,查询也可能使用的是表单上不同谓词组合,因此有必要对产生的目标查询进行调整或过滤。

图2 源表单

图3 目标表单

查询转换包含两个过程:

1)找出源查询在目标表单 T 上的联合查询 Q_t ,它是由源查询映射到目标表单上的所有查询构成的集合。

2)最小化代价处理,它是使用过滤器 σ 对 Q_t 进行过滤,找到其最小包含,以减少不相关的检索结果,尽量使过滤后的查询 $\sigma(Q_t^*)$ 与源查询尽可能接近。

事实上查询转换问题就是用源表单查询构造目标表单查询,然后对其进行过滤,得到一个语义最小包含源查询的查询集。

2.2 查询转换系统结构

文中设计的用户查询转换器由三部分组成:表单谓词识别,表单谓词映射和产生映射查询。其结构见图4。

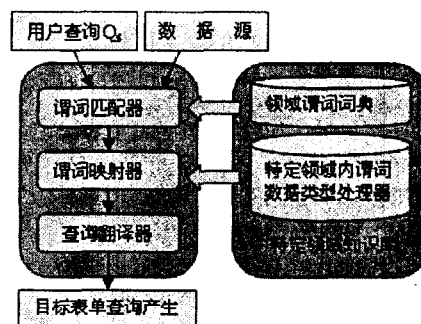


图4 查询转换器结构

(1)表单谓词识别。

其本质是识别谓词标签名及属性名中的同义词。在特定主题领域内,可以使用自动发现或手工编码的同义词词典的方法来完成谓词识别。由于 Deep Web 信息集成系统中数据源是动态选择的,因此对特定主题领域的词典可以根据已有的动态匹配技术^[2,3]来动态构造。

(2)谓词映射器。

对匹配于谓词词典的谓词,关键是如何确定该谓词在目标表单上的修饰语和值域。由于具有相同数据类型的谓词通常具有相同的映射模式,基于此发现文中使用了基于数据类型的搜索驱动的谓词映射方法。

(3)查询翻译器。

根据目标表单谓词组合方式映射用户查询,最终在目标表单上产生查询使其满足目标表单的语法约束。已有研究是基于范围谓词的查询重写技术^[4~6],并且它不针对特定主题领域的数据源。

2.2.1 表单谓词识别

谓词标签名与内部属性名是表单上最具有语义的元信息。谓词匹配器就是用来识别表单上谓词与主题领域谓词词典中的谓词是否是语义相关的。谓词匹配器是由特定主题领域词典定制的,它索引了某领域内经常使用概念的同义词。系统运行时,它用来对多个数据源上输入类型谓词的标签进行模糊匹配以判定它们是否表达的是相同的概念。谓词匹配器工作分两个步骤:

(1)预处理:预处理阶段执行谓词的标准化的,包括识别谓词标签名与属性名中的拼写错误,去除标签与属性名中的特殊符号等。

(2)核对同义词:由于谓词的值类型也具有一定的语义,同一个谓词在其值类型不同情况下通常具有不同的语义,因此也可以把谓词值类型考虑进来。如果输入类型谓词的值类型相同,此谓词中有超过 50% 的值与词典中一个谓词对象的值相同,同时它们的标签名互为同义词,则认为这个谓词和谓词词典中的对象是匹配的。对谓词值域类型的识别使用的是谓词映射过程中的类型识别器。

2.2.2 表单谓词映射

关于谓词映射已经存在的方法通常假设一个静态小规模规则集,如研究使用一个基于数据源的成对规则驱动的映射机制^[7]来映射谓词。以下是从源表单 S 到目标表单 T 谓词映射规则。

R1 [中国售价; FirCHNPrice; > = ; ¥ a], [中国售价; SecCHNPrice; < = ; ¥ b]

→映射:[价位; cost; 在其之间; ¥ (a, b)]

R2 [品牌; pinpai; 含有; \$ t] →映射:[品牌; brand; 含有; \$ t]

R3 [产地; local; 含有; \$ t] →映射:[产地; local; 含有; \$ t]

R4 [品牌关键词; model; 含有; \$ t] →映射:[系列; series l; 含有; \$ t]

然而这样的映射机制依赖于数据源的谓词模式特

征,同时成对的规则编码方式不适于数据源的动态更新。它存在如下缺点:

(1)基于数据源的映射只能映射数据源已经被预先配置的情况。

(2)成对规则编码方式不易于扩展到多个数据源上。

(3)由于它们是静态规则,因此数据源谓词查询范围一旦改变就需要重新编码规则,这使得维护规则非常费力。

为实现更加一般的可扩展性好的谓词映射机制,观察发现谓词映射不仅跟谓词模式结构有关,还与谓词值的数据类型相关。在此介绍了基于数据类型的搜索驱动的谓词映射机制,谓词映射器的结构见图 5。

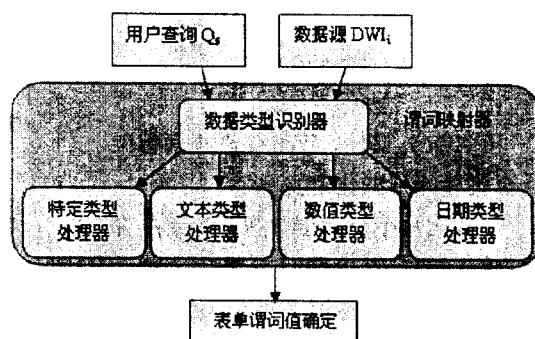


图 5 谓词映射器结构

谓词映射器由两部分组成:类型识别器和类型处理器。类型识别器判断出源谓词值的数据类型,然后将其转发到相应的类型处理器。类型处理器使用搜索方法映射特定类型的谓词,只要表单上的谓词具有相同的数据类型,就可以将用户查询转换到新增加的数据源接口上。基于数据类型的谓词映射提供了一个“平台”,它可以比较不同映射之间的语义关系,同时可以使用可扩展的搜索机制来映射谓词。替代硬编码书写映射规则的方式,可以编码映射知识作为每个谓词的评价器。这个评价器在一个特定数据类型的“平台”上实例化谓词查询的语义,这样谓词映射的语义可以通过谓词实例化来比较。寻找与源查询最相近的目标表单谓词映射成为一个搜索问题:在目标表单所映射的谓词搜索空间中搜索最小覆盖原查询的映射集。

当用户查询中谓词所指定的特定值与目标表单中提供的谓词值不匹配时,例如,用户查询中指定的查询是汽车价格不高于 60 000 元人民币,而目标表单段中含有范围谓词的时候。对于用户来说可以找到满足用户查询的最小范围,如可以选择目标表单谓词值为 100000,表示在 50000 到 100000 范围之内。然而对机器来说并非直接可以选定值,如图 6 和图 7 展示了两种普通的含有范围谓词的表单结构。

1. 选择购物车预算

2. 选择性能比重

	0	20	40	60	80	100
安全性	[Progress bar]					
动力性	[Progress bar]					
通过性	[Progress bar]					
经济性	[Progress bar]					
舒适性	[Progress bar]					
品牌	[Progress bar]					

图6 范围表单 I

汽车搜索

全部车型

价格搜索

最新报价

图7 范围表单 II

经过谓词识别,可以判断图6中的谓词“选择购车预算”和图7中的谓词“最新报价”都是与谓词词典中“价格”谓词匹配的。图6中的价格谓词是范围类型谓词的一种结构。在此表单中,价格谓词是由表单中两个元素共同构成的,即车价的上限值和下限值。在此结构中,价格谓词的上限值和下限值都被精确指定,但是它是由两个相互独立的元素构成。用户可以从上限元素和下限元素中分别选择一个值来构成一个价格范围。而图7价格谓词是范围类型谓词的另一种结构。在此表单中,提供了价格谓词成对的上下界限值,它使用一个独立的元素来构造。还有一种范围类型的表单谓词,它明确提供了一个边界值,而隐含了另一个边界值,例如谓词“行驶里程”,它就隐含表示了汽车行驶里程的最大值。当分析这类谓词时,系统首先通过分析把所提供的边界值委派给目标表单上一个上界值或下界值,然后利用系统提供的默认的上界值或下界值给另一个边界值。

2.2.3 产生映射查询

系统一旦确定了目标表单上相应谓词的值,就可以构造目标表单查询了。方法是从每个表单谓词中选择一组[属性名/值]对,把所有选择的[属性名/值]对连接在一起,然后将其追加到代表页面元信息的基本URL上,就构成了要提交的URL。产生目标查询的关键代码如下所示:

```
/* 建立查询请求串 */
```

```
URL BaseUrl = getParameterAttribute("action");
StringBuffer RequestPar = new StringBuffer("");
RequestPar.append(BaseUrl.toString());
RequestPar.append("?");
for(int i=0;i<VSize;i++){
    RequestPar.append(parametersNames.elementAt(i));
    RequestPar.append("=");
    RequestPar.append(parametersValues.elementAt(i));
    RequestPar.append("&");
}
RequestPar.deleteCharAt(RequestPar.length()-1);
Return RequestPar;
```

然而目标表单提供的谓词组合并非和用户查询中谓词组合完全匹配,为此设计了查询翻译器来获取最小映射查询集。

谓词翻译器以用户查询和目标表单对象为输入,输出最接近用户原始查询在目标表单上映射的查询。谓词翻译器的核心是类型处理器,每个类型处理器实现搜索驱动的查询映射机制。与一般的搜索算法相同,类型处理器需要三个关键部件:搜索空间,相近性估计和搜索策略。

(1)对任何搜索过程定义搜索空间是非常重要的,它隐含着搜索的复杂度。给定谓词模板 P ,定义实例空间 $I(P)$ (谓词所有可能实例)。同时定义搜索空间 $\Omega(P)$ 。 $\Omega(P)$ 中的每一项都可能是用户查询的映射结果。

(2)相近性估计:给定搜索空间 $\Omega(P)$,它覆盖了源表单查询的所有可能映射,与用户原始查询最相近的映射就是源查询的最小包含 C_{min} 。寻找 C_{min} 对于某些数据类型值的谓词比较容易,如对日期,数字类型谓词很容易判断,然而对于值为文本类型的谓词需要逻辑推理。为避免复杂的逻辑推理,开发了具体化的评价方法,后续内容将详细讨论。

(3)搜索算法:为寻找用户查询映射的最小包含 C_{min} ,先给出如下定义: S 表示用户查询谓词集, P 表示目标表单谓词模板, H 表示源查询映射的所有可能查询, R 代表所有可能映射查询 H 的最小包含 C_{min} 。具体算法如下所示:

```
H = ∅, R = ∅
for ∀ t ∈ Ω(P)
    if subsume(t, s): add t to H
for ∀ t' ∈ H
    if not t' ∈ H and t' ≠ t and subsume(t, t')
        add t to R
choose an x ∈ R 其中 x 具有最少数量的谓词
```

为提高算法效率,使用了一种贪婪算法:迭代获取原始用户查询的映射,从 $I(P)$ 上寻找一个实例集满足

最大覆盖源谓词 S,直到可以覆盖整个 S,若存在多个这样的集合,则选择映射中具有最小数量谓词个数的映射作为目标表单上的查询集。

查询翻译器以映射后的谓词组合来构造原始用户查询的最小包含查询集。查询翻译问题的研究有基于范围的查询重写技术,它可以产生原查询的最小包含映射。并将尽可能多的谓词转换到目标表单上,从而最大化地使用源表单上的谓词。

一旦在目标表单上产生了查询串,就可以将查询串转发到 Web 爬虫来下载结果页面。系统收集所获得的所有结果页面,然后将其发送给查询结果合并器,最终对结果进行处理后以统一风格展示给用户。

3 结束语

Deep Web 信息集成及 Deep Web 信息检索将是下一代分布式信息集成技术和分布式数据库集成技术发展的方向之一。文中提出了一个基于 Deep Web 的信息集成系统框架结构,并在此基础上实现了一个实验原型系统,重点介绍了用户查询转换。然而,如何模型化查询接口,以及如何选择查询接口和优化用户查询都将是 Deep Web 信息集成系统设计中有待深入解决的问题。

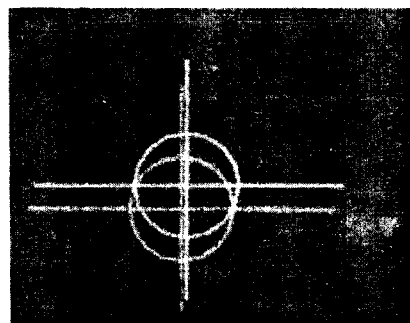
参考文献:

- [1] 黄晓冬. Invisible Web 研究综述[J]. 情报科学, 2004, 22

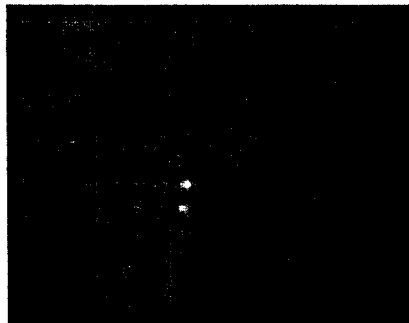
(9):1144-1147.

- [2] He Bin, Chang Kevin Chen - Chuan. Statistical Schema Matching across Web Query Interfaces[C]//In Proceedings of the 2003 ACM SIGMOD Conference. San Diego, California: [s. n.], 2003: 217-228.
- [3] He Bin, Chang Kevin Chen - Chuan, Han Jiawei. Discovering complex matchings across web query interfaces[C]//In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. New York: KDD, 2004: 148-157.
- [4] Papakonstantinou Y, Gupta A, Garcia - Molina H, et al. A query translation scheme for rapid implementation of wrappers [C]//International Conference on Deductive and Object - Oriented Databases. Singapore: Springer, 1995: 161-186.
- [5] Rajaranman A, Sagiv Y, Ullman J D. Answering queries using templates with binding patterns[C]//In Proceedings of the fourteenth ACM SIGACT - SIGMOD - SIGART symposium on Principles of database systems. California: [s. n.], 1995: 105-112.
- [6] Levy A Y, Rajaraman A, Ordille J J. Querying heterogeneous information sources using source descriptions[C]//In Proceedings of the Twenty - second International Conference on Very Large Databases. San Francisco: [s. n.], 1996: 251-262.
- [7] Chang K C - C, Garcia - Molina H. Approximate query mapping: Accounting for translation closeness[J]. The VLDB Journal, 2001, 10: 155-181.

(上接第 175 页)



(a) 色标图



(b) 相关图

图 5 色标及其相关匹配图

4 结束语

该视觉检测系统适用于目前国内大多数胶印机套印误差检测。引入机器视觉技术,对于套印误差检测这种带有高度重复性和智能性的工作而言,可以快速获取大量信息,实现智能化快速处理,具有直观性、非接触性、检测结果可靠等优点,从而大大提高印品的检

测质量和检测速度,降低了人工成本和管理成本。

参考文献:

- [1] 夏 军. 关于套印问题的研究[J]. 印刷杂志, 2002(6): 42-45.
- [2] Shapiro L G, Stockman G C. 计算机视觉[M]. 北京: 机械工业出版社, 2005: 6-7.
- [3] 韩玄武, 郑 莉. 胶印机工作原理与操作技术[M]. 北京: 化学工业出版社, 2004: 147-151.
- [4] 沈 洁, 杜宇人, 高浩军. 图像边缘检测技术研究[J]. 信息技术, 2005(12): 32-34.
- [5] 段瑞玲, 李庆祥, 李玉和. 图像边缘检测方法研究综述[J]. 光学技术, 2005, 31(3): 415-419.
- [6] 朱永松, 国澄明. 基于相关系数的相关匹配算法的研究[J]. 信号处理, 2003, 19(6): 531-534.