文章编号: 1001 - 9081(2008)06 - 1563 - 03

本体映射中一种改进的概念相似度计算方法

聂规划[1],左秀然[2],陈冬林[1]
(1. 武汉理工大学 经济学院,武汉 430070; 2. 武汉理工大学 管理学院,武汉 430070)
(zuoxiuran@163. com)

摘 要:本体映射是实现不同本体之间共享和交流的基础性工作。目前本体映射方法研究的重点主要集中在以自动化或半自动化方式实现映射和提高概念相似度计算的精度。本体映射的关键是不同本体概念间相似度的计算,单一的概念相似度计算方法往往不利于提高相似度的精度。针对以上不足提出了一种改进的概念相似度计算方法,并对其进行详细的描述,其中属性语义相似度计算方法改进了现有的基于属性计算语义相似度的方法,综合了数据类型属性和对象类型属性的语义相似度。经实例验证该方法有效且具有较高的精度。

关键词:本体;映射;相似度;语义
中图分类号: TP18 **文献标志码**: A

improved concept similarity comuting approach in Ontology mapping

NIE Gui-hua[1], ZUO Xiu-ran[2], CHEN Dong-lin[3]
(1. *College of Economics, Wuhan University of Technology, Wuhan Hubei* 430070, *China*;
2. *College of Management, Wuhan University of Technology, Wuhan Hubei* 430070, *China*)

Abstract: Ontology mapping is a preliminary work to realize the communication between different Ontologies. The current studies on ontology mapping approaches mainly focus on fulfilling mapping process automatically or semi-automatically and enhancing the precision of concept similarity. The key of ontology mapping is to comute the concept similarity, single concept similarity comuting approach is not in favor of improving precision of the result. in view of the above shortcomings, an improved concept similarity comuting approach was proposed and described in detail. The property semantic similarity of this approach improved the existing approaches to the similarities of both data type property and object type property. It is proved that this approach is effective and has high precision.

Key words: Ontology; mapping; similarity; semantic

0 引言

本体的应用,能够使用户和计算机更准确地基于语义进行交流而不仅仅局限于语法表达的数据。随着研究的深入,研究者们构建了越来越多的本体,这些本体之间存在异构,影响了它们之间的知识共享和重用。本体映射是解决本体之间异构问题的途径,本体映射过程中最关键的技术就是概念相似度的计算,目前关于本体映射中概念相似度计算方法的文献有很多:文献[1]基于概念词汇计算概念间的相似度;文献[2]对概念实例采用联合分布概率统计的方法,确定概念间语义相似度;文献[3]采用子概念间相似度计算概念间相似度。它们都采用单一的概念相似度计算方法,不利于提高计算结果的精度,因此在实际应用中应该综合多种概念相似度计算方法。

本文介绍了本体和本体映射的基本概念,提出了一种改进的概念相似度计算方法,此方法综合考虑了概念的语义相似度和属性的语义相似度,属性的语义相似度综合数据类型属性和对象类型属性的语义相似度。

1 本体及本体映射概述

由 T. R. Gruber提出,后被 R. Studer精化的本体的定义[3]2059是:一个本体是一个概念明确的、形式化的规范说明。Fensel对这个定义进行了分析,认为本体的概念包括四个主要方面[3]2059。

1)概念化。客观世界现象的抽象模型。

2)明确。概念及它们之间联系都被精确定义。

3)形式化。精确的数学描述。

4)共享。本体中反映的知识是其使用者共同认可的。

Gruber定义了一个典型的本体由五元组表示 $O = (C, E, R, F, A)$[3]2060。其中 C表示抽取出来的概念(类)的集合; I表示概念的实例; R表示定义在概念集合上的关系集合; F表示在概念集合上的函数集合; A表示公理集合,用于约束概念、关系、函数的一阶逻辑谓词集合。

本体映射是解决不同本体之间知识共享和重用问题的方法,其目的是找出不同本体中实体之间的语义关联,并且将其形式化地表达出来。Ehrig给出了本体映射的定义[3]2061:本体映射是指有两个本体 A、B,对于 A中的每个概念我们试图在本体 B中为它找到一个语义相同或相近的对应概念,对于概念 B中的每个概念或节点亦是如此。Ehrig也给出了一个形式化的本体映射函数 $map: O_{11} \rightarrow O_{12}$:

$map(e_{1i1}) = e_{2i2}$,如果 $\mathrm{sim}(e_{1i1}, e_{2i2}) > t$

其中 $\mathrm{sim}(e_{1i1}, e_{2i2})$是实体 e_{1i1}和 e_{2i2}的相似度, t是阈值,表示

收稿日期: 2007 - 12 - 24;**修回日期**: 2008 - 02 - 01。 **基金项目**:国家自然科学基金资助项目(70572079);国家科技支撑计划资助项目(2006BAH02A08);湖北省自然科学基金资助项目(2006ABA303)。

作者简介:聂规划(1957 -),男,河南周口人,教授,博士生导师,博士,主要研究方向:知识管理、人工智能、电子商务; 左秀然(1983 -),女,山东莒县人,硕士生,主要研究方向:管理信息系统、本体映射; 陈冬林(1970 -),男,湖北安陆人,副教授,博士,主要研究方向:电子商务智能推荐,语义网,网格技术。

将本体 O_{i1}中的实体 $e_{i1,j1}$映射到本体 O_{i2}中的实体 $e_{i2,j2}$,二者在语义上是等价的。

Ehrig and Staab总结了过去的工作,归纳出本体映射的 6个过程[3]2061。

1)特征提取。提取用于计算相似度的特征,如概念、属性的名称等。

2)选择用于映射的概念对。

3)进行相似度计算。

4)相似度整合。通常多种方法可以衡量本体实体之间的相似度,得出多种相似度值,因此要对各相似度进行综合考虑,从而得到一个整体上的相似度。

5)优化。第 4)步结束后,已经得到待映射的各个实体之间 的初始相似度,这时一般需要人工的干预,利用领域知识,对其进行调节。

6)迭代 1)～5),直到达到满意结果。

2 改进的综合概念相似度计算

2.1 综合概念相似度计算模型框架

从两个不同的本体中提取概念对,对这两个概念分别计算基于概念本身的相似度和基于属性的相似度,其中基于属性的相似度是综合数据类型属性相似度和对象类型属性相似度的结果,这样避免了现有的基于属性计算概念相似度的方法没有区分数据类型属性和对象类型属性的不足,对其进行改进,最后综合基于概念本身的相似度和基于属性的相似度得到两个概念的综合相似度值。这种综合概念相似度计算模型框架如图 1所示。

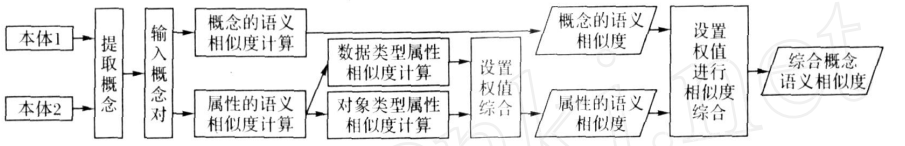

图 1 综合概念相似度计算模型框架

2.2 综合概念相似度计算方法

2.2.1 概念的语义相似度计算方法

对于概念的语义相似度计算,采用基于 WordNet的 Wu-Palmer概念语义相似度算法[4]。WordNet是一部能够表达概念关系的语义词典,它依据词义而不是词形来组织词汇信息。WordNet将英语词汇组织为一个同义词集合 (Synset),每个集合标明一个词汇概念,词汇关系在词语之间体现,语义关系在概念之间体现。由于语义关系是一种词义之间的关系,而词义可以用同义词集合来表示,因此很自然地把语义关系看作为同义词集合之间的一些指针,同义词集合之间是以一定数量的关系类型相关联的,这些关系包括同义关系、反义关系、上下位关系、部分与整体关系等。Wu-Palmer语义相似度算法基于长度定义相似度,引入了深度的制约条件,用于 ISA层级关系。对于两个概念 A和 B,用这种算法计算它们的概念语义相似度为:

$$S_1(A,B) = \frac{2 \times depth(lso(A,B))}{depth(A) + depth(B)} \quad (1)$$

其中, $lso(A,B)$是概念 A和概念 B的最近共同祖先概念, $depth(A)$和 $depth(B)$分别表示概念 A和概念 B在词典语义树中的深度。

2.2.2 属性的语义相似度计算方法

本体中概念的属性分为两种:数据类型属性 (DatatypeProperty)和对象类型属性 (ObjecttypeProperty)。

如果一个概念的实例通过一个属性与一般的数据类型相关联,那么这个属性就是数据类型属性,如:

```
<owl: DatatypeProperty rdf: ID = "price" >
  < rdfs: domain rdf: resource = "#article" / >
  < rdfs: range rdf: resource = "&xsd; demical" / >
< /owl: DatatypeProperty >
```

属性 price就是数据类型属性。

如果一个概念的实例通过一个属性与另一个概念的实例相关联,那么这个属性就是对象类型属性,如:

```
<owl: ObjectProperty rdf: ID = "course" >
  < rdfs: domain rdf: resource = "#Meal" / >
  < rdfs: range rdf: resource = "#MealCourse" / >
< /owl: ObjectProperty >
```

属性 course就是对象类型属性。

基于属性计算语义相似度方法的思路是对数据类型属性和对象类型属性分别计算语义相似度,之后为两种属性的语义相似度设置权值,综合两种语义相似度得到基于属性的语义相似度。

1)数据类型属性语义相似度计算。文献[5]介绍了一种基于数据类型的属性语义相似度的计算方法,本文的数据类型属性语义相似度也采用这种方法来计算。

步骤 1:将概念 A的数据类型属性按数据类型分类,这样概念 A的数据类型属性按照数据类型被分为若干个属性集合;

步骤 2:按步骤 1方法将概念 B的属性按数据类型分类;

步骤 3:对每一种数据类型构造概念 A和概念 B的属性相似度矩阵,相似度的计算采用 2.2.1节类似的方法;

步骤 4:求所有数据类型的语义相似度的平均值,记为 $SD(A,B)$。

2)对象类型属性语义相似度计算。对象类型属性语义相似度方法是计算对象类型属性所关联的概念的语义相似度。设概念 A的对象类型属性集合为 $attribute_A = \{a_1, a_2, \cdots, a_m\}$,概念 B的对象类型属性集合为 $attribute_B = \{b_1, b_2, \cdots, b_n\}$,其中 m, n分别为概念 A和概念 B的对象类型属性个数。概念 A的对象类型属性 $a_i (1 \le i \le m)$所关联的概念为 A_i,概念 B的对象类型属性 $b_j (1 \le j \le n)$所关联的概念为 B_j。

1)求出概念 A_i和概念 B_j的语义相似度作为概念 A的对象类型属性 a_i和概念 B的对象类型属性 b_j的相似度,记为 SO_{ij},计算方法采用 2.2.1节的方法。得到相似度矩阵:

$$SO' = \begin{bmatrix} SO_{11} & SO_{12} & \cdots & SO_{1n} \\ SO_{21} & SO_{22} & \cdots & SO_{2n} \\ \cdots & \cdots & & \cdots \\ SO_{m1} & SO_{m2} & \cdots & SO_{mn} \end{bmatrix} \quad (2)$$

2)遍历矩阵 SO取得相似度最大的 SO_{ij},将 SO_{ij}所属的行和列删除,在余下的矩阵中继续重复执行直到矩阵为空,得到最大相似度序列记为 $p_1, p_2, \cdots, p_k (k = \min(m, n))$。

3)综合得到对象类型属性语义相似度:

$$SO(A,B) = \frac{1}{k} \sum_{i=1}^{k} p_i \quad (3)$$

综合数据类型属性语义相似度和对象类型属性语义相似

度,得到属性的语义相似度为:

$$S_2(A,B) = w_1 SD(A,B) + w_2 SD(A,B) \quad (4)$$

其中 w_1、w_2 分别为数据类型属性相似度和对象类型属性相似度在整个属性相似度的权值,且 $w_1 + w_2 = 1$,一般可以取两个权值都为 0.5。

最后综合基于概念的语义相似度 $S_1(A,B)$ 和基于属性的语义相似度 $S_2(A,B)$,将它们的权重都设置为 0.5,得到两个概念 A和 B的综合相似度为:

$$sim(A,B) = 0.5S_1(A,B) + 0.5S_2(A,B) \quad (5)$$

3 实例验证

以两个概念 commodity和 merchandise的语义相似度计算为例,来说明本文所提出的方法,下面是两个概念的 OWL描述。

commodity的 OWL描述:

```
<owl:DatatypeProperty rdf:ID="name">
  <rdfs:domain rdf:resource="#commodity" />
  <rdfs:range rdf:resource="&xsd;string" />
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="code">
  <rdfs:domain rdf:resource="#commodity" />
  <rdfs:range rdf:resource="&xsd;string" />
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="price">
  <rdfs:domain rdf:resource="#commodity" />
  <rdfs:range rdf:resource="&xsd;float" />
</owl:DatatypeProperty>
<owl:ObjectProperty rdf:ID="hasprovider">
  <rdfs:domain rdf:resource="#commodity" />
  <rdfs:range rdf:resource="#provider" />
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hascategory">
  <rdfs:domain rdf:resource="#commodity" />
  <rdfs:range rdf:resource="#category" />
</owl:ObjectProperty>
```

merchandise的 OWL描述:

```
<owl:DatatypeProperty rdf:ID="name">
  <rdfs:domain rdf:resource="#merchandise" />
  <rdfs:range rdf:resource="&xsd;string" />
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="unit">
  <rdfs:domain rdf:resource="#merchandise" />
  <rdfs:range rdf:resource="&xsd;string" />
</owl:DatatypeProperty>
<owl:DatatypeProperty rdf:ID="price">
  <rdfs:domain rdf:resource="#merchandise" />
  <rdfs:range rdf:resource="&xsd;float" />
</owl:DatatypeProperty>
<owl:ObjectProperty rdf:ID="hassupplier">
  <rdfs:domain rdf:resource="#merchandise" />
  <rdfs:range rdf:resource="#supplier" />
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hascategory">
  <rdfs:domain rdf:resource="#merchandise" />
  <rdfs:range rdf:resource="#category" />
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hasmadearea">
  <rdfs:domain rdf:resource="#merchandise" />
  <rdfs:range rdf:resource="#area" />
</owl:ObjectProperty>
```

3.1 概念的语义相似度计算

在式(1)中令 A = commodity、B = merchandise,通过 WordNet词典查找概念 A和概念 B,得到 $lso(A,B)$ = *artifact*, $depth(A)$ = 6, $depth(B)$ = 7, $depth(lso(A,B))$ = 5, 所以 $S_1(A,B) = 10/13 = 0.77$。

3.2 属性的语义相似度计算

3.2.1 数据类型属性的语义相似度计算

对两个概念的数据类型属性集进行分类有 String类型:$A1$ = {name, code}, $B1$ = {name, unit}; Float类型:$A2$ = {price}, $B2$ = {price}。对 String类型的属性按式(1)计算语义相似度,得到相似度矩阵$\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}$,取平均值得 $SD_{string} = 1/2 = 0.5$。对 Float类型的属性按同样方法计算得 $SD_{float} = 1$。因此 $SD(A,B) = (SD_{string} + SD_{float})/2 = (0.5 + 1)/2 = 0.75$。

3.2.2 对象类型属性的语义相似度计算

概念 A的对象类型属性所关联的概念集合为 {provider, category},概念 B的对象类型属性所关联的概念集合为 {supplier, category, area},按公式(1)计算它们的语义相似度得到相似度矩阵$\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}$,取平均值得 $SO(A,B) = 1$。

因此概念 A和概念 B的属性的语义相似度 $S_2(A,B) = 1$。

最后综合 $S_1(A,B)$ 和 $S_2(A,B)$,求得 commodity和 merchandise的综合语义相似度 $sim(A,B) = (0.875 + 0.77)/2 = 0.823$。

4 结语

本文采用的方法改进了现有的概念相似度计算方法,综合考虑概念的语义相似度和属性的语义相似度,在计算属性的语义相似度时,针对现有计算属性的相似度方法没有综合考虑属性类型的不足,提出了综合数据类型属性和对象类型属性计算属性的语义相似度的方法。通过本文所采用的方法的计算,commodity和 merchandise的语义相似度为 0.823,如果只采用基于 WordNet的 Wu-Palmer概念语义相似度算法计算它们的相似度是 0.77, commodity和 merchandise都是指商品,它们在语义上具有极高的相似度,因此本文所采用的方法具有更高的精度。但本体映射是一个非常复杂的过程,整个过程中的每一个环节都会影响最终的映射结果,因此对本体映射的研究工作不应该只局限于对概念相似度计算方法的研究,还应该考虑到可能影响最终映射结果的其他因素。

参考文献:

[1] 何娟,高志强,陆青健,等. 基于词汇相似度的元素级本体匹配[J]. 计算机工程, 2006, 32(16): 185 - 187.

[2] DOAN A H, MADHAVAN J, DOMinGOS P. Learning to map between Ontologies on the semantic Web [C]// Proceedings of the 11th international Conference on World Wide Web, New York: ACM Press, 2002: 662 - 673.

[3] WONG A K Y, RAY P, WARAN N P. Ontology mapping for the interoperability problem in network management [J]. IEEE Journal on Selected Areasin Communication, 2005, 23(10): 2058 - 2068.

[4] BUDANITSKY A, GRAEME H. Evaluating WordNet-based measures of semantic distance [J]. Comutational Linguistics, 2006, 32(1): 13 - 47.

[5] 陈冬林,聂规划,刘平峰. 基于本体的 B2B电子商务 MAS模型及商品匹配算法[J]. 计算机工程与应用, 2007, 43(10): 199 - 201.