基于分簇的本体映射方法*

熊 颖,李海波,李 静
(华侨大学 计算机科学与技术学院,福建 厦门 361021)

摘 要:为了能够充分地挖掘、分享和重复利用本体中的知识,提出一种基于映射关系的分簇方法,先通过已存在的高质量的本体映射关系,对原本体和目标本体分别进行分簇,再挖掘出实体间潜在的关系。通过实验证明,采用改进的映射方法提高了本体映射的质量,采用具有完善实体关系的映射结果提高了检索系统的准确率和查全率。

关键词:分簇;本体映射;实体关系发现
中图分类号:TP393.4 **文献标识码**:A **文章编号**:1674-7720(2013)02-0070-04

Ontology mapping method based on clumping

Xiong Ying, Li Haibo, Li Jing
(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract: This paper proposed a clumping method based on mappings for fully digging, sharing and reusing of ontologies. It clumped the source ontology and target ontology respectively through the existing high quality mappings firstly, and then dug out the potential relationships among entities. The experiments show that the quality of ontology mapping has been improved by using the advanced mapping method, and the precision/recall of retrieval system has been improved by using the mapping results with well entities' relationships.

Key words: clumping; ontology mapping; discover entities' relationships

随着越来越多的本体被开发,以及持续性和高效性的知识访问需求不断提高,本体知识的充分挖掘、分享和重复利用已成为本体库优化的重要研究内容。由于本体的独立开发性,导致在相同或者重叠领域本体中实体的定义和实体间的关系有所不同,即本体间的互操作性较低。本体映射已成为当今本体研究中的热点,它是解决并促进本体间互操作性问题的重要方法。但是要更充分地挖掘、分享和重复利用本体知识,该方法还需要不断地改进和优化。本体映射过程中存在以下两个问题:

(1)基于特征低相似性进行本体映射的质量不高。目前对语义、词汇和结构特性相似度较高的本体进行映射,在一定范围内其映射质量是较高的,但是大部分本体的建模粒度都不相同,导致本体中实体表示的词汇和结构特性都不相同,即本体中实体的特征相似度较低,所以采用基于词汇和结构特征的相似性进行本体映射的质量不高。

(2)映射结果中实体关系不够完善。目前大量本体映射方法在建立了本体映射关系后不会对实体关系进行分析和处理,导致本体映射结果的实体关系不够完善,应用质量较低。

1 相关工作

目前大多数的本体映射方法(例如 ASMOV[1-2]和 RiMOM[3]等)是基于词汇和结构特征的相似性进行本体映射的,在一定范围内映射质量较高,但当两个本体的建模粒度不相同时,采用基于这些特征相似性进行本体映射的质量就较低。例如,石灰在原本体中包括氧化钙和氢氧化钙,在目标本体中包括煅烧石灰、熟石灰、石灰乳和消石灰,这两个本体中用不同的术语描述相同的信息石灰,采用传统的本体映射方法测量得到的映射准确度低于 1%,本体映射时就无法建立实体间高质量的映射关系。为解决这一问题,可以重复利用已存在的高质量本体映射关系,提高本体映射的质量。调查本体映射关

* 基金项目:福建省自然科学基金(2012J01272);福建省高校产学研科技重大项目(2010N5008);
中央高校基本科研业务费资助项目(JB-ZR1147);厦门市科技计划项目(3502z20110013);泉州市科技计划项目(2011G5)

系重复利用的目的在于利用多对一或者一对多的实体映射关系实现分簇的过程，从而获取准确的匹配信息。

本文以最新的 ASMOV 映射系统[4]为基础。它是一种半自动化本体映射过程，联合了元素级和结构级的相似度测量，使用本体中四种不同特征相似度的加权平均值作为实体间的总相似度，采用了语义验证要求遵守的规则来判断是否建立映射链接的技术，以确保建立的映射链接不包含语义矛盾，但是其映射匹配的准确率和查全率还有待提高。针对本体映射过程中存在的问题以及 ASMOV 在映射匹配质量方面的不足，提出了一种基于分簇的本体映射方法 OMMC(Ontology Mapping Method based on Clumping)，该方法有助于建立本体间高质量的映射关系，通过高质量的映射关系再进行实体间关系的再发现，从而提高了本体映射的应用质量。

2 基于分簇的本体映射

基于分簇的本体映射的流程是：先将原本体和目标本体分别进行分簇，再将分簇后的原本体和目标本体应用于 ASMOV 映射系统中的本体映射，主要包括分簇和建立映射链接两个模块。

2.1 分簇

定义 1 簇。利用已存在的高质量的本体映射关系，在多对一的映射场景下，一个本体 O 中的多个实体和另一个本体中的相同实体匹配，则将这多个实体看做是一个簇，本体 O 可划分为多个簇。

定义 2 划分。将一个本体划分成多个簇，这多个簇叫做本体的一个划分。

例如存在本体 $O=\{e_1,e_2,e_3,e_4\}$，本体 $O'=\{E_1,E_2\}$，它们之间的映射结果为：$O\rightarrow O'=\{e_1\rightarrow E_1,e_2\rightarrow E_1,e_3\rightarrow E_2,e_4\rightarrow E_2\}$，则本体 O 被划分为两个簇$\{e_1,e_2\}$和$\{e_3,e_4\}$，这两个簇是本体 O 的一个划分 $P_O=\{\{e_1,e_2\},\{e_3,e_4\}\}$。

对原本体和目标本体分别进行分簇，原本体和目标本体都被划分为多个簇，得到一个原本体的划分和一个目标本体的划分。以农业领域的本体 S 和本体 T 为例，原本体 S 是 PWP(Prism Web Pages)定义的一个中文本体，包含 1 028 个实体，S 和本体 PWP_1 已存在本体映射关系，通过它们之间的映射关系进行分簇，可将 S 划分为 196 个簇，S 中一个簇的映射关系如表 1 所示；目标本体 T 是 FAO(Food and Agriculture Organization)定义的一个中文本体，包含 2 420 个实体，T 和本体 FAO_1 已存在本体映射关系，通过它们之间的映射关系进行分簇，将 T 划分为 357 个簇，T 中一个簇的映射关系如表 2 所示。

表 1 S 中一个簇示例

S	PWP_1
包谷	玉米
玉茭	玉米
苞米	玉米
玉米	玉米
棒子	玉米

表 2 T 中一个簇示例

T	FAO_1
玉蜀黍	玉米
包果	玉米
玉米	玉米
棒子	玉米
粟米	玉米
番麦	玉米
珍珠米	玉米

2.2 建立映射链接

对原本体 S 和目标本体 T 分别进行分簇，得到了两个稳定性划分 P_S 和 P_T 后，直接进入本体映射匹配阶段，在 ASMOV 系统中，对本体 S 和目标本体 T 采用 OMMC 方法进行映射的过程如图 1 所示。

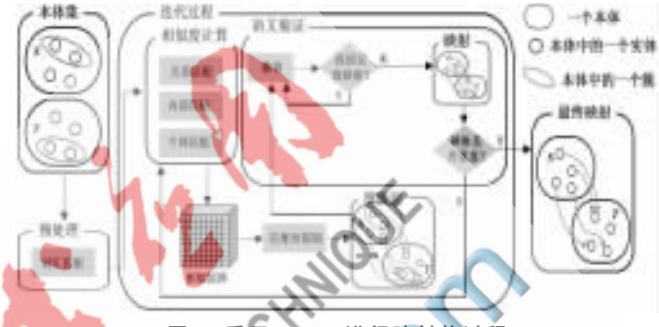


图 1 采用 OMMC 进行映射的过程

本体映射的核心模块是相似度计算，改进的 ASMOV 映射过程在相似度计算时要优先考虑在一个簇中的实体。假设 A 是 S 中的一个簇，B 是 T 中的一个簇，$\forall a\in A$，$\forall b\in B$，$\exists e\in A$，$\exists e'\in B$，当且仅当映射关系$<e\rightarrow e'>$成立，且 e 和 e' 间的相似度 $\tau(e,e')$ 是 e 的最大相似度时，建立 A、B 中实体间多对多的映射关系$<a\rightarrow b>$，且 a 和 a 间的相似度 $\tau(a,b)=\tau(e,e')$。在每一个迭代过程 n 中，采用一种综合的相似度计算方法，相似度 $\tau_n(e,e')$ 是 4 个相似度(词汇相似度 $S^L(e,e')$、关系相似度 $S_n^H(e,e')$、内部相似度 $S_n^R(e,e')$和个体相似度 $S_n^E(e,e')$的加权平均值，$M=\{L,E,H,R\}$表示由词汇 L、个体 E、关系 H 和内部 R 4 个方面组成的集合，w_m 是各个方面所占的权重，可以调节。$\tau_n(e,e')$的计算如式(1)所示。

$$\tau_n(e,e')=\begin{cases}\dfrac{\sum_{m\in M}(w_m S_n^m(e,e'))}{\sum_{m\in M}w_m}, & \text{实体 } e \text{ 和 } e' \text{的类型相同}\\ 0.0, & \text{其他}\end{cases}\tag{1}$$

关系相似度 $S_n^H(e,e')$是指结合实体的父类和子类分别相匹配计算的相似度。一个类或者属性可能包含多个父类和子类，$S_n^H(e,e')$是所有父类 Fat 和子类 Chi 相似度的加权平均值，其计算如式(2)所示。针对父类权重 W_{Fat} 和子类权重 W_{Chi} 的设置进行了改进，采用了长度优先算法[5]，如式(3)和式(4)所示，依据父类个数长度 Fat.size 和子类个数长度 Chi.size 的算法设置相应权重。一个类或者属性所包含的父类和子类的个数大小反映了其“独特性”，即个数越大，说明其父类或子类同义的词就越多，那么它们相似概率就越大、越独特。词汇相似度 $S^L(e,e')$及第 n 次迭代时父类相似度 $S_n^U(e,e')$、子类相似

度 $S_n^C(e,e')$、内部相似度 $S_n^R(e,e')$和个体相似度 $S_n^E(e,e')$的计算方法见参考文献[1]。

$$S_n^H(e,e')=W_{Fat}\times S_n^U(e,e')+W_{Chi}\times S_n^C(e,e') \quad (2)$$

$$W_{Fat}=\frac{\lg(Fat.size+1)}{\lg(Fat.size+1)+\lg(Chi.size+1)} \quad (3)$$

$$W_{Chi}=\frac{\lg(Chi.size+1)}{\lg(Fat.size+1)+\lg(Chi.size+1)} \quad (4)$$

运行改进后的 ASMOV 系统步骤如下：

(1)数据准备。准备好已经分簇的原本体和目标本体。

(2)预处理阶段进行词汇匹配。利用一个词库来计算概念、属性和个体的词汇相似度。

(3)进行相似度计算。包括外部关系、内部匹配和个体匹配相似度的计算，并将计算结果放入相似度矩阵中。

(4)从相似矩阵中提取两个本体中相似度最高的匹配对实体集，依据这些实体集找到对应的簇，建立簇中实体间多对多的映射关系，并放入预映射模块中。

(5)对预映射模块中的映射关系集进行语义验证，即通过一些已定义的规则进行验证并修剪无效的映射关系，且将连接无效映射关系的实体间相似度置零。循环执行步骤(3)~步骤(5)，直到本体 S 或 T 中所有簇都执行一遍迭代过程。

(6)提取最终的本体映射关系。

3 实体关系的再发现

在建立了高质量的本体映射关系后，连接每一条映射关系的两个实体间都可发现新的关系，主要包括父类关系发现、子类关系发现和等价类关系发现。

规则 1 父类关系发现是指若连接一条映射关系的两个实体的父类不同，那么这两个实体的父类可以合并，同时对合并后的父类消除重复，最后这两个实体得到了相同的新的父类集合，依此类推应用于每一条映射关系中。如图 2 所示，建立实体 C_4 和实体 E_2 的映射关系以后，C_4 和 E_2 的父类都为 C_1 和 E_1，若 C_1 与 E_1 重复，那么去除重复后 C_4 和 E_2 的父类都为 C_1 或者 E_1。

图 2 父类关系再发现

规则 2 子类关系发现是指若连接一条映射关系的两个实体的子类不同，那么这两个实体的子类可以合并，同时对合并后的子类消除重复，最后这两个实体得到了相同的新的子类集合，依此类推应用于每一条映射关系中。如图 3 所示，建立 C_4 和 E_2 的映射关系以后，C_4 和 E_2 的子类都为 C_5、C_6、C_7 和 E_4，若 E_4 与 C_5、C_6 和 C_7 其中一个重复，那么去除重复后 C_4 和 E_2 的子类都为 C_5、C_6 和 C_7。

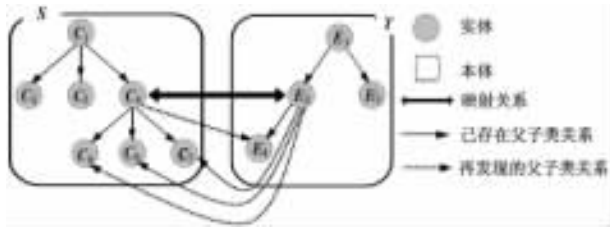


图 3 子类关系再发现

规则 3 等价类关系发现是指若连接一条映射关系的两个实体的等价类不同，那么这两个实体的等价类可以合并，同时对合并后的等价类消除重复，最后这两个实体得到除自身外的新的等价类集合，依此类推应用于每一条映射关系中。图 4 显示了等价类关系的再发现，由于 C_2、C_3 和 C_4 互为等价类，建立 C_4 和 E_2 的映射以后，那么 C_2、C_3、C_4 和 E_2 互为等价类；若 E_2 与 C_2、C_3 和 C_4 其中一个重复，那么去除重复的实体 E_2 后，E_2 的父类与子类的关系还需要传递到 C_2、C_3 和 C_4 中，即前面提到的父类与子类关系的再发现。

图 4 等价类关系再发现

在合并父类、子类及等价类关系时，以 T 为目标，且需互相说明彼此之间的关系，如果发生冲突，则调用以下冲突处理规则进行解决。

规则 4 类层次结构冲突处理[6]。以目标本体中类层次结构为基准，删除原本体的冲突结构，保证关系合并中类层次结构的完整性。例如在本体 S 中的 C_2 和 C_3 是等价类，在本体 T 中的 E_1 是 E_2 的父类，若 C_2 和 E_1 建立了映射关系，C_3 和 E_2 建立了映射关系，则先合并 E_2，然后合并 E_1，本体中 C_2 和 C_3 既是父子关系又是等价关系，此时就存在类层次冲突问题，以本体 T 层次结构为基准，把 C_2 和 C_3 的等价类关系删除。

4 实验评估

4.1 改进 ASMOV 前后映射质量比较实验

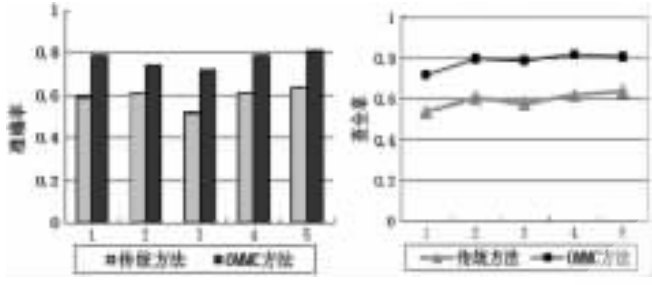
在 ASMOV 的测试场景中，逐步对本体 S 和本体 T 进行映射检测，获取的参数包括标准配对数 x、配对总数量 n 和配对总数 n 中准确的配对数 r，通过获取的参数值来计算匹配的准确率 P 和查全率 R，其计算公式为 $P=r/n$ 和 $R=r/x$。经过多次测试改进前后的方法，证明采用 OMMC 方法得到的映射匹配质量均明显提高，相对于传统方法，在相同条件下采用 OMMC 方法使得匹配的准确率和查全率均提高了约 0.2。比较结果如表 3 和图 5 所示。

4.2 关系再发现前后映射结果应用的质量比较实验

本文提出的分簇方法是在已存在的高质量的本体映射关系中发现本体内部结构，即将本体划分为若干个

表 3 映射质量比较

组号	S 和 T 映射			传统方法				OMMC 方法			
	S	T	x	r	n	P	R	r	n	P	R
1	200	500	120	65	110	0.59	0.54	86	109	0.79	0.72
2	400	1 000	360	219	357	0.61	0.61	288	388	0.74	0.80
3	600	1 500	523	301	489	0.62	0.58	415	577	0.72	0.79
4	800	2 000	935	578	948	0.61	0.62	771	981	0.79	0.82
5	1 028	2 420	1 366	898	1 413	0.64	0.66	1 101	1 365	0.81	0.81



(a)准确率 (b)查全率

图 5 映射匹配的准确率和查全率

簇。在映射匹配时采用以簇为单位替换以实体为单位的 ASMOV 方法，建立高质量的本体映射关系，然后对实体关系进行再发现，完善了映射结果的实体关系。将实体关系完善前后的映射结果应用于海量农业信息语义检索系统中，用多组请求信息分别进行检索，比较检索结果的准确率和查全率。

海量农业信息语义检索系统总体框架主要包括本体管理、数据获取、请求管理、请求信息匹配、海量农业信息处理及语义请求客户端 6 个部分。通过网络爬虫工具采集海量农业信息，并对爬下的网页进行信息的抽取和整理，抽取和整理后的网页可保存在海量农业信息数据库中作为检索时的资源库。为使实验能够更准确和更快速得出结论，选择了整理好的 10 万个网页作为资源库，运用该系统进行实验的具体步骤如下：

(1)清除本体库中已经存在的本体信息，将本体及映射结果添加到本体库中。

(2)将本体库中的本体信息与海量农业信息相关联，即运行信息标注与词频计算、倒排表建立和农业信息聚类 3 个模块，并将关联信息存入海量农业信息数据库中。

(3)通过配置文件管理接口设置配置文件信息，如本体库中等价类、父类和子类各自所占的权重等。

(4)启动系统服务器，在用户检索接口输入用户需要检索的信息。

(5)计算检索结果的准确率和查全率。

在建立高质量的本体映射链接后得到映射结果 M_1，在完善映射结果 M_1 中的实体关系后得到映射结果 M_2，将 M_1 和 M_2 分别应用于海量农业信息语义检索系统中，运行该系统进行实验，输入多组检索数据，计算检索结果的准确率和查全率，如图 6 所示。通过比较分析可知，对采用了 M_2 的系统进行检索，得到了较高的准确率和查全率，从而表明了完善映射结果中的实体关系对本体映射应用的重要性。

本文提出一种基于映射关系的分簇方法，首先通过各自已有的映射关系，对原本体和目标本体分别进行分簇，再采用改进的 ASMOV 映射系统，建立高质量的映射关系，并完善实体间的关系。通过对比采用 OMMC 方法和传统方法的 ASMOV 系统的映射质量，可知采用 OMMC 方法具有一定的优越性，即匹配结果更准确和全面；通过对比完善实体关系前后本体映射结果应用的质量，可知完善了实体关系的映射结果应用于检索系统中，提高了检索系统的准确率和查全率。

(a)准确率 (b)查全率

图 6 检索系统的准确率和查全率

参考文献

[1] Jérôme Euzenat, MEILICKE C, STUCKENSCHMIDT H, et al. Ontology alignment evaluation initiative: six years of experience[C]. Proceedings of the Journal on Data Semantics XV. Berlin Heidelberg: Springer, 2011: 158-192.

[2] JEAN-MARY Y R, SHIRONOSHITA E P, KABUKA M R. Ontology matching with semantic verification[J]. Web Semantics, 2009, 7(3): 235-251.

[3] Wang Zhichun, Zhang Xiao, Hou Lei, et al. RiMOM results for OAEI 2010[C]. Proceedings of the 5th International Workshop on Ontology Matching(OM-2010) collocated with the 9th International Semantic Web Conference(ISWC-2010). Shanghai: CEUR-WS, 2010: 195-202.

[4] ASMOV Results for OAEI 2007[EB/OL].[2012-06-30].http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-304/paper12.pdf, 2007.

[5] 张钋.基于语义的网络服务匹配机制的研究与实现[D].北京：清华大学，2005.

[6] 罗正海.面向语义 Web 服务的本体合并研究[D].大连：大连海事大学，2009.

(收稿日期：2012-09-29)

作者简介：

熊颖，女，1986 年生，硕士研究生，主要研究方向：语义 Web、信息检索等。

李海波，男，1972 年生，博士，副教授，硕士生导师，主要研究方向：工作流、服务计算技术等。

李静，女，1981 年生，博士，讲师，主要研究方向：语义 Web，信息检索，数据挖掘等。