第 38 卷 第 4 期
Vol.38 No.4

计 算 机 工 程
Computer Engineering

2012 年 2 月
February 2012

• 软件技术与数据库 •　　文章编号：1000—3428(2012)04—0052—03　　文献标识码：A　　中图分类号：TP311.5

基于本体的 Deep Web 数据源发现方法

李道申，刘 勇

(河南科技大学电子信息工程学院，河南 洛阳 471003)

摘 要：提出一种基于本体的 Deep Web 数据源发现方法，采用网页分类、表单内容分类、表单结构分类方式，确定符合某领域的 Deep Web 查询接口。在网页分类和表单内容分类中引入本体的半自动构建和自动扩展模块，在表单结构分类中添加启发式规则。实验结果证明，该方法能有效提高 Deep Web 数据源的查全率和查准率。

关键词：深网；本体；数据源；半自动构建；分类模型

Deep Web Data Sources Discovery Method Based on Ontology

LI Dao-shen, LIU Yong

(College of Electronic Information Engineering, Henan University of Science and Technology, Luoyang 471003, China)

【Abstract】This paper presents a Deep Web data sources discovery method based on ontology. It uses webpage classification, form structure classification and form content classification to find Deep Web querying interface in some fields. It proposes that semi-automatic construction and automatic extension of ontology are added to the webpage and form content classification, and heuristic rules are enriched in the form structure classification. Experimental results show that this method can improve the precision and recall of Deep Web database discovery effectively.

【Key words】Deep Web; ontology; data sources; semi-automatic construction; classification model

DOI: 10.3969/j.issn.1000-3428.2012.04.017

1 概述

随着万维网的发展，人们对网络信息的查询效率要求不断提高，传统的搜索引擎已不能满足人们的要求[1]。它只能找到带有超链接的相关网页，但隐藏在表单后的动态网页拥有海量信息，并且有质量高、专业性强、发展快的特点[2]。由于当今的搜索引擎爬虫不具备自动填充表单的能力，造成了绝大多数存储在 Web 中的高质量资源无法被检索。

Web 数据库具有异构性、分布性和自治性，因此，想要全面准确地检索某领域的深网(Deep Web)数据库是很困难的。目前，人们对 Deep Web 数据接口集成进行了大量研究，通过数据库查询接口的集成，实现统一的新查询接口，可通过这样的查询接口得到 Deep Web 信息[3-4]。本文提出一种基于本体的 Deep Web 数据源发现方法，以求全面、准确地发现某领域的 Deep Web 查询接口。

2 领域本体的半自动构建

本体构建思想是：由领域专家构建初始的领域核心本体，利用 WordNet 中的概念与初始核心本体概念比对，通过相似度计算抽取 WordNet 中相关的概念及其关系，将这些概念和关系用 is_a 规则组织成完整的本体，算法步骤[5]如下：

Step1 计算 WordNet 中的词汇 $c_i(i=1,2,\cdots,N)$ 与初始本体向量 $\boldsymbol{O}(o_1,o_2,\cdots,o_l)$ 中每个概念的相似度 $sim(c_i,o_j)$。

Step2 给定阈值 $K3$，如要判定某一词汇概念 $c_i(i=1,2,\cdots,N)$ 与本体 $\boldsymbol{O}(o_1,o_2,\cdots,o_l)$ 的相关度，可以采用 SSE 算法：$SSE(c_i,\boldsymbol{O})=\sum_{i=1}^{l}(1-sim(c_i,o_j))^2$。若 $SSE(c_i,\boldsymbol{O})<K3$，则将概念 c_i 放到预扩展概念集 E 中。

Step3 创建 is_a 关系规则。通过前 2 步的处理得到预扩展概念集 E。将 E 中的每一个概念 $e_k(k=1,2,\cdots,s)$，按照与本体 $\boldsymbol{O}$ 最大相似度以及在 WordNet 中的继承关系，添加到本体 $\boldsymbol{O}(o_1,o_2,\cdots,o_l)$，具体步骤可参考文献[5]。

3 Deep Web 查询接口表单分类

领域本体构建后，可采用网页分类、表单结构分类和表单内容分类，依次舍弃无关的网页，得到 Web 页面中特定领域的表单。

3.1 网页分类

网页分类采用基于本体的主题爬行技术来获得领域页面集合。可以将爬虫的爬行过程看作一个图搜索过程，而网页和链接分别是图的节点和边，爬虫从原始节点作为爬行起点，沿着边到达另一些节点。主题爬虫以由领域本体构建的主题向量来识别一个网页是否属于某个领域。

在主题分类之前，主题向量在领域本体概念中直接获得，设为 $\boldsymbol{O}=(o_1,o_2,\cdots,o_m)$，而网页的特征向量需要分析链接和网页内容。本文采用主题爬虫的最优优先策略：该算法的思想是以词频作为特征的权重，得到网页的特征向量模型 $\boldsymbol{D}$。网页特征向量构建算法步骤如下：

(1)设定网页中位置权重。网页权重系数向量假定为 $\boldsymbol{\omega}=(\omega_1,\omega_2,\cdots,\omega_n)$，$\boldsymbol{D}$ 的第 k 个特征项在不同位置出现的频率表示为向量 $\boldsymbol{f_k}=(f_{k1},f_{k2},\cdots,f_{kn}),k=1,2,\cdots,m$。

(2)网页特征向量可以表示为：

$\boldsymbol{D}=(d_1,d_2,\cdots,d_m)$

$$d_k=\frac{\boldsymbol{f}_k\cdot\boldsymbol{\omega}}{\sum_{k=1}^{m}\boldsymbol{f}_k\cdot\boldsymbol{\omega}}=\frac{\sum_{j=1}^{n}(f_{kj}\times\omega_j)}{\sum_{k=1}^{m}\sum_{j=1}^{n}(f_{kj}\times\omega_j)},\ k=1,2,\cdots,m$$

基金项目：国家自然科学基金资助项目(70671035)

作者简介：李道申(1986－)，男，硕士研究生，主研方向：Web 数据挖掘；刘 勇，教授

收稿日期：2011-07-19　**E-mail：**lidaoshen.good@163.com

基于本体的网页分类过程如图 1 所示。

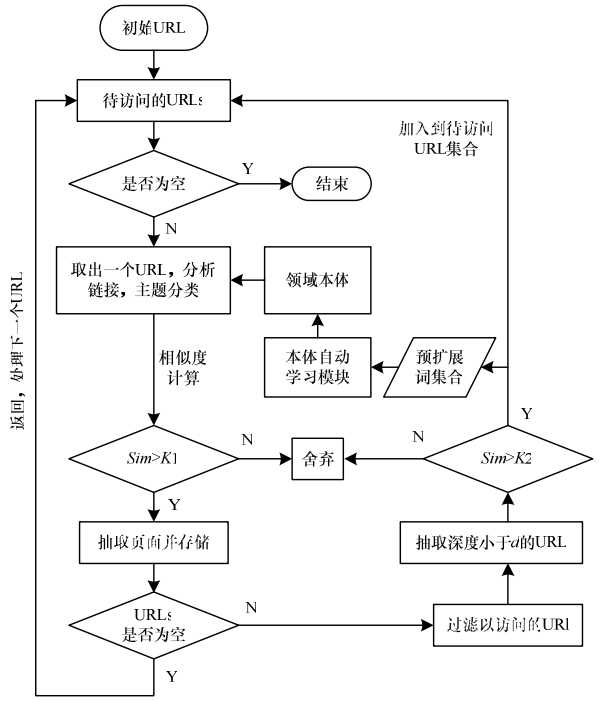

图 1　基于本体的网页分类过程

在图 1 中，主题分类模块利用领域本体计算所抓取的网页与主题相关度。如果相关，则对该网页所包含的超链接进一步抓取，本文主题分类模块的相似度计算公式如下[6]：

$$Sim(\boldsymbol{D},\boldsymbol{O})=\cos\theta=\frac{\boldsymbol{D}\cdot\boldsymbol{O}}{\|\boldsymbol{D}\|\cdot\|\boldsymbol{O}\|}=\frac{\sum_{k=1}^{m}d_k\times o_k}{\sqrt{(\sum_{k=1}^{m}d_k^{\,2})(\sum_{k=1}^{m}o_k^{\,2})}}$$

其中，$\boldsymbol{D}$ 表示页面特征向量 $\boldsymbol{D}=(d_1,d_2,\cdots,d_m)$，$d_i\ (1\leqslant i\leqslant m)$ 表示页面中第 i 个特征的权重；$\boldsymbol{O}=(o_1,o_2,\cdots,o_m)$ 表示主题特征向量，$o_i\ (1\leqslant j\leqslant m)$表示第 j 个主题词的权重；$\|\boldsymbol{a}\|$ 表示向量 $\boldsymbol{a}$ 的欧几里德范数，页面向量 $\boldsymbol{D}$ 和向量 $\boldsymbol{O}$ 的夹角 θ 越小，2 个向量的相似度越高。

而后做如下处理：

(1)如果利用上述算法得到的网页与主题的相似度大于事先给定的 $K1$，则当网页积累到一定数量时，进行页面保存并抽取网页中的 URL；如果某网页与主题的相似度小于 $K1$，舍弃该网页。

(2)对于与主题的相似度大于给定的 $K2$ ($K2>K1$)的网页，将对应的页面向量 $\boldsymbol{D}$ 中的概念加入预扩展词汇集合中，用来实现本体的自动扩展。同时抽取的网页中未访问的并且深度小于 d 的 URL，加到待访问的 URLs 中。

3.2　表单结构分类

利用网页分类模型只是得到了与某领域相关的 Deep Web 网页，然而网页中并不一定有查询表单。本文丰富了启发式规则，对这些网页表单进行过滤，得到含有查询表单的网页。网页查询表单启发式规则[7]如下：

(1)首先，判定网页中含有 Form 表单：

1)如果网页中不存在<Form>和</Form>标签，是非查询接口表单。

2)如果含有<Form>标签，其控件类型模式小于 3，那么为非查询接口表单。

3)不存在提交按钮，那么为非查询接口表单。

(2)判定表单为查询接口表单：

1)如果表单中包含 login、regist、register、registration、username、user 等词汇之一，则表单为非查询接口表单。

2)如果表单中出现 e-mail、mail、 EML、send、sender、addresser、addressee，则表单不是查询接口表单。

3)如果表单中出现 password 或某控件的类型 type 为 password，则不是查询表单。

4)如果表单各标签包含一个或多个 Search、Go、advanced search 等词汇，则为查询接口表单。

5)如果表单的参数 action 的值为 www.baidu.com 或者 www. google.com 时，则不是查询接口表单。

3.3　表单内容分类

经过前 2 步的分类，可以得到某领域内网页，并且这些网页带有查询接口，但是查询接口不一定是符合用户查询领域的查询接口，甚至相差很远。表单的内容分类模型利用查询接口的特征向量和主体特征向量的相似度来鉴定所给的查询接口是不是某领域的查询接口。表单内容分类模型如图 2 所示。

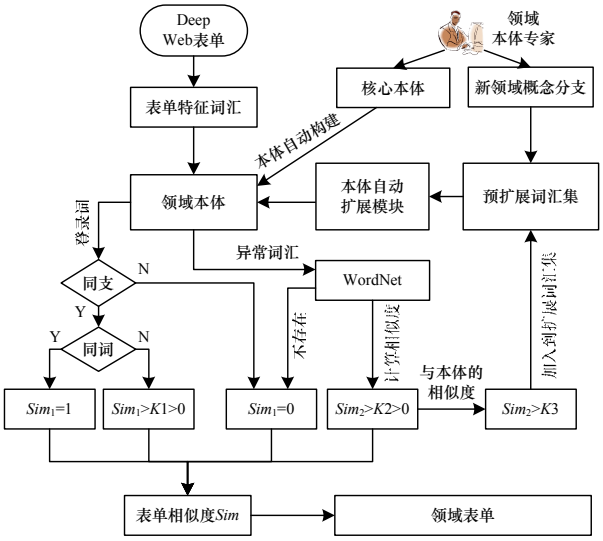

图 2　表单内容分类模型

主题特征向量 $\boldsymbol{O}=(o_1,o_2,\cdots,o_n)$ 可以通过本体概念转化而来，其中，o_i 是领域本体中具有代表性的概念。另一方面，通过对查询接口分析，利用从表单中可以得到表单标签及自由文本信息，构建查询接口特征向量 $\boldsymbol{D}=(d_1,d_2,\cdots,d_m)$，构建方法与网页特征向量的构建方法类同，这里不详细介绍。

由表单中的自由文本所构建的查询接口特征向量的各分量概念不一定相符，而且维数也不一定相同。如果概念 A、B 在领域本体中出现的词汇，本文定义为登录词汇。$Sim_1(A,B)$ 相似度计算方法为：

(1)如果 A 和 B 有 2 个或 2 个以上的公共上位词或存在继承关系，$Sim_1(A,B)=1$；

(2)除根节点以外，A 和 B 没有共同的上位节点，则 $Sim_1(A,B)=0$。

如果概念 B 是在领域本体中出现的词汇，而 A 不是，称 A 为异常词汇。这时可以借助 WordNet 工具来计算它们的相似度：

$$Sim_2(A,B)=\frac{2\times num(A,B)}{dep(A)+dep(B)}$$

其中，$num(A,B)$ 表示在 WordNet 中两者共同的上位词；

$dep(A)$ 表示 A 在 WordNet 中的深度。

因此，主题特征向量 $\boldsymbol{O}=(o_1,o_2,\cdots,o_n)$ 和查询接口特征向量 $\boldsymbol{D}=(d_1,d_2,\cdots,d_m)$ 的相似度可以分 2 个部分表示为(假设 $\boldsymbol{D}$ 有 k 个分量是登录词):

$$Sim(\boldsymbol{D},A)=\frac{\sum_{i=1}^{k}\max_{o_j\in O}(Sim_1(d_i,o_j))+\sum_{i=1}^{m-k}\max_{o_j\in O}(Sim_2(d_i,o_j))}{m}$$

在计算得到相似度之后，如果查询接口与主题的相似度大于 $K1$，就认为该查询接口是某特定领域的查询接口，否则就舍弃。

另外，在计算异常词汇与领域本体的相似度时，如果 $Sim(A,B)>K2$，说明该词汇与领域本体有一定的相似程度；给定更接近于 1 的阈值 $K3$，如果相似度大于阈值 $K3$，就将它加入到本体预扩展词汇集合中。当预扩展词汇集达到一定数量时，就通过本体扩展模块，对领域本体实现自动扩展。另一方面，在专家实时分析领域本体过程中，如果发现新的领域概念分支，也可以将概念加入到预扩展词汇集中，实现领域本体自动同步。

4 领域本体的自动扩展

本文在网页内容分类和表单内容分类中，加入了领域本体自动扩展模块，目的是自适应的实现领域本体的自动扩展。通过网页内容分类和表单内容分类已经得到预扩展词汇集 $E\{e_1,e_2,\cdots,e_s\}$，其中，$e_k(k=1,2,\cdots,s)$ 是扩展词汇集 E 中的概念。

由网页内容分类和表单内容分类过程可以看出，E 中的每一个概念 $e_k(k=1,2,\cdots,s)$ 都是在 WordNet 中存在的。因此，预扩展词汇集合的概念达到一定数量时，按照本体构建算法，只需将当前的领域本体设为 $\boldsymbol{O}(o_1,o_2,\cdots,o_l)$，对于每一个 $e_k(k=1,2,\cdots,s)$，直接转到本体构建算法 Step2 和 Step3，可以实现领域本体的自动扩展。

5 实验结果与分析

网页中不同位置词汇的权重参数设置如表 1 所示。

表 1 不同位置词汇的权重参数

位置	权重
title	4.0
h1~h3	2.5
h4~h6	2.0
tail	1.7
other	1.5

实验步骤如下:

(1)为验证该方法的可行性和有效性，本文利用 protégé 工具手工构建了几个领域的核心本体(Book、Movie、Job、Car、Hotel)，分别表示书籍、电影、工作、车、旅馆领域。

(2)确定原始网站首页，利用网络爬行技术从网站的首页开始爬取，由超链接进入下一步的爬取工作，并把爬取到的网页放入临时文件中备用，同时网页的深度增加 1，当网页深度达到 3 时就不再按照超链接进行下一网页的爬取。

(3)利用第 3 节中的网页分类，表单分类和表单内容分类对爬取的网页进行分析，去除相似度较低或者不含查询表单的网页，同时提取在与核心本体相似度超过给定阈值的词语，作为核心本体扩展词汇，加入到预扩展词汇集中。

(4)返回步骤(1)，进行下一轮的网页搜索，当本体扩展词汇达到一定数目，调用领域本体自动扩展模块进行核心本体的自动扩展。

本体自动加入扩展词汇前后的查全率及查准率比较如图 3、图 4 所示。可以看出，本文方法能够提高 Deep Web 数据源发现的查全率和查准率。

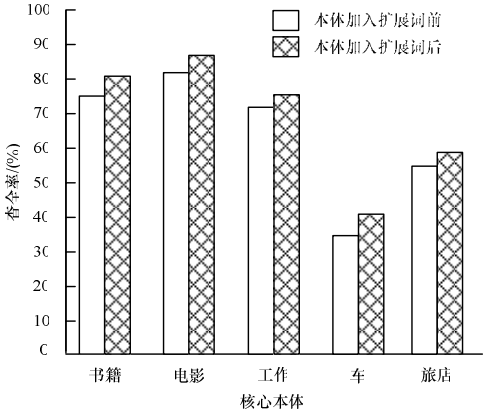

图 3 加入扩展词汇前后的查全率比较

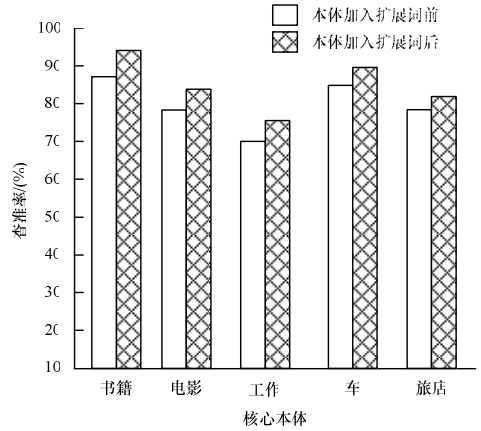

图 4 加入扩展词汇前后的查准率比较

6 结束语

Deep Web 数据源发现是 Deep Web 数据集成的关键。本文提出在网页内容分类和表单内容分类过程中加入领域本体扩展模块，健全了表单结构分类中的启发式规则。实验证明该方法是有效的。但是，表单结构分类利用的启发式规则针对的是 html 语言编写的 Web 文档，下一步将研究一种通用的 Web 文档分类方法，并建立一个较大的实验平台来验证模型的完备性和健壮性。

参考文献

[1] 马建华, 李赛红, 徐兰兰. 深层网中基于入口查询的表单填充策略[J]. 计算机工程, 2010, 36(7): 66-67.

[2] 刘 伟, 孟小峰, 孟卫一. Deep Web 数据集成研究综述[J]. 计算机学报, 2007, 30(9): 1475-1486.

[3] Yang Daowen, Liu Quan. The Discovery and Extraction of Query Interfaces Based on Deep Web[C]//Proc. of WCSE'09. [S. l.]: IEEE Computer Society, 2009.

[4] 王 辉, 刘艳威, 左万利. 使用分类器自动发现特定领域的深度网入口[J]. 软件学报, 2008, 19(2): 246-256.

[5] 周子力, 顾君忠. WordNet 的本体构建及其在安全领域应用关键技术研究[D]. 济南: 山东师范大学, 2009.

[6] 马 军, 宋 玲, 韩晓晖, 等. 基于网页上下文的 Deep Web 数据库分类[J]. 软件学报, 2008, 19(2): 267-274.

[7] 王 英, 左万利. Deep Web 数据集成技术研究[D]. 长春: 吉林大学, 2010.

编辑 顾姣健