基于数据场的大规模本体映射

仲 茜 李涓子 唐 杰 周立柱
(清华大学计算机科学与技术系 北京 100084)

摘 要 针对已有的本体映射方法在处理大规模本体映射任务时效率和有效性较低的问题,文中提出了一个基于数据场的本体映射算法.该算法首先使用高效的相似度算法,建立本体中元素对另一本体的初始相关度;然后,利用数据场势函数引入周围本体元素对当前元素的影响,修正初始相关度,并最终确定本体间的相关子本体;最后,利用针对性的方法对上述相关子本体进行更有效的映射.实验结果表明,该算法可以在提高映射结果质量的同时保证较高的映射效率.

关键词 数据场;势函数;本体;本体映射;语义 Web
中图法分类号 TP391 **DOI 号**: 10.3724/SP.J.1016.2010.00955

Data Field Based Large Scale Ontology Mapping

ZHONG Qian LI Juan-Zi TANG Jie ZHOU Li-Zhu
(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract As the cornerstone of ontology based data integration, data exchange and metadata management, ontology mapping, aiming to obtain semantic correspondences between two ontologies, has attracted wide attentions of researchers in community of the Semantic Web. However when getting the alignments between two large-scale ontologies, the existed mapping methods are not very effective and efficient due to neglecting of relevant sub-ontologies in those two ontologies. For addressing this important issue, in this paper, a data field based ontology mapping approach is proposed to improve the effectiveness and the efficiency of large scale ontology mapping tasks. At first, this approach employs a light weight similarity computing method to collect the initial relevance values between one ontology's elements to another ontology. Then, the potential functions of a data field are taken into account to revise the relevance of an ontology element according to its surroundings. Finally, the relevant sub-ontologies are found and extracted, and a fine-grained alignment approach is used to mapping between the extracted ontologies for better results. The experiments show that the proposed approach is able to effectively deal with the large scale ontology mapping issue with the satisfactory efficiency.

Keywords data field; potential function; ontology; ontology mapping; Semantic Web

1 引 言

本体作为语义 Web 的核心基础元素,已经广泛应用于语义数据集成、数据交换和元数据管理等领域.然而,由于本体定义的自发性,本体定义者自身的社会文化背景、对术语规范的使用习惯和对本体组织结构的理解不同,导致互联网上本体的分布性

收稿日期:2009-05-06;最终修改稿收到日期:2010-03-30.本课题得到国家自然科学基金(60973102, 60703059)、国家"九七三"重点基础研究发展规划项目基金(2007CB310803)和国家"八六三"高技术研究发展计划项目基金(2009AA01Z138)资助.**仲 茜**,男,1975 年生,博士研究生,主要研究方向为语义数据集成和本体映射. E-mail: zhongq05@mails.tsinghua.edu.cn. **李涓子**,女,1964 年生,博士,教授,博士生导师,主要研究领域为语义 Web、Web 服务和知识管理. **唐 杰**,男,1977 年生,博士,副教授,主要研究方向为大规模社会网络和机器学习. **周立柱**,男,教授,博士生导师,主要研究领域为数据库系统、数字图书馆和海量信息处理.

和异构性，极大限制了本体数据的共享与集成. 为了实现基于本体的语义互操作，就必须建立异构本体中元素（如概念、关系、实例等）之间的映射关系，这一过程称为本体映射. 目前，本体映射已成为语义 Web 领域中的一个研究热点.

针对本体映射问题，相关领域研究者做了大量的工作，提出了很多方法. 这些方法概括起来可分为 5 类：基于元素名称的方法[1-2]、基于本体结构的方法[3-4]、基于本体实例的方法[5-6]、基于推理的方法[7]和基于背景知识的方法[8-9]. 由于本体本身在术语规范、组织结构等上面差异，没有一种已知的本体映射方法对所有映射任务都适用，因此实际的本体映射工具[3,10]往往集成多种不同类型映射方法.

随着本体应用的深入，各专业组织定义的本体规模不断增大. 这些大规模本体往往涉及多个领域. 例如：GEMET 本体①包含 5280 个概念，涉及农业、空气、生物、气候、疾病等领域；AGROVOC 本体②包含 28439 个概念，涉及农业、森林、渔业、食品、环境等领域. 在包含这些本体的映射任务中，只有相关领域的概念才可能映射，其它无关概念若参与映射，不但会导致额外的资源开销，而且会干扰正确的概念映射，从而降低映射精度. 然而，目前的本体映射方法并不关心本体元素的所属领域，映射时必须考虑所有元素对的映射关系，这会导致大量资源开销和更多的错误映射. 例如：在本文实验中，使用基于多语言标签的编辑距离相似度方法对上述两个本体进行映射，耗时达 131840s，约合 1.5 天，是本文中方法的 3.4 倍，而综合映射效果却比本文方法低近 2%. 因此，这些多领域、大规模本体的广泛应用，在映射效率和映射精度等方面对当前映射方法提出了巨大的挑战.

本文的方法将主要针对大规模本体映射问题. 图 1通过一个多领域本体映射的例子，对上述问题进行说明. 图中将同一领域的概念及其关系用虚线框起.

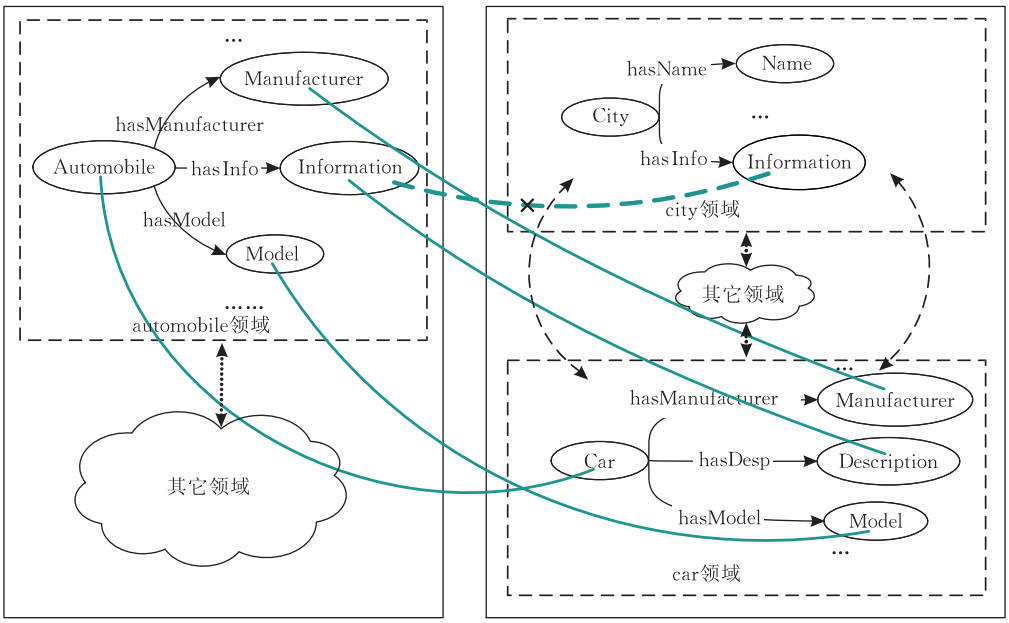


图 1 多领域本体映射示例

从图 1 中可以看到，Vehicle 本体中的 automobile 领域与 Travel 本体中 car 领域密切相关，而与 city 领域的相关度较弱. Vehicle 本体中 Information 概念应该与 Travel 本体中的 Description 概念建立映射关系. 但是若只追求映射的高效性而使用基于名称的策略，那么与 Vehicle. Information 映射的将是概念 Travel. Information. 当然，可以通过更有针对性的方法（如基于结构的映射方法），来修正这种错误. 但是，这类方法往往效率较低且需要占用较多的资源. 如 Similarity Flooding[4]算法，它根据周围节点的相似度对当前节点的相似度进行修正，可以解决图 1 中出现的错误映射，但是该算法采用迭代计算的策略，算法的效率较差，而且需要大量内存用于存储相似度传播图，当本体规模很大时，算法存在可用性的问题.

因此，针对大规模多领域本体的特点，如果能够以一种高效准确的、低资源消耗的方法发现两个本体中所有相关元素构成的相关子本体. 如图1中

① http://www.eionet.europa.eu/gemet
② http://www.fao.org/aims/ag_intro.htm

Vehicle 本体的 automobile 领域和 Travel 本体的 car 领域中元素组成的相关子本体.然后在相关子本体之间进行针对性的映射.一方面,可以降低映射规模以提高使用针对性映射方法的效率,且当针对性的方法效率较低时可以提高映射方法的整体效率;另一方面,可以克服不相关本体元素的影响以提高准确率.相关子本体发现问题主要面临两大挑战:(1)如何快速、准确、完整地发现相关子本体;(2)如何以一种高效的、低资源消耗的方式引入周围元素对当前元素的影响.

针对上述挑战,本文提出了一个基于数据场[11]的本体映射算法.这里的数据场源于物理场(如磁场、电场、重力场等),用于描述数据间的相互影响.该算法首先使用高效的相似度方法(如基于编辑距离的方法)建立当前本体中元素对另一本体的初始相关度.然后,利用数据场势函数度量本体中的每个元素所受到的周围其它元素的影响,修正初始相关度,并最终获得本体中的相关子本体;最后,利用针对性的算法对相关子本体进行更精确的映射.算法之所以采用数据场是因为,一方面数据场可以较准确地量化计算本体中元素间的相互影响,另一方面对于每个本体元素只需要一个额外的存储空间用于存放相关度的值,资源的开销少,而且通过估算场中数据的影响距离,可以简化计算、提高效率.利用 OAEI(Ontology Alignment Evaluation Initiative) 2007 Environment① 映射数据集进行实验,结果表明该算法可以在提高映射结果质量的同时保证较高的映射效率.在 OAEI 2008 FAO② 映射任务评测中,本文的方法取得了综合评价第 1 的成绩.前期研究成果发表在 SIGMOD 2009 国际会议论文集中[12],主要针对不平衡的本体映射问题.它利用某些映射中存在的本体规模不平衡的特点,在规模较大的本体中,通过高斯函数发现并抽取与规模较小本体大小相当的相关子本体,并最终在小规模本体和相关子本体间完成映射.而本文中的方法则面向更一般的大规模本体映射问题.针对本体的多样性,采用了更灵活的基于数据场的方法来发现本体间的相关子本体.由于无法直接确定相关子本体的大小,文中使用了基于统计的方法获得相关子本体的过滤阈值,并最终获得相关子本体.

本文第 2 节定义问题涉及的主要概念(如本体、本体映射和数据场);第 3 节详细描述基于数据场的本体映射算法;第 4 节通过实验验证文中方法的有效性和高效性;第 5 节介绍本研究领域的相关工作;第 6 节总结本文的工作并提出今后需进一步研究的内容.

2　相关概念

2.1　本体与本体映射

参考文献[13]给出本文中本体和本体映射的形式化定义.

定义 1.　本体主要包括概念(Concepts)、关系(Relations)、实例(Instances)以及公理(Axioms),可表示为 $O=(C,R,I,A^O)$,其中,C 表示概念集合,R 表示关系集合,I 表示实例集合,A^O 表示公理的集合.为了便于描述,本文将本体中的概念、关系、实例和公理统称为本体元素.

定义 2.　设 $M=(O_1,O_2,F_{\mathrm{Map}})$ 为本体 O_1 到本体 O_2 的映射,O_1 为源本体,O_2 为目标本体,F_{Map} 为建立 O_1 到 O_2 间本体映射的映射函数,定义为

$$F_{\mathrm{Map}}:\{e_{i1}\}\rightarrow\{e_{j2}\},$$

其中,$\{e_{i1}\}$、$\{e_{j2}\}$ 分别为 O_1、O_2 中元素的集合.本文只考虑本体概念之间的映射,这是多数本体映射任务所针对的问题.

2.2　数据场与势函数

场的概念最早是 1837 年由英国物理学家法拉第提出,用于描述物质粒子间的非接触相互作用[14].最初的场主要是指磁场、电场、重力场等物理场.在上述物理场中,通常利用矢量场强函数和标量势函数来描述粒子间的相互作用.

参照物理场,假设在给定 p 维空间 $\Omega\subseteq R^p$ 中,存在着包含 n 个数据元素的数据集 $D=\{X_1,X_2,\cdots,X_n\}$,其中,$X_i=(x_1,x_2,\cdots,x_j,\cdots,x_p)$,$x_j$ 为 X_i 第 j 维坐标.如果将 D 中的任意一个数据元素 X_i 视作一个具有一定质量的粒子,那么它就会对周围其它数据元素产生影响.这样 D 中数据元素的共同作用便可在 Ω 中形成一个虚拟的场,即数据场.关于数据场的详细描述请参阅文献[11].

与物理场类似,在数据场中也可以定义矢量场强函数和标量势函数.本文中主要使用了标量势函数.假设 F 为 D 中数据所产生数据场,函数 $\varphi_X(Y)$ 为其势函数,其中 $X\in D$,$Y\in\Omega$.它指出了数据元素 X 在 Y 处所产生的势值,$\varphi_X(Y)$ 必须满足以下条件:

(1) $\varphi_X(Y)$ 是一个连续、平滑、有界函数;

① http://oaei.ontologymatching.org/2007/environment/
② http://oaei.ontologymatching.org/2008/fao/

(2) $\varphi_X(Y)$具备各向同性;

(3) $\varphi_X(Y)$是一个关于距离 $\|X-Y\|$ 的减函数. 当 $\|X-Y\|=0$, $\varphi_X(Y)$取得最大值;当 $\|X-Y\|\to\infty$, $\varphi_X(Y)\to 0$.

原则上说满足上述 3 个条件的函数均可以作为数据场的势函数. 但在实际应用中,常常会参照使用相应物理场的势函数,最常用的有

(1) 拟重力场势函数

$$\varphi_X(Y)=m\Big/\left(1+\left(\frac{\|X-Y\|}{\sigma}\right)^k\right) \tag{1}$$

(2) 拟核力场势函数

$$\varphi_X(Y)=m\times e^{-\left(\frac{\|X-Y\|}{\sigma}\right)^k} \tag{2}$$

其中,$m\geq 0$ 表示 X 对 Y 的影响强度,可以理解为 X 的质量. $\sigma\geq 0$ 称为影响因子,它决定了元素的影响范围,当 σ 增大时,势函数值就增大.

图 2 对比了当 m, $\sigma=1$, $k=2$ 时,拟核力场与拟重力场的函数曲线.

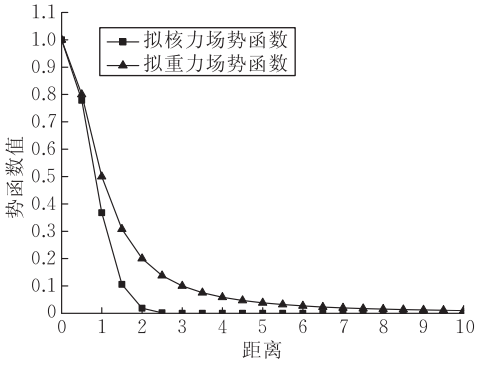


图 2 不同势函数曲线比较(m, $\sigma=1$, $k=2$)

由图 2 可知,随着距离的增加,拟核力场势函数曲线比拟重力场势函数曲线下降更为迅速. 通常将类似于前者的场称为短程场,类似后者的称为长程场.

确定势函数之后,可以计算 p 维空间 Ω 中任意一点 Y 关于 D 中的数据场 F 的势值,公式如下:

$$\varphi(Y)=\varphi_F(Y)=\sum_{X\in D}\varphi_X(Y) \tag{3}$$

3 基于数据场的本体映射

基于数据场的本体映射算法的基本思想是:首先分别确定待映射本体中的相关子本体,然后在相关子本体间建立映射. 具体过程为:先使用高效的相似度计算方法计算任两个本体概念的相似度,并在此基础上得到本体中概念与另一个本体的相关度;然后,根据相关度的值选择相关概念并获得相关子本体;最后,利用更有针对性的映射方法(此类方法往往较复杂且比较消耗资源)实现相关子本体间的映射. 采用上述映射流程的主要目的是为了同时兼顾较高的映射效率和映射精度. 具体地说,高效的相似度方法在面对较复杂的本体时准确性往往不足,例如:在图 1 的例子中基于名字的策略会产生 Vehicle.Information 与 Travel.Information 映射的错误. 要获得更有效的映射效果,一般需要使用更有针对性的方法,比如基于结构的方法,但是这类方法往往比较消耗资源,当本体的规模较大时映射效率较低. 针对上述问题,本文中引入基于数据场的方法,通过数据场发现相关子本体,(1) 可降低使用针对性映射方法时的映射规模,从而提高使用该方法的效率;(2) 可过滤掉无关本体元素,提高映射精度. 文中实验验证了上述结论.

3.1 映射算法概述

算法 1. OntologyMapping(O_1, O_2).

输入:参与映射的本体 O_1 和 O_2.

输出:映射结果 R_M.

1. $Relevance(O_1,O_2)\to R_1,R_2$.
2. $Choose(O_1,R_1)\to O_1'$, $Choose(O_2,R_2)\to O_2'$.
3. $Match(O_1', O_2')\to R_M$.

其中,$Relevance(O_1,O_2)$方法用以分别获得 O_1 中的概念与 O_2、O_2 中的概念与 O_1 的初始相关度集合,记为 R_1、R_2. $Choose(O_1,R_1)$和 $Choose(O_2,R_2)$方法根据第 1 步得到的初始相关度,利用数据场势函数计算并抽取 O_1 中与 O_2 和 O_2 中与 O_1 的相关子本体 O_1' 和 O_2'. $Match(O_1',O_2')$选择针对性映射方法建立 O_1' 与 O_2' 的映射关系,结果存入 R_M. 这里针对性的方法,并不限定为某个具体的映射方法,而是根据待映射本体的特点确定. 如在本文的实验中,根据 GEMET 和 AGROVOC 本体都是多语言本体的特点采用了基于多语言编辑距离的方法.

下面分别对算法 1 中的前两个步骤作详细说明.

3.2 本体概念与本体间的初始相关度

计算当前本体中概念与其它本体初始相关度的目的是为了大致确定相关子本体的位置,无需特别精确,后面会利用数据场对此相关度进行修正. 因此,为了提高效率,通常使用高效、资源需求少的方法(比如:基于概念名称中的文本相似度的方法). 算法 2 描述了初始相关度的计算过程.

算法 2.　Relevance(O_1, O_2).

输入：待映射本体 O_1, O_2

输出：初始相关度集合

$R_1=\{(c_{1i}, r_{1i})\mid r_{1i}$为概念$c_{1i}\in O_1$与$O_2$的初始相关度$\}$,

$R_2=\{(c_{2j}, r_{2j})\mid r_{2j}$为概念$c_{2j}\in O_2$与$O_1$的初始相关度$\}$.

1. R_1, R_2初值为空.
2. For each $c_{1i}\in O_1$.
3. 　(c_{1i},0)加入集合 R_1.
4. For each $c_{2j}\in O_2$.
5. 　(c_{2j},0)加入集合 R_2.
6. For each $c_{1i}\in O_1$.
7. 　For each $c_{2j}\in O_2$.
8. 　　计算概念 c_{1i}和c_{2j}的相似度 s_{ij}.
9. 　　If $s_{ij}\geq\alpha$ then
10. 　　$x=(c_{1i}, r_{1i})\in R_1$, $y=(c_{2j}, r_{2j})\in R_2$.
11. 　　$x.r_{1i}=x.r_{1i}+s_{ij}$.
12. 　　$y.r_{2j}=y.r_{2j}+s_{ij}$.
13. Return R_1, R_2.

相似度小于 α 的概念对视为不相关概念对，不将其相似度累加入概念与本体的初始相关度中. α 的值可采用经验值或通过基于用户反馈获得. 如前所述，算法 2 中一般采用资源需求少、效率高的方法来计算概念之间的相似度. 常用的有基于编辑距离的相似度和基于 WordNet 的相似度[15].

基于编辑距离(Edit-Distance)的相似度. 给定两个单词 w_i, w_j，基于编辑距离相似度计算公式如下：

$$Ed(w_i, w_j)=\frac{|\{op_i\}|}{\max(len(w_i), len(w_j))} \tag{4}$$

$$S_e(w_i, w_j)=1/(1+Ed(w_i, w_j)) \tag{5}$$

其中，$|\{op_i\}|$为将 w_i 修改为 w_j 所需增、删、改字符的最少步数. $len(w_i)$为单词 w_i 的字符数.

基于 WordNet 的相似度. 给定的两个单词 w_i, w_j，基于 WordNet 的相似度计算公式如下：

$$S_w(w_i, w_j)=\frac{2\times\log(p(s))}{\log(p(s_i))+\log(p(s_j))} \tag{6}$$

其中，s_i, s_j 分别是单词 w_i, w_j 在 WordNet 语义树中的代表节点，s 为 s_i, s_j 的最近公共父节点，$p(s)$为以 s 为根的子树节点占整个 WordNet 语义树节点的比例.

例 1.　考虑图 1 中的例子，相似度计算方法采用编辑距离和 WordNet 相结合的策略. 具体方法是：如果 S_e 和 S_w 有一个为 1，结果相似度就为 1；否则取二者的算术平均作为最终的相似度.

表 1 为完成相似度计算后的相似度矩阵.

表 1　例 1 中的相似度矩阵

	Auto	Manufacturer	Info	Model
Car	1.00	0.27	0.05	0.24
Manufacturer	0.27	1.00	0.29	0.31
Description	0.05	0.04	0.45	0.19
Model	0.37	0.31	0.24	1.00
City	0.26	0.23	0.23	0.22
Name	0.30	0.31	0.26	0.34
Information	0.14	0.23	1.00	0.24

取 $\alpha=0.3$，根据算法 2 得

relevant(Car, $O_{\text{Automobile}}$)=1,

relevant(Manufacturer, $O_{\text{Automobile}}$)=1.31,

relevant(Description, $O_{\text{Automobile}}$)=0.45,

relevant(Model, $O_{\text{Automobile}}$)=1.68,

relevant(City, $O_{\text{Automobile}}$)=0,

relevant(Name, $O_{\text{Automobile}}$)=0.95,

relevant(Information, $O_{\text{Automobile}}$)=1,

relevant(Automobile, O_{Travel})=1.67,

relevant(Manufacture, O_{Travel})=1.62,

relevant(Information, O_{Travel})=1.45,

relevant(Model, O_{Travel})=1.65.

对于相关度为 0 的概念(如 O_{Travel} 中的 City)，为了后面基于数据场的相似度传播，需给它一个很小的值，比如 0.01.

3.3　基于数据场的相关子本体抽取

通常本体可由有向图 $G(V, E)$来表示，V 表示 G 中节点(概念)的集合，E 表示 G 中边(属性)的集合，节点间通过相应的边建立语义关联. 如果定义 G 中两个概念节点的最短路径长度为两个节点的距离，那么随着距离的增加，当前概念对其它概念的语义影响将会逐渐减弱，这与数据场及其势函数特征基本一致. 因此，本文利用数据场势函数来描述本体中各概念间相关度的相互影响.

如前所述，本体中概念的组织形式存在着多样性，无法找到唯一的一个势函数对所有的本体都适用. 用户可以通过对本体片断的观察及势函数曲线的特点选择合适的势函数. 在多数情况下，本体中概念影响距离较短，与拟核力场等短程场类似，此时可采用式(7)中的势函数形式：

$$\varphi_{c_{1j}}(c_{1i})=r_{1j}\times e^{-(d(c_{1i}, c_{1j}))^2} \tag{7}$$

其中，c_{1i} 和 c_{1j} 为本体 O_1 中的两个概念，r_{1j} 为 c_{1j} 与本体 O_2 的初始相关度. $d(c_{1i}, c_{1j})$为 c_{1i} 和 c_{1j} 之间的最短路径长度即 c_{1i} 到 c_{1j} 的距离.

根据式(3)，c_{1i} 在本体 O_1 中的势值为

$$\varphi_{O_1}(c_{1i})=\sum_j\varphi_{c_{1j}}(c_{1i}) \tag{8}$$

考虑到 c_{1i} 与本体 O_2 的相关度不但与它的势值有关还与 r_{1i} 有关，于是最终的相关度的计算公式为

$$r'_{1i}=r_{1i}\times\varphi_{O_1}(c_{1i}) \tag{9}$$

其中，r_{1i} 为算法 1 计算出的初始相关度. 算法 3 显示了如何选择相关概念构成新的本体.

算法 3. Choose(O,R).

输入：本体 O 及 O 中概念的初始相关度集合 R

输出：本体 O 中与另一本体相关的子本体 O'

1. 估算影响半径 r
2. For each $c_i\in O$
3. 令 $\varphi=0$.
4. For each $c_j\in O$, $j\neq i$, $d(c_i,c_j)\leq r$
5. $\varphi=\varphi+\varphi_{c_j}(c_i)$
6. $x=(c_i,r_i)\in R$.
7. $x.r_i=x.r_i\times\varphi$.
8. 估算过滤阈值 γ.
9. For each $c_i\in O$
10. $x=(c_i,r_i)\in R$.
11. If $x.r_i\geq\gamma$
12. 将 c_i 加入 C'.
13. 抽取子本体 O'，其中包含 C' 中所有的概念并保持这些概念在 O 中的关系.
14. return O'.

由 2.2 节中势函数的性质(3)和图 2 中的函数曲线所示，随着距离的增加常用势函数的值迅速减小并趋于 0. 为了减少参与运算的节点数量，提高算法的效率，引入影响半径 r 描述势函数的最大作用距离，忽略距离参数大于 r 的势函数值. r 值的估算方法是：令相应势函数中的 m 值取 1，选择最后一个使势函数值大于某个小阈值 β 的 r 作为算法中的影响半径. 例如：当采用式(7)中的势函数时，一般可取 2 作为 r 的值，因为 $e^{-9}<1.24\times10^{-4}$(距离为 3)已经足够小了.

第 8 步估算过滤阈值的方法是：用“黑点”表示完成算法 2 后相关度非 0 的节点，用“白点”表示相关度为 0 的节点. 取一个区间宽度 w，然后从 0 开始每隔 w 统计该区间内黑点的比例 p，当 p 第一次大于等于某个值(如 1.0)时，就取此时区间的下界作为 γ.

例 2. 根据例 1 的计算结果，采用式(7)的势函数，r 的值取为 2. Automobile 本体中的概念都对应“黑点”因此都会被选中. Travel 本体中各概念的相关度(进行了归 1 化处理)为

relevant(Car, $O_{\text{Automobile}}$)=1

relevant(Manufacturer, $O_{\text{Automobile}}$)=0.42

relevant(Description, $O_{\text{Automobile}}$)=0.15

relevant(Model, $O_{\text{Automobile}}$)=0.53

relevant(City, $O_{\text{Automobile}}$)=0.01

relevant(Name, $O_{\text{Automobile}}$)=0.02

relevant(Information, $O_{\text{Automobile}}$)=0.02.

如果我们取 $w=0.1$、$p=1.0$，则 $\gamma=0.1$，从而将 car domain 中的概念选出，排除了 city domain 中的概念，修正了 Automobile.Information 映射的错误.

4 实验与结果分析

为了验证本文方法效果，在 OAEI 2007 Environment 任务数据集上，设计了 2 个实验. 实验 1 在映射本体规模和耗时等方面对算法 1 的各个阶段进行了对比；实验 2 与直接使用算法 1 中第 1 步或第 3 步的映射方法在映射性能上进行了对比. 最后，给出了利用本文的方法参加 OAEI 2008 FAO 本体映射任务的评测结果，在所有参赛的方法中，本文的方法取得了综合评价第 1 的成绩.

本实验的所有程序均通过 Java 编码实现. 实验的运行环境为 AMD Athlon 4000+CPU、4GB 内存、Windows XP 操作系统.

4.1 实验数据

采用 OAEI 2007 Environment 本体映射任务数据集，包括 3 个大规模本体.

(1) GEMET 本体. 由欧洲环境署负责开发，多语言(超过 20 种)本体，包含 5280 个概念，这些概念涉及农业、空气、生物、气候、疾病等领域；

(2) AGROVOC 本体. 由联合国粮农组织开发，多语言(10 余种)，包含 28439 个概念，涉及农业、森林、渔业、食品、环境等领域；

(3) NAL 本体①. 美国国家农业图书馆开发，由英语描述，包含 42327 个概念，涉及生物、食品、环境和健康等 30 多个领域.

4.2 实验方法

方法 1. 基于编辑距离相似度计算的本体映射方法，只考虑概念的英文标签；

方法 2. 对实验数据具有针对性的本体映射方法. 仍然使用编辑距离进行相似度计算，每个概念的所有语言标签都会被考虑；

方法 3. 本文提出的方法，使用方法 1 作为算法 1 中第 1 步的相似度计算方法，使用方法 2 作为算法 1 中第 3 步的针对性映射方法.

① http://agclass.nal.usda.gov/agt/

这里使用编辑距离是因为它的效率较高，且针对实验数据具有较好的效果.

4.3　实验评估

竞赛的组织者根据实验数据构建了 3 个本体映射任务分别是：GEMET-AGROVOC（GA，其中 GEMET 为源本体，AGROVOC 为目标本体，后同）、GEMET-NAL(GN)和 NAL-AGROVOC(NA). 由于本体的规模太大，很难提供一个完整的参考映射，因此组织者提供了一个基于采样的局部参考映射[①]用于评估，基本统计信息见表 2.

表 2　参考映射基本信息表

(a) GEMET-AGROVOC

名称	领域	映射数	评价目标
p_chem	化学	14	查准率
p_geo	地理	23	查准率
p_misc	杂项	28	查准率
p_tax	分类	21	查准率
p_nat	自然	35	查准率
p_risk	风险	21	查准率
r_agri	农业	61	查全率
r_geo	地理	87	查全率

(b) GEMET-NAL

名称	领域	映射数	评价目标
p_chem	化学	30	查准率
p_geo	地理	17	查准率
p_misc	杂项	29	查准率
p_tax	分类	15	查准率
p_nat	自然	23	查准率
p_risk	风险	30	查准率
r_agri	农业	61	查全率
r_geo	地理	77	查全率

(c) NAL-AGROVOC

名称	领域	映射数	评价目标
p_chem	化学	141	查准率
p_geo	地理	58	查准率
p_misc	杂项	231	查准率
p_tax	分类	10	查准率
r_anim	动物健康	10	查全率
r_rod	啮齿类	24	查全率
r_oaks	橡树	38	查全率
r_eur	欧洲	62	查全率
r_geo	地理	58	查全率

对最终的实验结果采用查准率(precision)、查全率(recall)、F1-Measure 和运行时间进行评价.

查准率 P. 映射结果中正确的映射数与发现的映射总数的比值.

查全率 R. 映射结果中正确的映射数与全部正确映射数的比值.

F1-Measure. 综合查准率和查全率，为映射结果给出一个总体评价，其计算公式为

$$F1=\frac{2}{1/P+1/R} \tag{10}$$

4.4　实验结果

(1) 实验 1

针对算法 1 中的各个步骤，在映射本体规模和耗时等方面进行了对比.

表 3 给出了对于本实验 3 个映射任务，原本体概念数 $\sharp O$ 与相关子本体概念数 $\sharp O'$ 的对比.

表 3　原本体与相关子本体概念数对比表

本体名称	$\sharp O$	$\sharp O'$	$\sharp O'/\sharp O$
GEMET	5280	4801	0.91
AGROVOC	28439	5157	0.18
GEMET	5280	5001	0.95
NAL	42327	6831	0.16
NAL	42327	36851	0.87
AGROVOC	28439	24630	0.87
平均	25348.67	13878.50	0.66

由表 3 中可以观察到，当源本体与目标本体规模相差较大时，较大本体对应相关子本体的规模相较原本体下降较多，大致与较小本体的规模相当.

方法 3 除了进行映射操作外，第 2 步还要从本体中抽出相关子本体. 表 4 对比了方法 3 中各步操作所消耗的时间.

表 4　方法 3 中各步骤耗时对比表

任务名	耗时/s			
	第 1 步	第 2 步	第 3 步	总时间
GA	17596	5	21304	38905
GN	49145	6	6389	55540
NA	200500	12	151258	351770

从表 4 可以看出，抽取子本体的时间远远小于进行映射的时间. 表明基于数据场的相关子本体抽取过程具有较高的效率.

(2) 实验 2

针对 4.2 节中的 3 个方法，根据 4.3 节中列出的评价指标分别进行评价.

由于 NAL 本体为单语言本体，因此对于 GN 和 NA 任务，方法 2 与方法 1 成为同一个方法. 表 5 对比不同方法在本实验的 3 个子任务中所消耗的时间.

表 5　不同方法耗时对比表

任务名	耗时/s			方法 3 第 3 步耗时/方法 2 耗时
	方法 1	方法 2	方法 3	
GA	17592	131840	38905	0.16
GN	49491	49491	55540	0.13
NA	201356	201356	351770	0.75

① http://oaei.ontologymatching.org/2007/results/environment/gold_standard/

表 5 的结果显示，由于 GA 任务中针对性的映射方法(方法 2)的效率较低，且从 AGROVOC 本体中抽取的新本体规模大大低于原本体，方法 3 的耗时远低于方法 2，为方法 2 的 30%，但要高于方法 1. 而在 GN 和 NA 任务中方法 2 与方法 1 相同，方法 3 与之相比需要消耗更多的时间. 即便如此，在 GN 任务中，由于对 NAL 本体的压缩率较高，方法 3 耗时仍与方法 2 相当(1.1 倍). 表 5 的最后 1 列在消耗时间上对方法 3 第 3 步和方法 2 进行了比较，由于压缩了映射规模，方法 3 可有效地提高使用针对性映射方法时的效率，特别是当待映射本体间规模或包含领域相差较大时，效果更为明显. 综合以上分析，从映射效率的角度考虑，方法 3 最适用于源本体与目标本体差异较大且针对性映射方法效率较低的本体映射任务.

图 3、图 4、图 5 分别为不同方法作用于相关采样数据集上的查准率、查全率和整体 $F1$-Measure. 其中，图 3、图 4 中 overall 的值由图中对应值按表 2 中所列的每个参考映射的映射数加权平均获得，图 5 中 $F1$-Measure 的值由对应 overall 的值根据式(10)计算而来，实验中主要通过上述数值对实验方法进行评价.

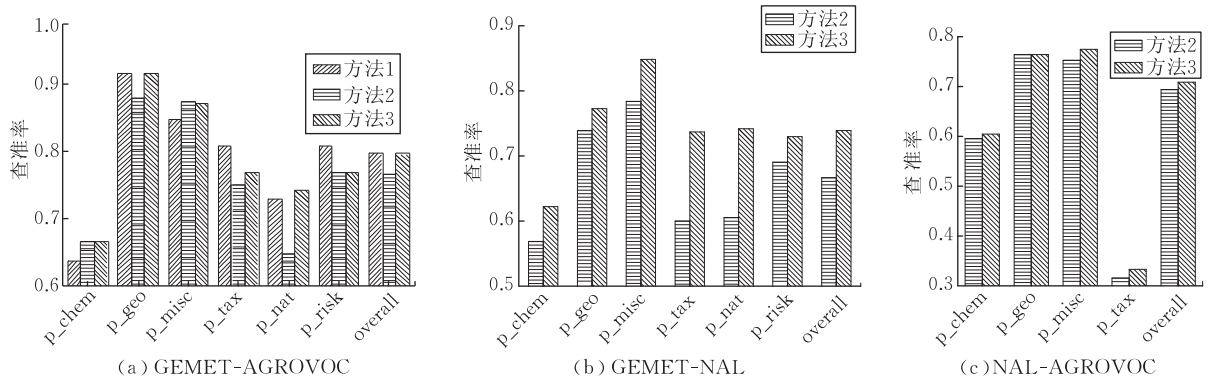


图 3 各方法基于评价查准率采样数据对比实验结果

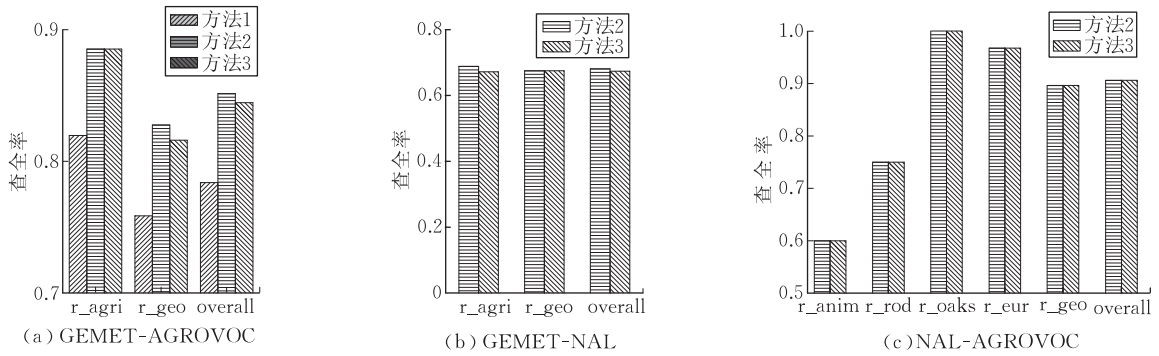


图 4 各方法基于评价查全率采样数据对比实验结果

由图 3 可以看到，方法 3 的查准率要比直接使用方法 2 高(注意在图(b)、(c)中方法 2 与方法 1 为同一方法)，这表明在方法 3 第 3 步使用针对性映射方法进行映射之前，已经过滤掉了不相关元素，从而提高了使用针对性方法时的准确率. 图 4 显示，方法 3 的查全率值与方法 2 的查全率值相当，在图(a)所示的映射结果中明显优于方法 1 的查全率，这表明基于数据场的相关子本体抽取方法可以保留绝大部分的相关元素，从而保证使用针对性映射方法时的召回率. 图 5 通过 $F1$-Measure 值给出了对参与实验的 3 个方法的整体评价，方法 3 的 $F1$-Measure 值在所有的 3 组映射任务中都是最高的. 另外，在 GA 任务中更具针对性的方法 2，由于所有语言标签都要考虑，在发现更多映射的同时也引入了较多错

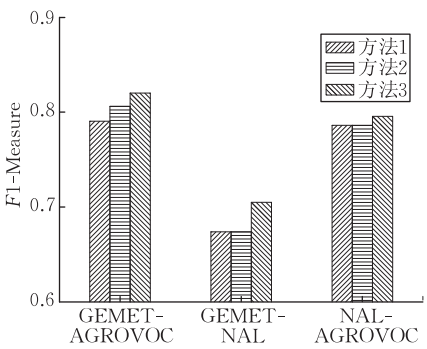
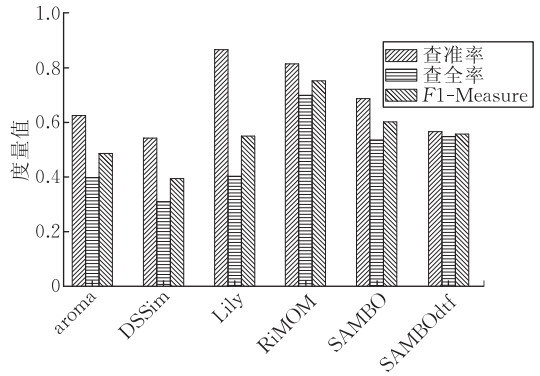


图 5 各方法整体 $F1$-Measure 对比实验结果

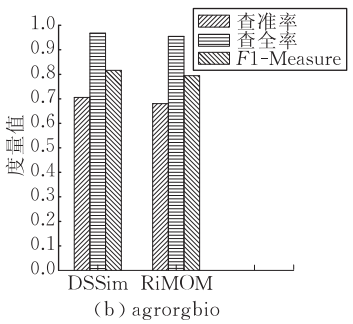
误，因此整体查全率高于方法 1，但整体查准率比方法 1 低，从整体映射效果看稍占优势.

4.5　OAEI 2008 FAO 任务评价结果

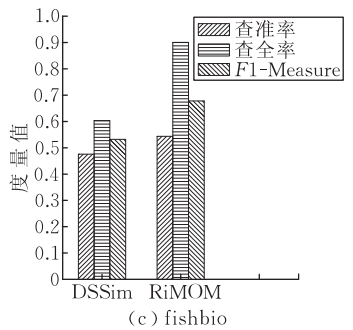
图 6 为 OAEI 2008 FAO 任务组织者给出的评价结果①. 该任务的数据集包含 3 个多语言本体：AGROVOC、ASFA 和 Fisheries 本体. RiMOM(Risk Minimization based Ontology Mapping)为采用本文方法的系统. 由图 6 所示，从 RiMOM 本身看，3 个子任务的结果是比较平均的，$F1$ 的值都在 0.65～0.8 之间. 从不同映射工具的对比看，RiMOM 在两个任务(agrafsa 和 fishbio)的综合的映射结果($F1$-Measure)是最好的，另外一个与最好的 DSSim 相当(约低 0.02).



(a) agrafsa



(b) agrorgbio



(c) fishbio

图 6　OAEI2008 FAO 任务评价结果

4.6　实验结论

上述实验结果表明：

(1) 基于数据场的本体映射算法可以有效地过滤不相关本体元素，消除这类本体元素对映射结果的影响，提高映射的准确率，同时该算法可以保留大多数相关本体元素，确保映射的召回率，进而提高整体的映射效果.

(2) 基于数据场的本体映射算法，可以有效压缩映射的规模，提高使用针对性映射方法时的映射效率. 特别是当待映射本体间规模或包含领域差异较大时，效果更为明显. 若针对性的映射方法效率较低，本文的方法还有助于提高整体的映射效率.

(3) 基于数据场的相关子本体抽取算法具有较高的效率.

5　相关工作

按建立映射关系所使用信息的不同，可将主要的本体映射方法分为基于元素名称、基于本体结构、基于本体实例、基于推理和基于背景知识的方法.

(1) 基于元素名称的方法. 在所有的本体映射方法中，基于本体元素名称的方法是最简单同时又是最基本的. 它主要利用本体元素的名称、标签和注释中的文本信息来建立元素间的映射关系. 可分为两类：一类基于本体名称中的字符序列，文献[1]中对比了主要的基于字符序列的方法，比如：基于编辑距离、Token 和 TDIDF 的方法. 另一类基于计算语言学，例如文献[2]提出的基于 WordNet 的方法.

(2) 基于本体结构的方法. 该类方法利用本体元素的结构信息来发现它们之间的映射关系. 典型的算法如：Similarity Flooding[4]，它根据待映射本体的图结构，构造相似度传播图，并利用迭代的方法对图中节点所代表的元素对的相似度进行传播和修正. 本文中提出的基于数据场势函数的相关度修正算法与该算法相似，但由于本文中的方法无需构建相似度传播图(对每个本体概念只需一个额外的存储单元以存放其相关度的值)，也无需迭代计算，因此具有较少的空间占用和时间消耗.

(3) 基于本体实例的方法. 此类方法利用本体中实例间的映射来确定本体概念或属性间的映射关系. 文献[5]根据概念间公共实例数量来衡量概念的映射关系. 文献[6]通过机器学习的方法，利用实例数据建立概念间的映射.

(4) 基于推理的方法. 一般而言纯推理的方法

① http://oaei.ontologymatching.org/2008/results/FAO-results.html

并不适用于本体映射,但是推理技术可对本体映射起一定的辅助作用.例如:文献[7]中实现了一个OWL Lite推理器,该推理器可以一种固定的顺序选用推理规则,同传统的数据映射技术相结合,提高了映射的性能.

(5)基于背景知识的映射方法.所谓本体的背景知识就是与本体相关的外部信息.文献[9]中提出了一个十分新颖的映射方法,它主要针对层次结构的目录型本体,利用Google搜索引擎计算某概念在其概念路径中的权重,并利用这些权重最终得到概念间的相似度.

6 结论和未来的工作

针对有效、高效的处理大规模本体映射的问题,本文提出了一个基于数据场的本体映射算法.该算法首先通过一种高效的相似度计算方法获得当前本体的元素与另一本体的初始相关度,然后利用数据场量化本体元素间的相互影响以修正初始相关度,进而实现相关子本体的发现与抽取.最后,选用性能较高却通常比较复杂的针对性映射方法在抽取后的本体上进行更为有效的映射.该算法并不限制使用任何具体的映射算法,具有较强的灵活性.实验结果表明该算法可以在提高映射结果质量的同时保证较高的映射效率.

目前,该算法采用用户反馈的方式对某些参数、阈值进行选择和设置(例如:数据场势函数),这对用户提出了较高的要求,用户对数据的熟悉程度会对映射效果产生一定的影响.今后,将利用本体自身结构和本体元素间的语义关系自动地完成上述设置,提高映射的效率和结果的质量.

参 考 文 献

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

[2] Budanitsky A, Hirst G. Evaluating WordNet based measures of lexical semantic relatedness. Computational Linguistics, 2006, 32(1): 13-47

[3] Do H-H, Rahm E. COMA—A system for flexible combination of schema matching approaches//Proceedings of the 28th International Conference on Very Large Data Bases(VLDB). Hong Kong, China, 2002: 610-621

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

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

[6] Wang S, Englebienne G, Schlobach S. Learning concept mappings from instance similarity//Proceedings of the 7th International Semantic Web Conference (ISWC). Karlsruhe, Germany, 2008: 339-355

[7] Udrea O, Getoor L, Miller R. Leveraging data and structure in ontology integration//Proceedings of the 26th International Conference on Management of Data (SIGMOD). Beijing, China, 2007: 449-460

[8] Aleksovski Z, Klein M, Kate W, Harmelen F. Matching unstructured vocabularies using a background ontology//Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW). Podebrady, Czech, 2006: 182-197

[9] Gligorov R, Aleksovski Z, Kate W, Harmelen F. Using Google distance to weight approximate ontology matches//Proceedings of the 16th International World Wide Web Conference (WWW). Beijing, China, 2007: 767-776

[10] Li J, Tang J, Li Y, Luo Q. RiMOM: A synamic multistrategy ontology alignment framework. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2009, 21(8): 1218-1232

[11] Li De-Yi, Du Yi. Artificial Intelligence with Uncertainty. Beijing: National Defense Industry Press, 2005(in Chinese)
(李德义,杜鹢. 不确定性人工智能. 北京:国防工业出版社, 2005)

[12] Zhong Q, Li H, Li J, Xie G, Tang J, Zhou L, Pan Y. A gauss function based approach for unbalanced ontology matching//Proceedings of the 28th International Conference on Management of Data (SIGMOD). Rhode Island, USA, 2009: 669-680

[13] Tang Jie, Liang Bang-Yong, Li Juan-Zi, Wang Ke-Hong. Automatic ontology mapping in semantic web. Chinese Journal of Computers, 2006, 29(11): 1956-1976(in Chinese)
(唐杰,梁邦勇,李涓子,王克宏. 语义Web中的本体自动映射. 计算机学报, 2006, 29(11): 1956-1976)

[14] Gan Wen-Yan, Li De-Yi, Wang Jian-Min. An hierarchical clustering method based on data fields. Acta Electronica Sinica, 2006, 34(2): 258-262(in Chinese)
(淦文燕,李德毅,王建民. 一种基于数据场的层次聚类方法. 电子学报, 2006, 34(2): 258-262)

[15] Lin D. An information-theoretic definition of similarity//Proceedings of the 15th International Conference on Machine Learning (ICML). Madison, Wisconsin, USA, 1998: 296-304

ZHONG Qian, born in 1975, Ph. D. candidate. His current research interests include semantic data integration and ontology mapping.

LI Juan-Zi, born in 1964, professor, Ph. D. supervisor. Her current research interests include Semantic Web, Web service and knowledge management.

TANG Jie, born in 1977, Ph. D., associate professor. His current research interests include large scale social network and machine learning.

ZHOU Li-Zhu, born in 1949, professor, Ph. D. supervisor. His current research interests include database system, digital library and massive information processing.

Background

In the infrastructure of the Semantic Web, ontology has become a dominant mechanism to represent the data semantics on the Web. A vast amount of data has been constructed through ontologies. Unfortunately, the heterogeneity among the ontologies brings a noticeable puzzle in sharing and integrating these semantic enriched datasets. Aiming to perform interoperation across the heterogeneous ontologies, ontology mapping have attracted much attention from research community.

The process of ontology mapping takes as input two ontologies and determines a set of relationships between concepts in the ontologies. Existing solutions utilize various techniques to attain satisfying mappinging results, such as name-based, structure-based, instance-based, external knowledge-based and reasoning-based methods. In addition, compound solutions which employ multiple techniques and aim to process various mapping scenarios are proposed. Such solutions include COMA, RiMOM, H-Match and Cupid etc. However, with increasing size of the ontologies, the large scale ontology mapping problem bring big challenges to the existing ontology mapping approaches. Focus on this issue, in this paper, we propose a data field based ontology mapping method. The motivation of our approach is to employ an efficient method to obtain the relevant sub-ontologies in each matched ontology and then use a special ontology mapping approach to perform the mapping process between the discovered sub-ontologies. Because of downsizing the mapping scales and filtering out the irrelevant ontology elements, the approach can improve both the effectiveness and the efficiency of the special ontology mapping approach.

This paper attributes to the project RiMOM (Risk Minimization based Ontology Mapping), which is supported by the National Natural Science Foundation of China (NSFC) under grant Nos. 60973102 and 60703059, the National Basic Research Program of (973 Program) China under grant No. 2007CB310803 and the National High Technology Research and Development Program (863 Program) under grant No. 2009AA01Z138. RiMOM is a tool for ontology mapping by combining different strategies, aiming at finding the "optimal" alignment results. The website of the project is http://keg.cs.tsinghua.edu.cn/project/RiMOM/. There are several papers in this project published on some international journals and conferences, such as TKDE, JoWS and SIGMOD. In OAEI(Ontology Alignment Evaluation Initiative) 2006/2007/2008/2009, RiMOM was ranked the top three among all the participated systems in many mapping tracks, such as the Benchmark, FAO, Directory, Anatomy, MLDirectory and Instance Matching.