中文深层网络的模式匹配和接口集成①

张晶星

(武汉大学 计算机学院, 武汉 430072)

摘 要: 目前国内外在深层网络方面的研究几乎都围绕英文环境进行, 还没有针对中文深层网络的研究. 提出了对中文深层网络进行模式匹配和接口集成的方法. 该方法首先创建一个用来存储同义词、超义词和子义词的字典, 然后使用基于规则的分词算法将从接口中抽取的属性分成词. 对于每一个属性, 从定义的字典中找到其对应的所有同义词、超义词和子义词, 生成一条相应的记录并存储到列表中, 再从每条记录中选取出现次数最多的属性作为联合接口的属性.

关键词: 中文深层网络; 分词算法; 词典; 模式匹配; 接口集成

Schema Matching and Interface Integration for Chinese Deep Web

ZHANG Jing-Xing

(School of Computer, Wuhan University, Wuhan 430072, China)

Abstract: Many researches about deep web focus on the deep web with English language, ignoring that with Chinese. In this paper, we present our work in schema matching and interface integration for Chinese deep web. We create a dictionary, which stores synonyms, hypernyms and hyponyms, at the very beginning. After interface extracting, we use Principle-based Segmentation algorithm to segment each attribute into words. Then, for each attribute, we look up the pre-created dictionary to find all its synonyms, hypernyms and hyponyms, form a record and store them in a list. Furthermore, we keep a counter for each attribute in the list to record times it appearing in the local interfaces. At last, we choose from each record a synonym with the largest count number as the attribute of union interface.

Key words: Chinese deep web; segmentation algorithm; lexical dictionary; schema matching; interface integration

深层网络也叫不可见网络, 是指存储在可被检索的网络数据库中的数据, 这些内容 “隐藏”在接口后面, 只能通过查询获取. 在深层网络中, 大量网络数据库通过它们的查询接口被动态访问. 调查和研究[1]表明深层网络有如下一些特点: (1)深层网络必须通过接口而非静态 URL 链接访问; (2)深层网络的规模远远大于表层网络; (3)深层网络中的大部分数据为结构化数据; (4)深层网络中数据的目录覆盖率非常低; (5)深层网络并不完全对检索 “隐藏”.

目前国内外在深层网络方面的研究几乎都围绕英文环境进行, 还没有针对中文深层网络的研究. 中文在很多方面不同于西方的语言. 中文是非字母形式的; 大量的汉字都是表意符号; 几乎每一个汉字都是一个有意义的词素. 在中文中, “词”由一个或多个字组成, 词的意思有时与组成它的字义在一定程度上相关(合成, 如 “大”和 “学”组成词 “大学”), 有时候却完全不同(表意词, 如 “和”和 “尚”组成 “和尚”). 中文中词的结构是非常灵活的: 一个长词能被任意地简化, 如 “武汉大学”可说成 “武大”; 新词不断地产生. 产生这类现象是因为每个词都有自己的意义, 它们可以独立地充当一些语言角色. 中文句子不像英文句子那样有空格作为词的分界, 中文并没有很好地定义词这一语言单位, 将一个中文句子进行分解得到的词串不一定唯一(即一个句子可以理解成多种意思), 从来不存在

① 基金项目:国家自然科学基金(60970018)

收稿时间:2012-04-23;收到修改稿时间:2012-05-29

一个被普遍接受的中文词典. 中文的这些特点使对中文的处理具有自身的特殊性和挑战性.

中国互联网络信息中心显示中文在线数据库数量在 2010 年达到 60 余万,预计 2015 年其数量将达到 90 万以上;中国人口众多,潜在信息需求很大,中文深网的研究特别是为用户提供对多个不同的中文深网提供统一的查询接口有着重要的意义. 例如, 如果能为用户提供一个能同时查询多个书店数据库的统一查询接口,将能快速准确地返回所需书目信息,用户便不用分别查看当当网和卓越网等书店来确定最合适的交易. 实现该目标可以分为两个步骤:首先将同领域的查询表单进行属性匹配,然后选取最具有表现力最容易为用户接受的属性组成统一接口. 本文研究中文深层网络的模式匹配和接口集成. 本文组织如下: 第 1 节给出了中文深层网络体系结构框架; 第 2 节提出了中文深层网络的模式匹配和接口集成系统实现的一些关键技术和算法; 第 3 节总结全文.

1 中文深层网络的体系结构框架

不管深层网络采用的是何种语言, 对它们的处理过程都是类似的. 以对英文深层网络的处理过程为例: 首先必须使用深层网络爬虫找到深层网络资源, 并且确定它们的接口(数据库的入口[2]. 然后, 对接口进行抽取获得各接口的属性[3], 并且实现各接口相关属性的匹配[4,5]. 接着, 对所有的本地接口集成为一个联合接口[6,7,8], 同时建立本地接口到联合接口的映射关系. 用户通过联合接口提交查询请求, 系统自动将该请求映射到每一个本地接口, 本地接口获得请求后查询其对应的底层数据库得到相应结果, 系统对各本地接口反馈回来的信息进行索引, 并通过联合接口展现给用户.

任意两种语言都存在差异. 与英文相比, 中文有自身的独特之处, 这一点已经在 1.2 部分介绍过了. 中文的所有这些特点中, 对访问过程影响最大的莫过于“词”. 中文没有空格作为词的分界, 因此中文中的“词”也不如英文中 “词”那么好区分, 必须引入一定的技术对中文句子进行分词. 本文采用了基于规则的分词算法和一个预先定义的字典来解决这一问题. 图 1 描述了中文深层网络的体系结构框架. 从图中可看出, 该框架有两个核心部分: 一个是句子分析器, 另一个是联合属性和映射产生器. 前者内部集成了基于规则的分词算法和一个预先定义的字典, 该字典用来生成与每个词相关的信息, 如词类、同义词、超义词和子义词, 分析器还能根据属性对这些词进行合并操作(如果必要的话). 后者接收属性的相关信息形成一个列表(一个属性对应列表中的一条记录). 为了优化最终的结果, 用户可对形成的列表进行干预. 对列表中的每一个属性, 产生器为其维护一个计数器记录它在本地接口出现的次数.

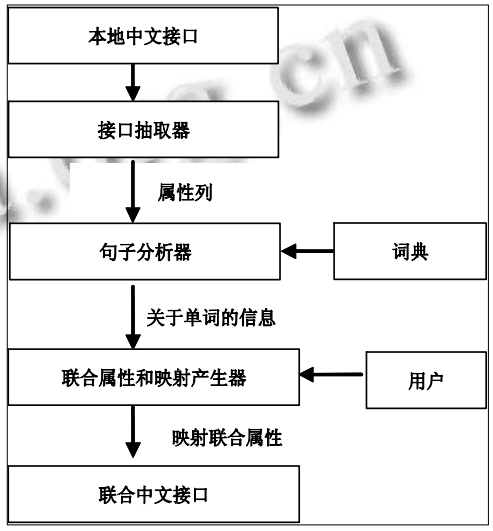

图 1 中文深层网络体系结构框架

2 中文深层网络的模式匹配和接口集成

2.1 形式化描述

定义 1. 表单用下面一个五元组表示:

$F = \{I, R, M, N, V\}$

其中:

I — 表单标识符, 用来唯一表示表单;

R — 表示提交表单对应的 URL 请求;

M — 提交表单数据的方式,POST 或 GET;

N — 正整数, 表示提交表单中属性的个数;

V — 属性名/值对集合, 用来表示表单中属性的详细信息,其个数为 N.

定义 2. 表单 A 与表单 B 集成之后,生成表单 C, 可以表示为:

$\{I^A, R^A, M^A, N^A, V^A\}$

$INT\{I^B, R^B, M^B, N^B, V^B\}$

$= \{I^C, R^C, M^C, N^C, V^C\}$.

简写为:A INT B=C(其中,符号 “INT”表示集成).

定义 3. 对于给定的一个阈值 ξ ,如果 Sim(Attribute A, Attribute B)$\geq \xi$ (Sim 为相似度方法,描述属性

之间的匹配程度),则认为属性 A 与 B 相匹配;否则认为属性 A 与 B 不相匹配.

本文的主要目标是通过对模式匹配和接口集成的处理使访问中文深层网络变得可行, 如图 1 所示. 图中的框架包含了一些组件, 其中接口抽取器在引文[2]中均有介绍. 本文采用了它们的基本思想: 使用接口表达式对每个接口进行抽取, 识别出接口中的每一个属性, 尤其是在后续步骤中要用到的各属性的标签.

2.2 词典

词典是分词、属性匹配和接口集成的基础, 因此, 它必须在最开始就建好. 本文在词典的建立过程中引入了一个非常重要的概念: 实体. 处理方法是将任何类型的词(名词、动词、形容词等)都看作是实体. 同时本文的辞典参考独立于域的在线英语字典 WordNet[9]的风格, 在字典中引入同义词集合,并使用语义分层组织汉字(名词、动词、形容词和副词). 具体地说, 在同义词集合中, 信息围绕逻辑分组进行组织. 每一个同义词集合由一系列的同义词和描述集合本身与其它同义词集合关系的指针组成. 一个词有可能不止在一个同义词集合中出现; 同一个集合中的所有词在一定程度上可以互换; 从词到同义词集合的映射按其出现的频率排序. 名词和动词基于同义词集合间的超义词/子义词关系被分层组织 ; 形容词被组织成包含同义词集主语和同义词集附属部分的簇. 每一个簇围绕反义词对组织(有时围绕三个互为反义的词组织). 反义词对(或三个互为反义的词)在簇的同义词集主语中说明; 许多同义词集主语拥有一个或多个同义词集附属部分.

2.3 句子分析器与接口属性匹配

词典建立后, 句子分析器就可以利用它来分析接口的属性, 从而产生词的分割. 图 2 描述了该组件的内部流程图. 该处理过程是一个循环的过程. 在每一次循环的最开始, 句子分析器读取来自接口抽取器的一个属性, 获得该属性的标签. 然后, 对标签进行规则化处理, 如去掉无实际意义的终止词, 如 “的”、“地”; 删除括号以及左右括号之间的内容等等. 接着, 使用基于规则的分词算法对经过正规化处理之后的标签进行分词, 产生词的分割. 对于属性的每一个词, 查询词典找到它所有的同义词、超义词、子义词等, 针对每一个属性组合这些词. 同时, 为每一个属性维护一个记录该属性在本地接口出现次数的计数器. 将属性的所有信息记录在一个列表中, 一个属性对应列表中的一条记录. 为了优化最终结果, 用户可以在这一步进行干预(编辑、增加或删除一些信息). 如果当前考虑的属性是某一接口的最后一个属性, 则结束分析过程, 进入下一处理步骤; 否则, 继续循环.

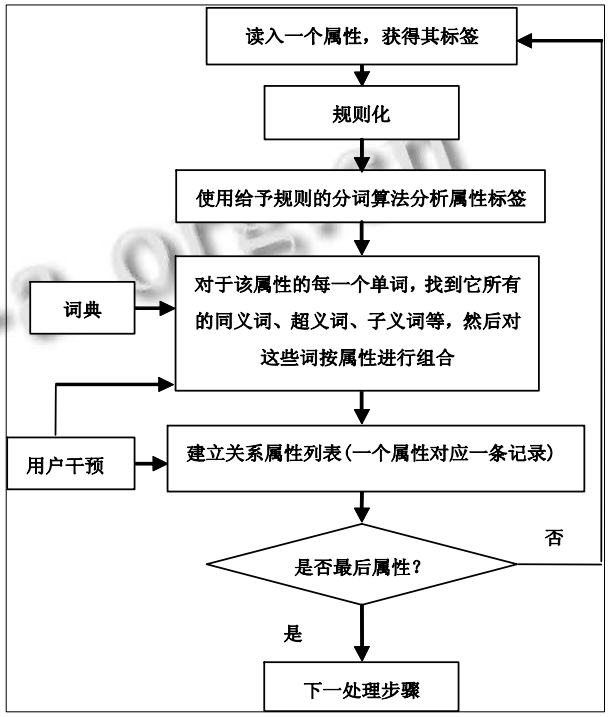

图 2

3 结语

本文首次研究了了中文深层网络的模式匹配和接口集成的问题, 在如下几个方面做出了贡献: (1)将中文深层网络纳入研究范畴. 目前国内外在深层网研究方面都针对英文深层网络进行研究. 尽管对深层网络的访问处理是相似的, 但是每种语言自身的特点还是会造成某些处理细节的差别. (2)给出了中文深层网络体系结构框架. (3)给出了中文深层网络模式匹配和接口集成的形式化描述. (4)给出了利用中文分词和同义词、超义词和子义词字典的中文深层网络模式匹配和接口集成方法.

参考文献

1 Chang KCC, He B, Li C, Zhang Z. Structured databases on the web: Observations and implications. SIGMOD Record, 2004,33(3):61−70.

(下转第 185 页)

度，增大图像反差. 图像融合结合两种边缘检测方法的特点，得到了很好的边缘图像. 实验结果表明，与著名的 Canny 算子和基于小波变换的边缘检测方法中任何一个方法相比，本文提出的边缘检测方法抗噪声能力强，并且有效地解决了抑制噪声和保留精细边缘之间的矛盾，实现了对图像进行更多有效细节上的边缘检测，达到了很好的边缘检测效果.

(a)原始 Lena 图像

(b) Canny 算子

(c) 基于小波变换的方法

(d) 本文方法

图 2 各种图像边缘检测方法实验结果

参考文献

1 郭显久.一种新的基于小波变换的边缘检测算法.大连水产学院学报,2005,20(2):158–162.

2 王俊卿,黄莎白,史泽林,于海斌.基于小波变换的图像边缘检测.系统工程与电子技术,2004,26(7):887–888.

3 邹福辉,李忠科.图像边缘检测算法的对比度分析.计算机应用,2008,28(1):215–219.

4 陈武凡.小波分析及其在图像处理中的应用.第 2 版.北京:科学出版社,2003.175–185.

5 Meer P, Ceor Cescu B. Edge Detection with Embedded Confidence. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2001,23(12):1351–1365.

6 万力,易昂,傅明.一种基于 canny 算法的边缘提取改善方法.计算技术与自动化,2003,22(1):24–26.

7 赵育良,赵友庚,李开端,李英杰.基于小波变换的复杂航空图像的边缘提取.光电工程,2002,29(4):57–60.

8 张红岩,张登攀.图像边缘检测二维小波算法研究与实现.中国工程科学,2003,5(4):61–64.

9 尹立敏,刘艳滢,顾蕊.一种可控的直方图均衡算法.微计算机信息,2005,3(21):147–149.

10 Abdukirim T, Hussain M, Niijima K, Takano S. The dyadic lifting schemes and the de-noising of digital image. International Journal of Wavelets, Multi-resolution and Information Processing, 2008,6(2):331–351.

(上接第 205 页)

2 Chang KCC, He B, Zhang Z. Toward Large Scale Integration: Building a MetaQuerier over Databases on the Web. Proc. of the Second Conference on Innovative Data Systems Research, 2005:44–55. http://www.sigmod.org/dblp/db/conf/cidr/cidr2005.html

3 He B, Chang KCC. Statistical schema matching across web query interfaces. Proc. of the 2003 ACM SIGMOD International Conference on Management of Data, 2003:217–228.http://www.sigmod.org/dblp/db/conf/sigmod/sigmod2003.html

4 He B, Chang KCC, Han J. Discovering Complex Matching across Web Query Interfaces: A correlation mining approach. Proc. of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2004: 148–157.

5 He H, Meng W, Yu CT, Wu Z. Automatic integration of Web search interfaces with WISE-Integrator. VLDB J. 13(3), 2004: 256–273.

6 Wu W, Yu CT, Doan A, Meng W. An interactive clustering-based approach to integrating source query interfaces on the deep web. Proc. of the ACM SIGMOD International Conference on Management of Data, 2004:95–106. http://www.sigmod.org/dblp/db/conf/ sigmod/sigmod2004.html.

7 He H, Meng W, Yu CT, Wu Z. Wise-integrator: An automatic integrator of web search interfaces for e-commerce. Proc. of the 31st International Conference on Very Large Data Bases, 2003:357–368. http://www.vldb2003.org/

8 http://www.cogsci.princeton.edu/~wn