基于概念相似度计算的多策略本体映射研究

王 凡,陈 健
(陕西师范大学 计算机科学学院,陕西 西安 710062)

摘 要:在信息过载时代,本体映射是本体相似度计算的关键步骤。针对当前本体映射中相似度计算没有充分利用本体富含的语义信息的问题,文中提出一种改进的多策略概念相似度计算方法。该方法从本体概念名称、属性、实例和层次结构四个方面讨论本体概念的相似度,并在计算实例相似度时,提出使用差异度与丰富度两个关键因子解决实例差异问题。最后采用 sigmoid 函数自动生成各策略结果对应的权重,合并映射结果,提高映射质量。文中采用了两组测试数据与 RIMOM 方法进行实验对比。实验结果表明,该方法能够有效提高映射结果的查全率和查准率。

关键词:本体;映射;语义相似性计算;信息增益
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2015)04-0038-05
doi:10.3969/j.issn.1673-629X.2015.04.010

Research on Multi-strategy Ontology Mapping Based on Conceptual Similarity Computing

WANG Fan, CHEN Jian
(School of Computer Science, Shaanxi Normal University, Xi'an 710062, China)

Abstract: Ontology similarity calculation is the pivotal to ontology mapping in the era of information overload. Considering that the current ontology mapping similarity calculation does not take full advantage of semantic information of each issue, an improved multi-strategy concept similarity computing approach was proposed to solve it. The method explores the similarity of ontology concept from four aspects, including ontology concept name, attributes, hierarchies and instances, and in calculating the instance similarity, two key factors, difference degree and richness are used to solve the instance differences problem. Finally, with the sigmoid function, the weight of each strategy are automatically generated and used to combine mapping results to improve mapping quality. Apply two group data to compare with the RIMOM. The experimental results show that the method can effectively improve the mapping results recall and precision.

Key words: ontology; mapping; semantic similarity computing; information divergence

0 引 言

本体在软件工程、信息检索、智能信息集成、自然语言处理[1]等前沿领域中扮演着越来越重要的角色。领域本体可以有效地组织管理领域中的知识,让知识可以更好的共享和重用。但是由于在世界范围内的本体构建还没有一个公认的标准,各个领域定义了自己的本体构建规范导致了本体异构性问题的存在。本体映射可以最小成本地解决异构性问题,本体映射是发现本体实体(概念、属性、关系和实例)之间的语义关系,并映射相应关系的过程,同时也是本体集成、本体合并与本体对齐的前提基础。

本体概念间的相似度计算是本体映射的关键与核心,主要分为两类:基于特征模型的相似度计算方法和基于信息内容的相似度计算方法。基于特征的相似度方法以 Tversky 模型为典型,该方法将概念的相同特征、差异特征的数量作为依据来衡量概念间的相似程度。基于信息内容的相似度方法由 Resnik 提出的为代表,该方法认为概念之间的相似性取决于它们共享信息的程度,共享信息程度越大越相似。这两种方法都有一定的缺点和不足。前者仅考虑了概念的公共特征集合,没有考虑特征之间的层次结构,所以概念的相似度计算结果准确度不高。Madhavan J 等运用形式

收稿日期:2014-05-19 **修回日期**:2014-08-24 **网络出版时间**:2015-02-23
基金项目:陕西省高等学校教学改革研究项目(13BY23);西安市工业应用技术研发项目(CXY1133(5))
作者简介:王 凡(1990-),男,硕士研究生,研究方向为本体映射、电子学习;陈 健,博士,副教授,CCF 会员,研究方向为电子服务、服务工程、软件工程。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150223.1233.013.html

概念分析的方法对第一类方法进行改进[2-3],但忽略了概念之间的语义关系,精度提高仍不明显。Yeung R 也提出了基于信息量的方法[4],该方法认为具有相同长度的连接边表示具有相同的语义距离,但是没有考虑到不同边的语义强度应该不同。Wong 在前人基础上提出结合关联规则的方法来获得更丰富的本体结构关系[5],但是没有充分利用这些结构信息来计算相似度。事实上,概念间的边仅表示节点间具有语义联系,并不能量化概念间的语义距离。

文中在 MD3 模型的基础上,提出一种改进的多策略概念相似度计算模型[6],解决了当前相似度计算不全面和计算量过大的问题。该模型基于本体与本体概念的特性,考虑多方面的因素,从本体概念名称、概念结构、概念属性、概念实例四个方面讨论概念相似度,使用 sigmoid 函数对权值变化得到各个相似度的对应权值,然后根据不同情况,运用适当的合并策略融合计算结果,最后得到一个合理的概念综合相似度计算结果。

1　本体及本体映射概述

本体(Ontology)最早是哲学上的概念,从哲学的范畴说,本体是客观存在的一个系统的解释或者说明,关心的是客观现实的抽象本质[7]。在人工智能界,本体最经典的定义由 T R Gruber 于 1993 年提出,"本体是概念化的明确的规范解释说明[8]"。其后 Studer 总结分析了这个定义,他将本体的概念分为以下四个方面。

(1)概念化模型:概念化模型是对现实客观世界事物抽象的模型,所抽象出的模型具有该事物的所有相关概念。

(2)明确:没有二义性,所使用的概念、概念关系、概念约束都有准确的定义。

(3)形式化:精准的数学描述,计算机对本体可理解。

(4)共享:本体中表达的知识是各自相关领域公认的概念集。

本体可由五元组形式化的表述为 $O=(C,F,R,A,I)$。其中,C 代表抽取出的类的集合;F 代表在类集合上的函数集合;R 代表定义在类集合上的关系集合;A 代表公理集合,是采用特定形式的逻辑断言(包括规则在内)的集合;I 代表类的实例集合。

本体映射是重用已存在的本体,通过一定的方法对其进行展开和组合,集成不同领域的本体用来实现一个更大的信息和知识池,从而支持新的交流与使用。本体映射的典型定义[5]:本体映射是指寻找两个本体之间的语义关联,对于两个异构本体 o_{i_1}、o_{i_2},对于 o_{i_1} 中的一个概念都可以在本体 o_{i_2} 中为其找到一个语义相同或相似的相应概念,本体 o_{i_2} 中的每个概念或节点同样如此。下面是本体映射形式化的描述:

(1) $\text{map}:o_{i_1}\rightarrow o_{i_2}$;

(2)如果 $\text{Sim}(e_{i_1j_1},e_{i_2j_2})>t$,则 $\text{map}(e_{i_1j_1})=e_{i_2j_2}$。其中,$t$ 是阈值,$e_{i_1j_1}$ 和 $e_{i_2j_2}$ 分别属于 o_{i_1} 和 o_{i_2},$\text{Sim}(e_{i_1j_1},e_{i_2j_2})$ 是实体 $e_{i_1j_1}$ 和 $e_{i_2j_2}$ 的相似度。当异构本体中两个元素的相似度大于给定阈值,则认为它们在语义上是相等的。

Ehrig 与 Staab 分析了前人的工作,整理出本体映射的 6 个步骤[9]。

(1)提取特征。在本体中提取概念名称、属性名称、概念实例等用于相似度计算的各种特征。

(2)选取用于映射的概念对。

(3)计算选取的概念对特征的相似度。

(4)整合相似度。衡量实体间的相似度往往使用多种不同的策略方法,产生多种相似度结果,所以要对每个相似度进行综合的考虑,从而得到一个合理的综合相似度。

(5)优化。第(4)步完成后,已经获得待映射的每个实体之间的最初相似度,此时可以利用领域相关知识对结果进行调节,例如两个概念的子概念相似,它们也可能相似。

(6)迭代。反复进行步骤(1)到(5),直到得到满意结果。

2　改进的概念相似度计算

2.1　概念相似度计算模型框架

文中方法首先在映射的异构本体中提取概念对,之后对提取的概念对分别从概念名称、特征属性、概念结构与概念实例四个方面计算概念相似度,最后使用 sigmoid 函数加权进行相似度合并,得到一个合理的概念综合相似度计算结果。其中特征属性的相似度计算是加权综合对象类型属性和数据类型属性两种属性相似度的结果,弥补了现有特征属性相似度计算中没有根据属性贡献度细分属性类型的不足,同时在计算特征属性相似度时还引入了信息增益的方法,缩小了概念范围,减少了概念相似度的计算量,并且差异度、丰富度因子的引入减少了实例少带来的映射误差。文中提出的综合相似度计算模型框架如图 1 所示。

2.2　相似度计算

2.2.1　概念名称相似度策略

概念名称相似度策略的理论依据是:如果表示 2 个概念名称的标志符是相同的或者相似的,那么它们的意义一般情况下也是相同或相似的。因此,通常可以使用知识库的方式对本体进行语言级的相似度判

断。普遍的方法是利用向量空间模型(Vector Space Model,VSM)来判断名称的相似性,它将相似度计算转换成了信息检索问题[2];Euzenat 等提出利用编辑距离的方法计算概念名称相似性[9]。但是这些方法都存在一定的缺陷:信息检索的方式会得到一些无法预测的结果(该方法只有在标识符相同或者部分相似的情况下才能使用);编辑距离的方法是将两个字串之间,由一个转成另一个所需的最少编辑操作次数(替换、插入和删除)作为衡量相似度的标准。但是这种方法忽略了同义异名的问题:两个字串的意义很相似,但是可能在拼写上完全不相同。

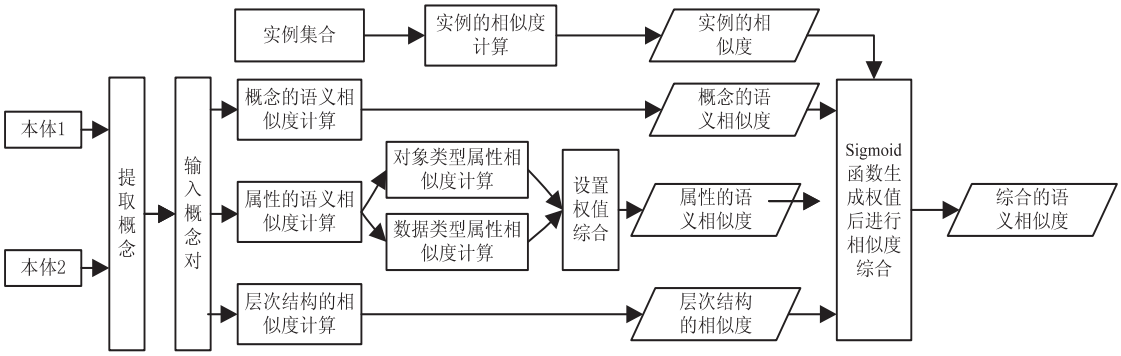

图 1 综合相似度计算模型框架

文中借鉴文献[10],采用一种基于 HowNet(知网)的改进方法来计算概念名称相似度。知网中的每一个概念都通过义原来描述,义原是知网中不能分割的最小单位。假设二个义原在由 Hyponymy 关系构成的层次体系树中有公共父节点,那么,可以获得两个义原间的语义相似度为:

$$\mathrm{Sim}(a,b)=\frac{2\times\log p(S_p)}{\log p(a)+\log p(b)} \tag{1}$$

$$p(s)=\mathrm{count}(s)/\mathrm{total} \tag{2}$$

其中,S_p 表示离义项 a,b 最近的共同祖先;$p(s)$ 是词典中节点 s 的子节点个数(包含自己)与树中的所以节点个数之比;total 是树中所有节点。一个概念往往由多个义原表述,所以先求出所有义原的相似度,再加权综合求出概念名称的相似度。公式如下:

$$\mathrm{Sim}_{\mathrm{name}}(A,B)=\sum_{i=1}^{m}w_i\times\max_{1\leqslant j\leqslant n}\mathrm{Sim}(a_i,b_i) \tag{3}$$

其中,m,n 是概念 A,B 的义原个数;w_i 是第 i 个义原占有的权重。

2.2.2 特征属性相似度策略

特征属性相似度策略理论依据是:如果 2 个概念的属性的公共属性越多,那么这 2 个概念越相似;如果概念的属性的域一样,那么这 2 个属性一样。属性包括属性名称、对象类型、实例数据等组成部分,衡量属性之间的相似度主要从这 3 个方面考虑。

属性名称、对象类型本身都是字符串,所以可以运用字符串匹配的方法来判断相似度。传统的字符串匹配方法有 Ngram、Levenshtein Distance[9]、Humming Distance 等。文中利用汉明距离[6]的方法来计算字符串之间的相似度。假定有 2 个字符串 s 和 t,衡量 s 和 t 之间的相似度公式如下:

$$\mathrm{Sim}_{\mathrm{str}}(s,t)=\frac{1-\left[\sum_{i=1}^{\min(|s|,|t|)}f(i)+\big||s|-|t|\big|\right]}{\max(|s|,|t|)} \tag{4}$$

其中,如果 $s[i]=t[i]$,那么 $f(i)=0$;否则 $f(i)=1$。因为所有概念的实例都对该概念的每一个相对属性分配了一个对应的值,所以对于属性实例数据的相似度可以采用基于概念实例的相似度计算方法求得。

设概念 A 和 B 的属性分别为 a_i 和 b_j,则 a_i 和 b_j 之间的相似度计算公式如下:

$$\mathrm{Sim}(a_i,b_j)=w_1\times s_1(a_i,b_j)+w_2\times s_2(a_i,b_j)+w_3\times s_3(a_i,b_j) \tag{5}$$

其中,w_i 表示权重,是属性名称、对象类型、实例数据对于计算属性相似度时的相对重要程度,其和等于 1。假定概念 A 和概念 B 一共求得 m 个 $\mathrm{Sim}(a_i,b_j)$,且设定了对应的权值 l_s,则 A 与 B 之间的属性相似度为:

$$\mathrm{Sim}_{\mathrm{attribute}}(A,B)=\sum_{i=1}^{m}l_s\mathrm{Sim}(a_i,b_j)\Big/\sum_{i=1}^{m}l_s \tag{6}$$

通常一个概念拥有很多属性,每个属性对概念的描述和作用各不相同,如果将概念的全部属性都计算在内,计算量毫无疑问会相当大。因此,可以先利用机器学习的方法计算出属性的信息增益[11],确定各个属性的优先级大小,最后选取优先级较大的属性来计算属性的相似度。

2.2.3 概念实例相似度策略

概念实例相似度策略的理论依据是:如果两个概念具有的实例完全一致,则认为它们二者是相同的;如果两个概念富含的相同实例越多,则这两个概念的相似程度越高。传统方法是采用文献[11]中利用计算

Jaccard 系数的方式,来计算概念 A 和 B 的实例相似度,然而这种方法没有考虑到实例个数的差异,当映射的本体间富含的实例数目差异较大时,得出的映射结果很容易失真,因此文中在文献[11]的方法上进行改进,增进丰富度和差异度 2 个关键因子。

丰富度:

$$\text{Richness} = \min\left\{1 - \frac{1}{\sqrt{\text{sum}_A + a}}, 1 - \frac{1}{\sqrt{\text{sum}_B + a}}\right\} \tag{7}$$

其中, sum_A , sum_B 分别是映射本体的实例集合; Richness 值随实例数目增大而增大,但随着实例数目增大, Richness 的增长速度会放平缓(2 个实例比 1 个实例可信度要高,但是 100 个实例和 90 个实例没有明显差别); a 的作用是使 Richness 在实例数目为 1 时不会太小。

差异度:

$$\text{Difference} = 1 - \frac{\min(\text{sum}_A, \text{sum}_B)}{\max(\text{sum}_A, \text{sum}_B)} \tag{8}$$

差异度反映映射本体之间实例丰富程度的差异,丰富程度相差越大,差异度越大。当差异度过大时,就算 A 的实例完全被映射($|\text{sum}_A \cap \text{sum}_B| = |\text{sum}_A|$),计算出来的实例相似度也达不到阈值。为了解决这个问题,当差异度较大时,分母用 $2 \times \min(|\text{sum}_A|, |\text{sum}_B|)$ 来取代 $|\text{sum}_A \cup \text{sum}_B|$ 。所以,改进后的基于实例的实体相似度计算公式为:

$$\text{Sim}_{\text{instance}}(A,B) = \begin{cases} \text{Richness} \cdot \text{JaccardSim}(A,B) & \text{Difference} \leqslant E \\ \text{Richness} \cdot \dfrac{|\text{sum}_A \cap \text{sum}_B|}{(|\text{sum}_A|, |\text{sum}_B|)} & \text{Difference} > E \end{cases} \tag{9}$$

2.2.4　概念层次结构策略

概念层次结构策略的理论依据是:如果两个概念具有相同或者相似的父概念和子概念,则认为这两个概念是相似的。本体的层次结构拥有着丰富的语义信息,对于衡量概念间的相似度起着相当重要的作用。所以,近年来很多相关的领域研究者都提出了基于层次结构的相似度计算方法。如谷志锋提出的通过匹配概念的语义邻居来计算其结构相似度的方法[12]。该方法用节点间的语义邻居包含的名称关系来计算相似度,但是这种方法忽略了其他层次信息,相似度计算结果并不准确。

文中采用西北大学徐茜改进的层次结构相似度算法[13],以基础相似度为基本,首先求得概念节点间的父概念的相似度(如公式(10)),再求得兄弟节点集合中各个兄弟节点的相似度(如公式(11))与子概念节点间的各个子概念节点的相似度(如公式(12)),之后综合父概念、子概念和兄弟概念的计算结果,最终得出概念的层次结构相似度。

$$\text{Sim}_{\text{parent}}(A,B) = \text{Sim}_{\text{base}}(A_i, B_j) \tag{10}$$

其中, A_i , B_j 分别代表概念 A 和 B 的父概念。

$$\text{Sim}_{\text{brother}}(A,B) = \sum_{i=1}^{m} w_i \sum_{j=1}^{n} w_j \, \text{Sim}_{\text{base}}(a_i, b_j) \tag{11}$$

其中, a_i 和 b_j 分别表示概念 A 和 B 的兄弟概念; m 和 n 则分别表示 a_i 与 b_j 的概念的数目。

$$\text{Sim}_{\text{son}}(A,B) = \sum_{i=1}^{k} w_i \sum_{j=1}^{l} w_j \, \text{Sim}_{\text{base}}(a_{si}, b_{sj}) \tag{12}$$

其中, a_{si} , b_{sj} 分别表示概念 A 和 B 的子概念; k 和 l 则分别表示 a_{si} 与 b_{sj} 的子概念的数目。

$$\text{Sim}_{\text{base}}(A,B) = \gamma_1 * \text{Sim}_{\text{name}}(A,B) + \gamma_2 * \text{Sim}_{\text{attribute}}(A,B) + \gamma_3 * \text{Sim}_{\text{instance}}(A,B) \tag{13}$$

其中, $\gamma_1 + \gamma_2 + \gamma_3 = 1$。

最后,概念结构相似度的计算公式如下所示:

$$\text{Sim}_{\text{structure}}(A,B) = \omega_1 * \text{Sim}_{\text{parent}}(A,B) + \omega_2 * \text{Sim}_{\text{brother}}(A,B) + \omega_3 * \text{Sim}_{\text{son}}(A,B) \tag{14}$$

其中, $\omega_1 \geqslant \omega_2 \geqslant \omega_3$ 且 $\omega_1 + \omega_2 + \omega_3 = 1$。

2.3　多策略相似度合并

为了得出一个合理的综合计算结果,需要将上述四种策略计算出的相似度进行合并,当前的合并方法多种多样,文中使用目前比较通行的 Composite 方法[14],使用的计算公式如下所示:

$$\text{Sim}(A,B) = \sum_{i=1}^{4} \varphi_i X_i \tag{15}$$

其中, X_1, X_2, X_3, X_4, X_5 分别是 $\text{Sim}_{\text{name}}(A,B)$, $\text{Sim}_{\text{instance}}(A,B)$, $\text{Sim}_{\text{attribute}}(A,B)$, $\text{Sim}_{\text{structure}}(A,B)$, $\text{Sim}_{\text{relation}}(A,B)$ 的权值; φ_i 通过 sigmoid 函数产生,sigmoid 函数是具有平滑作用的函数,它能使合并结果偏向于预测值高的策略[15]。文中采用的形式为:

$$\text{Sig}(X_i) = \frac{1}{1 - e^{-4(x-0.5)}} \tag{16}$$

权值的定义如下:

$$\varphi_i = \frac{\text{Sig}(X_i)}{\sum_{i=1}^{4} \text{Sig}(X_i)}$$

3　实验分析

3.1　实验设计

(1)实验数据。

为了对以上方法进行评估,文中在两个本体集 Catalog Ontology 和 Academic Department Ontology 上进行实验。前一个本体集包括 Mini Cornell 和 Mini Washington 两个本体,分别描述了康奈尔大学与华盛顿大学的课程体系。后一个本体集包括 Maryland University 和 Toronto University,分别描述了马里兰大学与多伦多大学大学的学院部门信息(如表 1 所示)。

表 1 马里兰大学与多伦多大学的学院部门信息

数据集	本体	概念	实例
Conurse	Cornell	34	1 526
CatalogI	Washington	39	1 912
Academic	Maryland	52	500
Department	Toronto	50	600

(2)评估标准。

文中采用信息检索领域查准率和查全率作为评价标准对实验结果进行评估,并有定义如下:

$$\text{Precision}=\frac{\text{已找出且正确的匹配对数}}{\text{已找出的匹配对数}}*100\%$$

$$\text{Recall}=\frac{\text{已找出的匹配总对数}}{\text{实际存在的匹配总对数}}*100\%$$

3.2 实验结果及分析

文中算法名暂定称为 NASI,该算法使用 Jena 语言实现,开发平台为 MyEclipse8.6。实验在 2.20 GHz 酷睿双核 CPU、2 G 内存、Windows 7 操作系统的计算机上进行,分别用 Falcon,RIMOM 和 NASI 对本体集 Catalog Ontology 和 Academic Department Ontology 进行 1:1 映射实验,并用 TempLoadRunner 工具进行计算量测试,实验结果如图 2 与图 3 所示(注:图中 C-W 代表 Cornell to Washington、M-T 代表 Maryland to Toronto)。

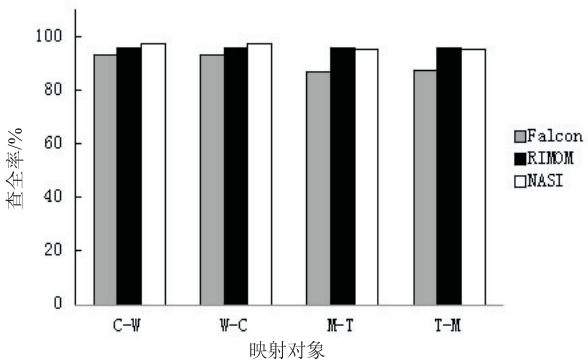

图 2 查准率

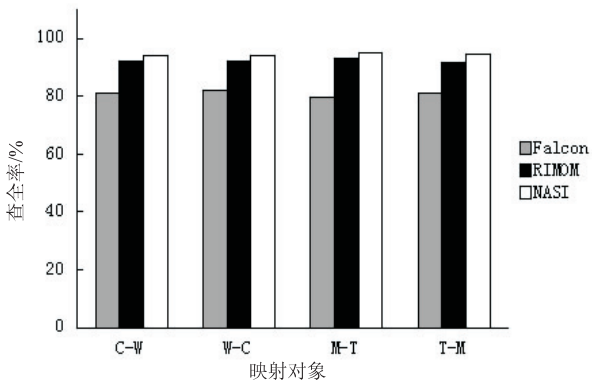

图 3 查全率

从实验结果可以看到,NASI 的映射精度显著高于 Falcon 算法,而与 RIMOM 算法基本相同。但在映射 Academic Department 这种实例差异大的本体集时,NASI 就明显优于 RIMOM 算法。这是因为 NASI 方法中,差异度与丰富度两个因子的引入,一定作用下减小了映射结果的失真,有效抵消实例差异带来的映射误差。此外,NASI 引入了信息增益的方法,缩小了概念范围,减少了概念相似度计算花费的时间,更有效率。综合来看 NASI 可以找寻到本体间的绝大多数映射关系,能够很好地完成本体映射任务。

4 结束语

文中提出了一种多策略的本体概念相似度计算方法,该方法首先计算概念的名称、实例、属性、结构四个方面的相似度,再使用 sigmoid 函数综合求到的相应权值,实现了权值与对应相似度的自适应结合。该算法还采用信息增益的方法,减少了计算量过大的问题,并引入了差异度和丰富度两个因子,减少了实例差异带来的映射误差,提高了映射准确度。但是,本体映射是一个相当复杂的过程,映射过程的每一步都会对最终结果产生不同的影响,概念相似度的计算只是其中的一步,接下来的研究重心是通过研究本体映射的其他步骤环节来提升映射中的准确度和减少映射中的复杂度。

参考文献:

[1] Klein M, Bernstein A. Searching for services on the semantic web using process ontologies[C]//Proceeding of the international semantic web working symposium. Amsterdam: IOS Press, 2001:159-172.

[2] Madhavan J, Bernstein P, Chen K, et al. Shenoy: corpus based schema matching[C]//Proceedings of the 2003 workshop on information integration on the web. Acapulco, Mexico: [s. n.], 2003.

[3] Reed M G, Syverson P F, Goldschlag D M. Anonymous connections and onion routing[J]. IEEE Journal on Selected Areas in Communications, 1998, 16(4): 482-494.

[4] Yeung R W, Zhang Z. Distributed source coding for satellite communication[J]. IEEE Transactions on Information Theory, 1999, 45(4): 1111-1120.

[5] Wong A K Y, Ray P, Parameswaran N, et al. Ontology mapping for the interoperability problem in network management [J]. IEEE Journal on Selected Areas in Communication, 2005, 23(10): 2058-2068

[6] Rodriguez M A. Assessing semantic similarity among spatial entity classes[D]. Maine: University of Maine, 2000.

[7] 邓志鸿,唐世渭,张 铭,等. Ontology 研究综述[J]. 北京大学学报:自然科学版,2002,38(5):730-738.

[8] Gruber T R. A translation approach to portable ontology specifications[C]//Proceedings of 2002 IEEE international sympo-

(下转第 47 页)

源码进行改动,在扫描数据集时,发现固定的栏目时直接过滤,该栏目 ID 将不参与候选集与频繁集的产生,当最大频繁集产生后再将固定栏目的 ID 添加到最大频繁项集中进行关联规则的挖掘,建立关联模型,程序运行的耗时结果如表 1 所示。

表 1　算法耗时比较

	原始算法 /ms	文中方案 /ms	结果
最小支持度和最小置信度分别为 10% 和 60%	157	141	优
最小支持度和最小置信度分别为 6% 和 60%	156	141	优

观察程序运行结果,发现候选集明显减少,从原算法的最多 10 个减少为 6 个,当最小支持度和最小置信度分别为 10% 和 60% 时,改进前和改进后的运行时间分别为 157 ms 和 141 ms。当最小支持度和最小置信度分别为 6% 和 60% 时,改进前和改进后的运行时间分别为 156 ms 和 141 ms。可见改进的 Apriori 算法的优化还是比较明显的。

4.2　Web 日志挖掘的结果分析

通过分析算法挖掘建立的模型,可以看出 b(企业概况)与 a(企业新闻)之间的关联性最强,其余栏目按照同 news 关联性由强到弱依次的顺序是 e(携手企业)、c(企业文化)、d(企业党建)。这给网站栏目的布局提出一定的建议:网站栏目在保留原有 a(企业新闻)第一的前提下,将 b(企业概况)和 e(携手企业)栏目次序前移至相邻位置,d(企业党建)栏目位置放在最后,这样对于访问者的阅读习惯是合适的。同时根据挖掘的企业也要加强企业文化自身的建设,这也是访问者关注和在意的热点。

5　结束语

文中结合企业实际应用对 Apriori 算法应用提出一种改进思路,并用企业网站产生的日志数据进行对比实验,证明了优化的有效性,并通过建立关联模型,分析各栏目之间的兴趣度,对网站的布局结构提出建议,对于协同门户和网站有一定的实际指导意义。

参考文献:

[1] Han Jiawei, Kamber M, Jian Pei. 数据挖掘:概念与技术[M]. 范　明,孟小峰,译. 北京:机械工业出版社,2012.

[2] 王飞跃. 开源情报与网络时代的国家安全[EB/OL]. 2007. http://www.libnet.sh.cn/tsgxh/hyzq/list.asp? id=2898.

[3] 王光宏,蒋　平. 数据挖掘综述[J]. 同济大学学报:自然科学版,2004,32(2):246-252.

[4] 韩家炜,孟小峰,王　静,等. Web 挖掘研究[J]. 计算机研究与发展,2001,38(4):405-414.

[5] Chen Ming-Syan, Park J S, Yu P S. Efficient data mining for path traversal patterns[J]. IEEE Transactions on Knowledge and Data Engineering,1998,10(2):209-221.

[6] 刘　兵. Web 数据挖掘[M]. 北京:清华大学出版社,2009.

[7] 阳小华,周龙镶. 基于用户访问模式的 WWW 浏览路径优化[J]. 软件学报,2001,12(6):846-850.

[8] Mobasher B, Dai Honghua, Luo Tao, et al. Integrating Web usage and content mining for more effective personalization[J]. Lecture Notes in Computer Science,2000,1875:165-176.

[9] Macclennan T Z H. 数据挖掘原理与应用[M]. 北京:清华大学出版社,2007:99-107.

[10] 易　芝,汪林林,王　练. 基于关联规则相关性分析的 Web 个性化推荐研究[J]. 重庆邮电大学学报:自然科学版,2007,19(2):234-237.

[11] 孙赵平,李龙澍. 基于关联规则的 Web 日志挖掘算法研究[J]. 电子技术,2010,47(8):11-13.

[12] 孟庆川,陈晓明. 基于关联规则 Web 日志挖掘算法的研究[J]. 信息技术,2010(3):96-98.

[13] 习慧丹. Web 日志挖掘探索[C]//第三届全国软件测试会议论文与移动计算、栅格、智能化高级论坛论文集. 出版地不详:出版者不详,2009:184-186.

[14] 曹　莹,苗志刚,张红霞. 基于改进的 Apriori 算法的学位预警应用研究[J]. 电脑开发与应用,2014,27(6):1-3.

(上接第 42 页)

sium on information theory. Los Alamitis:IEEE Computer Society,2002.

[9] Bouquet P, Euzenat J, Franconi E, et al. Specification of a common framework for characterizing alignment[EB/OL]. 2004. http://www.aifb.uni-karlsruhe.de/WBS/phi/kweb-221.pdf.

[10] 徐　猛,刘宗田,周　文. 一种基于知网语义相似度计算的应用研究[J]. 微计算机信息,2010,26(1-3):200-201.

[11] 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.

[12] 谷志锋,刘　勇,郭跟成. 基于相似度计算的本体映射优化方法[J]. 计算机工程,2008,34(19):56-57.

[13] 徐　茜,彭进业,李　展. 本体映射中一种综合的概念相似度计算方法[J]. 计算机工程与应用,2010,46(24):34-36.

[14] 马　驰,王念滨,张海燕. 基于相似度计算的本体映射框架[J]. 计算机工程,2009,35(12):61-63.

[15] Cai N, Yeung R W. A security condition for multi-source linear network coding[C]//Proceedings of 2007 IEEE international symposium on information theory. Nice:IEEE Computer Society,2007:561-565.

基于概念相似度计算的多策略本体映射研究

作者： 王凡， 陈健， WANG Fan， CHEN Jian
作者单位： 陕西师范大学 计算机科学学院,陕西 西安,710062
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015(4)