一种基于同义词词林的中文大规模本体映射方案

王 汀 邸瑞华 李维铭
(北京工业大学计算机科学与技术学院 北京 100124)

摘 要 本体映射是解决本体异构问题的重要途径和手段,中文知识是网络开放知识库的重要组成部分,但现有的中文本体映射系统在面对大规模本体映射任务时,显得效率较低且可用性不高,目前仍缺乏针对中文大规模本体映射的相关系统。为了解决中文大规模本体的映射问题,设计并实现了一个面向中文的大规模本体映射系统。首先,提出了一种基于拟核力场势函数的大规模本体压缩方法;其次,提出了一种基于同义词词林的中文概念等价关系确定算法;再次,实现了大规模中文本体映射的原型系统;最后,将本系统与相似度计算相关典型算法进行比较,证明其具备一定的可用性和较高的总体性能。

关键词 语义网,本体,本体映射,同义词词林,相关度计算
中图法分类号 TP301.6 **文献标识码** A

Tongyici Cilin-based Mapping Approach for Large-scale Chinese Ontology

WANG Ting DI Rui-hua LI Wei-ming
(College of Computer Science and Technology, Beijing University of Technology, Beijing 100124, China)

Abstract Ontology mapping is an important way to solve the problem of the ontology heterogeneous. The Chinese knowledge is an important part of the online knowledge base, but there is still lack of relevant system for Chinese large scale ontology mapping. While facing the large-scale ontology mapping task, current Chinese ontology mapping system has low efficiency and its availability is not high. In order to solve the Chinese ontology mapping problem, this paper put forward a kind of Tongyici Cilin-based ontology mapping architecture, firstly presented a data field-based method to compress the Chinese large-scale ontology, secondly purposed a Tongyici Cilin-based algorithm for the Chinese equivalent-class discovery, thirdly implemented a prototype system for Chinese large-scale ontology mapping, finally compared the system with the other typical similarity computing algorithm. The result proves that it has higher overall performance and usability.

Keywords Semantic Web, Ontology, Ontology mapping, Tongyici cilin, Similarity computing

1 引言

语义Web的愿景是建立"数据之网",以使机器能够理解网络上的语义信息[1]。本体作为语义Web的核心元素,是描述特定领域共享概念的形式化、规范化说明[2],是实现网络知识共享和语义互操作的关键。但是,由于不同本体之间存在异构性,导致了本体间的重用和共享变得困难。本体映射的任务就是要发现异构本体间的语义关联。

由于文化和背景原因,目前尚缺乏成熟的针对中文语言的本体映射系统。而随着语义网的发展,大规模的中文语言描述的本体也越来越多地被构建和共享出来。因此,本文主要探讨如何解决面向中文的大规模本体映射的问题。

2 相关工作

目前针对本体映射问题,国内外研究人员已经提出了多种映射方法和典型系统,例如:

文献[3]中列出了基于编辑距离和基于Token的几种典型元素级相似度计算算法,并对几种算法的性能进行了评测。

Melnik S等[4]提出了一种结构级本体映射算法:Similarity flooding,该系统利用本体的概念体系构造相似度传播图,并对概念之间的相似度进行传播和修正。

Zhong Qian等[5]开发了RiMOM系统,该系统基于本体实例、概念名称以及本体结构等特征的多策略映射方式,并通过引入普适的场论思想,使其适用于大规模本体的映射任务。但其缺乏针对中文特定的语言特点的优化。

Giunchiglia F等[6]提出基于语言学方法,并引入共享知识词典(如:WordNet),利用语言关系来进行语义关系发现。

文献[7]提出一种实例级的本体映射算法,它根据本体概念的公共实例数量来度量概念之间的相似度。

在中文本体方面,Wang Zhi-chun等学者[8]提出基于中文百科的分类体系抽取概念间的层次关系,获取含有Infobox的词条Web页面中的概念属性及百科词条实例,最终建立起

到稿日期:2013-08-15 返修日期:2013-12-23 本文受北京市科学技术研究院创新团队计划项目(IG201203C1),校企联合研究项目:IBM Shared University Research资助。

王 汀(1985-),男,博士生,主要研究领域为语义网、分布式计算,E-mail:S200807007@emails.bjut.edu.cn;邸瑞华(1947-),女,教授,博士生导师,主要研究领域为分布式计算、网格计算;李维铭 男,高级工程师,主要研究领域为分布式计算、网络通信技术。

基于百度百科和互动百科的两大中文大规模本体库。

Yidong Chen 等[9]提出利用中文百科 Infobox 中的属性-值对信息，自动提取良构的训练样本，进而基于统计学习模型从百科的非结构化文本中抽取海量的知识三元组，最终构建了一个面向开放域的中文知识库。

Li Jia 等[10]提出了一种基于知网(Hownet)的元素层概念相似度计算的方法，并实现了一个中文本体映射系统，该系统在面对大规模本体映射任务时，其适用性有待验证。Tian Jiu-le 等提出了一种基于同义词词林的中文词语语义相似度计算算法[11]，但其成果并未在语义网环境下应用。

总之，目前发布在 Web 上的中文大规模本体仍然较少，且存在较大的异构性，而现有的中文本体映射系统在面对大规模本体映射任务时，显得效率较低且可用性不高。目前仍缺乏针对中文的，且适应大规模本体映射任务的相关系统。为了解决中文大规模本体的映射问题，基于同义词词林设计并实现了一个面向中文的大规模本体映射系统。

3 问题定义

定义 1(知识三元组) 一个事实性陈述可以被表示为一个语义 Web 三元组(〈主语，谓语，宾语〉)，因此语义数据可以被形式化为一个三元组集合。知识三元组集合 $K\subseteq S\times P\times O$，其中，$S$ 是主语(Subject)集合，P 是谓词(Predicate)集合，O 是宾语(Object)集合。

定义 2(本体) 本体是对一个特定领域中重要概念的共享的形式化描述。一个本体模型可以由一个四元组来描述：$O=\{C,P,H^C,H^P\}$。其中 C 和 P 分别代表本体中的概念和属性集合，而 H^C 和 H^P 分别表达了概念和属性集合中元素之间的层次化语义关系。

定义 3(本体映射) 两个待映射本体 O_s 和 O_t，对于 O_s 中的每个概念 C_s，在本体 O_t 中找到与其语义相同或接近的对应概念 C_t，即：

定义映射函数 $map:O_s\rightarrow O_t$：

对于 $\forall C_s\in O_s$，$\forall C_t\in O_t$，若 $\mathrm{sim}(C_s,C_t)>t$，则有 $map(C_s)=C_t$。

其中 $\mathrm{sim}(C_s,C_t)$ 是待映射概念 C_s 和 C_t 的相似度，t 是阈值，表示当概念 C_s 与概念 C_t 的语义相似度大于 t 时，则将〈C_s，C_t〉作为已发现的概念映射对。

4 中文大规模本体映射系统

该系统主要由 3 大功能模块组成，分别是：本体概念初始相似度计算、本体压缩和确定性映射，下面将对每个模块作详细阐述。

4.1 基于编辑距离的初始相似度计算

编辑距离，又被称为 Levenshtein 距离，由俄罗斯科学家 Vladimir Levenshtein 提出。

在面对大规模本体的映射任务时，我们提出首先对待映射本体进行压缩。具体地，我们采用编辑距离算法首先进行概念集合之间的初始相似度计算。这是因为，在进行初始相似度计算时，往往考虑算法的高效性，而其精确性则被看作是次要因素。

也就是说，在获得待映射本体的初始相似度时，我们只考虑概念之间的字面相似性，而忽略其语义相关性。

特别地，对于两个概念 C_s 和 C_t，它们的编辑距离的值以及它们的相似度值可以由式(1)和式(2)来刻画：

$$\mathrm{EditDistance}(C_s,C_t)=\frac{|Do(C_s,C_t)|}{\max(L(C_s),L(C_t))} \tag{1}$$

其中，$|D_o(C_s,C_t)|$ 为待映射概念 C_s 和 C_t 的编辑操作次数，$L(C_s)$ 和 $L(C_t)$ 为待映射概念的字符长度。

$$\mathrm{SIMe}(C_s,C_t)=\frac{1}{(1+\mathrm{EditDistance}(C_s,C_t))} \tag{2}$$

其中，$\mathrm{SIMe}(C_s,C_t)$ 为待映射概念 C_s 和 C_t 的相似度。

具体的初始相似度描述见算法 1。

算法 1 InterlinkingValue(O_s,O_t)

```
Input:待映射源本体 Os 和目标本体 Ot
Output:Os 和 Ot的初始关联度哈希表 Map<Concept,Value>
1. get Concepts list from source Ontology Os,defined as List_Os
2. get Concepts list from target Ontology Ot,defined as List_Ot
3. for each Concept Cs in List_Os
4.    for each Concept Ct in List_Ot
5.        get interlinking value by using function EDvalue = EditDis-
          tance(Cs,Ct)
6.        Double curValue_Cs=Map_Os. get(Cs)
7.        Double curValue_Ct=Map_Ot. get(Ct)
8.        if EDvalue≥γ
9.        then Map_Os. put(Cs,curValue_Cs+EDvalue),Map_Ot. put
          (Ct,curValue_Ct+EDvalue)
10. for each Concept Cs in List_Os
11.   Double initialValue_Cs=Map_Os. get(Cs)
12.   if initialValue_Cs==0
13.   then Map_Os. put(Cs,0.06)
14. for each Concept Ct in List_Ot
15.   Double initialValue_Ct=Map_Ot. get(Ct)
16.   if initialValue_Ct==0
17.   then Map_Ot. put(Ct,0.06)
18. return Map_Os,Map_Ot
```

其中，γ 为待映射概念的初始相似度阈值，根据实验经验，系统中 γ 的取值为 0.45。也就是说，如果两个概念的编辑距离相似度小于 0.45，则该结果将不累加进入最终的初始相似度值。

对于某个概念最终的初始相似度值为零的情况，我们赋予其一个固定值：0.06。这样可使相似度的转移具有连续性。

4.2 大规模本体压缩算法

在面对大规模的本体映射任务时，传统的算法无论在时间还是空间复杂度方面都难以适应，因此需要相应的策略来对原始的待映射本体进行压缩。

数据场理论[12]的提出是基于物理学中的场论思想，将数域空间中数据之间的相互关系抽象为物质粒子之间的相互作用问题，最终形式化为场论的描述方法。该理论通过势函数来表达不同数据间的相互作用关系，从而体现出数据的分布特征，并根据数据场中的等势线结构来对数据集进行聚类划分。

4.2.1 势函数的定义

由于短程场能更好地反映出数据之间的相互作用情况，因此采用拟核力场势函数。

在本体映射问题中的具体定义如下：

已知本体概念集 $O=\{C_1,C_2,\cdots,C_n\}$，每个概念 C_i 的质

量为算法 1 得到的初始相关度值，也就是本体中的每个概念对于其他概念的影响程度，即：$M=\{initialValue\_C_1, initialValue\_C_2, \cdots, initialValue\_C_n\}$。

待映射本体 O 中，概念间的最短路径长度为 $\| C_i - C_j \|$，由于短程场的特性，因此我们定义概念之间的路径长度不大于 2。

则由数据场理论，我们得到概念 C_i 与 C_j 之间相互作用的势函数表达式，如式(3)所示：

$$\varphi_{C_j}(C_i) = initialValue\_C_i \times e^{-(\frac{\| C_i - C_j \|}{\delta})^k} \tag{3}$$

其中，$\delta \in (0, +\infty)$ 反映概念之间影响的粒度，也称为缩放因子，不妨取 $\delta=1, k=2$。这样，就得到了待映射本体 O 中每个概念 C_i 的势值函数表达式，如式(4)所示：

$$\varphi_O(C_i) = \sum_{j=1}^{n} \varphi_{C_j}(C_i) \tag{4}$$

最终得到待映射本体 O 中全部概念的势值集合 *potentialMap_O*$_s$ 和 *potentialMap_O*$_t$。势值集合的定义为：*potentialMap_O*$\langle C_i, \varphi_o(C_i) \rangle$。

4.2.2 本体压缩策略

为了对待映射本体 O 进行压缩，将 O 中的概念集合划分为两部分，称为：候选区和淘汰区。

具体地，对于执行算法 1 后得到的输出数据结构 *Map_O*$_t$ 和 *Map_O*$_s$，根据每个概念元素的键值统计出 *Map_O*$_t$ 和 *Map_O*$_s$ 中键值为 0.06 的概念总数称为 Range_Out，该变量即为淘汰区的区间大小。相应地，*Map_O*$_t$ 和 *Map_O*$_s$ 中键值不等于 0.06 的概念总数称为 *Range_Candidate*，该变量定义为候选区的区间大小。

对于势值集合 *potentialMap_O*$_t$ 和 *potentialMap_O*$_s$ 中的概念元素，根据键值进行降序排序，对于 $\forall C_i \in$ *potentialMap_O*，其排名用变量 $Rank_i$ 来标识，若 $Rank_i \in [1, Range\_Candidate]$，则概念 C_i 将被作为候选概念得到保留。同时，若 $Rank_i \in [Range\_Candidate+1, Range\_Out]$，则概念 C_i 将被淘汰。经过初步试验，该策略所产生的本体压缩比约在 35%至 60%之间。实验本体的相关信息见 5.1 节。

4.3 基于同义词词林的确定性映射

4.3.1 同义词词林简介

同义词词林是一个中文同义词典，它将每个词汇进行编码并以层次关系组织在一个树状结构中，树中的每个结点代表一个概念，而中文的概念共指关系识别，实际上可以抽象为中文同义词的识别相似度计算问题，因此同义词词林是最佳的选择。我们采用哈工大同义词词林扩展版[13]作为中文本体映射关系抽取的常识知识库。

同义词词林将词元组织为分层结构，自顶向下共有 5 层。每个层次都有相应的编码标识，5 层的编码从左至右依次排列起来，形成词元的词林编码。词语与词语之间隐含的语义相关度也随着层次的增加而提高。

下面以词元“物质”为例(词林编码为：Ba01A02=)，对词林编码格式进行解释，如表 1 所列。

表 1 词林编码示例

编码位	1	2	3	4	5	6	7	8
子编码	B	a	0	1	A	0	2	"=(=或#或@)"
含义	大类	中类	小类		词群	原子词群		同义\不等\孤立
层次	第 1 层	第 2 层	第 3 层		第 4 层	第 5 层		

4.3.2 确定性映射算法描述

根据同义词词林的结构特点，首先对待映射概念的词林编码进行解析，抽取出第 1 至第 5 层子编码，然后从第 1 层子编码开始比较。若出现子编码不同，则根据出现的层次来赋予该映射对相应的相似度权重。不同子编码出现的层次越深，则相似度权重越高，反之则越低。同时，每层的分支节点数的多少也对相似度有影响。

我们给出基于同义词词林的相似度计算：

$$SIM_T(C_s, C_t) = \lambda \times \frac{L_i}{|L|} \times \cos(N_T \times \frac{\prod}{180}) \times (\frac{N_t - D + 1}{N_t}) \tag{5}$$

由于本体映射任务更关注概念之间的语义相似性，因此我们引入调节参数：语义相关度因子 λ，通过 λ 来调节不同层级概念间语义相关性和语义相似性的关系以及控制处于不同层次分支的词元之间可能相似的程度，显然 $\lambda \in (0,1)$。λ 的值越大，表示不同层次之间的词元相似或等价的可能性越大，且不同层次的语义相关性对于最终概念相似度的影响越大，反之则越小。

特别地，在面对中文本体映射任务时，由于更突出概念间的语义相似度，因此 λ 的取值不宜过高。

其中，$L=\{1,2,3,4,5\}$，对于 $\forall L_i \in L$，L_i 为第 i 层所代表的层次数。概念相似度权重系数为 $\lambda \times (L_i / |L|)$。$N_T$ 为词元 C_s 和 C_t 在第 i 层分支上的节点总数，D 为词元 C_s 和 C_t 的编码距离。

例如，当 $\lambda=0.7$ 时，具体每层分支的权重分配如下：

①若待映射概念对的第 1 层子编码不同，则权重系数为 0.14；

②若待映射概念对的第 2 层子编码不同，则权重系数为 0.28；

③若待映射概念对的第 3 层子编码不同，则权重系数为 0.42；

④若待映射概念对的第 4 层子编码不同，则权重系数为 0.56；

⑤若待映射概念对的第 5 层子编码不同，则权重系数为 0.70。

特别地，若待映射概念对的 5 层编码均相等，且词林编码最后一位为“=”号，则相似度函数 SIM_T 的返回值为 1.0。显然，函数 SIM_T 的值域为(0,1]。

5 实验结果及分析

5.1 实验数据

相比于国际通用的本体及其映射任务，例如：OAEI(Ontology Alignment Evaluation Initiative)等国际组织发布的多领域标准本体及其映射的基准评测指标，现有开源的中文大规模本体仍然较为匮乏。

因此，本文采用中文网络开放百科知识库作为实验数据源，通过使用爬虫工具包 HTMLParser[14]，将中文网络开放百科中的分类体系进行解析，并将其以中文字符三元组的形式组织起来，最终形成了待映射本体概念集合。

具体地，本文分别对百度百科[15]和互动百科[16]的开放分类页面进行爬取和解析，构建了包含有 1380 个概念的百度百科本体框架以及含有 29263 个概念的互动百科本体框架。

其中，百度百科分类体系中的顶层分类包括：人物、科学、历史、体育等12大类；而互动百科包括人物、技术、热点话题等13大顶级分类。

5.2 评价指标

本文采用对属性值识别的准确率(Precision)、召回率(Recall)和F-measure作为最终的评价标准。其中：

Precision(P)＝输出的正确映射对数/输出的映射对总数×100%

Recall(R)＝输出的正确映射对数/标准结果中的映射对总数×100%

F-measure($F1$)＝$2\times P\times R/(P+R)\times 100\%$

对于大规模中文本体映射任务，选取百度百科和互动百科本体概念集中三大顶层分类：人物、科学和社会子类中的正确映射对，并以此作为评价算法效率的参考映射。

具体信息见表2。

表2 Baidu-Hudong本体参考映射统计

顶层分类	参考映射对数	评价指标
人物	57	查准率和查全率
科学	62	查准率和查全率
社会	60	查准率和查全率

5.3 实验结果

将本文系统与编辑距离相似度算法和同义词词林相似度算法进行比较，查准率评测结果如图1所示，查全率评测结果如图2所示，F1值如图3所示。

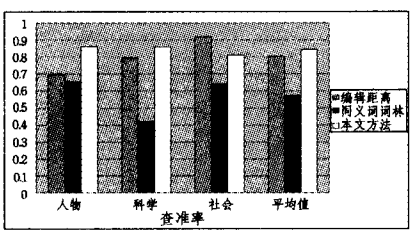

图1 查准率评测结果

从图1中可以看到，本文系统在查准率方面比编辑距离相似度算法平均提高约4%，这主要是由于我们引入了结构级映射，从而在一定程度上避免了单纯的编辑距离算法所带来的高匹配、低精度的现象。同时本系统的查准率也高于文献[11]中的算法，这是因为本体映射问题更注重概念之间的共指关系识别，而文献[11]算法过分关注词语之间的语义相关度，而这就导致了在进行词语相似度计算时引入较大误差。

在查全率方面，由于引入同义词词林作为知识库，因此查全率也高于编辑距离相似度算法，结果如图2所示。

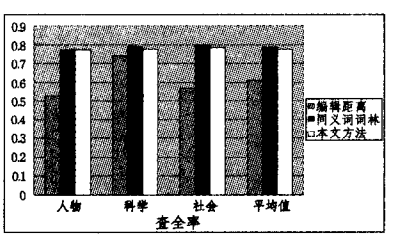

图2 查全率评测结果

最后，在总体性能上(F1值)，本文系统比编辑距离算法和同义词词林相似度算法平均高出约11%和15%，如图3所示。

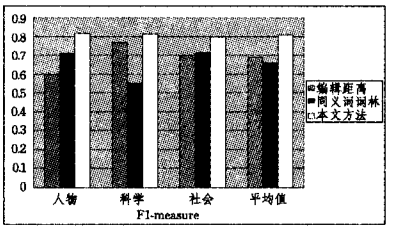

图3 系统总体性能

结束语 现阶段缺乏成熟的中文大规模本体映射系统，设计并实现了一个基于同义词词林的中文本体映射原型框架，该系统解决了大规模本体映射系统的可用性问题。它着眼于现有中文大规模本体的特征，进行元素级和结构级映射。今后将根据不同中文本体的特征，考虑引入实例级以及概念定义相似度的映射参数，以进一步提高中文映射系统的健壮性和准确性。

参 考 文 献

[1] Berners-Lee T. Semantic Web Road map [OL]. http://www.w3.org/DesignIssues/Semantic.html,1998

[2] Borst W N. Construction of Engineering Ontologies for Knowledge Sharing and Reuse[D]. Enschede: University of Twente, 1997

[3] Cohen W, Ravikumar P, Fienberg S. A comparison of string distance metrics for name-matching tasks[C]//Proceedings of the IJCAI Workshop on Information Integration on the Web (IIWeb). Acapulco, Mexico, 2003: 73-78

[4] Melnik S, Garcia-Molina H, Rahm E. Similarity flooding: A versatile graph matching algorithm and it s application to schema Matching[C]//Proceedings of the 18th International Conference of Data Engineering(ICDE). San Jose, California, 2002: 117-128

[5] Zhong Q, Li H, Li J, et al. A gauss function based approach for unbalanced ontology matching[C]//Proceedings of the 28th International Conference on Management of Data (SIGMOD). Rhode Island, USA, 2009: 669-680

[6] Giunchiglia F, Yat skevich M. Element level semantic matching [D]. Italy: Dept. of Information and Communication Technology University of Trento, 2004

[7] Isaac A, Meij L, Schlobach S, et al. An empirical study of instance-based ontology matching[C]//Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference(ISWC/ASWC). Busan, Korea, 2007: 253-266

[8] Wang Z, Wang Z, Li J, et al. Knowledge extraction from chinese wiki encyclopedias[J]. Journal of Zhejiang University-Science C, 2012, 13(4): 268-280

[9] Chen Yi-dong, Chen Li-wei, Xu Kun. Learning Chinese Entity Attributes from Online Encyclopedia[C]//APWeb. 2012: 179-186

[10] 李佳，祝铭，刘辰，等. 中文本体映射研究与实现[J]. 中文信息学报，2007，21(4)：27-33

[11] 田久乐，赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报，2010，28(6)：602-608

[12] 李德毅，杜鹢. 不确定性人工智能[M]. 北京：国防工业出版社，2005

[13] 梅家驹，等. 同义词词林[M]. 上海：上海辞书出版社，1993

[14] http://htmlparser.org/

[15] http://baike.baidu.com/

[16] http://www.hudong.com/