

一种用于深层网接口集成的模式匹配方法

陈丽君^{1,2}, 林怀忠²

(1. 浙江越秀外国语学院网络传播学院, 浙江 绍兴 312000; 2. 浙江大学计算机学院, 杭州 310027)

摘 要: 针对已有证据理论(DS)方法在深层网接口集成方面的局限性, 设计一种基于概念词与语义异构模型的深层网模式匹配方法。通过提取概念词对概念词模型进行预处理, 识别并组合成组属性, 使 $m:n$ 的复杂匹配转变为 $1:1$ 的简单匹配, 提高系统执行速度。在语义异构模型中引入属性实例, 将挖掘语义异构的同义属性问题, 转化为对属性间各特征相似值的计算、综合评测和选取问题。实验结果表明, 该方法在匹配效率和准确率上较 DS 方法有较大改进。

关键词: 深层网; 概念词; 语义异构; 模式匹配; 接口集成

Pattern Matching Method for Deep Web Interface Integration

CHEN Li-jun^{1,2}, LIN Huai-zhong²

(1. College of Network Communication, Zhejiang Yuexiu University of Foreign Languages, Shaoxing 312000, China;

2. College of Computer, Zhejiang University, Hangzhou 310027, China)

【Abstract】 By anglicizing the limitations of existing evidence theory method for Deep Web interface integration, a Deep Web pattern matching method based on concept word and semantic heterogeneity model is proposed. The method preprocesses pattern through extracting concept word, discriminates and combines group attributes to convert $m:n$ complex matching into $1:1$ simple matching for improving implement efficiency. By introducing instance into semantic heterogeneity model, the problem of mining semantic heterogeneity synonymy attributes is resolved by computing, synthetic evaluating, and selecting similarity values of attribute features. Experimental results indicate that compared with evidence theory method, the efficiency and accuracy of the method is improved obviously.

【Key words】 Deep Web; concept word; semantic heterogeneity; pattern matching; interface integration

DOI: 10.3969/j.issn.1000-3428.2012.12.012

1 概述

与表层网(Surface Web)相比, 深层网(Deep Web)资源的信息量更大、质量更优、内容更丰富、主题更专一, 而且近年来增长迅速^[1], 因此, 研究如何挖掘并利用 Deep Web 的资源具有重要的现实意义。表层化^[2]和接口集成均为访问 Deep Web 的有效途径, 由于接口集成具有精确、专业、深入等优点而深受欢迎。模式匹配是接口集成的关键问题之一。

模式匹配即获取不同模式间的语义关联关系, 可以用公式 $M(S_1, S_2)=R$ 表示。匹配函数 M 在 Deep Web 中具体表现为对两模式 S_1 、 S_2 属性相似程度的度量和计算。模式匹配的关键是要做好评判依据选择、评判方法运用 2 项工作。目前有关 Deep Web 模式匹配研究按照评判依据不同可分为 3 类: 基于实例信息^[1,3], 基于使用信息^[4]和基于接口模式信息^[5-8]。在评判方法方面, 主要涉及数据挖掘、统计归纳、信息检索、机器学习等技术。为此, 本文分析研究文献[7]中的证据理论(Dempster Shafer, DS)方法并加以改进, 提出一种基于概念词与语义异构模型的 Deep Web 模式匹配方法 CSM(Deep Web Pattern Matching Based on Concept Word and Semantic Heterogeneous Model)。

2 CSM 系统框架

2.1 相关定义

定义 1(概念词和同概念词匹配关系) 概念词是指可以单独作为属性名并能表征属性本质含义的词, 它一般有以下 2 个特征: (1)能反映该属性的本质含义; (2)具有领域代表性。如属性名“max price”中的“price”。称概念词相同且含义相

同的匹配关系为同概念词匹配关系。

定义 2(语义异构和语义异构匹配关系) 语义异构是指:

- (1)具有相同含义的属性在不同接口模式的属性名不同;
- (2)相同的属性名在不同接口模式的属性含义不同。前者如属性“邮编”, 有的用“zip code”, 也有的用“post”; 后者如属性名“type”, 有的表示车型, 也有的表示燃料类型等。称属性名不同但含义相同的匹配关系为语义异构匹配关系。

2.2 系统框架

CSM 方法借鉴启发式思想、证据理论及图论建立模型, 将匹配关系分为同概念词匹配关系和语义异构匹配关系 2 种, 并交予不同子模型处理, 以此提高效率 and 准确率。CSM 方法的系统框架见图 1, 主要由概念词匹配模型(Concept-word Matching Model, CMM)和语义异构匹配模型(Semantic-heterogeneous Matching Model, SMM)组成。在 CMM 模型中, 先由概念词提取器(Concept-word Extractor, CE)从一系列同领域初始模式集(ISS)提取各属性概念词, 建立与 ISS 一一对应的概念词模式集(CSS), 接着利用概念词匹配器(Concept-word Matcher, CM)将 CSS 中满足同概念词匹配关系的属性聚集在一起形成一个全局属性, 最后生成全局属性粗糙集

基金项目: 国家科技支撑计划基金资助项目(2009BAH43B02); 浙江省公益性技术应用研究计划基金资助项目(2010C33151); 浙江越秀外国语学院科研基金资助项目(B11006)

作者简介: 陈丽君(1979—), 女, 讲师、硕士, 主研方向: 深层网, 数据库技术; 林怀忠, 副教授、博士

收稿日期: 2011-10-12 **E-mail:** 99637841@qq.com

(Global-attribute Rough Set, GRS)并输出。在SMM模型中, 首先由语义匹配器(Semantic Matcher, SM)、词形匹配器(Morphology Matcher, MM)、实例匹配器(Example Matcher, EM)从语义、词形、实例3个方面计算GRS中各全局属性间的相似值, 然后借鉴证据理论不确定推理方法对相似值进行综合评价, 建立全局属性匹配对集(Global-attribute Matched-pair Set, GMS), 接着借助图论以及模式属性间的负相关性, 用匹配对筛选器(Matched-pair Filter, MF)筛选出满足语义异构匹配关系的属性并合并, 最后输出求精后的全局属性集(Global-attribute Set, GS)。

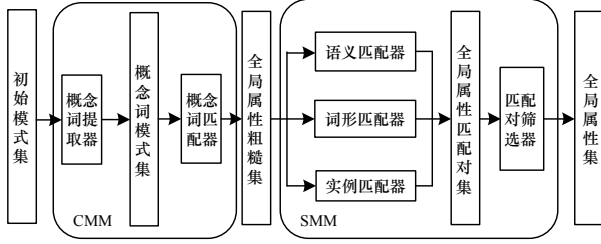


图1 CSM方法的系统框架

3 概念词匹配模型

已有研究表明, 一个领域的基本属性会以相近的表现形式反复出现于各接口模式, 而这种相近性主要表现为: 属性名包含相同的表征属性本质含义的词, 即概念词^[6]。CMM模型的主要任务是在输入的初始模式集ISS中发现同概念词匹配关系。设 $ISS=\{S_1, S_2, \dots, S_n\}$, 其中, $S_i=\{a_{i1}, a_{i2}, \dots, a_{im}\}$ 代表一个接口模式, a_{ij} 为模式 S_i 的第 j 个属性; $GRS=\{G_1', G_2', \dots, G_u'\}$, 其中, G_k' 代表初步的全局属性。CMM模型的基本思想为: 对ISS中的每一个 S_i , 分析并获取其每个属性 a_{ij} 的概念词, 然后将符合同概念词匹配关系的属性聚集形成全局属性粗糙集GRS并输出。

3.1 概念词提取器

概念词提取器的功能是从给定的属性名中提取概念词, 过滤噪声以消除其他词带来的干扰。在具体处理时, 可将属性名看作有序的词序列, 其处理过程如下:

- (1)扫描初始模式集ISS, 获得该领域的属性集 $A=\{a_i\}$, 将 a_i 按所包含词数升序排序, 集合 $CenterwordSet=\Phi$ 。
- (2)遍历属性集A, 若属性名 a_i 的词数 $|a_i|=1$, 或 a_i 的任何非空真子集 a_i' 都不是A的元素, 则将 a_i 放入 $CenterwordSet$ 。
- (3)从ISS读取模式属性 a_{ij} , 设置临时容器 $TempC=\Phi$ 。
- (4)按词数从小到大依次取 a_{ij} 的非空子集 a_{ij}' , 若 $a_{ij}' \in CenterwordSet$, 则将 a_{ij}' 放入 $TempC$, 并将 a_{ij}' 从 a_{ij} 删除, 同时从 a_{ij} 删去在词数上已达不到要求的孤立词(或词组)。如此重复直到 a_{ij} 为空。最后将 $TempC$ 中的元素按原 a_{ij} 的顺序整理即可获得与属性 a_{ij} 对应的概念词 C_{ij} 。

- (5)重复第(3)步和第(4)步直到遍历完ISS所有属性。若在同一模式中遇概念词相同且类型一致的属性, 则将其组合共用一个概念词。最终建立与ISS一一对应的概念词模式集 $CSS=\{S_1', S_2', \dots, S_n'\}$, 其中, S_i' 的第 j 个属性为概念词 C_{ij} 。

3.2 概念词匹配器

概念词匹配器的目标就是要发现同概念词匹配关系, 并将它们聚到一起形成一个初步的全局属性。概念词匹配器的具体处理过程包括:

- (1)设置 $GRS=\Phi$ 。
- (2)概念词匹配表达式 $CMExp=Cword \times Ctype \times Cexample$,

仅当两属性具有相同概念词($Cword=1$), 且类型兼容($Ctype=1$), 同时实例交集不为空或其中任一属性的实例可为任意值($Cexample=1$)时, $CMExp=1$, 其他情况 $CMExp=0$ 。

(3)从CSS中读取 C_{ij} , 若在GRS中存在使 $CMExp=1$ 的属性 G_k' , 则将 C_{ij} 并入 G_k' , 同时修改 G_k' 的实例等信息; 若不存在, 则向GRS添加 C_{ij} (此时 C_{ij} 就成为 G_k' , C_{ij} 的实例就是 G_k' 的实例)。如此重复直到读完CSS所有模式的所有概念词, 最终输出全局属性粗糙集GRS。

4 语义异构匹配模型

CMM模型主要考虑各模式属性的同概念词匹配关系, 实际上, 在不同模式之间还存在异名同义与同名异义的语义异构现象。SMM模型的主要任务是在全局属性粗糙集GRS中进一步挖掘语义异构匹配关系。SMM的基本思想为: 对GRS中的每个属性 G_i' , 计算它与其他属性 G_j' 的相似值 V_{ij}' ($i \neq j$), 将 V_{ij}' 降序排序, 从中选取符合语义异构匹配关系的属性合并, 最后输出全局属性集 $GS=\{G_1, G_2, \dots\}$, 其中, G_i 为求精后的全局属性。

4.1 各种匹配器及其组合

查询接口模式通常包含多种信息, 每种信息从一个侧面描述查询接口模式特征。本文选用语义、词形和实例3种信息作为模式匹配的评判依据, 对应的3种匹配器为:

(1)语义匹配器

语义匹配器通过“理解”属性名的含义来评判属性相似性, 并借助WordNet进行度量。两属性名 a_1, a_2 的语义相似值计算公式 $Sim_{wn}(a_1, a_2)$ 可参见文献[7]。

(2)词形匹配器

词形匹配器由属性名的“外形”来判断相似性, 它采用编辑距离(也称Levenshtein距离)进行测量。两属性名 a_1, a_2 的词形相似值计算公式 $Sim_{ed}(a_1, a_2)$ 可参见文献[7]。

(3)实例匹配器

属性实例是对属性语义的深入描述, 使用实例进行模式匹配有助于增强匹配精度。实例匹配器通过计算两相同类型属性实例交集大小来度量属性相似值。两同类型属性 a_1, a_2 的实例相似值计算公式 $Sim_{dt}(a_1, a_2)$ 定义为:

$$Sim_{dt}(a_1, a_2) = \begin{cases} |a_1 \cap a_2| / |a_1 \cup a_2| & a_1, a_2 \text{ 为数值型} \\ |a_1 \cap a_2| / \min(|a_1|, |a_2|) & a_1, a_2 \text{ 为字符型} \\ 1 & \text{其他} \end{cases}$$

其中, 若 a 为数值型则 $|a|$ 表示 a 的范围大小; 若 a 为字符型则 $|a|$ 表示 a 的实例数。

仅依赖某一种信息评判2个模式是否匹配, 会使匹配结果带有片面性和不准确性, 因此, 采用组合不同匹配器的输出值进行综合计算, 其公式定义如下:

$$Sim(a_1, a_2) = \frac{(Sim_{ed}(a_1, a_2) + Sim_{wn}(a_1, a_2)) \times Sim_{dt}(a_1, a_2)}{2}$$

其中, $Sim(a_1, a_2)$ 为两属性 a_1 和 a_2 的相似值, $Sim(a_1, a_2)$ 值越大, 说明 a_1 和 a_2 同义的可能性越大。最后, 借鉴证据理论不确定推理方法^[9]并结合式 $Sim(a_1, a_2)$, 对2个模式的相似值进行综合评测, 可生成如下所示的全局属性匹配对集GMS:

$$GMS(S, T) = \begin{bmatrix} Sim(s_1, t_1) & Sim(s_1, t_2) & \dots & Sim(s_1, t_m) & Sim(s_1, null) \\ Sim(s_2, t_1) & Sim(s_2, t_2) & \dots & Sim(s_2, t_m) & Sim(s_2, null) \\ \vdots & \vdots & & \vdots & \vdots \\ Sim(s_n, t_1) & Sim(s_n, t_2) & \dots & Sim(s_n, t_m) & Sim(s_n, null) \end{bmatrix}$$

其中, S 和 T 均为模式; s_i 和 t_j 分别是 S 和 T 的属性; 空值null是一个特殊的属性; $Sim(s_i, null) = [(1 - Sim(s_i, t_j))]$ 是对属性

s_i 与其他属性 t_j 不相似程度的度量。

4.2 匹配筛选器

匹配筛选器的作用如下：

(1)负相关属性处理。设 $S_{G_i'}$ 是全局属性粗糙集 GRS 中属性 G_i' 的所有来源模式集, $S_{G_j'}$ 是属性 G_j' 的所有来源模式集, 若 $S_{G_i'} \cap S_{G_j'} \neq \emptyset$, 则根据文献[8]的定义, G_i' 和 G_j' 为负相关属性。例如对于属性 zip 有 $S_{zip}=\{01, 02, 03\}$, 属性 body style 有 $S_{body\ style}=\{01, 05\}$, 由于 $S_{zip} \cap S_{body\ style}=\{01, 02, 03\} \cap \{01, 05\}=\{01\} \neq \emptyset$, 因此 zip 与 body style 为负相关属性。本文将负相关属性的相似值直接定义为 0, 即有 $Sim(G_i', G_j')=0$ 。

(2)同义属性的确立。本文借用图论思想进行确立: 首先将每个属性 G_i' 看作图 G 中一个点 v_i , 然后从大到小搜索全局属性匹配对集 GMS 中各属性间的相似值, 若存在 G_i' 和 G_j' 同时满足: $Sim(G_i', G_j') > Sim(G_i', null)$ 且 $Sim(G_i', G_j') > Sim(G_j', null)$, 则认为 G_i' 和 G_j' 是同义属性, 并在 v_i 和 v_j 间建立一条边 e_{ij} 。如此搜索直到搜索不到满足条件的属性为止。最后, 在全局属性粗糙集 GRS 中合并在图 G 中存在边的属性, 可以得到求精后的全局属性集 GS 。至此整个匹配过程结束。

5 实验结果与分析

实验数据来自 UIUC^[10] 的 Automobile, Book, Hotel 3 个领域, 经数据清洗、规范化处理后保存为 XML 格式。采用信息检索的评测指标查准率、查全率和 F-measure。表 1 是输出部分同义属性及相似值。图 2 是对 CSM 方法在 Automobile、Book 和 Job 3 个领域的性能测试结果。实验结果表明, CSM 方法在 Automobile 和 Book 领域的匹配效果相对要好些, 主要原因是 Job 领域数值型属性的语义异构比较突出, 而数值型实例数据的区分度小, 对模式匹配的贡献不明显。

表 1 同义属性列表(部分)

序号	同义属性	相似值
1	zip, post code	0.581
2	transmission, transmission type	0.701
3	body style, type	0.524

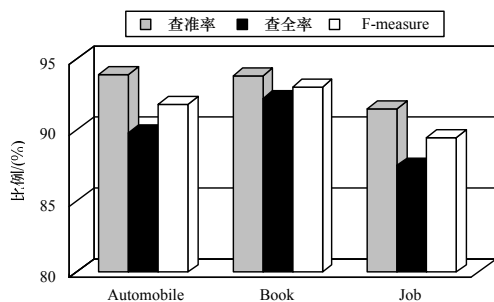


图 2 CSM 方法的查准率、查全率和 F-measure

图 3 和图 4 分别是 CSM 方法与 DS 方法在性能和效率上的比较。

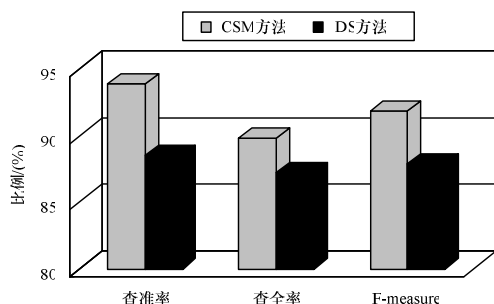


图 3 CSM 方法与 DS 方法的性能比较

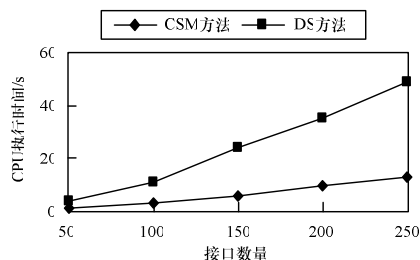


图 4 CSM 方法与 DS 方法的效率比较

实验结果表明, CSM 方法引入属性实例作为评判依据之一, 能有效弥补 DS 方法难以识别专业名词的缺陷、减少因过度依赖属性名导致的匹配丢失现象, 改善了匹配性能。在匹配效率上, CSM 方法将匹配过程分为概念词、语义异构两阶段, 并根据阶段和任务有选择地使用匹配器, 从而降低了系统时间复杂度, 提高了匹配效率, 且随着模式规模加大, 这种提升更加明显。

6 结束语

本文提出的 Deep Web 模式匹配方法, 通过引入属性实例、分阶段选用匹配器, 在效率和性能方面都比 DS 方法有明显提高。下一步将研究 SMM 模型, 优化模式匹配效果, 同时考虑中文 Deep Web 站点的模式匹配。

参考文献

- [1] Wang Jiying, Wen Jirong, Ma Weiying, et al. Instance-based Schema Matching for Web Databases by Domain-specific Query Probing[C]//Proc. of the 30th International Conference on Very Large Data Base. Toronto, Canada: [s. n.], 2004.
- [2] Madhavan J, Ko D, Kot L. Google's Deep Web Crawl[C]//Proc. of the 24th International Conference on Very Large Data Base. Auckland, New Zealand: [s. n.], 2008.
- [3] Wu Wensheng, Doan A, Yu C. WebIQ: Learning from the Web to Match Deep Web Query Interfaces[C]//Proc. of the 22nd International Conference on Data Engineering. Atlanta, USA: [s. n.], 2006.
- [4] Elmeleegy H, Ouzani M, Elmagarmid A. Usage-based Schema Matching[C]//Proc. of the 24th International Conference on Data Engineering. Cancun, Mexican: [s. n.], 2008.
- [5] Dong Yongquan, Li Qingzhong, Ding Yanhui, et al. A Query Interface Matching Approach Based on Extended Evidence Theory for Deep Web[J]. Journal of Computer Science and Technology, 2010, 25(3): 537-547.
- [6] He Bin, Chang K C C. Statistical Schema Matching Across Web Query Interfaces[C]//Proc. of the ACM SIGMOD International Conference on Management of Data. San Diego, USA: ACM Press, 2003: 217-228.
- [7] Hong Jun, He Zhongtian, Bell D. An Evidential Approach to Query Interface Matching on the Deep Web[C]//Proc. of the 24th International Conference on Very Large Data Base. Auckland, New Zealand: [s. n.], 2008.
- [8] He Bin, Chang K C C, Han J. Discovering Complex Matching Across Web Query Interfaces: A Correlation Mining Approach[C]//Proc. of the 10th International Conference on Knowledge Discovery and Data Mining. Washington D. C., USA: [s. n.], 2004.
- [9] 王宏生, 孟国艳. 人工智能及其应用[M]. 北京: 国防工业出版社, 2009.
- [10] The UIUC Web Integration Repository[EB/OL]. (2010-08-17). <http://metaquerier.cs.uiuc.edu/repository>.

编辑 陆燕菲

