第34卷 第15期 计 算 机 工 程 2008年8月
Vol.34 No.15 Computer Engineering August 2008
·软件技术与数据库· 文章编号：1000—3428(2008)15—0057—03 文献标识码：A 中图分类号：TP311

基于 Top-*k* 映射的本体匹配方法

王 颖，刘 群，张 冰
(哈尔滨工程大学计算机科学与技术学院，哈尔滨 150001)

摘 要：针对本体之间的异构问题，提出一种基于 Top-*k* 映射的本体匹配方法。该方法是对现有匹配方法的一种扩展，它以相似度计算为基础，从元素级和结构级计算 2 个概念之间的相似度，并在匹配过程中同时产生 *k* 个映射而不是一个最佳映射。实验结果表明，该算法在查全率和查准率方面都有很好的表现，并且其查准率要优于 GLUE 方法。

关键词：本体；本体匹配；相似度；Top-*k* 映射

Ontology Matching Method Based on Top-*k* Mapping

WANG Ying, LIU Qun, ZHANG Bing
(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001)

【Abstract】 Aiming at the heterogeneity problem between ontology, an ontology matching method based on Top-*k* mapping is put forward. This is an extension of existing methods. It takes similarity computation as a basis of the algorithm by calculating the similarity of two concepts at element and structure level. Moreover, it can generate *k* mapping simultaneously not one best mapping. Experimental shows that the method has advantages in the recall and precision, and excels GLUE in precision.

【Key words】 ontology; ontology matching; similarity; Top-*k* mapping

1 概述

本体作为一种领域知识概念化和模型化的方法已经获得广泛认可。随着本体应用领域的增多，如何解决本体间的互操作成为一个比较棘手的问题。近几年，国内外的研究学者和实验室对本体匹配的方法和过程进行了大量的研究，并且已经开发出来一些具体的本体匹配系统，比较著名的有 Similarity Flooding, S-Match, GLUE 等。通过匹配各种各样的本体，解决了语义 Web、Agent 通信、Web 服务集成、P2P 数据库集成及个人信息共享等中存在的一些问题。然而，大多数匹配系统只能产生 2 个本体元素之间的一个最佳映射，在匹配相似程度较低的复杂本体时，精确度仍然不高。因此，本文提出了一种基于 Top-*k* 映射的本体匹配方法。该方法避免了阈值选取的不足，生成 Top-*k* 映射，也就是 *k* 个相似度最高的映射，提高了匹配的精确度。

2 基本概念

为了叙述方便，下面介绍本文用到的概念和定义。

概念 1(本体) Gruber 定义了一个典型的本体由五元组表示[1]：$O=(C,I,R,F,A)$。其中，C 代表概念集合；I 表示概念的实例；R 为定义在概念集合上的关系集合；F 为定义在概念集合上的函数集合；A 表示公理集合。

概念 2(相似度) 一般用相似度来度量本体之间概念的匹配程度。当 2 个概念具有某些共同特征时，则定义它们是相似的。形式上，相似度计算满足[2]：

(1) $sim(x,y)\in[0,1]$，相似度的计算值为[0,1]区间中的实数。

(2) $sim(x,y)=1$，当且仅当 $x=y$。如果 2 个对象完全相同，则相似度为 1。

(3) $sim(x,y)=0$，如果 2 个对象没有任何共同特征，那么其相似度为 0。

(4) $sim(x,y)=sim(y,x)$，相似关系是对称的。

定义 1(映射函数) 本体 O_1 和 O_2 的映射是指 2 个本体之间概念的对应关系(相等或语义相似)。本文将映射函数定义如下：

$$map:\{e_{i1}\}\rightarrow\{e_{j2}\}\quad e_{i1}\in O_1,\ e_{j2}\in O_2$$

$$map(\{e_{i1}\})=\{e_{j2}\}$$

本体 O_1 中的概念集合 $\{e_{i1}\}$ 被映射为 O_2 中的概念集合 $\{e_{j2}\}$；概念集合可以包含本体中的一个概念、多个概念或者为空。

3 相似度计算

3.1 元素级相似度

使用概念名称计算相似度是发现映射关系最直接也是最基本的方法。本文从词形和词义两个方面考察概念名称的相似度，使用编辑距离来计算词形相似度。编辑距离是由 Levenshtein 提出的一个用来衡量 2 个字符串(后来扩展到语句)差别的方法[3]。它用一个动态规划算法计算把一个字符串转换成另一个字符串所需要对字符进行的最小操作数，包括对字符的插入、删除、替换。计算式如下：

$$sim_{dis}(e_{i1},e_{j2})=1-\frac{distance(e_{i1},e_{j2})}{\max(|e_{i1}|,|e_{j2}|)}$$

其中，$sim_{dis}(e_{i1},e_{j2})$ 表示 2 个概念之间的编辑距离相似度，e_{i1} 和 e_{j2} 分别表示本体 O_1 和 O_2 的概念，$distance(e_{i1},e_{j2})$ 表示两个字符串的编辑距离。

词形相似度计算忽略了一个问题，即 2 个概念可能在拼

作者简介：王 颖(1982－)，女，博士研究生，主研方向：本体，语义 Web；刘 群，教授、博士生导师；张 冰，博士研究生

收稿日期：2007-08-30 **E-mail**：yingwang@hrbeu.edu.cn

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

写上完全不同，但其意义却可能很相似，因此，还要从词义角度计算概念之间的相似度。需要借助于外部词典 WordNet。WordNet 与传统词典的不同之处在于它是依据词义而不是词形来组织词汇信息。它将所有词组织在树状的层次结构中，其中每一个节点 s 表示一个词义，节点中保存了多个同义词或者短语，每个单词或短语又可以存在于多个语义节点中(即表明该单词有多个词义)。可以利用路径长度的方法计算相似度，如果从一个节点到另外一个节点的路径越短，表明他们相似的程度越高，反之越小。Lin 等人提出利用概率的方法计算 2 个概念的相似度[4]。

$$sim_{wn}(e_{i1},e_{j2})=\frac{2\times\lg(p(s))}{\lg(p(s_1))+\lg(p(s_2))}$$

其中，$sim_{wn}(e_{i1},e_{j2})$ 表示利用 WordNet 得到的相似度；$p(s)=count(s)/total$ 表示 WordNet 中词义节点 s 及其子节点所包含的单词个数在整个辞典中所占的比例，$total$ 是 WordNet 的单词总数。另外 $e_{i1}\in s_1$，$e_{j2}\in s_2$ 表示单词 e_{i1} 和 e_{j2} 分别位于节点 s_1 和 s_2 中。节点 s 是 s_1 和 s_2 的公共祖先节点。

通过上述分析，综合 2 种方法的优点，给出计算元素级相似度表达式如下：

$$sim_{el}(e_{i1},e_{j2})=\alpha\cdot sim_{dis}(e_{i1},e_{j2})+\beta\cdot sim_{wn}(e_{i1},e_{j2})$$

其中，$\partial,\beta\in[0,1]$，$\partial+\beta=1$。

3.2 结构级相似度

事实上，在概念间的层次结构、语义邻居关系中蕴涵了大量的潜在语义，对相似度的影响很大，映射过程中必须予以考虑。本体中的概念是分层的，因此本体也可以看成一棵概念树，树中每个结点代表一个概念，树中也存在子概念、父概念和兄弟概念。以树的一些性质、一些领域公理和领域专家所定义的一些规则为依据，定义了如下一些启发规则：

(1)如果 2 个概念的属性都相同，那么这 2 个概念可能是相同的。

(2)如果 2 个概念具有相同的实例，那么这 2 个概念可能是相同的。

(3)如果 2 个概念的子概念在一定程度上也相似，那么这 2 个概念是相似的。

(4)如果 2 个概念的父概念相似，那么这 2 个概念也可能相似，并且这 2 个概念的部分子概念也可能相似。

(5)如果某个概念的兄弟概念结点与某一概念 X 相似，那么该概念与概念 X 也可能相似。

依据上述规则，当 2 个概念没有共享的信息(属性、实例，子概念，父概念，兄弟概念)时，可以将其相似值定义为 0。如果所有的信息均相同，其相似值为 1。本文采用如下公式：

$$sim_i(e_{i1},e_{j2})=\frac{\sum_{a\in A}\sum_{b\in B}sim(a,b)}{|A|\cdot|B|}$$

其中，A 和 B 分别表示个概念 e_{i1} 和 e_{j2} 的属性集、实例集、子概念集、父概念集或兄弟概念集；$|A|$ 和 $|B|$ 表示集合中的元素个数。则 2 个概念 e_{i1} 和 e_{j2} 的结构级相似度为

$$sim_{str}(e_{i1},e_{j2})=\sum_{i=1}^{5}w_i sim_i(e_{i1},e_{j2})$$

其中，w_i 为权值。

3.3 相似度合并

通过权重法将元素级相似度和结构级相似度进行合并得到综合相似度，即

$$sim(e_{e1},e_{j2})=w_{el}\times sim_{el}+w_{str}\times sim_{str}\ \text{且}\ w_{el}+w_{str}=1$$

4 Top-k 映射

一个匹配系统要获得最优映射需面对 2 个问题：(1)匹配系统应能为用户提供正确的映射；(2)应该避免不正确的映射。把正确的映射同不正确的映射分开是一项困难的工作。阈值是一种经常使用的技术，依赖事先设定的阈值，任何相似度高于阈值的元素对被认为是匹配的；反之，则不匹配。尽管如此，阈值也只能使用在分界线比较清晰的情况之下。当相似程度无法明显地区分出来，阈值的选取就变得很困难。阈值选取过大，可能无法得到正确的匹配结果，可靠性下降；阈值选取过小，则会带来许多不必要的计算，速度下降。因此，本文提出了基于 Top-k 映射的本体匹配方法。

Top-k 映射可以递归定义如下[5]：

对任意 $i>0$，令 M_i^* 表示第 i 个最佳映射。当 $k=1$ 时，第 k 个最佳映射 M_1^* 为候选映射 M 中相似度最高的映射。$k>1$ 时，若已知 $k-1$ 个最佳映射 M_1^*，M_2^*，…，M_{k-1}^*，则第 k 个最佳映射 M_k^* 可以被定义为候选映射 M 中除 $M_1^*,M_2^*,\cdots,M_{k-1}^*$ 之外的最佳映射。因此，Top-k 映射定义为，对任意 $M_i\subseteq M$，使得 $M_i\notin\{M_1^*,M_2^*,\cdots,M_k^*\}$ 且 $M_i\leqslant\min_{1\leqslant j\leqslant k}M_j^*=M_k^*$。

基于 Top-k 映射的本体匹配方法是对目前本体匹配方法的一种扩展，即产生 Top-k 映射而不是只得出一个最佳映射(可以看作是 Top-1 映射)。这种方法避免了阈值选择的困难，在匹配过程中同时产生 k 个最佳映射进行下一次迭代，直到没有新的映射出现为止，从而得到 2 个本体之间的有效映射。

每次迭代选取相似度最高的 k 个映射进行下一次匹配，而一些不相关或者较差的映射就被过滤掉，因此，节省了匹配过程的执行时间。在迭代过程中，当第 i 个最佳映射降级为第 $i+1$ 个映射时，匹配器强制舍弃至少一个相似度最小映射，同时保持 Top-k 的全局置信度。所以，映射生成过程也可以被看作是匹配器迭代地舍弃最小相似度映射的过程。k 值的具体值可以根据匹配本体的复杂程度来确定。

5 算法描述及实验分析

5.1 具体算法

输入：2 个本体 O_1 和 O_2

输出：2 个本体之间的映射表

$Matrix_{el}(e_{i1},e_{j2})=Matrix_{str}(e_{i1},e_{j2})=Matrix(e_{i1},e_{j2})=\Phi$；

$MapList=\Phi$，$E_1=O_1$，$E_2=O_2$；//初始化相似矩阵及映射表

For (Change(MapList))　//当没有新映射产生时，迭代完成。

{

For each $e_{i1}\in E_1$

{

For each $e_{j2}\in E_2$

{

$Matrix_{el}(e_{i1},e_{j2})=sim_{el}(e_{i1},e_{j2})$；

$Matrix_{str}(e_{i1},e_{j2})=sim_{str}(e_{i1},e_{j2})$；

$Matrix(e_{i1},e_{j2})=sim(e_{i1},e_{j2})$；　//计算相似度

}

Descending ($Matrix(e_{i1},e_{j2})$，k);　//相似度降序排列

If ($sim(e_{i1},e_{j2})!=0$)

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Add((e_{i1}, e_{j2}) , MapList);//将 Top-k 映射存入映射表中

}

E_1 = First(MapList) , E_2 = Second(MapList) ;

//映射表中元素进入下一次迭代

}

Return Map_{i1}

5.2 实验分析

本实验的硬件条件是 CPU P4 3.0 GB，内存 512 MB；软件工具有 JDK 1.4.2, Jena 2.2, WordNet 2.1, eclipse 3.0；操作系统为 Windows XP。

为了对算法进行评估，本文在 3 个数据集上作了实验，它们的统计数据如表 1 所示。Course CatalogI 和 Course CatalogII 数据集中的本体分别描述了 Cornell 大学和 Washington 大学的课程体系，其中第 1 个数据集是课程的迷你版本，它们之间很相似，第 2 个数据集本体结构比较大并且相似程度不高。Company Profiles 数据集中的本体分别描述了 Standard.com 和 Yahoo.com 公司的商务信息，这 2 个本体的概念和实例较多，映射也较多。

表 1 测试集的统计数据

数据集	本体	概念/个	实例/个	深度/层	映射/个
Course CatalogI	Cornell	34	1 526	4	34
	Washington	39	1 912	4	37
Course CatalogII	Cornell	176	4 360	4	54
	Washington	166	6 957	4	50
Company Profiles	Standard.com	333	13 634	3	236
	Yahoo.com	115	9 504	3	104

一般使用信息检索领域查全率(Recall)和查准率(Precision)作为评价匹配算法的主要准则。

从表 2 的实验结果和图 1 中与其他方法的比较结果不难发现，基于 Top-k 映射的本体匹配方法在查全率和查准率方面都有很好的表现，并且其查准率与 GLUE 算法相比有所提高。

表 2 实验结果

数据集	映射	查全率/(%)	查准率/(%)
Course CatalogI	Cornell to Washington	92.1	92.1
	Washington to Cornell	93.7	94.5
Course CatalogII	Cornell to Washington	82.3	81.6
	Washington to Cornell	81.5	75.5
Company Profiles	Standard.com to Yahoo.com	83.3	83.3
	Yahoo.com to Standard.com	84.7	81.4

对于相似程度较高的测试集 Course CatalogI，2 个匹配算法的查准率都很高，而对其他两个结构较大、相似程度较低的本体，本文提出的方法查准率较高。主要原因在于，基于 Top-k 映射的本体匹配方法不但从元素级计算相似度，而且考虑到本体自身的结构，计算结构级相似度，并且对于复杂的本体，它能够利用每次选取 k 个最佳映射的方法来提高匹配的精确度。

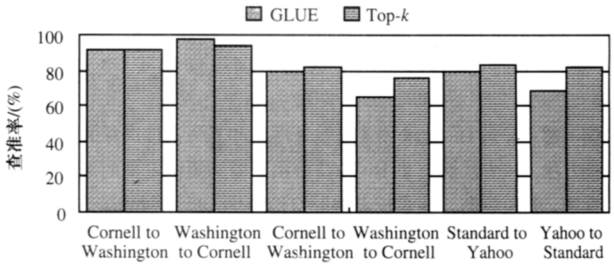

图 1 实验结果比较

6 结束语

本体匹配是解决本体异构问题的主要方法。目前研究者已经提出了多种本体匹配技术，但由于存在本体匹配的结果精度不高、本体匹配的过程无法自动完成等缺点，因此本文提出了一种基于 Top-k 映射的本体匹配方法，对目前的匹配方法进行了改进。该方法以相似度计算为基础，并生成 Top-k 映射，避免了阈值选择的困难。如何进一步提高算法的效率是下一步研究的重点。

参考文献

[1] Gruber T R. A Translation Approach to Portable Ontologies[J]. Knowledge Acquisition, 1993, 5(2): 199-201.

[2] Ehrig M, Sure Y. Ontology Mapping — — An Integrated Approach[C]//Proceedings of ESWS'04. Heraklion, Greece: [s. n.], 2004: 4-6.

[3] Levenshtein V I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals[J]. Cybernetics and Control Theory, 1966, 10(8): 707-710.

[4] Pantel P, Lin D. Discovering Word Senses from Text[C]// Proceedings of the 2002 ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada: [s. n.], 2002: 613-615.

[5] Gal A. Managing Uncertainty in Schema Matching with Top-k Schema Mappings[J]. Journal on Data Semantics, 2006, 40(6): 99-101.

(上接第 56 页)

(2)按部队实体报告：按照参与训练的部队编制结构，报告各级部队的任务行动成绩。这种报告形式反映的是参演的某个具体部队建制的训练状况。

4 结束语

合同战术训练是一个“复杂的巨系统”，它的评估需要多种软件、硬件技术的融合，合同战术训练评估系统的实现是一个螺旋上升、逐步完善的过程。本文分析了合同战术训练的流程，建立了系统的体系结构，明确了系统框架、组件的功能和接口，探讨了系统开发中数据模式、实现方法等关键技术，初步实现了一个合同战术训练评估原型系统，并在相关单位得到了应用，达到了预期的目的。

参考文献

[1] 马开城, 张 波, 刘智慧. 合同战术实兵演习系统设计[J]. 军事运筹与系统工程, 2003, 17(4): 36-39.

[2] 宋祥斌, 张宏军, 郝玉龙. 战场综合防护计算机评估系统[J]. 计算机工程, 2001, 31(2): 206-208.

[3] 王月平, 赵志强. 任务空间概念模型研究的若干问题探讨[J]. 军事运筹与系统工程, 2003, 17(2): 21-24.

[4] Waite W F. Object-oriented Representation Schemas for CMMS [EB/OL]. (2000-10-02). http://www.odu.edu.

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net