领域本体中基于多维特征的语义相似度算法研究*

Semantic Similarity Algorithm Based on Multidimensional Features in Domain Ontology

王惠敏　聂规划　付　魁

(武汉理工大学经济学院电子商务系　武汉　430070)

摘　要　在电子商务与现代物流系统的集成中,信息资源之间存在着语义的区别,电子商务本体和物流本体之间概念的映射显得非常重要。概念相似度计算是领域本体映射中的一个"瓶颈问题"。提出综合概念的语义距离特征、语义信息特征和概念属性的语义特征来计算概念的语义相似度。本体概念语义特征的结合采用了更符合概念特性的非线性组合方式。将该算法应用于已建立的电子商务本体和物流本体的映射系统中,取得了较好的概念匹配结果。

关键词　本体　本体映射　语义相似度　多维特征

中图分类号　TP391

本体是共享概念的规范、精确的描述[1]。它作为一种能在语义和知识层次上描述信息的概念模型建模工具,就像一部规范的词典,给通信各方提供公共的概念解释,用于表达信息源的语义、识别和建立概念间的语义关联,以达成语义一致。领域本体描述的是特定领域中的概念及概念之间的关系,它为工作在不同领域之中或者应用不同操作平台的人员能够进行语义层次上的知识的共享和互操作。

在不同领域中概念可能使用不同的术语。当两个本体需要交流或交换信息时,它们必须事先达成一致,也就是必须在两个不同领域的本体间实现映射。例如,在电子商务与现代物流系统的集成中,信息资源之间存在着语义的区别,两个异构的电子商务本体和物流本体之间概念的映射就显得非常重要。本体间的语义映射过程实际上是将一个本体中的概念、属性和关系映射为另一本体中的概念、属性和关系的过程,是本体间概念和关系取得一致性的一个规范说明。本体映射算法以两个本体作为输入,然后为这两个本体的各个元素(概念、属性或者关系)建立相应的语义关系。相似性提取是本体映射的一个重要步骤,它主要是进行概念相似度的计算,并产生一个概念的相似矩阵。当其相似度大于某个阈值时,就认为这两个概念之间存在一定的映射关系。

目前,已有许多研究者关注本体间概念相似度的计算方法。文献[2]主要从概念的名称、属性、结构等方面综合考虑概念的相似度。文献[3]提出的算法将概念相似度计算分为两层:一层是利用概念之间的距离来计算概念的初始相似度;另一层是其在概念初始相似度的基础上,计算概念通过非上下位关系体现出的相似度。综合二者得到领域本体中概念的实际相似度。文献[4]提出的方法充分利用本体特点来计算相关概念之间的相似度。文中概念间相似度的计算,主要是基于按照概念间结构层次关系组织的语义词典的方法,根据在这类语言学资源中概念之间的上下位关系以及其它一些因素,如语义重合度、语义距离、层次深度、调节因子等多种因素来计算领域内部概念之间的语义相似度。文献[5]提出了一种综合的相似度计算方法。首先根据两个概念名称相似性过滤出最相关的概念,减少相似度的计算;然后基于概念实例、基于概念属性、基于概念关系计算概念相似度,并进行综合。概念和属性的相似采用字符串相似度计算方法进行判定。

概念相似度计算是领域本体映射中的一个"瓶颈问题",在本文中我们提出了一个领域本体中基于多维特征的语义相似度算法,该算法综合了本体概念的语义距离特征、本体概念的语义信息内容特征和概念的属性特征,并且本体概念语义特征的结合采用了非线性的组合方式,这更符合概念的特性。

1　语义相似度计算方法

1.1　传统的语义相似度算法　语义相似度的量化计算,现今主要有两类方法:其一是基于语义距离的模型,其二是基于信息的模型[5]。

基于语义距离的模型是根据概念在层次结构中的位置来

基金项目:国家科技部科技支撑计划项目"电子商务与现代共性支撑体系与应用示范工程"(编号:2006BAH02A08);国家自然科学基金项目"基于知识网络的电子商务智能推荐系统研究"(编号:70572079)。

作者简介:王惠敏,女,1971 年生,博士,讲师,研究方向为商务智能、电子商务推荐系统和知识管理与知识工程;聂规划,男,1957 年生,博士,教授,博士生导师,主要研究方向为商务智能、信息资源管理和知识管理与知识工程;付　魁,男,1977 年生,博士,讲师,研究方向是中文领域本体学习、商务智能、知识管理与知识工程。

计算语义相似度的。该类方法中具有代表性的算法主要有 Hfirst - St - Onge 语义相关度算法、Leacock - Chodoraw 语义相似度算法和 Wu - Palmer 语义相似度算法。Hfirst - St - Onge 语义相关度算法的基本假设是:当两个词在 WordNet 同义词集中有一条较短的路径相连时,在语义上就具有相对较大的语义相关度,而语义相似度和语义相关度成正比关系。Leacock - Chodoraw 算法和 Wu - Palmer 算法考虑到当路径长度相同时,越靠近根节点语义相似度越小,因此引入了深度的制约条件。

基于信息理论的方法首先由 Resnik 提出,其基本原理是两个概念共享信息越多,语义上越接近,语义相似度越大[6]。一个概念的信息内容取决于在语料库中该概念及其子概念出现的频率。本体中每个概念都是从其祖先节点细化得到,该概念包含其祖先节点的所有信息内容。如果两个概念同时共同拥有一个祖先节点,它们就共同拥有该祖先节点的所有信息内容。该类方法中具有代表性的算法主要有 Resnik 语义相似度算法、Jiang - Conrath 语义距离算法和 Lin 语义相似度算法[5]。Resnik 算法的基本假设是:两个概念的语义相似度,由他们共同拥有的那部分概念所决定。根据两个概念的公共祖先节点的最大信息量来衡量两个概念的语义相似度。Jiang - Conrath 语义距离算法相当于给定了共有祖先后,利用子节点的条件概率来计算语义距离。

1.2　**融合语义距离和语义信息的概念相似度算法**　从传统的语义相似度算法中分析可知,概念间的语义相似度不仅由概念间的路径长度确定,而且也应由概念的深度和概念的信息量确定。基于语义距离的方法仅考虑了路径长度和概念深度特征,而基于信息容量的方法仅利用了来自语料库的信息,考虑了概念的局部信息密度。

因为语义相似度受概念的路径长度、概念深度及概念信息的综合影响,本文综合考虑这些方面的因素,提出新的语义相似度算法。该算法主要分为三个部分。

1.2.1　概念的路径长度对相似度的影响。概念的语义相似度受概念的路径长度的影响主要表现为路径长度越大,其相似度就越小;相反,路径长度越小,其相似度就越大。当路径长度为 0 时,其相似度为 1;当路径长度相当大时,其相似度为 0。研究表明,相似度看作是路径长度的非线性函数比较合理。在本文中,我们将概念相似度看作是路径长度的单调递减函数。

$$\mathrm{sim}(c_1, c_2) = e^{-\alpha l} \tag{1}$$

l 为两概念间的最短路径长度,α 是调节因子,指数函数的选择是确保相似度值在 0 和 1 之间。

1.2.2　概念深度及概念的信息容量对相似度的影响。在一个本体中,处于本体树中较上层的概念具有一般性的语义,概念间的相似度比较低;相反,处于较低层的概念具有比较具体的语义,概念的相似度较高。

概念信息容量的计算可以通过统计的方法,准备足够量的领域资料,一个概念在该特定领域的文集中出现的概率越大,说明该概念越抽象,它所携带的信息容量就越少;反之,信息容量就越多。

Resnik 相似度算法既考虑了概念的深度信息,也考虑了概念的信息容量[7]。本文中概念的深度及信息容量对相似度的影响采用 Resnik 的计算方法。

$$\mathrm{sim}(c_1, c_2) = IC(lso(c_1, c_2)) \tag{2}$$

$$\mathrm{sim}(c_1, c_2) = \log p(lso(c_1, c_2)) \tag{3}$$

式中,$lso(c_1, c_2)$ 是 (c_1, c_2) 的最近公共祖先,$p(lso(c_1, c_2))$ 是其公共祖先在特定本体库中出现的概率。

1.2.3　基于概念语义信息的概念相似度算法。我们已知,概念的相似度随概念的深度和信息容量的增加而增加。因此,本文考虑将概念的深度信息和信息容量对相似度的影响采用单调递增函数来体现。由于概念的语义相似度值处于 0 和 1 之间,本文选取了指数函数。基于概念语义信息的概念相似度计算方法为:

$$\mathrm{sim}(c_1, c_2) = e^{-\alpha} e^{-\beta(IC_{max} - IC(lso(c_1, c_2)))} \tag{4}$$

式中,$e^{-\alpha}$ 代表两概念的最短路径对相似度的影响,$e^{-\beta(IC_{max} - IC(lso(c_1, c_2)))}$ 代表概念的深度和信息容量对相似度的影响,β 为调节因子,IC_{maX} 的选取是为了使概念相似度处于 0 和 1 之间。

2　基于多维特征的语义相似度算法

本文提出的基于多维特征的语义相似度算法综合考虑了概念的语义信息,如概念的深度特征、概念的长度特征、概念的信息容量特征和概念的属性特征。概念的属性对概念的描述具有十分重要的作用。如果两个概念的属性都相同,那么这两个概念是相同的;如果两个概念具有相似的属性,那么这两个概念也是相似的。

2.1　**概念语义的初始相似度**　概念语义的初始相似度是对概念相似度的一个预定值,通过概念在本体中的上下位关系信息来体现,可按本文给出的公式(4)计算。

2.2　**概念属性的语义相似度**　本体中的每个概念也可用它的属性特征描述,本体中所有概念的属性特征可以组成一个集合:$\{P_1, P_2, \cdots, P_n\}$,本体中的概念及其属性可形成一个概念 - 属性矩阵 $X(m, n)$。第 i 行第 j 列的元素 $x_{i,j}$ 代表概念 i 的第 j 个属性。我们假设如果两个概念具有相似的属性,则这两个概念可能是相似的。考虑到概念的属性对概念的描述能力是不同的,因此在计算两个概念的语义相似度时,分别计算这两个概念在同一个属性 P_j 上的语义相似度 S_j,然后确定 S_j 的权重 W_j,根据下述公式计算两个概念的语义相似度。

$$\mathrm{sim}_p(c_1, c_2) = \sum_{j=1}^{n} S_j W_j \tag{5}$$

其中,S_j 可根据改进的语义相似度算法来计算。

2.3　**综合的语义相似度算法**　综合概念的语义信息相似度和概念属性的语义信息相似度,概念的语义相似度定义为:

$$\mathrm{sim}(c_1, c_2) = k \cdot \mathrm{sim}_c(c_1, c_2) + \lambda \cdot \mathrm{sim}_p(c_1, c_2)$$

k 和 λ 为权重因子,$0 \leq k, \lambda \leq 1$,具体取值由应用确定。

3 实验及结果

本文将提出的语义相似度算法应用于本课题组正在进行的电子商务与现代物流语义集成的技术研究,概念属性矩阵来自于本课题组正在研究的电子商务本体与物流本体。具体实现过程描述如下:

for (电子商务本体中的概念及其所有的属性)
  for (选择物流本体中与相匹配的概念)
    寻找两概念间的直接公共节点;
    计算两概念间的最短路径;
    获得公共节点的信息熵;
    计算两概念的初始相似度;
    if 相似度 < T(T 为阈值)
      计算两概念的属性相似度;
      综合初始相似度和属性相似度形成两概念的相似度;
    end
  end
  选取与电子商务本体中概念具有最大相似度值的物流本体中的概念;
end

该算法已经成功应用于本课题组正在研究的电子商务与物流本体的映射系统中,实验结果表明利用本文提出的概念相似度算法能够从电子商务本体和物流本体中获得比较理想的概念匹配结果。

4 结　论

本文提出了一种基于概念的语义距离特征、语义信息特征和概念属性的语义特征的概念相似度算法,由于该算法比较全面地考虑了概念的特征,我们将其应用于已建立的电子商务本体和物流本体的映射系统中,取得了较好的概念匹配结果,为解决电子商务与现代物流系统中的语义冲突奠定了基础。

参 考 文 献

1 T R Gruber. A Translation Approach to Portable Ontology Specifications[J]. Knowledge Acquisition,1993,5(2):199 - 220

2 徐德智,肖文芳,王怀民. 本体映射过程中的概念相似度计算[J]. 计算机工程与应用,2007,43(9):167 - 169

3 陈　杰,蒋祖华. 领域本体的概念相似度计算[J]. 计算机工程与应用,2006,33(2):163 - 166

4 李　鹏,陶　兰,王弼佐. 一种改进的本体语义相似度计算及其应用[J]. 计算机工程与设计,2007,28(1):227 - 229

5 Alexander Budanitsky, Graeme Hirst. Evaluating WordNet - based Measures of Lexical Semantic Relatedness[J]. Computational Linguistics. 2006,32(1):13 - 47

6 P Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence[C],1995:448 - 453

7 曹泽文,钱　杰,张维明等. 一种综合的概念相似度计算方法[J],计算机科学,2007,34(3):174 - 175

(责编:贺晓利)

(上接第 24 页)间具有相关性,但这些指标含义不同、各自独立,不可能相互代替,只能是相互补充、相互参照;c. h 指数是累积指标,适于测度科研人员的中、长期科研绩效;d. 为提高 h 指数的区分度,有效发挥其评价功能,提高评价结果的客观公正性,可将 h 指数及其衍生指数与其它传统计量指标相结合,以构建多维评价体系,对评价对象做出更全面的评价。

h 指数在评估小论文集合时具有很大潜力,是微观科研评价中的一项革命性的指标,它丰富了科学计量学和文献计量学的指标体系,提升了科学计量学和文献计量学方法在科学评价方面的影响[10],可以为同行评审提供充分而可靠的信息。但目前实证研究还比较欠缺,因此,为了更深入地揭示 h 指数的应用前景及其局限性,有必要在各学科领域开展系统性的实证研究,为将来可能的 h 指数和类 h 指数的推广应用做好理论与实践准备。

参 考 文 献

1 Hirsch J E. An Index to Quantify an Individual's Scientific Research Output[J]. Proceedings of the National Academy of Sciences of the USA, 2005, 102(46):16569 - 16572

2 万　锦,花平寰,赵呈刚. 中国部分重点大学 h 指数的探讨[J]. 科学观察,2007,2(3):9 - 16

3 Anthony F J, Van R. Comparison of the Hirsch - Index With Standard Bibliometric Indicators and With Peer Judgment for 147 Chemistry Research Group[J]. Scientometrics,67(3):491 - 502

4 Cronin B, Meho L. Using the H - Index to Rank Influential Information Scientists[J]. Journal of The American Society for Information Science and Technology,2006,57(9):1275 - 1278

5 Oppenheim C. Using the H - Index to Rank Influential British Researchers in Information Science and Librarianship[J]. Journal of the American Society for Information Science and Technology, 2007,58(2):297 - 301

6 邱均平,缪雯婷. h 指数在人才评价中的应用 —以图书情报学领域中国学者为例[J]. 科学观察,2007,2(3):17 - 22

7 张学梅. 用 h 指数对我国图书情报学界作者进行评价[J]. 图书情报工作,2007,51(8):48 - 50

8 Rousseau R. Hirsch 指数研究的新进展[J]. 科学观察,2006,1(4):23 - 25

9 刘俊婉,苏新宁,邓三鸿. 经济学研究现状 —基于 CSSCI 的评析[J]. 经济学家,2004(4):73 - 80

10 赵基明,邱均平,黄　凯等. 一种新的科学计量指标 - h 指数及其应用述评[J]. 中国科学基金,2008(1):23 - 32

(责编:刘影梅)