

基于语义网的高效信息查询方法

夏美翠^{1*}, 时鸿涛^{2,3}

(1. 青岛农业大学 档案馆, 山东 青岛 266109; 2. 青岛农业大学 网络管理中心, 山东 青岛 266109;

3. 中国海洋大学 信息科学与工程学院, 山东 青岛 266100)

(* 通信作者电子邮箱 qiaonan@qau.edu.cn)

摘要: 为了提高 Web 信息检索的准确率, 提出一种基于语义网的高效信息查询方法。首先从本体库中提取目标资源与查询关键字之间的语义路径, 通过分析语义路径所包含的属性的权重和识别能力, 分别计算每个语义路径的权重; 然后, 根据资源与查询关键字之间的语义路径的权重、数量和特异性, 分别计算每个资源与各关键字之间的语义相关性, 并结合关键字的涵盖范围和识别能力综合计算每个资源与关键字集之间的语义相关性; 最后, 以该相关性为依据对所有资源进行排序和输出。实验结果表明, 与 OntoLook、tf * idf 和 TMSubtree 三种语义网查询算法相比, 基于语义网的高效信息查询方法的平均正确率分别提高了 69.0、25.0 和 21.0 个百分点; 平均召回率分别提高了 77.1、28.3 和 24.3 个百分点; 平均 *F* 测度值分别提高了 72.4、26.4 和 22.4 个百分点。实验结果表明: 该方法不仅能够有效提升语义查询的准确率, 而且对隐性信息也有很好的查询效果。

关键词: 语义网; 本体; 语义关系路径; 权重计算; 特异性

中图分类号: TP391.3 **文献标志码:** A

Efficient information search method based on semantic Web

XIA Meicui^{1*}, SHI Hongtao^{2,3}

(1. Archives, Qingdao Agricultural University, Qingdao Shandong 266109, China;

2. Network Center, Qingdao Agricultural University, Qingdao Shandong 266109, China;

3. School of Information Science and Engineering, Ocean University of China, Qingdao Shandong 266100, China)

Abstract: In order to improve the accuracy of Web information retrieval, an efficient information search method based on semantic Web was proposed. Firstly, all semantic paths between the target resources and the query keywords were extracted from the ontology library, and the weight of each semantic path was calculated by analyzing the weight and identification power of attributes included in it. Then, based on the weights, the number and the specificity of the semantic paths between resources and query keywords, as well as the semantic correlation between each resource and each keyword were calculated; and combining with the coverage and identification power of each keyword, the semantic correlation between each resource and the keyword set was calculated. Finally, on the basis of the correlation, all the resources were sorted and output. The experimental results show that compared with three different semantic Web search algorithms, including OntoLook, tf * idf and TMSubtree, the proposed method improved the average precision of 69.0, 25.0, 21.0 percentage points, respectively; average recall of 77.1, 28.3, 24.3 percentage points, respectively; and average *F*-measure of 72.4, 26.4, 22.4 percentage points, respectively. These results prove the proposed method can not only effectively improve the accuracy of semantic search, but also have good query results for indirect information.

Key words: semantic Web; ontology; semantic relationship path; weight calculation; specificity

0 引言

随着互联网信息技术的发展, Web 数据量越来越大, 如何从海量的 Web 数据中高效、快速、准确地检索到用户所需要的信息已经成为目前信息检索领域的研究热点。传统的基于关键字的查询方法由于缺少知识表示和语义处理的能力, 导致查询结果的准确率偏低^[1-2]。语义网是一种新型的网络体系结构, 它能够为网络中的源文档添加语义信息, 并且具有良好的概念层次结构和对逻辑推理的支持, 因而在 Web 信息检索中得到了广泛的应用^[3-4]。

近年来, 基于语义网的关键字查询技术受到了很多学者的关注, 许多查询方法被相继提出^[5]。文献[6]提出了一种

名为 OntoLook 的语义查询算法, 该算法通过构建一个基于关键字及其之间语义路径的概念关系图, 实现了对关键字查询过程中语义关系的识别, 然而这一方法缺少对语义路径权重的分析以及对查询结果的排序。为了克服这一缺点, 文献[7]提出了一种新的语义网信息检索框架, 该框架使用 tf * idf 算法对每个语义路径赋以权重, 并使用向量空间模型对查询结果进行排序。虽然这一算法的查询性能较 OntoLook 算法有较大幅度的提高, 但该算法对语义路径权重的分析粒度较粗, 并且没有考虑语义路径之间的差异。此外, 基于不同数据模型的语义网查询方法也被广泛研究, 文献[8-9]使用关系模型中的主、外键进行语义关系模型的构建, 并实现了语义信息的查询和排序; 文献[10]将图模型的概念与语义网进行融

收稿日期: 2015-05-05; 修回日期: 2015-07-08。 **基金项目:** 国家自然科学基金资助项目(60933011); 山东省教育厅项目(621326)。

作者简介: 夏美翠(1962-), 女, 山东莱阳人, 副研究馆员, 主要研究方向: 语义网、查询扩展; 时鸿涛(1981-), 男, 陕西西安人, 高级工程师, 博士研究生, 主要研究方向: 语义网、本体、大数据计算。

合,提出了一种基于语义链接结构的查询方法。文献[11-12]主要研究了基于XML模型的语义网查询方法,其中文献[12]提出一种基于最紧致匹配子树(Tightest Matched Subtree, TMSubtree)的查询算法,并取得了较好的查询效果。然而,这些方法虽然解决了语义路径权重的量化问题,并实现了对查询结果的排序,但却没有考虑语义路径之间的特异性以及关键字的涵盖范围和识别能力等问题。

从总体上看,基于语义网的关键词查询技术主要集中于语义路径的权重量化和查询结果的排序两个方面。目前的研究存在一些不足:首先,对于语义路径权重的计算不够准确,一些研究中的语义路径权重仍需要领域专家手动设置,不但影响系统效率,而且准确度较低;其次,对于语义路径的特异性没有进行识别,导致查询结果中包含有大量不相关信息;最后,对于关键字的涵盖范围和识别能力没有分析和考虑,导致对查询结果的排序不够合理。

本文针对以上问题提出了一种基于语义网的高效信息查询方法。该方法能够自动实现对语义路径权重的准确计算,同时充分考虑了语义路径的特异性,并对关键字的涵盖范围和识别能力进行了量化,从而保证了信息查询的准确性和高效性。

1 本体数据模型

本体是描述概念及概念之间语义关系的数据模型,它能够通过概念之间的关系来描述概念的语义,本体通常由Schema及其实例组成,其结构如图1所示。

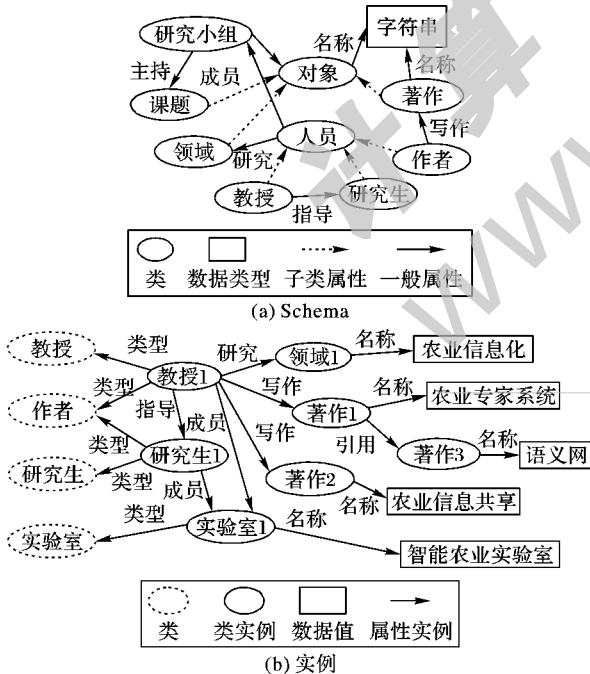


图1 本体样例

本文中不考虑本体之间的隐性语义关系,因此所有本体被表述在一个包括RDF特征、对象属性、数据类型属性以及反向属性的OWL-Lite子集中。由于本文的语义查询方法仅依赖于基本的本体语言特征,因而该方法具有更好的通用性和实用性。

针对本文的本体数据模型有如下定义:

定义1 Schema S 被定义为三元组 $\langle C, D, P \rangle$ 。其中: C 是类集, D 是数据类型集, P 是属性集。所有类、属性和数据类型

都通过统一资源标识符(Uniform Resource Identifier, URI)被准确表示,并且对于任意 $d \in C, r \in C \cup D$,有属性 $p(d, r) \in P$,其中 d 和 r 分别被称为属性 $p(d, r)$ 的领域和范围。

定义2 基于Schema $S = \langle C, D, P \rangle$ 的实例图被定义为一个有向图 $G = \langle V, E \rangle$ 。其中: V 是类实例集, E 是属性实例集。在实例图中,每一个资源表示一个类的实例。令 $[c]$ 表示类 $c \in C \cup D$ 的所有实例的集合,对于每个 $v \in V$,当 $v.type = c$ 时,有 $v \in [c]$ 。令 $[p(d, r)]$ 表示属性 $p(d, r) \in P$ 的所有属性实例集合,对于每个属性实例 $e(v_i, v_j) \in E$,当 $e.type = p, v_i.type = d, v_j.type = r$ 时,则 $e(v_i, v_j) \in [p(d, r)]$,其中 v_i 和 v_j 分别为 e 的主体和客体。

定义3 语义路径 sp 是Schema $S = \langle C, D, P \rangle$ 中的一个属性序列 $p_1(d_1, r_1) p_2(d_2, r_2) \cdots p_m(d_m, r_m)$,其中 $p_i(d_i, r_i) \in P$ 并且 r_i 和 d_{i+1} 是相同的类或具有相同的父类。

定义4 设语义路径 $sp = p_1(d_1, r_1) p_2(d_2, r_2) \cdots p_m(d_m, r_m)$,则 $ip = e_1(s_1, o_1) e_2(s_2, o_2) \cdots e_m(s_m, o_m)$ 为 sp 的一个语义路径实例,当 $e_i(s_i, o_i) \in [p_i(d_i, r_i)]$ 并且对于所有 e_i 有 $o_i = s_{i+1}$ 时,称 s_1, o_m 分别是 ip 的源和目的。

定义5 查询语句 Q 被定义为二元组 $\langle T, K \rangle$,其中 T 是类集, K 是关键字集。对于一个给定的Schema $S = \langle C, D, P \rangle$ 以及一个基于 S 的实例图 $G = \langle V, E \rangle$,语义搜索就是查找 $Q = \langle T, K \rangle$ 的资源集 R ,其中 $T \in C$ 。对于每个资源 $r_i \in R$,需要在 G 中至少有一个从资源 r_i 到数值为 s 的语义路径实例,其中 $r_i.type \in T$ 且数值 s 包含关键字 $k_i \in K$ 。

2 语义搜索引擎

本文的语义搜索引擎主要由语义关系提取模块、本体遍历模块、语义查询排序模块、输入和输出5部分组成,整个查询框架如图2所示。

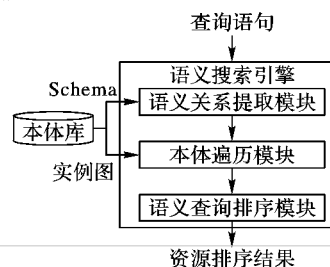


图2 语义搜索引擎

其中各部分的解释如下:

输入 输入为一条语义查询语句 $Q(T, K)$,其中 T 是查询资源所属的类, K 是一个关键字集合。

语义关系提取模块 该模块负责查找目标类型与查询关键字类型之间可能存在的各种语义路径。它能够从Schema中查找所有从类 T 到与关键字集合 K 相关联的类或数据类型之间的语义路径,从而能够将基于关键字的查询方式扩展为基于Schema的语义查询方式。

本体遍历模块 该模块负责查找与查询语句 Q 相匹配的所有资源(本体实例)。它能够根据语义关系提取模块中得到的语义路径遍历整个本体实例图,并返回与查询语句 Q 中的关键字集 K 相关联的资源集 R ,其中每个资源 $r_i \in R$ 通过实例图中的语义路径实例 $ip(r_i, k_i)$ 能够到达 K 中的某个关键字 k_i 。

语义查询排序模块 该模块负责计算每个资源与查询

关键字集的语义相关性,并按照其相关性大小对资源进行排序。它能够根据每个资源 $r_i \in R$ 与各查询关键字 $k_i \in K$ 之间语义路径实例的权重和数量、资源 r_i 对关键字集 K 的涵盖范围 $Cov(r_i, K)$ 、每个关键字 k_i 的识别能力来联合计算 r_i 与关键字集 K 之间的语义相关性 $Rank(r_i, K)$, 并根据 $Rank(r_i, K)$ 的大小对整个资源集 R 进行排序。

输出 输出为一个以 $Rank(r_i, K)$ 降序排序的资源序列。

在整个语义搜索引擎中,语义关系提取模块和本体遍历模块分别对应于传统语义查询系统中的查询扩展组件和查询匹配组件,而语义查询排序模块则是语义搜索引擎的核心,它负责对所有结果资源进行语义相关性计算和排序,从而提高查询结果的准确率。

3 语义相关性算法

3.1 语义路径的权重

3.1.1 属性的权重

一个语义路径通常由多个属性组成,因此语义路径的权重能够通过对其每个属性的权重计算来获得,在本文中,属性的权重将按照该属性的信息量和识别能力来计算。

根据信息理论,一个事件 x 的发生所产生的信息内容能够被量化为:

$$I(x) = -\lg pr(x) \quad (1)$$

其中: $pr(x)$ 是事件 x 的发生概率。对于一个基于 Schema S 的实例图 G ,属性 $p(d, r)$ 的发生概率 $pr(p(d, r))$ 为:

$$pr(p(d, r)) = \frac{|sub(p(d, r))|}{|N|} \quad (2)$$

其中: $sub(p(d, r))$ 为所有从本体类 d 到本体类或数据类型 r 的属性实例的数量, N 为所有属性实例的数量。因此,属性 $p(d, r)$ 所产生的信息量可表示为:

$$I(p(d, r)) = -\lg pr(p(d, r)) \quad (3)$$

此外,一个属性的识别能力能够通过信息理论中的互信息(Mutual Information, MI)来定义,因此属性 $p(d, r)$ 在领域 d 和范围 r 之间的互信息度量公式可表示为:

$$MI(p(d, r)) = \sum_{o \in r} \sum_{s \in d} pr(s, o) \cdot \lg \left(\frac{pr(s, o)}{pr(s)pr(o)} \right) \quad (4)$$

其中: s 为本体类 d 的实例, o 为本体类或数据类型 r 的实例, $pr(s, o)$ 为从 s 到 o 的属性的发生概率, $pr(s)$ 为以 s 为主体的属性的发生概率, $pr(o)$ 为以 o 为客体的属性的发生概率。

为了便于属性权重的计算,需要对式(3)~(4)进行零均值化处理,其均值化公式表示如下:

$$\begin{cases} \hat{I}(p(d, r)) = \frac{I(p(d, r)) - \min_{p \in P} I(p)}{\max_{p \in P} I(p) - \min_{p \in P} I(p)} \\ \hat{MI}(p(d, r)) = \frac{MI(p(d, r)) - \min_{p \in P} MI(p)}{\max_{p \in P} MI(p) - \min_{p \in P} MI(p)} \end{cases} \quad (5)$$

其中: P 为所有属性集合,计算后的 $\hat{I}(p(d, r))$ 和 $\hat{MI}(p(d, r))$ 分别属于区间 $[0, 1]$ 。

通过对式(5)进行合并,属性 $p(d, r)$ 的权重公式可表示为:

$$w(p(d, r)) = \alpha \cdot \hat{I}(p(d, r)) + \beta \cdot \hat{MI}(p(d, r)) \quad (6)$$

其中: $0 \leq \alpha, \beta \leq 1$, 且 $\alpha + \beta = 1$ 。

3.1.2 语义路径权重

尽管一个语义路径由多个属性组成,但语义路径的权重

并不能由所组成属性的权重相加来计算,实际上,当语义路径所包含的属性越多,源与目的之间的相关性就越小,因此语义路径的权重也就越小。此外,源与目的之间语义路径实例的数量也会影响该语义路径的识别能力,因此也需要对这一因素进行考虑。

综合各种因素,本文的语义路径 sp 的权重公式可表示如下:

$$W(sp) = \left(\prod_{p(d, r) \in sp} w(p(d, r)) \right) \cdot \delta^{length(sp)-1} \quad (7)$$

其中: $p(d, r)$ 是 sp 所包含的属性, $w(p(d, r))$ 是属性 $p(d, r)$ 的权重, δ 为区间为 $(0, 1)$ 的衰减指数, $length(sp)$ 为语义路径 sp 所包含属性的数量。

在实际应用过程中,本文设定语义路径的实例 ip 具有与其对应的语义路径 sp 的权重,即

$$W(ip) = W(sp) \quad (8)$$

通过以上公式,语义路径的权重在没有领域专家参与的情况下也能够被准确计算。

3.2 语义相关性算法

3.2.1 资源与关键字的语义相关性

在本体实例图中,资源与关键字之间关联的语义路径实例数量越多,它们之间的语义相关性也越大。因此,资源与关键字之间的语义相关性计算公式可表示如下:

$$R(r_i, k_i) = \sum_{ip \in IP(r_i, k_i)} W(ip) \quad (9)$$

其中: $IP(r_i, k_i)$ 是从资源 r_i 到关键字 k_i 的所有语义路径实例的集合, $W(ip)$ 表示语义路径实例 ip 的权重。

然而式(9)却存在一个问题,虽然3.1.2节中规定语义路径中属性数量越多,路径的权重越小,但在本体实例图中,包含有较多属性的语义路径往往具有大量的语义路径实例,这将导致与关键字不相关的资源可能具有较高的语义相关性。为了克服这一缺点,需要对式(9)进行修改以反映语义路径的特异性,修改后的公式如下所示:

$$R(r_i, k_i) = \sum_{ip \in IP(r_i, k_i)} (W(ip) \cdot spec(ip)) \quad (10)$$

其中: $spec(ip)$ 表示语义路径实例 ip 的特异性,计算公式如下所示:

$$spec(ip) = \prod \frac{1}{degree(s_i, p_i)} \quad (11)$$

其中: $degree(s_i, p_i)$ 是以 s_i 为主体的属性 p_i 的语义实例数量。

3.2.2 资源对关键字涵盖范围

资源对关键字涵盖范围是指与资源相关联的查询关键字的数量。由于资源所关联的查询关键字越多,该资源与关键字集之间的关联程度越高,本节提出了一个用于度量资源的关键词涵盖范围的算法。

对于资源集 R 和查询关键字集 K , 本文将资源 $r_i \in R$ 映射到一个 $|K|$ 维空间点 $[0, 1]^{|K|}$, 其中 r_i 的每个坐标表示它与相应的关键字 k_i 之间的相关性,该相关性能够通过式(6)来计算。因此,资源 r_i 的关键字的涵盖范围的计算公式如下所示:

$$Cov(r_i, K) = 1 - \left[\frac{\sum_{1 \leq i \leq |K|} (1 - NR(r_i, k_i))^p}{|K|} \right]^{\frac{1}{p}} \quad (12)$$

其中: p 为调节参数; $NR(r_i, k_i)$ 为零均值化后的从资源 r_i 到关键字 k_i 的属性的相关性。 $NR(r_i, k_i)$ 的计算公式如下所示:

$$NR(r_i, k_i) = \frac{R(r_i, k_i)}{\max_{r_m \in A} R(r_m, k_i)} \quad (13)$$

3.2.3 资源与关键字集的语义相关性

为了计算资源与关键字集之间的语义相关性,首先引入关键字识别能力的概念。一个关键字的识别能力能够通过该关键字的逆向资源频率 irf 来计算^[13]。关键字 k_i 的逆向资源频率计算公式如下:

$$irf(k_i) = \log \frac{|DV|}{|DV_{k_i}|} \quad (14)$$

其中: $|DV|$ 为本体实例图中所有类实例和数据值的数量, $|DV_{k_i}|$ 为本体实例图中包含关键字 k_i 的类实例和数据值的数量。通过对 $irf(k_i)$ 的均值化,关键字 k_i 的识别能力 $D(k_i)$ 可表示为:

$$D(k_i) = \frac{irf(k_i)}{\max_{k_m \in K} irf(k_m)} \quad (15)$$

关键字的识别能力表示一个关键字的权重,从而反映了一个关键字的重要性。由于在本体实例图中,通过语义路径实例与识别能力较强的关键字相连的资源具有更高的相关性,因此关键字的权重能够决定资源的语义相关性。

通过合并式(13)和式(15)最终得到了资源与关键字集的语义相关性公式:

$$Rank(r_i, K) = 1 - \left[\frac{\sum_{1 \leq i \leq |K|} (D(k_i) \cdot (1 - NR(a, k_i)))^p}{\sum_{1 \leq i \leq |K|} D(k_i)^p} \right]^{\frac{1}{p}} \quad (16)$$

式(17)能够准确地计算出资源与用户查询语句之间的语义相关性,从而提高语义查询的准确性。

4 实验与分析

实验选取青岛农业大学副教授以上职称人员的科研信息作为数据资源,并通过语义关系将这些数据资源构建为本体形式,构建后的本体库共有7984个类实例和数据值(包括人员、论文、著作、课题、实验室等)以及31413个属性实例(包括发表、编著、主持、参与等)。为了有效评估本文方法,为查询实验设计了5个不同的查询语句,表1显示了这些查询语句的表达式、查询目的以及相关资源数量。

表1 语义查询语句

编号	查询语句	查询目的	相关资源数量
Q1	<教授, {'玉米', '栽培'}>	查询研究玉米栽培的人员	14
Q2	<教授, {'花生', '产量', '施肥'}>	查询研究花生产量和施肥的人员	11
Q3	<文章, {'玉米', '根系', '灌溉'}>	查询玉米根系和玉米灌溉的文章	34
Q4	<著作, {'玉米', '栽培'}>	查询玉米栽培的著作	7
Q5	<课题, {'玉米', '选育'}>	查询玉米选育的课题项目	8

对于每条查询语句的查询结果,仅对前10项结果进行评价,并使用信息查询领域中基本的评价指标进行评价:

$$\text{准确率: } P = \frac{|A \cap RA|}{|A|}, \text{ 其中 } A \text{ 为查询结果的资源集合,}$$

RA 为本体库中相关的资源集合。

召回率: $R = \frac{|A \cap RA|}{N}$, 其中, 当 $|RA| \leq 10$ 时, $N = |RA|$; 当 $|RA| > 10$ 时, $N = 10$ 。

$$F \text{ 测度值: } F = \frac{2PR}{P+R}。$$

为了准确评价本文方法的性能,实验中将本文方法与 OntoLook 算法^[6]、tf * idf 算法^[7], 以及 TMSubtree 算法^[12] 进行对比,实验结果如表2~4所示。

表2 正确率比较

编号	本文方法	OntoLook	tf * idf	TMSubtree
Q1	1.000	0.300	0.600	0.800
Q2	0.900	0.200	0.500	0.500
Q3	1.000	0.200	1.000	1.000
Q4	0.600	0.100	0.300	0.300
Q5	0.750	0.000	0.600	0.600
平均值	0.850	0.160	0.600	0.640

表3 召回率比较

编号	本文方法	OntoLook	tf * idf	TMSubtree
Q1	1.000	0.300	0.600	0.800
Q2	0.900	0.200	0.500	0.500
Q3	1.000	0.200	1.000	1.000
Q4	0.857	0.140	0.428	0.428
Q5	0.937	0.000	0.750	0.750
平均值	0.939	0.168	0.656	0.696

表4 F 测度值比较

编号	本文方法	OntoLook	tf * idf	TMSubtree
Q1	1.000	0.300	0.600	0.800
Q2	0.900	0.200	0.500	0.500
Q3	1.000	0.200	1.000	1.000
Q4	0.706	0.120	0.353	0.353
Q5	0.833	0.000	0.667	0.667
平均值	0.888	0.164	0.624	0.664

对五个语义查询的实验结果显示:与 OntoLook 算法、tf * idf 算法和 TMSubtree 算法相比,本文方法的平均正确率分别提高了69.0、25.0和21.0个百分点;平均召回率分别提高了77.1、28.3和24.3个百分点;平均F测度值分别提高了72.4、26.4和22.4个百分点。三种评价指标本文方法均优于其他三种算法。

本文方法之所以具有更高的准确性,其主要原因在于该方法不仅能够准确地计算语义路径的权重,而且能够识别和量化语义路径的特异性以及关键字的覆盖范围和识别能力,从而保证了语义查询和排序的准确率。相比之下,OntoLook 算法因缺少对语义路径权重的分析和对查询结果的排序,因而查询效果最差;tf * idf 算法由于对语义路径权重计算粒度较粗,且未能对语义路径的特异性进行识别,因而导致查询结果不准确,并且无法查询隐性信息;TMSubtree 算法虽然较好地实现了语义路径权重的量化和查询结果的排序,但由于其没有识别语义路径的特异性,也没有考虑关键字的覆盖范围和识别能力,进而导致对于查询结果的排序不够准确。

5 结语

本文提出了一种基于语义网的高效信息查询方法,该方

法能够在语义查询过程中对本体中的语义路径权重进行准确计算,同时能够对语义路径的特异性以及关键字的覆盖范围和识别能力进行识别和量化,从而提高了语义查询和排序的准确率。实验结果表明,该方法比 OntoLook 算法、tf * idf 算法和 TMSubtree 算法具有更高的查询准确率,同时对隐性信息也具有更好的查询效果。

后续工作包括基于该方法的软件系统的开发及其在信息工程领域的应用。

参考文献:

- [1] MA S, ZHAO W, YUAN C, *et al.* Research on critical technologies of semantic retrieval based on rule reasoning [J]. *Acta Electronica Sinica*, 2013, 41(5): 977–981. (马森, 赵文, 袁崇义, 等. 基于规则推理的语义检索若干关键技术研究[J]. *电子学报*, 2013, 41(5): 977–981.)
 - [2] JIN Y, WANG Z. Research on semantic Web retrieval model based on reasoning and key technologies [J]. *Computer Engineering and Design*, 2013, 34(7): 2585–2589. (金燕, 王志华. 基于推理的语义网检索模型及关键技术研究[J]. *计算机工程与设计*, 2013, 34(7): 2585–2589.)
 - [3] DU X, LI M, WANG S. A survey on ontology learning research[J]. *Journal of Software*, 2006, 17(9): 1837–1847. (杜小勇, 李曼, 王珊. 本体学习研究综述[J]. *软件学报*, 2006, 17(9): 1837–1847.)
 - [4] YE Y, OUYANG D. New research advances in technologies of semantic Web search [J]. *Computer Science*, 2010, 37(1): 1–5. (叶育鑫, 欧阳丹彤. 语义 Web 搜索技术研究进展[J]. *计算机科学*, 2010, 37(1): 1–5.)
 - [5] LI H, QU Y. Keyword-based search on semantic Web data: the state of the art [J]. *Computer Science*, 2011, 38(7): 18–23. (李慧颖, 瞿裕忠. 基于关键词的语义网数据查询研究综述[J]. *计算机科学*, 2011, 38(7): 18–23.)
 - [6] LI Y, WANG Y, HUANG X. A relation-based search engine in semantic Web [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(2): 273–282.
 - [7] CASTELLS P, FERNANDEZ M, VALLET D. An adaptation of the vector-space model for ontology-based information retrieval [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(2): 261–272.
 - [8] LIU F, YU C, MENG W, *et al.* Effective keyword search in relational databases [C]// *SIGMOD 2006: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*. New York: ACM Press, 2006: 563–574.
 - [9] LUO Y, LIN X, WANG W, *et al.* SPARK: top-*k* keyword query in relational databases [C]// *SIGMOD 2007: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. New York: ACM Press, 2007: 115–126.
 - [10] VIDAL J, LAMA M, OTERO-GARCIA E, *et al.* Graph-based semantic annotation for enriching educational content with linked data [J]. *Knowledge-Based Systems*, 2014, 55(1): 29–42.
 - [11] THEOBALD M, SCHENKEL R, WEIKUM G. An efficient and versatile query engine for TopX search [C]// *Proceedings of the 31st International Conference on Very Large Data Bases*. New York: ACM Press, 2005: 625–636.
 - [12] ZHOU J, CHEN Z, TANG X, *et al.* Fast result enumeration for keyword queries on XML data [C]// *DASFAA 2012: Proceedings of the 17th International Conference on Database Systems for Advanced Applications*. Berlin: Springer-Verlag, 2012: 95–109.
 - [13] SALTON G, McGIL M J. Introduction to modern information retrieval [M]. Hightstown: McGraw-Hill, 1986: 1140–1151.
-
- (上接第 2904 页)
- [2] LI D, LIAO X, FAN F, *et al.* A focused network crawler with topic knowledge automatically growing [J]. *Computer Applications and Software*, 2014, 31(5): 30–33. (李东晖, 廖晓兰, 范辅桥, 等. 一种主题知识自增长的聚焦网络爬虫[J]. *计算机应用与软件*, 2014, 31(5): 30–33.)
 - [3] LU Y, LI Y. Improvement of text feature weighting method based on TF-IDF algorithm [J]. *Library and Information Service*, 2013, 57(3): 89–94. (路永, 李焰峰. 改进 TF-IDF 算法的文本特征项权重计算方法[J]. *图书情报工作*, 2013, 57(3): 89–94.)
 - [4] QIU Y, ZHAO B, LIN M, *et al.* Improved *k*-means clustering algorithm combined semantic similarity of short text [J/OL]. [2015-05-01]. *Computer Engineering and Applications*, <http://www.cnki.net/kcms/detail/11.2127.TP.20150624.1129.028.html>. (邱云飞, 赵彬, 林明明, 等. 结合语义改进的 *k*-means 短文本聚类算法[J/OL]. [2015-05-01]. *计算机工程与应用*, <http://www.cnki.net/kcms/detail/11.2127.TP.20150624.1129.028.html>.)
 - [5] HUANG C, YIN J, HOU F. A text similarity measurement combining word semantic information with TF-IDF method [J]. *Chinese Journal of Computers*, 2011, 34(5): 857–862. (黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度度量方法[J]. *计算机学报*, 2011, 34(5): 857–862.)
 - [6] SUN Z, ZHENG Q, YUAN J, *et al.* Semantic retrieval based on shallow semantic analysis technology [J]. *Computer Science*, 2012, 39(6): 107–110. (孙志军, 郑烜, 袁婧, 等. 基于浅层语义分析技术的语义检索[J]. *计算机科学*, 2012, 39(6): 107–110.)
 - [7] SCHUBERT F, LI H. Chinese word segmentation and its effect on information retrieval [J]. *Information Processing and Management*, 2004, 40(1): 161–190.
 - [8] CHENG X, LI Y. An ontology-based semantic extraction method of heterogeneous data [J]. *Computer and Modernization*, 2014(6): 2–6. (成欣, 李扬. 一种基于本体的异构数据语义抽取方法[J]. *计算机与现代化*, 2014(6): 2–6.)
 - [9] YU J J Q, LI V O K. A social spider algorithm for global optimization [EB/OL]. [2015-04-10]. <http://arxiv.org/pdf/1502.02407v1.pdf>.
 - [10] CHEN Y, CHEN Y, YANG Y, *et al.* Design and research on search strategy of focused crawler based on genetic algorithm [J]. *Journal of Chengdu University of Information Technology*, 2011, 26(5): 534–537. (陈悦, 陈运, 杨义先, 等. 基于遗传算法的聚焦爬虫搜索策略设计与研究[J]. *成都信息工程学院学报*, 2011, 26(5): 534–537.)
 - [11] YU H. Page feature extraction technology research [J]. *Journal of Shandong University of Technology: Science and Technology*, 2011, 25(2): 108–110. (于洪波. 网页特征提取技术研究[J]. *山东理工大学学报: 自然科学版*, 2011, 25(2): 108–110.)
 - [12] HE F, HE Y, LIU N, *et al.* A microblog short text oriented multi-class feature extraction method of fine-grained sentiment analysis [J]. *Acta Scientiarum Naturalium Universitatis Pekinensis*, 2014, 50(1): 48–54. (贺飞艳, 何炎祥, 刘楠, 等. 面向微博短文本的细粒度情感特征抽取方法[J]. *北京大学学报: 自然科学版*, 2014, 50(1): 48–54.)