基于加权的冗余相似本体实例发现的研究

卢传耀,徐 敏
(南京航空航天大学 计算机科学与技术学院,江苏 南京 210016)

摘 要:在开放式的环境下一般由不同的组织和人员对一个领域的本体库的知识实例进行维护和添加,这就可能出现重复描述的实例的问题,会出现对同一对象的不同实例描述,甚至是相互矛盾的,从而出现多义性。这就要通过本体的匹配发现这些重复描述的实例,并把它们合并。但是合并前需要找出这些重复的实例,文中通过对实例属性及其值的加权并通过文中提供的算法来查找出这些冗余的实例。通过实验结果可以发现,此方法可以为解决这个问题提供一种很好的解决方法。

关键词:冗余实例;对象-属性;信息熵;重复度;相似度
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2014)09-0011-05
doi:10.3969/j.issn.1673-629X.2014.09.003

Research on Ontology Instances Found Redundancy Based on Entropy Weighted

LU Chuan-yao, XU Min
(College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: In the open environment, the instances of knowledge for a domain are maintained and added by different organizations and personnel, which may occur in case repeated description problem, there will be different instances of the same object, even contradictory. This will need to find these repeated instances through the ontology matching, and merge them. However, it is necessary to find these repeated instances before merging, the instance attributes and their values are weighted and through the provided algorithm to check out these redundant instances in this paper. The experimental results show that this method is a good solution.

Key words: redundant instances; object-attribute; information entropy; repeatability; similarity

0 引 言

随着语义网的有力发展,本体的使用迅速增加,其规模也迅速扩大,对同一对象用不同的实例记录描述的情况经常出现,甚至是对同一对象的相互矛盾的描述。因此查找出这些重复实例的方法应运而生,在语义网的发展中有着重要的作用。

本体给出的方法是基于实例的属性和属性值的加权计算,通过各种算法的最终比较有效地发现重复的实例。

1 相关工作

目前研究比较热的本体匹配方法[1]主要分为元素级匹配和结构级匹配。元素级[2]匹配方法主要是基于语法和基于实例的方法。基于语法根据概念的定义及其同义词进行匹配,基于实例考虑的是概率间的相似性判定。它们使用的技术主要是利用概念定义的概念类型、属性、唯一性和可选性的约束技术,利用实例的名称、标签、注释等信息字符串的比较之间的相似度技术,还会利用 WordNet、HowNet 等外部语义资源[3]来进行本意匹配。结构级匹配方法主要是基于本体内部结构[4]的方法,包括本体的从属关系、本体的属性集、属性之间的关系,通过内部结构来计算概念的相似性,它主要用于相同或类似领域的概念[5]。

文中仅针对元素级的基于实例的相似性判定实例间是否是描述同一对象。主要通过实例的 URI,属性名称和属性值的字符串之间的相似度比较的技术。

收稿日期:2013-11-07 **修回日期**:2014-02-13 **网络出版时间**:2014-07-17
基金项目:江苏省自然科学基金(NS2013087);南京航空航天大学基础研究项目(NS2013087)
作者简介:卢传耀(1988-),男,硕士,研究方向为人工智能、语义网;徐 敏,副教授,研究方向为人工智能、语义网、软件工程。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140717.1231.040.html

2 冗余实例发现

2.1 问题描述

冗余实例发现要对源实例和目标实例进行本体映射来建立一种实例间的映射关系[6],如定义 1。

定义 1:本体映射可描述为一个三元组<n_1,n_2,R>,其中 n_1 为源实例,n_2 为是目标实例,R 表示两个实例之间的语义关系,它包括以下四种关系:= 等价、⊥ 不相交、⊆ 蕴含和 ⊇ 蕴含于。

这里的实例映射的目的是为了发现等价实例和基于蕴含关系查找出冗余实例。但是基于蕴含关系的是基于两个实例的本体之间的属性的蕴含关系及属性值的蕴含关系。但是在多数情况下,这种蕴含关系是相当稀少的,因为维护本体库的机构或人员会在下次维护本体库的时候对某些本体的属性进行增加或减少。这对本体概念之间的属性蕴含关系是一种破坏。所以用上面的定义来查找冗余实例还不是完全满足需要的。所以下面将给出基于熵加权的冗余实例发现的方法。

由 RDF 定义[7]可知,RDF 数据是以形如(S,P,O)的三元组形式来描述和组织数据的,其中 S 表示主语,P 表示谓语,O 表示宾语。主语表示 Web 的某一资源,谓语表示主语的某一属性或与其他资源的关系,宾语表示谓语的属性值或与之有关系的另一资源。(S,P,O)三元组中有以下三种形式的数据:统一资源标识符(URI)、字面值(Literal)、空节点(Blank Node)。在一个本体库中 URI 应该是唯一的,它标记所描述的对象的唯一性。即可以通过相同的 URI 来找到等价的实例。

2.2 基于熵加权的实例相似性计算方法

在 RDF 的定义中,所有的概念和属性的重要性通常都是同等的。但是在实际应用中,用户通常会根据自己的需求对某些属性感兴趣而添加或减少一些属性的使用。因此文中将对属性进行加权来区别它们的重要性,还有根据属性值的重复度来提取出识别度较高的属性。文中主要是基于文本相似性来发现冗余实例的,文献[3,8-10]对基于文本相似性和词典相似性的方法来进行本体匹配进行了相关的研究。

定义 2:定义三元组 $K=(S,P,I)$ 为本体库的背景关系,集合 S 表示本体库中所有的实例集合,P 表示所有属性集合,I 是二元关系,$I \subseteq S \times P$。如果 $sIp, s \in S \wedge p \in P$,则说明实例 s 有属性 p。

运用上面的定义,给出第一个权值属性出现的频率的计算公式,如定义 3 所示。

定义 3:定义本体库三元组 $K=(S,P,I)$。其中 S 表示实例集合,$S=\{s_1,s_2,\cdots,s_m\}$;P 表示属性集合,$P=\{p_1,p_2,\cdots,p_n\}$;属性集合对应的信息熵出现频率为 $W=\{w_1,w_2,\cdots,w_n\}$ $(0 \leqslant w_i \leqslant 1,\ 0 \leqslant i \leqslant n)$,其中 w_i 为属性 p_i 的出现频率。计算公式如下:

$$w_i = \frac{1}{m} \cdot \sum_{i=1}^{n} f(s_i,p_i),\ I \subseteq S \times P \tag{1}$$

其中,函数 $f(s_i,p_i)=\begin{cases}1, 若\ s_i I p_i\ 成立\\ 0, 若\ s_i I p_i\ 不成立\end{cases}$,$I \subseteq S \times P$。

定义 4:三元组 $K=(S,P,I)$,其中属性集合 $P=\{p_1,p_2,\cdots,p_n\}$,对应的属性值重复度的集合为 $V=\{v_1,v_2,\cdots,v_n\}$。属性重复度 $v_i(0 \leqslant i \leqslant n)$ 的计算公式如下:

$$v_i = \frac{属性\ p_i\ 不同值的个数}{属性\ p_i\ 值的总数} \tag{2}$$

对属性值的比较使用字符串相似度,文中使用文献[11]中用到的 Humming Distance 方法。设两个字符串 s 和 t,则它们的相似度为:

$$\mathrm{Sim}(s,t) = 1 - \frac{\sum_{i=1}^{\min(|s|,|t|)} f(i) + ||s| - |t||}{\max(|s|,|t|)} \tag{3}$$

最后对两个实例进行相似性的计算,给出如下定义。

定义 5:本体库三元组 $K=(S,P,I)$,有任意两个实例 s_1 和 s_2,$s_1 \neq s_2 \wedge s_1, s_2 \in S$ 且 P_1、P_2 分别为 s_1 和 s_2 的属性集合,P_1、$P_2 \subseteq P$。s_1 和 s_2 的相似度计算公式[12]如下:

$$\mathrm{Sim}_{\mathrm{instance}}(s_1,s_2) = \frac{\sum_{p_i \in P_1 \cap P_2, 1 \leqslant i \leqslant n} w_i \times \mathrm{Sim}(L(s_1,p_i),L(s_2,p_i))}{\sum_{i=1}^{n} w_i} \tag{4}$$

其中,函数 $L(s_1,p_i)$ 表示实例 s_1 中属性 p_i 的字符串形式的值。

3 实验及分析

3.1 实验数据

文中实验的数据来自网络的 DBLP[13],它是由德国特里尔大学的一个团队开发和维护的,提供了计算机领域高质量的科学文献搜索服务,并且只储存这些文献的相关元数据,如标题、作者、发表日期等,不提供全文下载。和一般情况不同,DBLP 并没有使用数据库而是使用 XML 存储元数据。该数据为截止到 2012 年 2 月份的全部元数据,遵循 ODC-BY 1.0 数据开放协议供用户公开使用。DBLP 收录的文献类型有 article, inproceedings, proceedings, book, incollection, phdthesis, mastersthesis, www 共 8 类,提供的属性描述信息为 author, editor, title, booktitle, pages, year, address, journal, volume, number, month, url, ee, cdrom, cite, publisher,

note,crossref,isbn,series,school,chapter。

3.2　数据解析与处理

DBLP 数据是用 XML 语言按照 RDF 规范来存储元数据的一个本体库。它是一个很大的文件,实验处理的 rdf 文件是截止到 2006 年的 DBLP 本体库大小为 564 M,而到 2012 年 2 月的本体库大小为 1.13 G。如果直接对这些数据文件进行操作实验的话,效率肯定是很低的,而且对普通内存只有 2 G 的电脑来说,这个文件实在太大了,内存肯定不够用。所以要借助一些现有的工具来辅助实验。也许数据库是个不错的选择。对于正在增长的本体库来说,肯定涉及到大量数据的问题。现在有关系型数据库和 NoSQL 数据库可以选择,因为处理大量数据对事务的要求很低,而且实验主要关注的是它们的访问读取速度,因此 NoSQL 很符合要求,所以实验选择 NoSQL 面向文档的 MongoDB[14] 数据库作为实验数据解析后的存放介质。

解析工具选择 dom4j 按 XML 语法进行解析,解析后的属性值都是以字符串的形式给出。最后存储到 MongoDB 数据库中。下面将给出 MongoDB 数据库存储的模式设计,因为下面的实验都是围绕着这个设计模式来进行数据分析的。

通过 rdf 的 dtd 文件描述和先期对文件扫描得到本体库所有的属性集为{address, booktitle, cdrom, chapter,cite,creator,crossref,date,editor,identifier,isbn, journal,month,note,number,pages,publisher,rating,reviewid,school,series,title,type,volume,year},这个集合的属性是有顺序的,从第一个开始编号为 0,1,…,24。这是为了方便后面实验的统计并能提高效率。

MongoDB 数据库存储的模式设计如图 1 所示。

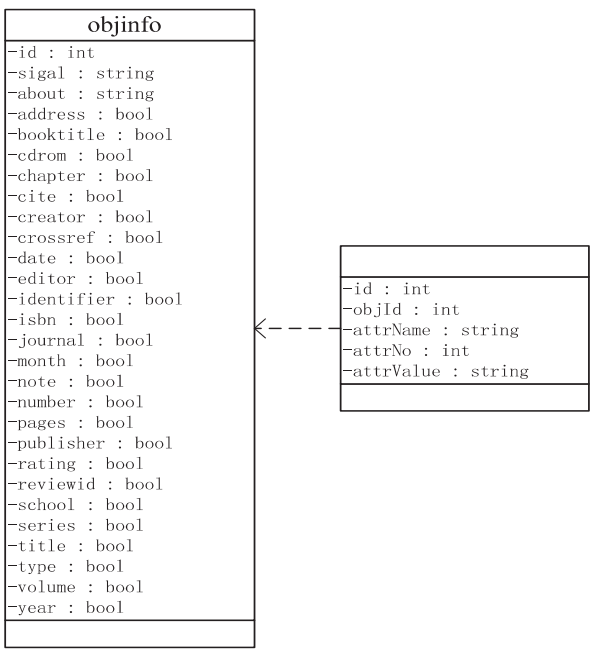

图 1　MongoDB 数据库存储的模式

objinfo 为实例集合 sigal 是长度为 25 的 0、1 字符串,其对应位上的字符为 1 说明实例有相应编号的属性,同时对应的属性字段设为 true。sigal 是用来标记实例所拥有的属性。attrinfo 是实例的属性值集合,其 objId 说明属性值是哪个实例的,attrName 和 attrNo 记录了属性名称和属性在上面属性集合的编号。

3.3　同一对象的简单查找

由 3.1 的问题描述中可知,在 RDF 数据的三元组的同一标识符(URI)在同一本体库中的描述应该是唯一的。在本实验数据中一个实例是通过 rdf:Description 的 rdf:about 属性值来表示,即在实验数据中如果两个实例 about 的值相等,则表明它们描述的是同一个对象。

通过实验对实例集合简单的比较发现有 URI 值相等的实例。表 1 将列出相等实例数比较多的实例 URI。

表 1　相同 URI 的实例实验部分结果

数量	URI
34	http://www.elsevier.nl/locate/entcs/volume30.html
25	http://dx.doi.org/
24	http://www.elsevier.nl/locate/entcs/volume29.html
23	http://www.elsevier.nl/locate/entcs/volume59.html
17	http://www.elsevier.nl/locate/entcs/volume42.html
14	http://www.elsevier.nl/locate/entcs/volume31.html
13	http://www.elsevier.nl/locate/entcs/volume58.html
12	http://www.elsevier.nl/locate/entcs/volume57.html
5	http://www.kuenstliche-intelligenz.de/archiv/2004_3/KI-Tagung-2004-web.pdf
4	http://elj.warwick.ac.uk/jilt/BookRev/98_1y2k/default.htm

从实验结果可以看到在本体库中重复实例的存在性和对相似实例的查找且合并是有意义的。

3.4　基于熵加权的实例相似性计算方法的数据分析

经过解析,实验的数据有 151 655 个实例,有属性 1 686 351 个。如果对任意两个实例进行相似度对比,这个工作量是相当大的,效率肯定很低。随着实例数据的增加,这个任务可以说是不可能完成的。为了减少比较的实例的数量,缩小查找对比的范围,此次实验提出通过属性出现的频数和属性值的重复度来建立识别属性集合。

计算步骤如下:

1)利用公式(1)计算出各属性出现的频数,实验结果如表 2 所示。

从实验结果数据可以看出,每个属性的出现频数是不同的。在频数集中要选取频数大于 0.01 的属性,

因为频数越小,说明拥有此属性的实例越少。说明它的价值越小。按此原则可以建立属性集合 $A=\{$cdrom, creator, date, identifier, journal, number, pages, title, type, volume, year$\}$。

表 2 属性出现的频数

属性编号	属性名称	出现的频数
0	address	0.0
1	booktitle	0.001 0
2	cdrom	0.127 1
3	chapter	0.0
4	cite	0.006 0
5	creator	0.993 0
6	crossref	0.000 6
7	date	1.0
8	editor	0.0
9	identifier	1.0
10	isbn	0.0
11	journal	0.998 7
12	month	0.001 2
13	note	0.0
14	number	0.916 4
15	pages	0.935 6
16	publisher	0.0
17	rating	0.0
18	reviewid	0.0
19	school	0.0
20	series	0.0
21	title	1
22	type	1
23	volume	0.997 1
24	year	0.999 8

2)利用公式(2)计算出各属性的重复度,实验结果如表 3 所示。

由表 3 结果可以得出,每个属性的重复度是不同的,而且相互的差异很明显。对于那些属性值重复度很高的属性,说明此属性很可能是实例的一个自动标记的内容,对文中的目标识别率贡献很低。所以实验中将重复度大于 0.9 和等于 0 的属性去掉,保留重复度在 0 到 0.9 之间的属性建立集合 $B=\{$cdrom, creator, editor, pages, publisher, reviewed, series, title$\}$。

3)建立识别属性集合 $D=A\cap B$,得到集合 $D=\{$cdrom, creator, pages, title$\}$。

通过上面的识别属性集合 D,得到识别实例集合。实验结果如表 4 所示。

从实验结果看,遍历的实例集合与本体库的 151 655个实例来说,体积已经有很大的缩小,为两两遍历提供了可能性。下面将对此识别实例集合计算出实例间的相似度。

表 3 各属性的重复度

属性编号	属性名称	属性的重复度
0	address	0
1	booktitle	0.961 0
2	cdrom	0.097 0
3	chapter	0.0
4	cite	1.0
5	creator	0.590 6
6	crossref	0.989 7
7	date	0.994 7
8	editor	0.303 0
9	identifier	0.0
10	isbn	0.0
11	journal	0.998 5
12	month	0.972 5
13	note	0.0
14	number	0.998 0
15	pages	0.736 8
16	publisher	0.666 7
17	rating	0.950 8
18	reviewid	0.015 9
19	school	0.0
20	series	0.666 7
21	title	0.020 1
22	type	1.0
23	volume	0.967 4
24	year	0.999 6

表 4 识别实例集合

识别属性集合	实例个数
{cdrom, creator, pages, title}	1 728

3.5 计算出实例间的相似度

通过公式(4)计算得出实例对间的相似度,通过对实验数据的分析发现多对实例间的相似度是一样的。即可以通过相似度将这些实例对作为一个集合。同一集合的实例对就是要找的冗余实例。下面整理出相似度比较高的实例对集合,给出集合的相似度和集合的实例对个数,如表 5 所示。

表 5 冗余实例集合的相似度及集合数目

实例集合相似度	集合数目
0.898 9	1
0.894 5	61
0.800 7	49
0.798 6	2
0.794 6	6
0.794 3	71

这里就不一一列举。从实验结果看,文中的方法是可行的。实验达到了文中的目标,即找出了冗余的实例。

4　结束语

文中的方法和实验没有用到本体库的语义和层次关系,因为此次实验主要关注的是实例之间的冗余,可能冗余实例是跨概念和跨层次的,所以简单地对实例进行两两相似度的计算来找出冗余的实例。但是文中没有机械地将所有的实例进行相似度的计算,这样做是不可行的,因为数据太大。所以提出按照信息上的加权即属性出现的频数和属性字符串值的重复度来计算出一个可识别的属性集合,通过可识别的属性集合来得到可识别的实例集合。从实验结果看,它大大缩小了识别实例集合规模。由实验结果可以说明,文中的方法可以找出相似度较高的实例,是行之可效的;但是还是比较粗放,而且可识别实例集合还不是最优的。下面的工作是如何改进上面的方法使可识别实例集合最优并使集合实例间的重复度最大,并且将这些冗余的实例整合为唯一的实例。

参考文献:

[1] 滕广青,毕　强. 国外本体协调研究前沿进展及热点分析[J]. 中国图书馆学报,2012(1):113-120.

[2] 沈亦军,吕　刚. 基于实例相似度的本体映射方法研究[J]. 重庆科技学院学报(自然科学版),2012,14(3):170-172.

[3] Leacock C,Chodorow M. Combining local context and WordNet similarity for word sense and WordNet similarity for word sense identification[M]//WordNet:an electronic lexical database. Cambridge,MA:MIT Press,1998.

[4] 李文超,杨妮妮. 基于本体的语义相似性研究[J]. 科学技术与工程,2012,20(21):5328-5330.

[5] Elmeleegy H,Elmagarmid A K,Lee J. Leveraging query logs for schema mapping generation in U-MAP[C]//Proceedings of the ACM SIGMOD international conference on management of data. Athens,Greece:[s. n.],2011.

[6] 徐德智,黄　旭. 一种基于冗余消除的本体映射后处理方法[J]. 计算技术与自动化,2012,31(3):88-91.

[7] RDF 教程[S/OL]. 2013-09-11. http://w3school. com. cn/rdf/index. asp.

[8] 李　华,苏　乐. 基于关联规则的本体相似度的综合计算方法[J]. 计算机应用,2012,32(9):2472-2475.

[9] Bucccella A,Cechich A,Gendarmi D,et al. Building a global normalized ontology for integrating geographic data sources[J]. Computers & Geosciences,2011,37(7):893-916.

[10] 刘秀磊,廖建新,朱晓民,等. 本体匹配中基于词义组合的词法分析算法[J]. 电子学报,2012,40(8):1624-1630.

[11] 曹泽文,钱　杰,张维明,等. 一种综合的概念相似度计算方法[J]. 计算机科学,2007,34(3):174-175.

[12] 刘宏哲. 一种基于本体的句子相似度计算方法[J]. 计算机科学,2013,40(1):251-256.

[13] The DBLP computer science bibliography[S/OL]. 2013-09-11. http://www. informatik. uni-trier. de/ ~ley/db.

[14] Chodorow K,Diroff M. MongoDB 权威指南[M]. 程显峰,译. 北京:人民邮电出版社,2011.

(上接第 10 页)

[2] Natu M,Sethi A S. Active probing approach for fault localization in computer networks[C]//Proc of 4th IEEE/IFIP workshop on end-to-end monitoring techniques and services. [s. l.]:IEEE,2006:25-33.

[3] 褚灵伟,邹仕洪,程时端,等. 概率和噪声环境下基于主动探针的 Internet 服务故障管理[J]. 中国科学:E 辑,2008,38(10):1733-1746.

[4] Lin A. A model-based automated diagnosis algorithm[M]//Methodology and Tools in Knowledge-Based Systems. Berlin:Springer,1998.

[5] Huang Xiaohui, Zou Shihong, Wang Wendong, et al. Fault management for Internet services: modeling and algorithms[C]//Proc of IEEE international conference on communications. Istanbul:IEEE,2006:854-859.

[6] 张顺利,邱雪松,孟洛明. 网络虚拟化环境下的服务故障诊断算法[J]. 软件学报,2012,23(10):2772-2782.

[7] Liao J,Zhang C,Li T,et al. A quasi-optimal probabilistic fault localization algorithm in communication networks[J]. Chinese Journal of Electronics,2011,20(1):151-154.

[8] 李　晶,朱　敏. 一种基于事件驱动的 SOA 故障疑似集选择算法[J]. 计算机应用与软件,2011,28(5):181-183.

[9] 杜晓丽,朱程荣,熊齐邦. 一种基于依赖图的故障定位算法[J]. 计算机应用,2004,24(B12):67-69.

[10] 范贵生,虞慧群,陈丽琼,等. 基于 Petri 网的服务组合故障诊断与处理[J]. 软件学报,2010,21(2):231-247.

[11] Bagchi S,Kar G,Hellerstein J. Dependency analysis in distributed systems using fault injection:application to problem determination in an e-commerce environment[C]//Proc of 12th international workshop on distributed systems: operations & management. [s. l.]:[s. n.],2001:15-17.

[12] 唐　渊,金可音,周　昆,等. Web 服务失败分类法[J]. 湖南工业大学学报,2009,23(2):73-76.

[13] 刘　丽,况晓辉,方　兰,等. Web 服务故障的分类方法[J]. 计算机系统应用,2010,19(8):258-263.

[14] Ide J S,Cozman F G,Ramos F T. Generating random Bayesian networks with constraints on induced width[C]//Proc of European conference on artificial intelligence. [s. l.]:[s. n.],2004:323-334.

基于加权的冗余相似本体实例发现的研究

作者： 卢传耀， 徐敏， LU Chuan-yao， XU Min
作者单位： 南京航空航天大学 计算机科学与技术学院,江苏 南京,210016
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(9)