文章编号:1671 - 6833(2016)03 - 0016 - 05

基于 WordNet 的本体查询方法研究

陈淑鑫, 张凌宇

(齐齐哈尔大学 计算中心,黑龙江 齐齐哈尔 161006)

摘 要: 不同的本体之间普遍存在着很多语义互操作的问题,如:语义冲突和结构异构,这些问题严重地影响了本体之间的知识共享和重用;同时也给本体查询服务带来了很大的困难,为此,提出一种基于 WordNet 的本体查询框架 OQ-WordNet. 该框架首先使用语义词典库 WordNet 来精确地计算不同本体(源本体)之间概念的相似度;然后通过本体集成和本体映射方法为所有源本体生成一个目标本体,并建立它们之间的语义映射关系;最后 OQ-WordNet 采用本体查询语言 SPARQL 来实现本体之间的查询功能.

关键词: 本体;WordNet;本体查询;本体映射;本体集成

中图分类号: TG335.58 **文献标志码**: A doi:10.13705/j.issn.1671 - 6833.2016.03.004

0 引言

在实现不同信息系统之间的知识共享或者交换的过程中,很多研究者都将本体[1]视为一种重要的知识库模型. 由于不同的本体构建者对相关领域知识的理解是不同的,而且他们不可能使用完全相同的概念来组织本体内的知识,因此不同本体之间普遍存在异构的现象. 本体之间的异构性会严重地影响本体查询方法[2]的性能,从而导致知识平台无法提供有效的查询检索服务. 为此,需要在设计本体查询方法的过程中,引入解决本体异构性方面的研究(如:本体映射和集成),以解决异构本体之间的查询问题.

为了实现异构本体之间的查询,很多研究者都建议构建一个标准的本体知识库框架[3],并在此框架下提供一系列可靠的本体查询理论基础,以供本体查询任务作为参考. 还有些研究者更专注于对本体集成的研究[4],他们建议先使用本体映射方法[5]为异构本体先建立语义关联,然后将这些本体集成到一个本体中,最后在这个本体中实现本体查询过程中所需要的各种行为.

对于很多领域来说,知识以及知识结构的更新速度是非常快的. 这使得构建以及维护一个上层领域本体集合变得非常困难. 为此,笔者将本体映射和集成方法引入到本体查询过程中,并提出一种基于 WordNet 的本体查询方法. 该方法首先使用 WordNet 语义词典库来提高计算异构本体之间相似度的精确度,然后根据相似度的计算结果来建立本体之间的映射关系,即为语义相同或者相似的概念建立映射关联,并且将异构的本体(源本体)集成到一个共享的本体(目标本体)中,最后根据目标本体与源本体之间的映射关系,实现本体查询的功能.

1 相关工作

目前,很多知识工程领域的研究者都在关注消除本体异构性的方法. 为此,笔者针对本体映射、本体集成和本体查询展开了深入的研究. 对于本体映射的研究来说,很多研究者都采用多种类型的相似度计算方法来提高映射的查全率和查准率. SM-Context[6]在使用多种相似度计算方法的时候还将概念周围的语言环境作为计算的一个重要因素,从而提高了语义映射的精确度. RIMOM[7]是使用了一种风险最小化模型来建立本体映射. 在本体映射的基础上,很多研究者通过本体集成来有效地整合本体之间共享的、冗余的知识信息. 文献[8]对各类本体集成模式进行了全面的研究. 文献[9]提出一种适用于生物工程领

收稿日期:2015 - 10 - 27;**修订日期**:2015 - 12 - 10

基金项目:国家自然科学基金资助项目(61204127);黑龙江省自然科学基金资助项目(F201334,F2015024);齐齐哈尔大学青年资助项目(2014k - M08);黑龙江省高校科技成果产业化前期研发培育项目(1254CGZH04).

作者简介:陈淑鑫(1978—),女,黑龙江哈尔滨人,齐齐哈尔大学副教授,主要从事人工智能、数据挖掘及语义 Web 方面的研究,E-mail:shuxinfriend@126.com.

域内本体的集成方法 SAMBO. 文献[10]提出了一种基于视图的本体集成框架 OIS-View,并将本体集成任务交给多个视图模块来完成.

关于本体查询的研究一直是本体研究领域内的一项重大课题. 文献[11]将关系数据库中的数据表转换成 RDF 数据形式,并使用 SPARQL 来间接地查询数据库中的数据. 文献[12]将 SPARQL 查询转换成 SQL 查询,并实现关系数据库的查询功能. 文献[13]使用斯坦福大学开发的 Parser 分析器来解析用户的自然语言查询,然后使用 SPARQL 实现本体查询. 文献[14]使用 SPARQL 解决了海量 RDF 的查询问题.

2　背景知识

2.1　本体定义

在计算机研究领域内,本体是一个明确的、形式化的、共享的概念模型,可被定义成一个四元组 $O=(C, P, I, Z)$.

C:概念集合. C 中的概念也可称为类,由三元组构成:$C=(c, A^c, V^c)$. 其中,c 是概念的名称;A^c 是描述概念的属性标识集合;V^c 是属性域集合.

P:属性集合. 描述概念的特征,由两部分组成:属性标识集合 A^c 和属性值集合 V^c. 其中 A^c 包含了属性所有的标示符;V^c 包含了属性的所有取值.

I:实例集合. 实例可由集合 A^c 中的属性以及 V^c 中的属性值来描述,即 $I=(id, v)$. 其中,id 是集合 A^c 中的具体元素名称;v 是集合 V^c 中具体的值.

Z:公理集合. 约束概念、属性和实例之间的隶属关系以及表示形式.

2.2　相似度

假设 x、y、z 是实体,$\mathrm{sim}(x,y)$的值表示 x 和 y 之间的相似度,那么相似度的形式化定义如下:

①$\mathrm{sim}(x,y)$ [0,1];

②$\mathrm{sim}(x,y)=1$ 表示 x 和 y 等价($y=x$);

③$\mathrm{sim}(x,y)=0$ 表示两个对象不相交,即它们之间没有共同特征;

④$\mathrm{sim}(x,y)=\mathrm{sim}(y,x)$,相似度的对称性;

⑤$\mathrm{sim}(x,z)<=\mathrm{sim}(x,y)+\mathrm{sim}(y,z)$,相似度满足三角不等式.

3　基于 WordNet 的本体查询框架

为了查询不同本体之间的数据,笔者提出了一种基于 WordNet 的本体查询框架 OQ-WordNet (Ontology Query Based on WordNet). 该框架是由多个功能模块组成,如图 1 所示.

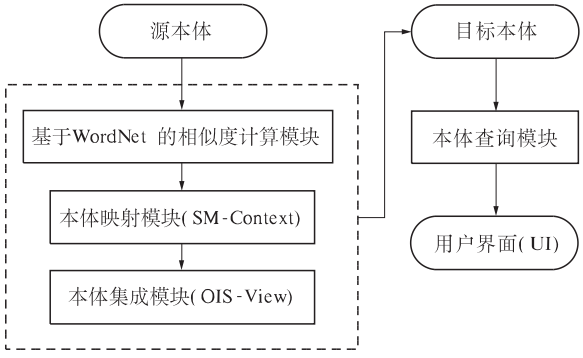

图 1　本体查询框架 OQ-WordNet

Fig. 1　The ontology query framework: OQ-WordNet

首先,OQ-WordNet 使用语义词典库来计算不同本体(源本体)的概念之间相似度;其次,使用本体映射模块来建立本体之间的语义关联;再次,使用本体集成模块,将这些本体之间共享的、可重用的、冗余的概念集成到一个标准本体,称之为目标本体;最后,使用查询模块(SPARQL 模块),以问答或者提示的方式来辅助使用者来生成查询表达式,并完成在目标本体和源本体中查询数据的操作.

3.1　基于 WordNet 的相似度计算

在现实中,本体中的大多数概念都是由多个词组成,而这些单词在 WordNet 中被称之为"原子概念". 因此,为了计算概念之间的相似度,需要依次比较原子概念之间的相似性. 使用 WordNet 词典库来计算原子概念之间的相似度时,可以测量它们在 WordNet 中的深度,以及它们最小公共概念的深度,然后再将原子概念之间的相似度定义成它们最小公共祖先的信息量.

从宏观上来说,如果两个原子概念存在于 WordNet 中同一个同义词集合,那么它们相似的可能性很大. 例如,单词 learner 出现在两个名词性的同义词集合{learner, scholar, assimilator}和{apprentice, learner, prentice};单词 student 出现在两个名词性的同义词集合{student, pupil, educate}和{scholar, scholarly person, bookman, student}. 这样,scholar 是 student 同义词集合和 learner 同义词集合的公共词语,student 和 learner 之间存在着相似关系. 如果我们继续在 student 同义词集合和 learner 同义词集合中查找同义词,与 student 和 learner 相似的词语的数量将会很大,这意味着 student 和 learner 之间的相似度很大. 根据

上面的分析,给出计算本体中概念之间的相似度计算公式,如下所示:

$$\mathrm{sim}(C,C') = \frac{\sum_{i=1,j=1}^{m,n}\mathrm{sim}(w_i,w'_j)}{\max(\mathrm{size}(C),\mathrm{size}(C'))}. \tag{1}$$

其中,概念 C 和 C'的原子概念分别为$\{w_1, w_2, \cdots, w_m\}$和$\{w'_1, w'_2, \cdots, w'_n\}$,函数 size 返回原子概念集合的大小,$1 <= i <= m$,$1 <= j <= n$. 如果原子概念 w_i 和 w'_j在一个同一词集合中,它们之间的相似度为 1;否则可以使用信息量来计算它们之间的相似度,详见公式(2). 其中函数 LCA 可以找到 WordNet 中 w 和 w'的最小公共祖先节点,函数 IC 可以计算概念的信息量.

$$\mathrm{sim}=(w,w') = \frac{\mathrm{IC}(\mathrm{LCA}(w,w'))}{\mathrm{IC}(w)+\mathrm{IC}(w')}. \tag{2}$$

3.2 本体映射与集成

在计算概念之间相似度的基础上,本体查询框架 OQ-WordNet 采用笔者在文献[6]和[10]中给出的本体映射(SM-Context)和本体集成方法(OIS-View),实现重新组织目标本体和源本体所包含的概念信息的任务.

3.2.1 SM-Context

为了充分地考虑概念位置对建立语义映射过程的重要性,方法 SM-Context 使用描述逻辑来表示概念以及语境(Context). 这样,接下来的语义映射任务就可以通过谓词推理来完成. 图 2 给出了 SM-Context 的映射过程.

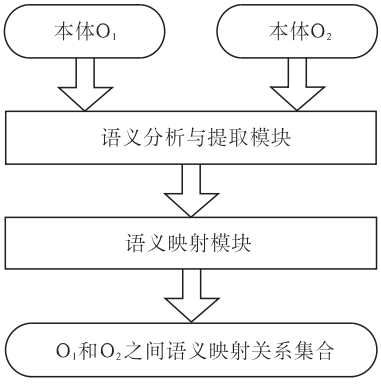

图 2 SM-Context 映射过程

Fig. 2 The mapping process of SM-Context

简要地介绍下语义映射模块建立映射关系的规则. 首先,假设 S_i、l_i、L_i 分别是概念 c_i 的结构、名称和标签,其中标签可能是概念的名称,也可能是概念的内容. 那么可以给出下面的规则来建立不同本体之间的概念关系.

(1)当关系 $c_1 \equiv c_2$,满足下面条件之一即可.

①$\mathrm{sim}(c_1, c_2) = 1$;

②$\mathrm{sim}(L_1, L_2) > 0$ && $c_1 \cap c_2$;

③$\mathrm{sim}(L_1, L_2) > 0$ && $\mathrm{sup}(c_1) \equiv \mathrm{sup}(c_2)$父概念等价;

④$\mathrm{sim}(L_1, L_2) > 0$ && $\mathrm{sub}(c_1) \equiv \mathrm{sub}(c_2)$子概念等价;

(2)当关系 $c_1 \supseteq c_2$,满足下面条件之一即可.

①$\mathrm{sim}(L_1, L_2) > 0$ && $\mathrm{sim}(S_2, S_1) = 1$ && $\mathrm{sim}(S_1, S_2) < 1$;

②L_1 是 L_2 的下义词 && $\mathrm{sim}(S_2, S_1) = 1$ && $\mathrm{sim}(S_1, S_2) < 1$;

③L_2 是 L_1 的上义词 && $\mathrm{sim}(S_2, S_1) = 1$ && $\mathrm{sim}(S_1, S_2) < 1$;

(3)当关系 $c_1 \subseteq c_2$,同关系 $c_2 \supseteq c_1$.

3.2.2 OIS-View

方法 OIS-View 将视图概念应用于本体集成. 该方法包含 3 个视图模块:集成视图(IV,Integrated View)、更新维护视图(UMV,Update Maintenance View)和整合视图(MV,Merged View). 图 3 给出了 OIS-View 的集成过程.

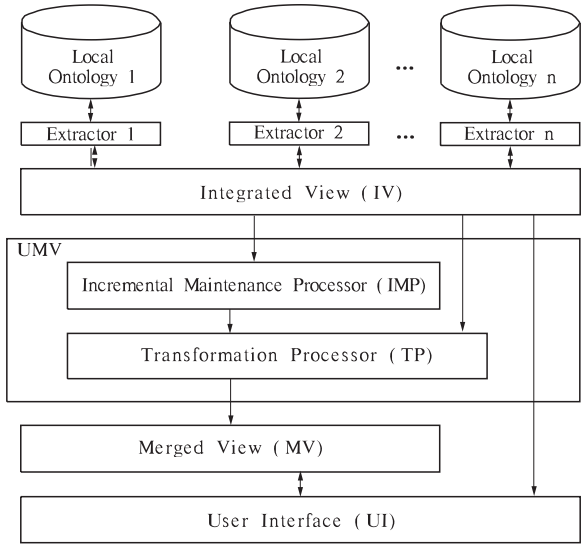

图 3 OIS-View 的集成过程

Fig. 3 The integrating process of OIS-View

在集成过程中,IV 的作用是根据映射结果从各个源本体中找出语义相同或者相似的概念. MV 统计出源本体所共享的概念集合,并将冗余的概念合并到目标本体中. MUV 的任务则是以日志的形式记录下所有关于本体的操作,并将修改内容返回给各个源本体,以保证目标本体和源本体之间的数据一致性. 在 MV 中,概念的集成规则如下所示:

(1)如果 $c_1 \equiv c_2$,那么删除 c_1;

(2)如果 $c_1 \supseteq c_2 \wedge \exists c_1 \equiv c_3$,其中 c_2 是 c_3 的子类,那么删除 c_1;

(3)如果 $c_1 \supseteq c_2 \wedge \neg \exists c_1 \equiv c_3$,其中 c_2 是 c_3 的子类,那么 c_2 是 c_1 的子类;

(4)如果 $p_1 \backsimeq p_2$,那么删除 p_2(例如:job $\backsimeq$ occupation);

(5)如果 $p_1 \subseteq p_2$,那么删除 p_1(例如:age $\subseteq$ birthday).

3.3　本体查询

本体查询框架 OQ-WordNet 中的查询模块采用本体查询语言 SPARQL 来实现. 具体来说,OQ-WordNet 通过 SPARQL 来完成两个任务:①接收并响应用户界面提出的查询请求;②通过本体映射关系对原始查询表达式进行重写,从而实现目标本体和源本体之间的知识查询.

SPARQL 查询是一个四元组(Key, DS, GP, SM),其中:

(1) Key 是用户输入的查询关键词,该项不能为空;

(2) DS 表示要查询的对象,即某个具体的本体,如果该项为空,则需要对目标本体以及所有的源本体进行查询;

(3) GP 表示图模型集合,本体内所有元素都被表示成 RDF 三元组,即 < $Subject$, $Predicate$, $Object$ > ,这样 RDF 图可以存储于 GP 中;

(4) SM 表示解决方案序列修饰符,主要作用是对无序的查询结果进行重新排列.

假设,用户使用 SPARQL 写出来的一条查询为($Person$, SO_i, $RDF\text{-}G_j$, $Order_{\mathrm{Age}}$). 其中 SO_i、$RDF\text{-}G_j$ 的下标表示集合中元素的编号,$Order_{\mathrm{Age}}$ 中的下标表示排序的字段,那么,框架 OQ-WordNet 可自动地解析这条查询,定位到相应的查询目标并完成相应的查询功能.

如果查询涉及到 2 个或者 2 个以上本体时,框架 OQ-WordNet 还需要基于本体之间的语义映射关系,实现查询重写的功能. 以查询($Person$, $Order_{\mathrm{Age}}$)为例,OQ-WordNet 先根据查询关键词 $Person$,从映射集合中找出相关的本体映射关系;然后再从映射关系中找到与 $Person$ 语义等价的概念以及概念所在的本体;最后使用找到的概念和本体对查询进行重写. 利用重写后的查询,OQ-WordNet 可以从目标本体和源本体中找出所有与 $Person$ 相关的 RDF 三元组,并使用 Age 信息对这些三元组进行排序.

4　实验结果

本实验采用 OAEI (Ontology Alignment Evaluation Initiative)所提供的数据集 conference 作为实验对象. 该数据集中共有 16 个关于“会议”的本体模型. 由于这些模型是由不同组织、机构所设计,它们之间都存在着异构性,这使得用户无法在多个本体之间查询知识. 为了说明的需要,笔者先选取数据集 conference 中的两个本体(Confious 和 OpenConf)作为实验对象. 本体 Confious 和 OpenConf 所包含的概念数量分别为 57 和 62. 至于在其他本体之间的查询,方法 OQ-WordNet 有着类似的性能表现,在此就不列出具体的数据分析.

在实验的过程中,查询关键词(select ?)和查询对象(from ?)可能有 3 种情况:目标本体、本体 Confious 和本体 OpenConf. 因此,笔者采用了 9 个查询语句($Q_1 \sim Q_9$)来全面地检测方法 OQ-WordNet 的查询响应时间. 例如:Q_2 是以目标本体中的概念作为关键词,本体 Confious 为查找源. 运行结果如表 1 所示.

表 1　查询的响应时间

Tab. 1　The response time of query

语句	关键词	查询对象	响应时间/ms
Q_1	目标本体	目标本体	634
Q_2	Confious	目标本体	906
…	…	…	…
Q_9	OpenConf	OpenConf	671

结果表明,方法 OQ-WordNet 的时间复杂度为 $O(n^2)$. 关键词和查询对象是相同本体时,查询相应时间大约在 600 ~ 650 ms 之间;关键词和查询对象是不同本体时,查询相应时间大约在 900 ~ 1 000 ms 之间.

5　结论

笔者提出了一种基于 WordNet 本体查询框架 OQ-WordNet,该框架首先利用 WorldNet 计算本体库中包含的项与待比较词语之间的语义相似度;然后根据计算结果再完成本体映射和集成任务,并为某一领域内的源本体生成一个目标本体;最后,使用 SPARQL 语言来实现该框架在目标本体和源本体之间的查询功能.

参考文献:

[1]　张凌宇, 马志晟, 陈淑鑫. 一种基于多种类型匹配

器的本体映射方法[J]. 郑州大学学报(工学版), 2015, 36(3): 11-15.
[2] 李韧. 基于Hadoop的大规模语义Web本体数据查询与推理关键技术研究[D]. 重庆: 重庆大学计算机学院, 2013.
[3] LANGE C. Ontologies and languages for representing mathematical knowledge on the semantic web[J]. Semantic web, 2012, 4(2): 119-158.
[4] DUONG T H, NGUYEN N T, JO G S. A Method for integration of wordNet-based ontologies using distance measures[C]//Internal Conference on Knowledge-Based Intelligent Information and Engineering System, KES 2008, Berlin, Heidelberg: Springer-Verlag, 2008: 210-219.
[5] HAASE P, HORROCKS I, HOVLAND D, et al. Optique system: towards ontology and mapping management in OBDA solutions[C]//Proceedings of the Second International Workshop on Debugging Ontologies and Ontology Mappings. New Jersey: Citeseer, 2013: 21-32.
[6] 张凌宇, 陈淑鑫, 张新. 一种基于上下文的语义映射方法[J]. 计算机应用研究, 2014, 31(10): 2990-2993.
[7] LI Jianzi, TANG Jie, LI Yi, et al. RiMOM: A dynamic multi-strategy ontology alignment framework[J]. IEEE Transaction on knowledge and data engineering, 2009, 21(8): 1218-1232.
[8] SOFIA P H, MARTINS J P. Methodology for ontology integration[C]//Proceedings of the First International Conference on Knowledge Capture. New York: ACM, 2001: 131-138.
[9] LAMBRIX P, TAN H. SAMBO-a system for aligning and merging biomedical ontologies[J]. Journal of web semantics, science, services and agents on the world wide web, 2006, 4(3): 196-206.
[10] 张凌宇, 陈淑鑫, 李敬有. 基于视图的本体集成系统框架的研究[J]. 计算机仿真, 2014, 31(7): 238-242
[11] LAUSEN G, MEIER M, SCHMIDT M. SPARQLing constraints for RDF[C]//Proceedings of the EDBT. New York: ACM, 2008: 499-509.
[12] HERT M, REIF G, GALL H C. Updating relational data via SPARQL[C]//Proceeding of the 2010 EDBT/ICDT Workshops. New York: ACM, 2010: 1-8.
[13] 张宗仁, 杨天奇. 基于自然语言理解的SPARQL本体查询[J]. 计算机应用, 2010, 30(12): 3397-3400.
[14] 汪璟玢, 方知立, 张燕琴. 面向分布式的SPARQL查询优化算法[J]. 计算机科学, 2014, 41(7): 227-231.

Study on Ontology Query Based on WordNet

CHEN Shuxin, ZHANG Lingyu

(Computer Center, Qiqihar University, Qiqihar 161006, China)

Abstract: There are many problems of semantic interoperability between different ontologies, such as semantic confliction and structure heterogeneous. These problems seriously influenced knowledge sharing and reusing between ontologies, and made it difficult to the service of ontology query. In view of this, this paper proposes framework of ontology query based on WordNet, which is called OQ-WordNet. This framework firstly calculates similarities for concepts from different ontologies (i.e., source ontologies) precisely by the lexicon WordNet. Then, OQ-WordNet applies the methods of ontology mapping and integration to generate a new ontology, called target ontology, and to create semantic mappings for these ontologies. Finally, OQ-WordNet achieves the function of ontology query by SPARQL that is a kind of ontology query language.

Key words: ontology; WordNet; ontology query; ontology mapping; ontology integration