一种基于 HBase 的本体复用新方法

许 闯[1],刘 鹏[1],刘志忠[1],鲍爱华[1],杨盛祥[2]
(1.解放军理工大学 指挥自动化学院,江苏 南京 210007;
2.73281 部队,江苏 淮安 223001)

摘 要:基于 HBase 的本体复用方法,是通过统计概念间关系在不同本体出现的频度来产生其可信度,最后形成带有统计信息和领域信息的大型概念空间。利用当前流行的开源云存储技术 HBase 表来存储具有海量信息的此概念空间,实现高效的检索功能,最终实现本体的复用。此方法规避了繁琐的异构本体映射过程,具有较好的适用性。并且所融合的本体数量越多,可信度越高,并且这种方法自动化程度高,不需要反复的人工参与,所以实用性比较强。

关键词:本体复用;统计;概念空间;HBase
中图分类号:TP393.08 **文献标识码**:A **文章编号**:1673-629X(2012)06-0057-04

A Novel Method of Ontology Reuse Based on HBase

XU Chuang[1],LIU Peng[1],LIU Zhi-zhong[1],BAO Ai-hua[1],YANG Sheng-xiang[2]
(1. Institute of Command Automation,PLA University of Science & Technology,Nanjing 210007,China;
2. 73281 Troops,Huaian 223001,China)

Abstract:The method of ontology reuse based on HBase,after statistic learning on concept relations,the frequency of different ontologies appeared in concept relations reveals certainy factor and helped to build a large-scale concept relations network including the statistic information and domain categories. Using the popular open source cloud storage technology HBase to store the concept of space can achieve efficient search function,and ultimately ontology reuse. This approach avoids the tedious process of heterogeneous ontology mapping,has good applicability,and the greater the number of ontology reusing is,the higher the credibility is. With high degree of automation,this approach does not require complicated human intervention,so its usefulness is stronger.

Key words:ontology reuse;statistic;concept of space;HBase

0 引 言

语义网研究的先驱 Gruber 提出,本体是概念模型的明确的规范说明[1]。随着信息技术的高速发展,本体在计算机相关行业的许多方面都体现出了重要性,在智能信息检索、异构信息集成、数字图书馆等领域,本体技术都有了比较广泛的使用。本体的构建是一个相对主观和分布的过程,不同领域的知识工程师在构建其特定的领域本体时,由于知识结构、语言背景或需求侧重点的差异,本体的描述内容往往存在很大的差异性(例如同义异构或同构异义的问题),因此本体的异构性是广泛存在而且不可避免的。除此之外,这种异构性同样会出现在同一领域内的本体中,甚至出现于同一工程师于不同时期构建的本体之中。另外,跨领域本体、无法准确判定领域的混合本体的存在,也使得本体的异构问题变得越来越复杂。

另外,本体的异构性影响了信息的共享和互操作,例如,语义 Web[2] 应用中,不同应用系统间的交互非常普遍和频繁,本体异构使系统间的信息交互存在巨大障碍;局部本体带有强烈的领域性和主观性,即使数量再多,也无法形成对世界的客观完整描述,那么在指导应用时,就会偏离客观。

所以为了解决本体异构的问题,急需寻找好的本体共享和集成的途径方法。目前,国外的相关研究较为活跃,但在本体共享和集成领域涉及的相关概念比较混杂,没有一个公认的定义,方法形态各异。主要有本体映射(Ontology-mapping)[3]、本体合并(Ontology-merging)[2]、本体集成(Ontology-integration)[4]、本体对齐(Ontology-alignment)[5]等,这些概念可以统称为本体复用。所谓本体复用,是指从已有的本体中获取知识,让它们能够得到重复利用。目前,本体复用技术已经广泛应用于地理空间信息系统、生物信息学等领域。

收稿日期:2011-10-30;**修回日期**:2012-01-30
基金项目:江苏省自然科学基金(BK2010130)
作者简介:许 闯(1985-),男,河南商丘人,硕士研究生,CCF 会员,主要研究领域为语义网;刘 鹏,博士,教授,博士生导师,主要研究领域为网格计算与云计算。

1 本体复用研究现状

Pinto 认为本体复用过程有本体集成和本体合并两种形式[6],强调以复用本体的主题领域来区分本体集成与合并的含义。而 Noy、Stumme 等人则认为本体合并和本体集成两者是等同的。Sowa 认为为了能够使多个异构本体交互,而在这些本体的实体间建立映射和处理映射,以达到本体对齐或者合并,多个本体组合为一个一致的本体的过程可以称为本体集成。奥尔登堡大学提出了按照集成程度把本体集成分为映射、连接与融合等 3 种类型,本体合并的结果是创建一个包含所有源本体信息的本体,而与之相比,本体连接是描述两个不同本体中实体的相应关系的映射的集合。虽然这些概念没有统一而权威的定义,且英文术语的翻译也存在差异,但这些概念的核心意义都指向本体复用。

本体异构的直接原因是本体之间的不匹配。明确这些因素是解决本体异构问题的基础。所谓的不匹配可以分为两个层次:语言层上的不匹配和模型层上的不匹配,Kitakami 和 Visser 将这 2 个层次分别称为非语义(non-semantic)和语义层次(semantic)。Visser 等人又将语义层次上的不匹配区分为概念化不匹配和解释不匹配两种情况。

有些异构性问题可以通过转换来解决,比较容易,主要集中在语言层,例如:语法不匹配、逻辑表达不匹配等;但是有些异构性问题比较复杂,很难解决,主要集中在概念层。近年来,国内外研究者针对此问题做了大量研究工作,主流的方法是通过寻找、构建本体间一对一的映射关系来消除异构性[7]。理论研究主要集中在对集成模型的研究,同时,研究者也开发了多种本体集成/映射的工具。

Fernández-Breis 和 Martínez-Béjar 提出了本体集成协作框架[8]。其集成算法基于分类特征和对两个本体中同义概念的识别。OISS[9]是 Calvanese 等人提出的一个面向本体集成系统的形式化框架,该框架中本体采用描述逻辑来书写,本体间的映射以合适的、基于查询的机制来表达,将本体中的一个概念映射到一个视图。Madhavan 等人提出了面向本体映射的框架[10],其特点在于:

(1)可以在以不同表示语言书写的模型之间实现映射,无须进行语言形式转换。

(2)可以表示不完备的或者缺失部分信息的映射。

OntoMapO[11]是 Kiryakov 等人提出的访问和集成上层本体的框架,事实上是一项本体映射服务。Kent 提出了一个面向本体结构的支持本体共享的框架 IFF[12],框架基于 Barwise 和 Seligman 所提出的信息流理论。

Noy 等人[10]认为可以将本体集成/映射分为四个层次:

合并(merging):合并两个本体为一个新的本体,代表性的工具有 iPROMPT,Chimaera,OntoMerge 等;

转换(transformation):通过定义转换函数将一个本体融入另一个,如:OntoMorph[;

对齐(alignment):通过寻找相关概念对在两个本体概念中定义映射,如:AnchorPrompt, GLUE, FCA-Merge;

连接(articulation):只在两个源本体的相关部分定义映射规则,如:ONION。

2 算法设计

文中提出一种新的思路来解决本体异构问题,实现本体复用。这是一种引入云计算技术的统计式机器学习的方法,通过自动学习快速集成符合一定规范的、来自不同领域的本体。为将主要精力集中在机器学习研究上,特规定只融合符合 OWL Lite 规范的本体。其融合过程类似"大鱼吃小鱼":母体(大鱼,称之为融合概念空间)通过不断"吞食"并"消化"比自己小的子体(小鱼,即被融合的本体)而拥有越来越丰富的概念知识。将这个过程称为本体融合(Ontology-Fusion)。

此过程主要包括以下内涵:

(1)本体融合是本体复用和共享的过程,主要针对的是本体异构问题;

(2)本体融合的目的是达到多个本体信息的重用和互操作,通过将多个不同本体所表达的知识融为一体,形成统计式概念空间,为用户提供趋于客观的概念描述;

(3)本体融合是以融合概念空间为母体的基于统计的机器学习过程,此概念空间具有不符合 owl 标准规范的特殊结构。

所谓融合概念空间,是一张巨大的概念关系网,所融合子体(本体)中的概念以及概念间的关系都要在此概念空间中记录。融合概念空间初始为空,可不断扩展,它不符合 owl 规范,但是可以表示为若干顶点(即概念)和若干条连接顶点的边(即概念间的关系)组成的逻辑关系图,这点与一般的符合 owl 规范本体文件相同。不同的地方在于,每条概念关系强度的统计信息记录在融合概念空间里的一条边上,具体包括:该关系(边)在所有子体中出现的总次数和换算出来的强度,以及按照需要,在某些特定领域里该关系出现的次数及强度。在数学上,每一条边所携带统计信息可以表示为:$\{(N,\lambda) \mid [N_i,\lambda_i]i, i = 1\cdots n\}$。$N$ 表示该关系在所有子本体中出现的总次数,λ 代表使用机

器学习方法根据N、每一个(N_i,λ_i)及全局情况计算出来的关系强度因子。N_i和λ_i分别代表该关系在第i个领域的出现次数以及计算出来的强度。$[N_i,\lambda_i]i,i=1\cdots n$记录从领域1到$n$每一个$(N_i,\lambda_i)$的情况。因此,可以形象地理解,关系(边)在融合概念空间的关系网里有的粗、有的细,还可以根据需要"进到任意一条边(关系)里面"观察该边(关系)在特定领域的粗细程度。

以融合概念空间为母体融合一个子体(被融合对象本体)的过程为:

(1)取出子体中的任意一个概念C,找到它在母体中的位置。如果找不到,则在母体中创建这个概念;

(2)取出子体中与C有关的任意一个关系R(设R的另一端是概念D);

(3)在母体中定位D,如找不到则创建D;

(4)在母体中增加C与D的关系R'的统计信息。如果子体有明确的领域信息,还要增加R'的领域统计信息;

(5)继续第2步直到子体中与C有关的关系都被统计;

(6)继续第1步直到子体中的概念都被取完;

(7)最后根据母体此次融合前后的状态变化,修改所有概念间关系的强度因子,对统计式信息进行机器学习和全局优化,从而达到融合子体所携带知识的效果。

3　数据检索

上述本体自动融合方法会遇到海量本体样本及分析结果的快速存储与检索问题,因此在分析过程中融合概念空间需要存储海量本体样本中的概念及其关联,并且对这些关联的强度进行高效存储。除了存储以外,对现有融合概念空间的数据检索也对融合效率有着非常高的影响,这从本课题所提出的融合过程即可看出。

融合概念空间的数据存储与管理有其独特的要求:

首先,融合概念空间需要海量存储支持。由于样本数目较少通常会带来较为严重的数据集偏斜问题,同时融合概念空间所包含概念的广度与深度也会存在很大不足,因此就需要对目前互联网上广泛存在的各领域的异构本体进行统一分析,这就对融合概念空间的存储提出了较高的要求。在存储融合概念空间时,不仅要对异构本体中所出现的每类概念进行存储,还需要对概念之间的对应关系与强度分析数据进行关联存储,而且这种关联是双向的,并且随着分析领域的不断增多,关联强度数据也会不断丰富。因此,融合概念空间需要海量数据存储与管理的支撑,并且这是常见的数据库技术所不能满足的。

其次,融合概念空间的存储需要具有高可扩展性。本课题所提出的本体自动融合方法本质上是一种统计学习的方法,统计学习的本体样本越多,分析的结果也会越准确。因此,本课题所提出的本体融合过程是一个持续演进的过程,且随着时间的推移,融合概念空间的广度与精度会不断提高,其价值也会逐步体现。但是,这种持续演进的方式对于分析结果的存储也带来了挑战,每个概念所关联的概念集是持续增加的,因此融合概念空间的存储要能够很好地支持扩展,从而为不断增长的概念及关联提供有效的存储与管理。

第三,融合概念空间需要具备高效的数据检索能力。为了能够得到更为完整的融合概念空间,需要通过统计学习的方法将新的本体融合进去,而这个融合进程离不开快速的数据检索。在这个过程中,需要快速检索融合概念空间中是否已经包含待融合本体中当前正在分析的概念,此后还需要对关联概念进行检索,以确定其是否包容于融合概念空间。最后,检索分析融合概念空间是否已经包含待分析的概念关联,并进一步对关联强度进行计算。在这个过程中,随着融合概念空间的不断增大,对其检索性能的要求也越来越高,否则本体融合效率难以得到保证。

从上述分析可以看出,由于融合概念空间持续演进的特性,传统的数据库系统已经难以满足融合概念空间的存储管理需求。对于这个问题,将在本体融合过程中引入当前流行的云计算技术,通过分布式结构化表服务HBase来解决融合概念空间的存储管理问题。

HBase - Hadoop Database,是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统,对于任何应用,HBase在逻辑上都将所有数据存储在一张表中,但是为了提高扩展性和检索性能,HBase在物理存储上,将逻辑上的"一张表"划分为若干子表存储,每个子表都存储了部分行和部分列,然后再通过子表元数据对其存储数据进行描述。通过这样的机制,用户在查询指定的行和列的时候,就可以通过元数据快速定位该数据位于哪一个子表,然后再通过子表文件查询,获取对应的数据。由于数据被划分为子表存储,且子表可以灵活的增加,因此在HBase中每一行都可以存储任意列的数据,同时表中的行数也可以不受限制的增长。通过这种机制,HBase既可以保持很好的扩展性,又能够实现数据的高速查询。

HBase的上述特性恰恰适用于本体融合概念空间的存储。由于每个概念的关联概念在融合概念空间中都是持续增长的,因此可以将概念作为行键存储,而与

其相关的概念则作为该行的列键(可能有很多)存储,对应的数据内容则为两个概念的关联及其统计数据(强度)描述。由于HBase的行可以任意增长,因此融合概念空间在融合新增的本体时,待融合本体的概念可以任意添加到融合概念空间中;由于HBase每一行都可以包含任意多的列,因此待融合本体中新增的关联也可以任意添加到融合概念空间中。通过HBase来对融合概念空间进行映射,可以快速查询一个概念是否出现在融合概念空间中,也可以查询一个概念是否存在于待分析概念的关联概念中。同时,由于HBase不对其存储的数据做任何的限制,因此可以自行定义概念关联及其强度的数据结构,并且可以不受限制地更新数据。

4 结束语

基于HBase的本体复用的优点是所融合的本体数量越多,融合概念空间越大,可信度越高;机器式学习自动化程度高,所以不需要繁复的人工参与;既保持了领域内信息,又能更好地实现跨领域融合,具有较好的适用性。近年来,基于统计式学习辅助解决人工智能问题研究取得了突破性进展,在手写输入识别、语音识别、自然语言理解等众多领域都有大量实用系统出现。因而,基于HBase进行本体融合是值得尝试的。

参考文献:

[1] 于 琦,周 勇.一种基于本体的异构数据集成[J].计算机技术与发展,2008,18(2):35-36.

[2] 张丽坤,蒋 波.基于本体的语义Web研究[J].计算机技术与发展,2007,17(6):116-119.

[3] Algergawy A,Schallehn E,Saake G. A Sequence-based Ontology Matching Approach[C]//Proceedings of the 18th European Conference on Artificial Intelligence, Workshop on Contexts and Ontologies. Patras,Greece:[s. n.],2008.

[4] Barwise J,Seligman J. Information Flow:the Logic of Distributed Systems[D]. Cambridge:Cambridge University,1997.

[5] Mcguiness D L,Fikes R,Rice J,et al. An environment for merging and testing large ontologies[C]//Proceedings of the 7th International Conference on Principles of Knowledge Representation and Reasoning. San Francisco,Cal.,USA:Morgan Kaufmann,2000:483-493.

[6] Pinto H S,Perez A G,Martins J P. Some Issues on Ontology Integration[C]//Proceedings of IJCAI99's Workshop on Ontologies and Problem Solving Methods:Lessons Learned and Future Trends. [s. l.]:[s. n.],1999.

[7] 赵国增,郭恒川.基于本体的异构数据共享研究[J].计算机技术与发展,2010,20(10):40-42.

[8] Fernández-Breis J T,Martínez-Béjar R. A Cooperative Framework for Integrating Ontologies[J]. International Journal of Human-computer Studies,2002,56(6):662-717.

[9] Calvanese D,Giacomo G D,Lenzerini M. A framework for ontology integration[C]//Proceedings of the 1st International Semantic Web Working Symposium (SWWS). [s. l.]:[s. n.],2001:303-316.

[10] Madhavan J,Bernstein P,Domingos P,et al. Representing and Reasoning about Mappings between Multiple Domain Models[C]//Proceedings of the AAAI Conference. [s. l.]:[s. n.],2002.

[11] Kiryakov A,Simov K,Dimitrov M. OntoMap:Portal for Upper-level Ontologies[C]//Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS2001). Ogunquit,Maine,USA:[s. n.],2001.

[12] Kent R. The information flow foundation for conceptual knowledge organization[C]//Proceedings of 6th International Conference of the International Society for Knowledge Organization. [s. l.]:[s. n.],2000.

(上接第56页)

of the VLDB 2007. New York:ACM Press,2007:411-422.

[7] Sidirourgos L,Goncalves R,Kersten M,et al. Column-store support for RDF data management:not all swans are white [C]//Proc. of the VLDB Endowment. [s. l.]:[s. n.],2008:1553-1563.

[8] Lv Bin,Du Xiaoyong,Wang Yan. Selectivity Estimation of Correlated Properties in RDF Data for SPARQL Query Optimization[C]//Fifth International Conference on Semantics, Knowledge and Grid. [s. l.]:[s. n.],2009:176-183.

[9] Vermeij M,Quak W,Kersten M,et al. MonetDB:a novel spatial column-store DBMS[C]//Academic Proceedings of the 2008 Free and Open Source for Geospatial Conference. [s. l.]:[s. n.],2008:193-199.

[10] 易雅鑫,宋自林,尹康银.RDF数据存储模式研究及实现[J].情报科学,2007,25(8):1218-1243.

[11] Library catalog data[EB/OL]. [2007-01-09]. http://simile. mit. edu/rdf-test-data/barton.

[12] 鲍 文,李冠宇.本体存储技术研究[J].计算机技术与发展,2008,18(1):146-150.