2011 年 9 月 郑州大学学报(工学版) Sep. 2011
第 32 卷 第 5 期 Journal of Zhengzhou University (Engineering Science) Vol. 32 No. 5

文章编号:1671 - 6833(2011)05 - 0080 - 04

基于概念格理论的语义相似度模型研究及验证

张小红

(河南财政税务高等专科学校 信息工程系,河南 郑州 451464)

摘 要:将语义相似度计算模型定义为域、概念、属性组成的三维空间模型,并结合领域本体集,从概念格理论的角度考虑了该模型对语义相似度计算的影响.该模型通过对不同的向量加不同的权值来调节其对语义相似度计算的贡献,使计算结果达到最优,从而提高语义相似度计算的准确度.实验结果表明,与单方面计算相似度的方法相比,该方法能有效地提高语义相似度计算的查全率和查准率.

关键词:语义相似度,概念格,领域,概念,属性

中图分类号:TP391 **文献标志码**:A

0 引言

针对网上海量的信息,怎样快速准确地检索相关信息已经成为当今信息领域的研究热点.由于概念是组成信息的最小单位,所以概念之间语义相似度对信息检索十分重要.在语义相似度计算的过程中,本体占有重要的地位,然而本体的建立一直没有一个统一的规范来进行约束,由此引发了诸如系统异构、结构异构、语义异构等许多问题.本体映射的研究则正是为了解决这些异构问题,而本体的语义相似度计算是本体映射的关键环节.但是目前的本体映射存在相似度的计算方法不完善、计算量过高、概念相似度的计算过于片面等问题[1-4],不能很好地反映语义相似度的相关因素.刘群的基于 HowNet 的词汇语义相似度计算模型[5]给出了计算相似度的一般方法,但是在计算过程中缺少语义相似度模型的层次体系结构.笔者针对上述问题,从概念格理论的角度出发,抽取出三维空间模型,利用领域知识库(本体)来计算语义相似度,探讨其计算模型,并且对基于概念格的语义相似度计算模型验证.

1 语义相似度空间模型的建立

概念格也称为 Galois 格,由 R. Wille 于 1982 年首先提出[6].假设给定形式背景(context)为三元组 $T=(U,P,R)$,其中 U 是对象集合,P 是描述符(属性)集合,R 是 U 和 P 之间的一个二元关系,则存在唯一的一个偏序集合与之对应,并且这个偏序集合产生一种格结构,这种由背景(U,P,R)所诱导的格称为一个概念格.

根据概念格的基本理论,抽取出语义相似度的 3 个特征,建立三维空间模型.其中形式背景对应相似度计算的域特征,对象的集合对应相似度计算的概念特征,属性的集合对应相似度计算的属性特性.从而结合领域本体知识库,构建语义相似度的三维空间模型.其中 X 坐标表示语义相似度的域信息,反映了领域知识中本体之间的语境信息;Y 坐标表示语义相似度中的概念信息,反映了知识库中概念之间相似度的计算;Z 坐标表示语义相似度的属性信息,反映了知识库中概念之间的属性的相似关系,如图 1 所示.

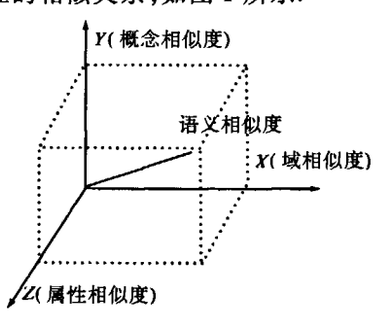

图 1 语义相似度计算模型

Fig. 1 Semantic similarity computing model

收稿日期:2011 - 05 - 04;修订日期:2011 - 06 - 25

基金项目:河南省科技攻关计划项目(102102210532)

作者简介:张小红(1962 -),女,河南新乡人,河南财政税务高等专科学校副教授,主要从事软件工程研究,E - mail: zosha@ sina. com.

2　域相似度

域相似度的计算主要是确定层次知识树中跨本体的相似程度,如果域相似度的值大于某个给定阈值,则层次知识树中两个跨本体之间具有相似关系.研究发现,概念的实例信息和概念的层次结构信息在一定程度上都反映了跨本体之间的概念间的相似关系,而每一个单独的信息对跨本体之间的概念相似度的影响都不是全面的,故对于域相似度应从两方面全面的考虑,实现对域相似度计算信息的完整性.

假设有知识库两个领域本体 O_i、O_j($i,j=1,2,\cdots,n$),从两个领域本体中分别取出一个概念 c_1 和 c_2,其语义相似度用 $sim_{dom}(c_1,c_2)$ 表示.则根据上面对域相似度计算的分析,我们分解成两个关键部分:基于概念的实例的相似度和基于概念层次结构的相似度,且其分别用 $sim_{inst}(c_1,c_2)$,$sim_{stru}(c_1,c_2)$ 表示,它们每个部分所占权重分别为 w_{inst},w_{stru},它们由经验值给出,且满足 $w_{inst}+w_{stru}=1$,因此域相似度的计算总公式可表示为[7]

$$sim_{dom}(c_1,c_2)=w_{inst}\times sim_{inst}(c_1,c_2)+w_{stru}\times sim_{stru}(c_1,c_2) \tag{1}$$

2.1　概念的实例相似度

概念的实例信息在一定程度上反映了跨本体之间的概念的语义关系.如果两个概念的实例的集合有很多重叠的部分,那么我们就说这两个概念间具有丰富的相似关系.

利用机器学习方法来计算实例的联合分布概率.采用朴素贝叶斯(Naive Bayes)的学习技术来训练学习器,同时通过合并不同匹配器匹配结果,产生的是原子级的1:1对应的映射关系[1].从而得到实例相似度 $sim_{inst}(c_1,c_2)$ 的矩阵.对于一个实例,利用 Jaccard 系数[7]来计算概念相似度,这一点源于 GLUE[1] 系统的思想,用机器学习方法计算一对概念($A\in O_1$,$B\in O_2$)的联合分布从而求得 $P(A,B)$,$P(A,\overline{B})$,$P(\overline{A},B)$,然后求得两个概念间的相似度:

$$sim_{inst}(A,B)=\frac{P(A\cap B)}{P(A\cup B)}=\frac{P(A,B)}{P(A,B)+P(A,\overline{B})+P(\overline{A},B)} \tag{2}$$

式中:$P(A,B)$ 表示从一个层次本体库中的实例空间中随机选取一个实例属于概念 A,并且同时属于概念 B 的实例在实例空间中所占的比重;$P(A,\overline{B})$ 表示从一个层次本体库的实例空间中随机选取一个实例属于概念 A,但是不属于概念 B 的实例在实例空间中所占的比重;$P(\overline{A},B)$ 表示从一个层次本体库的实例空间中随机选取一个实例属于概念 B,但是不属于概念 A 的实例在实例空间中所占的比重.

2.2　概念结构相似度

在层次本体库的描述中主要考虑最常见的两种语义关系:概念之间部分与整体的关系,即层次本体库中的 part-of 关系;概念之间的继承关系,即层次本体库中的 is-a 关系.

我们在考虑两种常见的语义关系的基础上综合考虑语义半径和路径距离的概念的范围内查找与该概念有语义关系的所有邻居,从而得到一个集合,这样由来自不同本体的概念,分别得到两个相关集合,从而求得概念结构相似度计算公式[8]:

$$sim_{stru}(A,B)=\frac{|a\cap b|}{|a\cap b|+\alpha(A,B)|a-b|+(1-\alpha(A,B))|a-b|}\quad(0\le\alpha\le1). \tag{3}$$

式中:a 和 b 分别表示概念 A 和 B 的描述集合(同义词集、特征集);$|a\cap b|$ 指集合 a 和 b 的交集的元素个数;$|a-b|$ 表示属于集合 a 而不属于集合 b 的元素个数.比例因子 α 满足:

$$\alpha(A,B)=\begin{cases}\dfrac{depth(A)}{depth(A)+depth(B)}, & depth(A)\le depth(B)\\ 1-\dfrac{depth(A)}{depth(A)+depth(B)}, & depth(A)>depth(B)\end{cases} \tag{4}$$

式中:$depth(A)$ 表示从概念 A 到根(root)的最短路径距离.

3　概念相似度

概念间相似度的计算,主要是基于概念间层次关系组织的语义词典,根据语义距离的因素,来计算层次知识树内概念之间的语义相似度[9].通常情况下,直接对概念之间的相似度进行计算是比较困难的,通常可以先计算概念之间的语义距离,然后再转换成概念之间的相似度.

3.1　语义距离

两个概念之间的语义距离,是指在层次知识树中连接这两个节点的通路中的最短路径所跨的边数.本文中用 $Dist(c_1,c_2)$ 来表示概念 c_1 与 c_2 之间的语义距离.一个概念与其本身的距离为0.一般而言,两个概念的距离越大,其相似度越低;反之,其相似度越大.这样相似度和距离之间可以建立一种简单的对应关系.满足以下条件:①两个概

念距离为0时,其相似度为1;②两个概念距离为无穷大时其相似度为0;③两个概念的距离越大,其相似度越小.由此可见它们之间是一种单调递减的关系.

由此可知概念相似度与语义距离的关系为:$sim_{con}(c_1,c_2) \propto \frac{\partial}{Dist(c_1,c_2)+\partial}$.式中,$\partial$是一个可调节参数,由领域专家给出.

3.2 概念相似度的计算

根据上面层次知识树中概念之间的距离因素,可知知识树中概念相似度的计算可表达为:

$$sim_{con}(c_1,c_2)=\frac{\beta\times\partial}{Dist(c_1,c_2)+\partial} \tag{5}$$

式中:β是概念相似度计算的可调节系数,由领域专家给出,且$\beta\in(0,1]$.

4 属性相似度

在层次知识树中,如果两个概念所拥有的相同属性名称越多,那么说明这两个概念也就越相似,由它们构成的属性相似度就应该越大[10].假设在层次知识树中,任意给出两个概念c_1,c_2,则关于属性相似度的计算公式如下[9]:

$$sim_{attr}(c_1,c_2)=x[f(attr(c_1)\cup attr(c_2))-f(attr(c_1)-attr(c_2))-f(attr(c_2)-attr(c_1))]/f(attr(c_1)\cup attr(c_2)) \tag{6}$$

式中:$attr(c_1)$和$attr(c_2)$分别表示概念c_1和概念c_2的属性的集合,它们的属性集分别为$attr(c_1)=\{a_1,a_2,\cdots,a_n\}$,$attr(c_2)=\{b_1,b_2,\cdots,b_m\}$;$f(attr(c_1)\cup attr(c_2))$表示概念$c_1$和概念$c_2$所有属性的集合的个数;$f(attr(c_1)-attr(c_2))$表示概念$c_1$拥有而概念$c_2$没有的属性集;$x$表示可调节系数由领域专家给出,且$x\in(0,1]$.

5 语义相似度关键因子及归一化计算

根据语义相似度三维空间模型,语义相似度的计算可表述为

$$sim(c_1,c_2)=W_1\times sim_{dom}(c_1,c_2)+W_2\times sim_{con}(c_1+c_2)+W_3\times sim_{attr}(c_1,c_2) \tag{7}$$

其中W_1,W_2,W_3为关键因子系数,由领域专家给出,满足$W_1>W_2>W_3$,这样体现了关键因子对相似度的贡献力度.

为了保证让语义相似度的取值范围为[0,1],提出的语义相似度计算的归一化公式如下:

$$sim(c_1,c_2)=1-\mu^{sim(c_1,c_2)} \tag{8}$$

式中:μ为归一化因子,取值为大于1的正实数.当μ取值越大,计算结果趋近1的速度越快.

6 试验及其结果

6.1 实验数据及过程

为计算的方便,构建知识库$O=\{O_1,O_2\}$,分别为描述计算机方向的领域本体,同时为了充分验证语义相似度的各个关键因子,假设所选的概念都是来自层次本体库中不同的领域,实验步骤如下:

Step1,对知识库O进行预处理,形成一个层次知识树;

Step2,分别从O_1,O_2中任意选取两个概念c_1,c_2,满足$c_1\in O_1,c_2\in O_2$;

Step3,对域相似度$sim_O(c_1,c_2)$进行计算,如果$sim_{dom}(c_1,c_2)=0$,则令$sim_{con}(c_1,c_2)=0$,$sim_{attr}(c_1,c_2)=0$,并且转到Step6;

Step4,对概念相似度$sim_{con}(c_1,c_2)$进行计算,如果$sim_{con}(c_1,c_2)=0$,则令$sim_{attr}(c_1,c_2)=0$转到Step6;

Step5,对属性相似度$sim_{con}(c_1,c_2)$进行计算;

Step6,确定关键因子系数W_1,W_2,W_3,令$W_1=0.5$,$W_2=0.3$,$W_3=0.2$,并且计算语义相似度$sim(c_1,c_2)$;

Step7,对比几种相似度计算,给出结论.

6.2 实验分析

通过表1的4组计算结果可知:语义相似度在充分考虑域相似度、概念相似度、属性相似度关键因子的前提下,通过对关键因子加权值的方法对相似度计算的贡献,可以过滤掉不相关的领域本体,减少了计算的复杂性,对于域相似度大于所设定阈值的领域本体中,从语义的概念和属性两个特征进行考虑,分别计算概念相似度和属性相似度,这种综合的基于概念格理论的语义相似度计算方法虽然在计算复杂性上比单方面计算相似度高,但是该方法能够有效地提高信息检索的效率,更加有效地反映了人类的思维方式,在基于知识库的信息检索方面有较好的研究价值.

7 结论

基于概念格理论,在层次知识库下,对语义相似度计算模型构建三维空间模型,把语义相似度的计算模型分为域相似度、概念相似度、属性相似度3个关键因子,并对3个关键因子进行详细的分析,最后得出关键因子加权值来说明对语义相似度的贡献,并进行归一化处理,实验结果验证,

概念格理论下的语义相似度的计算方法能有效地提高了信息检索质量和效率.但是这种方法也存在一些不足,在计算的过程中要不断的对相似度计算公式的权值进行人工的设置,这样增加了很多人为因素,有可能会对计算结果产生误差,所以下一步的工作就是通过不断的实验迭代,找到比较合适的权值,为语义相似度的计算提供最优的结果.

表1　相似度计算比较

Tab.1　Similarity comparison

概念 c_1	概念 c_2	域相似度	概念相似度	属性相似度	归一化语义相似度
Organization	Education Organization	0.763	0.652	0.425	0.662 1
Research Assistant	Assistant	0.415	0.874	0.264	0.522 5
Course	Work	0.301	0.368	0.137	0.288 3
Person	Employee	0.124	0.763	0.429	0.376 7
Article	Publication	0.017	0.354	0.762	0.267 1
Article	Paper	0.325	0.695	0.856	0.542 2
Publication	Paper	0.013	0.253	0.312	0.144 8
Student	Undergraduate Student	0.865	0.584	0.750	0.757 7

参考文献:

[1] DOAN A H, MADHAVAN J, DOMINGOS P, et al. Learning to map between ontologies on the semantic web [C] // Proceedings of the 11th International World Wide Web Conference. Honolulu, Hawaii, USA, 2002:662 - 673.

[2] MITRA P, WIEDERHOLO G, KERSTIN M. A Graph-oriented model for articulation of ontology interdependencies[C] // Conference on Extending Database Technology 2000, Konstanz, Germany, Mar. Berlin:Springer-Verlag, 2000:86 - 100.

[3] PENG Yun, ZOU Yo-yoing, LUAN Xiao-cheng at al. Semantic resolution for E-Commerce[C] // Proceeding of AAMAS 2002, Bologna, Italy, 2002:219 - 230.

[4] 郑丽萍,李光耀,梁永全,等.本体中概念相似度的计算[J].计算机工程与应用,2006,42(30):25 - 27.

[5] 李峰,李芳.中文词语语义相似度计算——基于《知网》[J].中文信息学报,2007,21(4):99 - 105.

[6] WILLE R. Restructuring lattice theory: An approach based on hierarchies of concepts[C] // RIVAL I. Ordered Sets Boston 1982, Boston, 1982:445 - 470.

[7] BISSOM G. Why and how to define a similarity measure for object based representation systems[C] // Towards very Large Knowledge Base. 1995:236 - 246.

[8] 周莉.本体间相似度计算及映射方法的研究[D].大连:大连海事大学信息科学技术学院,2008.

[9] 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.

[10] 黄果,周竹荣,周亭.基于领域本体的语义相似度计算研究[J].计算机工程与科学,2007,29(5):112 - 116.

Research of Semantic Similarity Computing Model and Validation Based on Concept Lattice

ZHANG Xiao-hong

(Computing Information Project Department, Henan Junior College of Finance & Taxation, Zhengzhou 451464, China)

Abstract: In this paper, the computing model of semantic similarity is defined as a three-dimensional model consisting of domain, concept and property. At the same time, according to the domain-specific ontology set, from the perspective of the theory of concept lattice, the paper studies the impacts of this model on the semantic similarity computing. This method assigns different weights to different features to adjust the contributions of each feature to the semantic similarity computing model in order to achieve optimal results and improve the accuracy. Experiments show that this method, compared with the methods of other computing similarity, can effectively improve the recall and precision of semantic similarity calculation.

Key words: semantic similarity; concept lattice; domain; concept; attribute