

语义集成中的不确定性本体映射模型

黄黎^{1,2}, 方巍^{1,2}, 崔志明^{1,2}

HUANG Li^{1,2}, FANG Wei^{1,2}, CUI Zhi-ming^{1,2}

1.江苏省计算机信息处理技术重点实验室, 江苏 苏州 215006

2.苏州大学 智能信息处理及应用研究所, 江苏 苏州 215006

1.Jiangsu Key Laboratory of Computer Information Processing Technology, Soochow University, Suzhou, Jiangsu 215006, China

2.Institute of Intelligent Information Processing and Application, Soochow University, Suzhou, Jiangsu 215006, China

E-mail: szhuangli@gmail.com

HUANG Li, FANG Wei, CUI Zhi-ming. Ontology mapping model with uncertainty in semantic integration. Computer Engineering and Applications, 2009, 45(28): 140-144.

Abstract: Semantic Web is developing with high speed and features of dynamic and heterogeneous. Resolving the heterogeneity in semantic information is the key of information integration. Ontology is a form of knowledge expression for the semantic Web information, which enhances the semantic information for Web. To resolve semantic heterogeneity, and realize interoperation, it is in need of establishing mappings between heterogeneous ontologies. However, establishing absolute precise ontology mappings for huge heterogeneous ontologies is unpractical. There is usually some degree of uncertainty in ontology mappings. This paper proposes a new ontology mapping framework, ontology mapping with uncertainty in semantic integration, researches various ontology features, integrates several mapping strategies, and introduces uncertain matching in each mapping strategy. Experiment shows that this new approach has a reliable performance, and is more general and scalable.

Key words: semantic Web; ontology integration; ontology mapping; uncertainty matching

摘要: 语义 Web 的高速发展使其具有动态性和异构性特征, 解决语义信息的异构性问题成为实现信息集成的关键。本体作为一种语义 Web 的知识表示形式, 增强了 Web 的语义信息。因此, 为了解决语义异构性, 实现数据间的互操作, 必须建立异构本体间的映射关系。然而, 为庞大的异构本体建立完全精确的本体映射是不现实的, 本体映射中存在一定程度的不确定性。提出了一种新型的本体映射框架——语义集成中的不确定性本体映射。从不同方面研究本体特征, 集合了多种映射策略, 并引入了各映射策略中不确定性匹配的解决方案。实验证明, 该方法具有可靠的实验性能, 并且具有很好的通用性和可扩展性。

关键词: 语义 Web; 本体集成; 本体映射; 不确定性匹配

DOI: 10.3778/j.issn.1002-8331.2009.28.042 **文章编号:** 1002-8331(2009)28-0140-05 **文献标识码:** A **中图分类号:** TP18; TP311

1 引言

Web 信息量的不断扩大已远远超出了计算机的处理能力, 使 Web 资源无法得到有效的利用。语义 Web 的诞生是对传统 Web 的扩展, 使 Web 信息易于被计算机理解, 实现计算机之间的智能交互^[1]。本体(Ontology)作为描述语义 Web 的建模手段, 形式化地描述了领域知识, 为计算机处理提供了语义信息, 并成为语义 Web 的核心。最广泛的本体定义是由 Gruber 提出的, 即“本体是概念模型的明确的规范说明”^[2]。

本体为知识表示提供了一种概念化和领域化的描述方式, 在一个较大的领域范围内对知识做出了统一、明确的定义, 便

于计算机理解和互操作。为了实现语义信息的共享, 越来越多的领域相应定义属于本领域的本体。由于 Web 本身的异构性, 即使是同一领域甚至同一机构对各自本体的描述也是各不相同的, 因此, 容易造成本体描述信息的异构性。而资源的异构性和自治性为信息检索造成了巨大的障碍, 亟待一种统一有效的资源管理方式。一方面, 本体描述逻辑为信息检索提供了智能化手段, 便于机器的理解, 可以为用户提供更丰富的资源。另一方面, 本体作为一种语义资源的描述手段, 其本身存在的异构性也为语义信息集成带来了新的挑战。本体之间的异构常常表现为语法上的不匹配和语义上的不匹配, 解决本体异构的关键

基金项目: 国家自然科学基金(the National Natural Science Foundation of China under Grant No.60673092); 2005 年度教育部科研重点项目(the Key Project of Chinese Ministry of Education under Grant No.205059); 2006 年江苏省“六大人才高峰”项目(the “Six Talent Peak” Project of Jiangsu Province under Grant No.06-E-037); 2006 年度江苏省软件和集成电路业专项经费项目(the Specialized Fund Project for the Software and IC of Jiangsu Province in 2006 under Grant No.[2006]221-41)。

作者简介: 黄黎(1982-), 女, 硕士生, 主要研究方向: Deep Web, 数据集成, 语义 Web; 方巍(1975-), 男, 博士生, 主要研究方向: 数据挖掘, 数据集成, Deep Web, 语义 Web; 崔志明(1961-), 男, 博士生导师, 教授, 主要研究方向: 智能化信息处理, 计算机网络与数据库应用。

收稿日期: 2008-05-26 **修回日期:** 2008-08-15

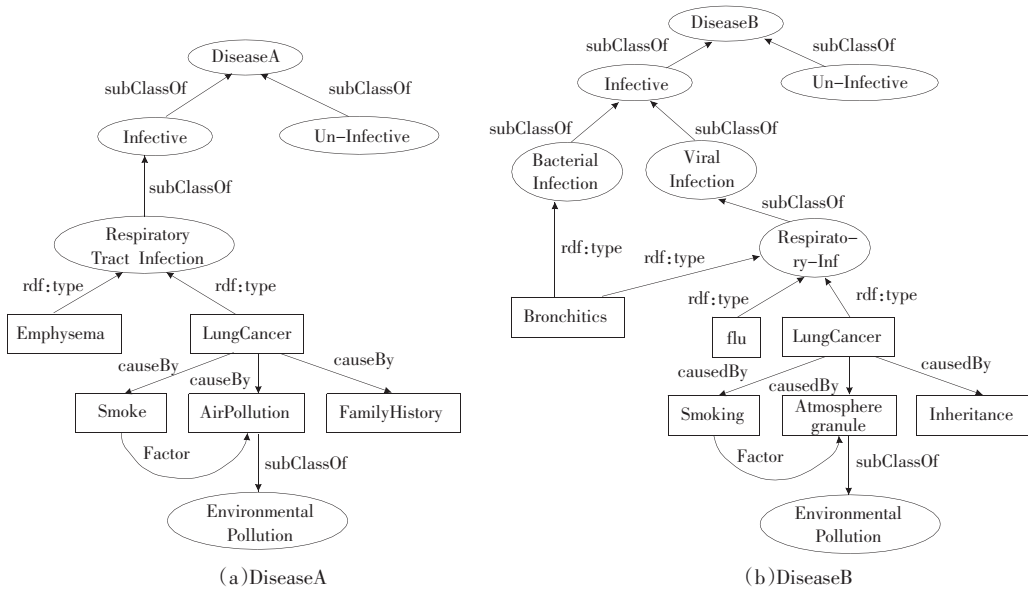


图1 异构本体实例

就是建立本体之间的映射。

目前已经有很多自动或半自动的本体映射方法,这些方法都基于启发式推理和机器学习技术,它们本身不能提供完全精确的本体映射。本体映射的建立需要领域专家的参与,根据领域知识建立本体之间的映射关系,但是对大规模的异构本体,描述清晰的映射关系就显得越来越困难。因此,建立完全准确的本体映射关系是不可能的。在分布式语义信息集成的过程中,信息量的巨大和人们领域知识的不完备性,使本体映射的建立存在不确定性。其次,任何机构都无法胜任语义标注的重任,语义标注本身的不完善也会带来本体映射的不确定性。要实现异构本体的集成,就必须解决本体映射的不确定性,真实反应现实语义 Web 中的不确定性问题。不确定性本体映射是实现数据空间的关键步骤,能够为搜索提供更加合理的响应结果^[3]。现有的本体映射方法强调一种精确的匹配,缺乏可扩展性,为了提高集成的质量,亟待一种不确定性本体映射方法来解决语义集成中的不确定问题。

该文将不确定性引入语义集成的本体映射中,重点讨论语义上的不匹配问题,旨在提高语义集成的通用性和可扩展性,改善查询结果响应的质量。该方法为语义集成提供了一种新的描述框架——不确定性本体映射(UnOM),准确地描述了本体映射的特征,有效地解决了本体的异构性和本体映射的不确定性,使语义集成更加可靠,更有可扩展性。在第2章中,首先介绍本体的相关术语,分析本体的异构性,并对本体映射进行形式化描述;第3章是对不确定性本体映射模型进行总体概括,引入基于概率的实体映射模式以刻画建立本体映射时存在的不确定性;第4章分别对不确定性本体集成的多策略进行详细阐述,并对多个映射策略进行合并;第5章是对不确定性本体映射模型的实验验证和分析;第6章介绍该领域相关研究;最后是对全文的总结与展望。

2 本体问题定义

2.1 本体形式化描述

本体提供了对领域知识的共同理解,使概念和关系在共享的范围内具有公认的、明确的、唯一的定义,便于人机交互以及

机器之间的交互。不失一般性,给出一种本体结构的形式化描述^[4]: $O=\{C,R,I,A\}$ 。其中 C 是概念的集合(concept), c 表示概念($c \in C \times C$); R 是关系的集合(relation), r 表示关系($r \in R \times R$); I 表示实例集合, i 表示实例($i \in I$),是对本体概念的特化; A° 表示本体公理(axiom),表示永真断言。简而言之,本体学习就是本体概念获取、概念之间关系的获取,以及公理的获取。

本体概念是对特定领域中一个或一类事物的概括描述。从语义角度来说,概念是对个体的抽象,概念的详细解释由概念中的若干属性来描述,概念与概念之间的差异也就体现在其包含的属性之间的区别。

在本体中,关系刻画了实体之间的联系,分为分类关系和非分类关系。分类关系表示实体之间的父类和子类的上下位关系,而非分类关系反映的是除上下位关系以外的其他关系。关系集合 R 中的每个关系 $r_i(e_i, e_j)$ 反映的是实体 e_i 和 e_j 之间的二元关系,表示实体之间的联系。因此,本体之间的关系可以用树或图模型来形式化描述实体之间的关联,通过层次结构反映概念之间的分类,利用关联边表示属性之间的关系。该文将本体概念、属性和关系元素统称为实体元素 $e(e_i \in C \cup R)$,作为研究对象。

2.2 本体映射

在分布式语义信息环境中,不同领域、不同组织定义大量的本体,其目的是实现知识的共享,但是由于本体自身构建的主观性和分布性,造成本体在表示语言和表示模型上存在差异,产生了本体的异构性问题。本体之间的不匹配是造成本体异构的直接原因,它包括语法层上的不匹配和语义层上的不匹配。语法层的异构是指数据类型、格式的差异;语义层的异构则是指在一定领域内专用的词汇意义的共享和交流。主要表现为:(1)不同的数据源使用多种术语(词汇)表示同一概念;(2)同一概念在不同的信息源中表达不同的含义;(3)相同名字在不同本体中表示不同的概念;(4)各数据源使用不同的结构来表示相同(或相似)的信息;(5)各数据源中的概念之间存在着各种联系,但由于本体的异构性使这种隐含的联系无法体现出来^[5]。图1描述了医学领域的两个异构本体 DiseaseA 和 DiseaseB,本体 DiseaseA 中的“Infective”包含“Respiratory Tract

Infection”,而本体 DiseaseB 中的“Infected”因划分粒度不同,包含“Bacterial Infection”和“Viral Infection”;本体 DiseaseA 中的“Respiratory Tract Infection”和本体 DiseaseB 中的“Respiratory-Inf”,虽然形式不同,但都表示相同的概念。

解决本体异构的关键途径是本体集成和本体映射。传统的本体映射是建立相似实体之间的联系,它形式化为一种朴素的二元关系 $map(O_1, O_2)$,反映出本体 O_1 中所有实体在本体 O_2 中寻找相似语义实体并建立对应关系的过程。本体映射中的实体可以包含概念、属性或者关系,传统的映射本身只能反应出实体 e_1 和 e_2 之间可能存在二元映射关系($e_1 \in O_1, e_2 \in O_2$),而无法表达二者之间映射关系的可靠程度。Zedeh^[6]提出的模糊集为模糊理论奠定了基础,随着自然语言处理和信息论的发展,模糊逻辑可以很好地应用于语义本体的映射中,能较完备地表述语义映射的不确定性。因此,定义了一种基于概率的不确定性本体映射模型,准确地刻画本体映射的语义匹配中的不确定性,使整个本体集成框架更加合理,易于实现框架的可扩展性。因此,将不确定性本体映射形式化定义为如下形式:

$$map = Umap(\{e_i\}, \{e_j\}, O_1, O_2, u) \quad (1)$$

其中, $e_i \in O_1, e_j \in O_2$, 分别表示本体中的实体元素。本体映射的种类包括 1:1, 1:n, m:n 等。因此,本体中实体映射的关系比较复杂,在语义匹配不确定性的前提下,使用了基于概率模型下的变量 u 来表征实体间关系的相似程度,从而满足本体语义映射的需要。

3 不确定性本体映射模型

所提出的不确定性本体映射模型是一种多映射策略有机组合和迭代的过程,从本体的数据本身和结构特性出发,改变原有的简单映射方式,引入不确定逻辑,使整个映射策略更加合理,更有利于异构本体的集成。不确定性本体映射模型包括如下几个过程:实体映射预定义、多策略映射建立、多映射关系整合及反馈,如图 2 所示。

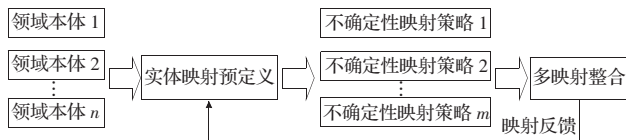


图2 不确定性本体映射模型

实体映射预定义需要领域专家的参与,由用户和专家预先对领域本体中的各个实体建立简单的映射关系。映射规则可以是基于实体数据本身的,按照语义和语法的相似性,实现将源本体实体映射到目标本体实体上的过程。也可以是基于本体结构层次的,按照本体构造的层次关系,通过实体的上下文结构反映出实体间的映射关系。

预定义的本体映射集为多策略不确定性映射提供了输入集合,该策略集合了多种不确定性映射方法,包括基于概念语义的不确定性映射、基于实体内容的不确定性映射和基于本体结构的不确定性映射。在每种映射策略中,都综合了相对复杂的不确定性映射方法,改变了传统的朴素二元映射关系,为最终的映射合并提供更加合理的候选映射的预测值。

综合考虑多个映射策略产生的候选映射关系。为了提高整个映射过程的效率,为映射的建立设定相似度阈值,筛选出大

于该阈值的映射关系作为候选映射。然后,将所有候选映射综合得到整体上的映射预测值。

综合映射预测值还不是最终的映射结果,需要人工的参与和反馈,利用逐步获得的领域知识和推理对其进行优化,直到发现满意的映射结果为止。

4 不确定性本体映射策略

在语义集成的不确定性本体映射框架中,运用了多个本体映射策略,从不同方面对本体特征进行描述,包括本体概念的语义特性、本体的内容特征和本体的结构特征等多个方面,尽可能地发掘可能存在的映射关系,从不确定性角度表述映射过程。

4.1 基于概念语义的不确定性映射

不同的领域本体有其特定的本体定义规范,表达能力各不相同,实体定义也千差万别,因此出现很多同义词、同形异义词、上位词或下位词等。通过同义词之间在语义上的相似性,可以建立异构本体中同义概念之间的等价映射关系。目前有许多通用语义辞典为词汇相似度计算提供依据,WordNet 就是其中较为常用的一个。WordNet 根据语义组织词汇,因此异构本体词汇间的同义、反义、上下位关系都可以由 WordNet 方便地获得。显然,两个实体词汇越接近,它们的相似度越大。

WordNet 只是一个通用的语义辞典,它并不能涵盖所有的领域本体,因此在任意两个词汇之间建立语义映射关系都存在一定程度的不确定性,需要用基于概率的方法进行规约化。利用概率统计方法,可以尽可能地依据辞典内容推理获得词汇映射间的不确定性程度,记为 $uc(w_1, w_2)$ ($uc(w_1, w_2) \in [0, 1]$), w_1 和 w_2 分别表示两个不同的词汇。如果 w_1 和 w_2 可以在 WordNet 中找到相应关系,则 $uc(w_1, w_2)=1$ 。该文以 WordNet 为依据,将概念语义的不确定性映射匹配定义为:

$$Sim_{uc}(e_1, e_2) = \frac{2 \times \log P(e)}{\log p(e_1) + \log p(e_2)} \times uc(e_1, e_2) \quad (2)$$

其中, e 是 e_1 和 e_2 的公共祖先节点, $P(e)$ 表示辞典中节点 e 及其子节点包含的单词数量占整个辞典词汇量的比例, $uc(e_1, e_2)$ 表示词汇 e_1 和 e_2 之间匹配的不确定程度。本体概念的属性和关系的不确定性映射关系同样可以用这种方法来度量。

4.2 基于实体内容的不确定性映射

基于实体内容的不确定性映射主要考虑运用实体本身内容的相似程度来衡量,通过不同概念之间的编辑距离和差异替换来计算两者间的相似度,编辑距离越短,则两个实体越相似。对于复杂词汇、短语以及缩略词等,首先需要利用领域知识对词语进行标准化,为映射的建立提供更加确切的词汇集合,建立异构词汇集合之间的综合映射关系。将基于实体内容的不确定性映射匹配定义为:

$$Sim_m(e_1, e_2) = \frac{\sum_{w_1 \in e_1, w_2 \in e_2} \max sim(w_1, w_2) + \sum_{w_2 \in e_2, w_1 \in e_1} \max sim(w_1, w_2)}{|e_1| + |e_2|} \quad (3)$$

其中 $w_1 \in e_1, w_2 \in e_2$ 分别表示实体 e_1 和 e_2 中标准化后的词汇片断,很自然地就产生了两个词汇集合之间的映射关系,构造一个映射矩阵,在源词汇和目标词汇的相似度预测中选择最大的作为映射匹配值,即 $\max sim(w_1, w_2)$ 。

4.3 基于本体结构的不确定性映射

基于本体结构的不确定性映射利用本体结构中的上下文结构建立映射关系。本体通常以树或图的形式表示,它恰恰能反映出概念之间的一种上下位层次关系,即父类与子类的关系。当两个实体具有相同或相似的上下文结构时,就不难判断这两个实体之间必然存在一定程度上的相似关系,可以在两者之间建立映射。无论是实体本身还是实体的上下文环境中的映射建立都可能存在匹配的不确定性,上位词和下位词所建立的映射关系中的不确定性,以及实体本身的不确定性匹配都会影响实体映射关系的产生,因此从上下文结构出发,综合考虑实体的映射关系。

实体通过其特性约束(如:rdfs:subClassOf,rdfs:superClassOf,rdfs:subPropertyOf,rdfs:superPropertyOf),利用其上位词(hypernyms)和下位词(hyponyms)反应出父类与子类之间的包含关系。例如图1中,本体DiseaseA中的AirPollution和本体DiseaseB中的Atmosphere granule分别具有相同或相似的上下文结构,均属于LungCancer,并且都具有Smoke的导致因素,所以可以推断二者是相似的。实体映射间的不确定性具有传递性,它会影响到相邻节点的匹配度量值。将基于本体结构的不确定性映射匹配描述如下:

$$Sim_{struct}(e_1, e_2) = Sim_{hyper}(e_1, e_2) + Sim_{hypon}(e_1, e_2) + Sim(e_1, e_2) \quad (4)$$

其中, $Sim_{hyper}(e_1, e_2)$ 和 $Sim_{hypon}(e_1, e_2)$ 分别表示实体 e_1 与 e_2 之间的上位词和下位词的不确定性匹配相似度,其计算方法可以利用4.1节中的概念语义不确定性匹配方法, $Sim(e_1, e_2)$ 表示实体本身不确定性匹配相似度,可以利用4.2节中的实体内容不确定性匹配方法来度量。

4.4 多不确定性映射策略整合

将多个不确定性映射策略进行合理整合,才能更加恰当地反映出异构本体之间真实的映射关系。通过如下公式将上述几种不确定性映射策略进行合并:

$$HSim(e_1, e_2) = \lambda_1 Sim_{lex}(e_1, e_2) + \lambda_2 Sim_{lin}(e_1, e_2) + \lambda_3 Sim_{struct}(e_1, e_2) \quad (5)$$

其中 $\lambda_i(i=1, 2, 3)$ 是为每个映射策略分配的权重, $\lambda_i \in [0, 1]$, $\sum \lambda_i = 1$ 。权重的确定可以在实验中不断调整,以获得最佳的映射关系。

5 实验验证

通过具体实验,将提出的新模型与其他本体映射策略进行比较。实验表明,该不确定性本体映射框架在语义集成中更加有效,具有通用性和可扩展性。

5.1 实验建立

为了验证不确定性本体映射的高效性,从Bremen大学索引的实际本体库(<http://www.fb10.uni-bremen.de/anglistik/lang-pro/webospace/jb/info-pages/ontology-root.htm>)中选择了2个关于医学领域的异构本体信息作为数据源。在本体信息中,包含了1047个类、3823个属性以及5983个实例等信息,分别反映了医学本体的领域知识。

首先,由领域专家对异构本体数据源的子集按照语义和结构的相似性对实体建立简单的本体映射关系,为多映射策略的实现提供候选映射依据。从领域专家提供的映射依据可以分析得出,大约有4%的实体具有相同的名称,25%的概念相同,

19%的属性蕴含,以及10%的属性等价。然后依据各个不确定性本体映射方法,从不同方面进行推理映射,建立本体间更完备的映射关系。针对复杂概念实体进行分词、标准化处理,在原有映射关系的基础上计算不确定性的概率,重新合并本体映射结果。

实验基于神经网络技术,通过学习异构本体中实体间的相似度以及实体间的不确定性计算,建立本体间的映射关系。将异构本体数据源中多种实体信息组成的训练集作为神经网络的输入数据,并且为训练器预先设定不确定性概率阈值。在训练过程中,不断为输入信息加权,并提供给隐藏层。每个输出单元取前一层单元输出的加权和作为输入,实验中选择了Sigmoid函数作用于加权输入,将神经元的输入范围映射到(0,1)区间上。给定单元的净输入 x ,则该单元的输出 O_x 可计算为:

$$O_x = \frac{1}{1 + e^{-x}}$$

在实验中,为了使实验取得更好的效果,将激励函数进行如下修改:

$$O_x = \frac{1}{1 + e^{-5(x-\alpha)}}$$

其中 x 是映射预测值, α 是预测阈值,只有大于 α 的预测才是可用的,在实验中将 α 设为0.5。

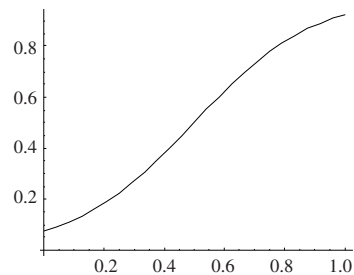


图3 Sigmoid激励函数

对于隐藏层次的选择没有明确的规定,需要在实验中不断测试。在实验中,选择了三层神经网络模型,也可以达到比较好的实验效果。

最后对测试数据进行验证,获得相似的实体对及其相似度。

5.2 实验参数设置

在提出的不确定性本体映射模型中,引入了多个参数,包括不确定性的概率阈值 λ_i 、各策略的分配权重 $\lambda_i(i=1, 2, 3)$ 。概率阈值的确定有助于提高匹配效率,将低于阈值的不确定性映射实体对过滤,建立比较精确高效的映射关系。

在实验中,先将 λ_i 初始化为0.5,随着模型的不断反馈迭代逐渐修正 λ_i 的值。在多不确定性本体映射方法的集成框架中,分别从三个方面考虑本体映射关系的建立,即语义、内容和结构。在实验的初始化阶段,等同考虑语义、内容以及结构映射策略的权重,即 $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$,同样通过迭代反馈修正 λ_i 的值。

5.3 实验结果评价

实验结果使用准确率(Precision)和召回率(Recall)进行性能评估,分别反应所有被识别的映射结果中的正确识别映射结果的比例,以及所有正确映射结果中的正确识别映射结果的比例。

$$Precision = \frac{m_{fi}}{m_{fi} + m_{fn}} \quad Recall = \frac{m_{fi}}{m_{fi} + m_{nd}}$$

其中, m_{fi} 表示被识别的正确的映射数量, m_{fi} 表示识别的错误的映射数量, m_{ni} 表示没有被识别的正确的映射数量。

为了验证提出的不确定性本体映射模型(UnOM)的性能, 将该方法与最小风险本体映射模型(RiMOM)进行比较, 实验结果如表 1 所示。

表 1 不确定性本体映射实验结果比较 (%)

映射模型	映射	P	R
RiMOM	(DiseaseA, DiseaseB)	86.43	88.46
	(DiseaseB, DiseaseA)	82.37	85.71
UnOM	(DiseaseA, DiseaseB)	88.23	90.12
	(DiseaseB, DiseaseA)	86.96	87.44

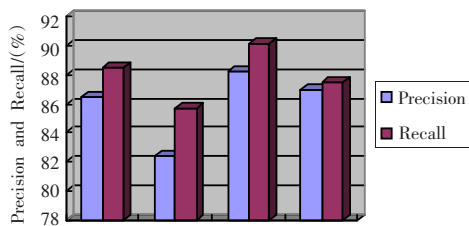


图 4 实验结果比较

实验结果表明, 不确定性本体映射模型取得的效果略优于最小风险本体映射模型, 其准确率和召回率平均都达到了 87% 以上, 表明该映射模型是可行和有效的。经研究分析, 在实验过程中, 本体各种特征对于映射的建立所做的贡献是不同的, 即各种映射策略的权重是不相等的。通过用户的参与反馈, 和整个模型的迭代, 最终获到最优权重分布: $\lambda_1=0.3, \lambda_2=0.2, \lambda_3=0.5$ 。对于不同的数据源, 其不同特征的映射策略权重是不同的, 需要通过实验的检验来修正。

总之, 提出的不确定性本体映射模型为语义集成提供了一种有效和通用的自动映射策略, 它是一种实用的异构本体集成方法, 对信息集成研究和实践都具有很高的应用价值。

6 相关工作

本体集成通常采用一种全局本体-局部本体的集成方式^[7], 由于其缺乏通用性和可扩展性, 无法解决本体的异构性和动态性带来的语义集成难题。本体集成的关键就是本体映射, 很多学者在本体映射方面做了很多研究。Melnik 等人提出了 SF 算法, 该算法基于图的相似性传播思想, 利用两个概念的邻居节点间的匹配推断两个概念之间匹配可能性^[8]。G.Stoilos 等人提出的 I-Sub 算法, 采用了一种基于字符串比较的方法, 从术语学的角度进行本体匹配^[9]。RiMOM 是基于最小风险本体映射模型开发的本体匹配系统, 它采用贝叶斯决策理论, 将映射发现问题转化为风险最小化问题^[10]。基于机器学习的映射方法通常利用成熟的机器学习方法, 如 Bayes、SVM 以及神经网络等。

以上方法从不同方面考虑本体映射的可利用因素, 但是它们仍然是在传统映射方式上的一种扩展, 没有考虑到语义 Web 的动态性和分布性, 忽略了映射过程中可能存在的不确定性。因此, 在对本体特征研究的基础上, 引入了不确定因素, 承认本体映射本身的不确定性。在基于概率模型的基础上, 构造了不确定性本体映射框架, 更全面合理地解决了异构环境中的本体集成问题。

7 总结与展望

主要针对本体集成的关键问题——本体映射进行详细阐述, 提出了一种新型的不确定性本体映射框架。详细分析了本体的概念、属性、关系等多种实体的特征, 从不同的角度采用不同的不确定性映射策略解决了语义 Web 中本体映射的不确定问题, 并最终形成一种合理的本体映射解决方案。实验表明, 不确定性本体映射框架比较真实地反映了异构本体环境下的实体关系, 并且比传统的本体映射方法获得了更好的集成效果, 为用户提供了更加合理和可靠的响应结果。

不确定性本体映射的提出顺应了数据空间架构的发展趋势, 希望逐渐改变传统的关系数据库等对数据管理所带来的诸多局限。同时, PayGo 模式提出了一种循序渐进的改善语义映射的方式, 以改变传统的全局映射的不足。在本体集成研究过程中, 数据空间架构和 PayGo 模式都为本体映射带了很大程度的改进, 这也需要在未来的研究中付出更多的努力。

参考文献:

- [1] Berners-Lee T, Fischetti M, Dertouzos M L. Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor[M]. USA, San Francisco: Harper, 1999.
- [2] Gruber T R. A translation approach to portable ontology specifications[J]. Knowledge System Laboratory, 1993, 5(2): 199-220.
- [3] Halevy A Y, Franklin M J, Maier D. Principles of dataspace systems[C]//Proc of PODS, 2006.
- [4] Maedche A. Ontology learning for the semantic Web[J]. IEEE Intelligent Systems, 2001, 16(2): 72-79.
- [5] Sheth A, Larson J. Federated database systems for managing distributed heterogeneous and autonomous database[J]. ACM Computing Surveys, 1990, 22(3): 183-236.
- [6] Zadeh L A. Fuzzy sets[J]. Information and Control, 1965, 8(3): 338-353.
- [7] Calvanese D. A framework for ontology integration[C]//Cruz I, Decker S, Euzenat J. The Emerging Semantic Web - Selected Papers from the First Semantic Web Working Symposium. Netherlands: IOS Press, 2002: 201-211.
- [8] Melnik S, Garcia-Molina H, Rahm E. Similarity flooding: A versatile graph matching algorithm and its application to schema matching[C]//Proceedings of the 18th International Conference on Data Engineering, San Jose, California, USA, 2002: 117-128.
- [9] Ehrig M, Euzenat J, Hess A, et al. D2.2.4: Alignment implementation and benchmarking results[EB/OL]. <http://knowledgeweb.semanticweb.org/semanticportal/deliverables/D2.2.4.pdf>.
- [10] 唐杰, 梁邦勇, 李涓子, 等. 语义 Web 中的本体自动映射[J]. 计算机学报, 2006, 17(9): 1837-1847.
- [11] Udrea O, Getoor L, Miller R J. Leveraging data and structure in ontology integration[C]//SIGMOD, Beijing, China, 2007.
- [12] Dong X, Halevy A Y, Yu C. Data integration with uncertainty[C]//VLDB, 2007: 687-698.
- [13] Mitra P, Noy N, Jaiswal A. Ontology mapping discovery with uncertainty[C]//Proceeding of the 2nd International Semantic Web Conference, Galway, Ireland, 2005: 537-547.