一种基于子图近似同构的 e-Learning 学习资源本体匹配方法*

习海旭[1], 于 枫[2,3], 王 直[3], 宋爱波[2], 王晓跃[1]
(1. 江苏理工学院 计算机工程学院, 江苏 常州 213012; 2. 东南大学 计算机科学与工程学院, 南京 211189; 3. 江苏科技大学 计算机科学与工程学院, 江苏 镇江 212003)

摘 要: 针对 e-Learning 学习资源本体异构问题,提出一种基于子图近似同构的本体匹配方法。该方法对现有本体匹配方法进行扩展,综合编辑距离、层次关系等特征,计算本体的结构级相似性,以点、边有序交替匹配来判断实体的有向图近似同构问题,实现本体匹配判定。演示算法处理过程,给出算法时间复杂度理论分析,说明其有效性。

关键词: 本体匹配; e-Learning 学习资源本体; 子图同构; 时间复杂性
中图分类号: TP391.6 **文献标志码**: A **文章编号**: 1001-3695(2014)02-0417-05
doi:10.3969/j.issn.1001-3695.2014.02.023

e-Learning resource ontology matching algorithm based on subgraph approximate isomorphic

XI Hai-xu[1], YU Feng[2,3], WANG Zhi[3], SONG Ai-bo[2], WANG Xiao-yue[1]
(1. *School of Computer Technology, Jiangsu University of Technology, Changzhou Jiangsu* 213012, *China*; 2. *School of Computer Science & Technology, Eastsouth University, Nanjing* 211189, *China*; 3. *School of Computer Science & Technology, Jiangsu University of Science & Technology, Zhenjiang Jiangsu* 212003, *China*)

Abstract: For the heterogeneity problem among ontologies, this paper presented an algorithm based on sub-graph approximate isomorphic here. It was an extension of existing methods in ontology matching. Under the comprehensive application of features such as edit distance and hierarchical relations, it calculated the similarity of graph structures between two ontologies. Lastly, it determined the ontology matching on the condition of sub-graph approximate isomorphism based on the alternately mapping of nodes and arcs in the describing graphs of ontologies. It used an example to demonstrate this ontology matching process and analyzed the time complexity to explain its effectiveness.

Key words: ontology match; e-Learning resource ontology; sub-graph isomorphism; time complexity

0 引言

20 世纪 90 年代开始,计算机网络与多媒体技术的发展为教育提供了发展的新活力,不仅教育模式、手段、范围等发生巨大变革,更实现了全球优秀教育资源的共享与传播,这种以计算机网络技术为支持的教育模式常被称为 e-Learning[1]。然而,由于 Internet 是一个高度开放、异构、分布的信息空间,以 URL 技术进行学习资源检索时并不理解信息真实含义,目标学习资源常常被淹没于大量无用冗余信息中,不能实现高效的 e-Learning 学习资源发现。为加强信息语义特征,URL 技术的发明者 Berners-Lee 提出通过本体(ontology)表示共同认可的、可共享的知识,对概念及概念间关系进行严格定义以确定概念的含义[2]。本体技术支持下的 e-Learning 中,按照学习资源元数据标准,对学习资源进行描述,建立学习资源本体,应用本体的相似性计算与匹配来支持 e-Learning 学习资源发现,能够防止学习者在网络学习环境中迷失学习方向,提高学习的效率和准确性。

相似性是判断两个 e-Learning 学习资源本体是否匹配的基本条件。然而,当前 e-Learning 环境中,学习资源本体常常是由不同创建者应用不同数据规范、建模方法和技术创建的,同一领域主题的学习资源本体常常存在很大差异,直接影响着 e-Learning 学习资源发现的效率,如何有效地解决异构学习资源本体的匹配问题,即语义 Web 中所称的本体匹配,是 e-Learning 学习所面临的一项挑战。

目前,国内外学者已提出了很多本体匹配方法,主要有基于语言学的、基于结构的、基于实例的等,并开发出各种本体匹配工具,如美国 Stanford 大学开发的 ONION[3]、美国 Washington 大学开发的 GLUE[4]、德国 Karlsruhe 大学开发的 FAOM[5]等,其中 PROMPT 是基于语言学的,GLUE 和 QOM 是基于机器学习方法的。但是,目前的本体匹配方法当应用于学习资源本体匹配时,仍存在如下问题:

a)基于语言学的方法难以解决学习资源本体的匹配问

收稿日期: 2013-05-13; **修回日期**: 2013-07-09 **基金项目**: 江苏省现代教育技术研究课题(2011-R-18859,2013-R-25582);全国教育信息技术研究"十二五"规划课题(116230340)

作者简介:习海旭(1981-),男,江西吉安人,讲师,硕士,主要研究方向为教育技术学、语义网(just_xhx@126.com);于枫(1974-),女,副教授,博士研究生,主要研究方向为下一代网络、Petri 网理论及应用;王直(1963-),教授,博士,主要研究方向为语义网、自动化控制;宋爱波(1969-),男,副教授,博士,主要研究方向为语义 Web、服务计算;王晓跃(1980-),男,讲师,硕士研究生,主要研究方向为语义网.

题。原因在于:当前学习资源本体元数据标准和规范各不相同,如存在 IEEE 下属的学习技术标准委员会 LTSC 提出 LOM、国际图书馆计算机中心 OCLC 提出的都柏林元数据核心集 DCMS、IMS 全球学习联盟发布 LRM 等。不同的元数据规范决定了不同的学习资源本体描述语言,难以定义科学的语义距离,不能解决学习资源本体匹配问题。

b)现有结构匹配方法不能满足学习资源本体匹配需求。现有基于结构的本体匹配方法大多只着眼于本体自身的层次结构,较少考虑其他关系对本体匹配的影响。e-Learning 学习资源本体的匹配需要综合考虑各种关系构成的整体结构相似性,从而不能应用现有树状结构相似匹配判定方法。

c)基于实例的匹配方法受限于机器学习技术自身的复杂性、计算性能、正确性、优化等问题的困扰,在实际本体匹配应用中的有效性尚待考查,因此也不是一种可作为优选技术的学习资源本体匹配方案。

总之,在研究和分析已有本体匹配方法之后,本文提出一种 e-Learning 学习资源本体匹配方法,该方法在综合概念编辑距离、层次架构与其他关系相似性基础,对本体的有向图中交替进行点、边匹配,从而以子图近似同构来判定本体匹配。该方法以结构整体相似性为判断标准,有助于加强 e-Learning 学习资源本体的匹配效率,提高资源发现能力。

1 本体匹配研究现状

e-Learning 学习资源本体匹配是发现不同学习资源间映射关系的关键技术,对 e-Learning 学习资源本体的检索、集成、重用等有重要支撑作用。国外学者自 20 世纪 90 年代开始针对本体匹配的研究,已形成很多著名的本体匹配系统。在本体匹配方法上,文献[6]按匹配时的信息粒度、输入类型为标准,总结出如图 1 所示的本体匹配方法分类图,其中元素级是指基于本体的单个实体信息而不考虑实体间关联性,结构级是指将本体各实体信息作为一个整体结构。

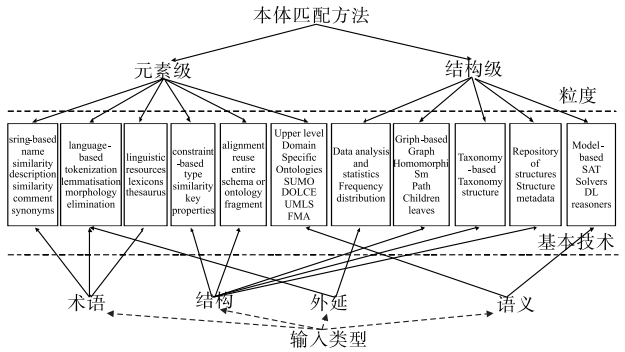


图 1 本体匹配方法分类

在匹配技术上,主要有:

a)基于字符串的匹配技术[7]。将本体的文体描述信息作为字符串来处理,利用字符串匹配方法计算本体文本间的相似度,利用编辑距离衡量字符串 S_1 和 S_2 间相似度的公式为

$$\mathrm{sim}_{\mathrm{Edit}}(S_1,S_2)=\frac{\max(|S_1|,|S_2|)-\sum_i \mathrm{oper}_i}{\max(|S_1|,|S_2|)} \tag{1}$$

其中:$|S_1|$ 和 $|S_2|$ 分别是字符串 S_1 和 S_2 的长度;iper_i 表示插入、删除、替换及相似字符调换等操作。

b)基于上层本体或领域本体的匹配技术。上层本体与领域无关,可作为待匹配本体共同认可的外部知识且可用来发现待匹配本体间的语义关系,常见的上层本体如 Cyc 本体、SUMO、DOLCE 等。领域本体包含领域内的共同背景知识,可用于消除一词多义现象,如生物医学领域的 FMA、UMLS、OBO 等。

c)基于结构的匹配技术。通常,本体被表示为树状层次结构或有向标记图结构,相似性度量或借助于 Tversky 模型,或以对象结构上的关系来计算。一般地,基于相似度的本体匹配系统架构可概括为如图 2 所示。

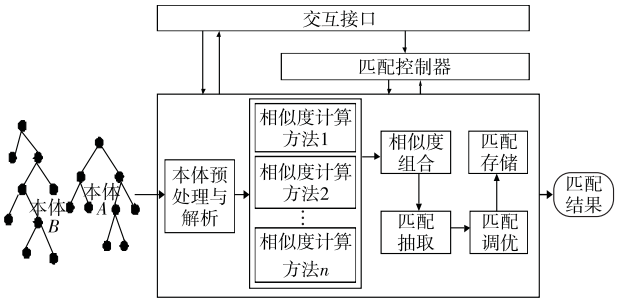


图 2 基于相似度的本体匹配系统架构

这类本体匹配技术中,结构相似特征集的抽取、相似度计算是其关键要素之一。提取结构特征信息的不同,相似性度量与计算方法就不同。例如,SF(similarity flooding)[8]的结构匹配方法不考虑模式信息,基于图节点相似性的传递性来判定本体匹配,即:若不同模式中两元素邻接节点相似,则此两元素节点相似。Cupid[9]的结构匹配阶段,叶节点相似性取决于语言学和数据类型的相似及邻近节点的相似,非叶节点相似性通过计算以其为根的子树的相似性来获得。Anchor-PROMPT[10]方法中,本体被看做有向标记图,以 anchors 间的定长路径作为抽取的结构特征,通过遍历被 anchors 限定子图中的路径,标记路径中相同位置节点相似值来表征语义上的相似。ASCO[11]方法中,节点的邻接关系和概念层次路径为被抽取为本体结构特征,结构相似性是度量和计算概念在邻接结构和路径的相似比例再取加权和。上述基于结构的本体匹配方法中,本体结构特征的相似度传播是判定匹配的重要因素,但当前方法在结构相似度计算上过多依赖于邻接节点相似度,相似性传播大多需要遍历全图,计算量大且带有盲目性,尚待进一步深入研究。

2 本体的图表示与相似性

2.1 本体的有向图表示

本体的形式化定义很多,本文采用文献[12]的定义。

定义 1 本体可定义为五元组 $O=(C,I,P,Hc,R,A^0)$,其中,C 为概念集,I 是实例集,P 是概念属性集,Hc 是概念间的层次关系集,R 是概念间其他关系集,A^0 是本体公理集。

对 $\forall r\in R$,其定义域、值域分别记为 $r.\mathrm{dom}$、$r.\mathrm{ran}$。

$$r.\mathrm{dom}=\{c_i\mid c_i\in C\wedge c_i \overset{r}{\mapsto}\},\ r.\mathrm{ran}=\{c_j\mid c_j\in C\wedge \exists c \overset{r}{\mapsto} c_j\}$$

定义 2 设本体 $O=(C,I,P,Hc,R,A^0)$ 的有向标记图表示为 $G(O)=(V,E,L_V,L_E,\mu,\eta)$,其中:

a)节点集 $V=C$,边集 $E=V\times V$;

b)$\mu:V\to L_V$ 是从节点集到节点标记集的映射函数;

c)$\eta:E\to L_E$ 是从边集到边标记集的映射。

例如,当 $\mu:V\to L_V$ 为节点赋标记为本体概念,$\eta:E\to L_E$ 为有向实线弧赋以概念间层次关系,有向虚线弧赋概念间 R 关系,则如图 3 就可看做是对一个本体的描述。

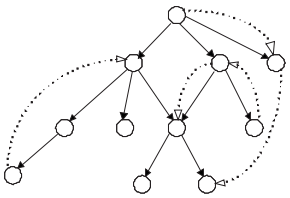


图 3　本体的有向图表示

2.2　相似性

本体相似性的一个重要指标是语义相似度,如概念的编辑距离、节点基距离、实例的概率相似、结构相似等,领域学者已提出了多种语义相似度计算方法,如编辑距离计算如式(1)所示,不再阐述。节点间基距离的计算公式可如下所示:

$$\mathrm{dist}(A,B)=1-\frac{2m}{n_1+n_2} \tag{2}$$

其中:n_1、n_2 分别是节点 A 在本体 O_1、B 在在本体 O_2 中词的个数,m 为重叠词的个数。

实例的概率相似可表示为

$$\mathrm{sim}(A,B)=\frac{P(A\cap B)}{P(A\cup B)}=\frac{P(A,B)}{P(A,B)+P(\overline{A},B)+P(A,\overline{B})} \tag{3}$$

其中:$P(A,B)$是实例同时属于概念 A 和 B 的概率,$P(\overline{A},B)$是实例属于概念 B 但不属于概念 A 的概率,$P(A,\overline{B})$是实例属于概念 A 但不属于概念 B 的概率。

基于结构的本体匹配中,图的匹配是一个 NP 完全问题,难以直接应用图结构的匹配来解决本体匹配,因此这类方法常常是通过对本体结构的相似性进行计算和匹配来实现的。一般的指导思想是:通过图中相邻元素的相似性来推测元素自身的相似度,即若节点的邻居节点相似,则节点本身也相似,其核心是相似度传播。最典型的两个基于结构的本体匹配算法 SF 和 GMO,其核心思想都是:拥有相似父/子概念的概念可能相似,拥有相似属性的概念可能相似。其中,similarity flooding 算法的相似度传播只考虑已匹配概念对邻居节点的传播,GMO 则是对全体全局进行相似度传播。

3　基于子图近似同构的本体匹配算法

3.1　e-Learning 学习资源本体匹配问题

本体匹配是解决 e-Learnin 学习资源本体异构的一种有效途径,通过计算分析不同学习资源 ontology 间的相似度判断其语义关系,实现本体的语义兼容或映射。在匹配粒度上有概念—概念、属性—属性、概念—属性间的匹配等,若两个本体 A 和 B,对于 A 中的概念,在 B 中找到一个与它语义相同或相近的对应概念,且 B 中的每个概念亦如此,则 A 与 B 是概念—概念匹配的。本文中,e-Learning 学习资源本体的匹配是指发现不同本体的实体间(概念、属性、关系等)整体语义对应的过程,形成化描述为:

定义 3　一个 e-Learning 学习资源本体匹配(ontology matching)是一个语义对应,表示为四元组:

$$OM(A,B)=\langle e_1,e_2,\mathrm{rel},\mathrm{sim}\rangle \tag{4}$$

其中:e_1、e_2 分别是本体 A、B 的实体(概念、属性、实例、公理等);rel = {⊆,⊇,⊥,=}是实体间的语义关系集(⊆,⊇,⊥,=分别表示语义的包含、不包含、无关、等价);sim = [0,1]是实体间的语义对应程度度量值。

3.2　基于子图近似同构的本体匹配方法

3.2.1　整体框架

基于子图近似同构的 e-Learning 学习资源本体匹配方法(ontology matching based subgraph isomorphism, SIOM)的整体框架如图 4 所示。由图可知,SIOM 是一个顺序匹配器,主要含锚点选择与子图抽取、子图结构相似度计算、子图近似同构判定和基于相似同构子图的本体匹配四大步骤。

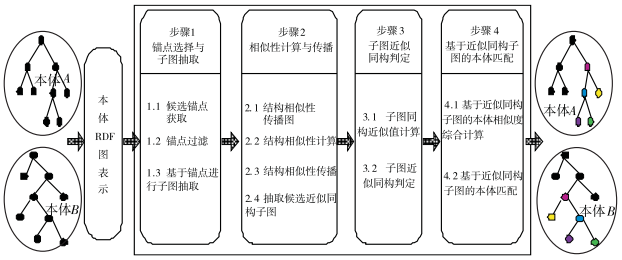


图 4　基于子图近似同构的本体匹配框架

3.2.2　锚点选择与子图抽取

所谓锚点,本文指待匹配本体 A 与 B 中第一对可确定相似的概念,表现在本体的有向标记图上第一对确定匹配的节点。定义如下:

定义 4　锚点。给定两个待匹配本体 A、B,其对应的图结构表示分别为 $G(A)$、$G(B)$,若对于 $G(A)$ 中的一个节点 $x\in C_A$,在 $G(B)$ 中存在节点 $y\in C_B$,有

a) $OM(x,y)$,即概念 x 与 y 可匹配;

b) $\forall I(x)\in I_A, P(x)\in P_A, Hc(x)\in Hc_A, R(x)\in R_A, A_A^0(x)\in A_A^0, \exists I(y)\in I_B, P(y)\in P_B, Hc(y)\in Hc_B, R(y)\in R_B, A_B^0(y)\in A_B^0$,有

$$OM(I(x),I(y))\wedge OM(P(x),P(y))\wedge OM(R(x),R(y))\wedge OM(Hc(x),Hc(y))\wedge OM(A_A^0(x),A_B^0(y)) \tag{5}$$

则称$\langle x,y\rangle$为 A、B 的一对锚点,x 与 y 称为锚点概念。

根据锚点在本体层次结构中所处位置的不同,可分为如下九种情况:a)x 与 y 分别为 $G(A)$、$G(B)$的根节点;b)x 为 $G(A)$的根节点,y 为 $G(B)$的中间节点;c)x 为 $G(A)$的根节点,y 为 $G(B)$的叶节点;d)x 为 $G(A)$的中间节点,y 为 $G(B)$的根节点;e)x 为 $G(A)$的中间节点,y 为 $G(B)$的中间节点;f)x 为 $G(A)$的中间节点,y 为 $G(B)$的叶节点;g)x 为 $G(A)$的叶节点,y 为 $G(B)$的根节点;h)x 为 $G(A)$的叶节点,y 为 $G(B)$的中间节点;i)x 为 $G(A)$的叶节点,y 为 $G(B)$的叶节点。

定义 5　给定本体 O,x 为 O 的锚点概念,则由锚点导出的本体 O^x 可表示为五元组 $O^x=(C^x,I^x,P^x,Hc^x,R^x,A_x^0)$,其中:

a) $C^x=\{c\in C\mid (cHcx)\vee(xHcc)\vee(cRx)\vee(xRc)\}$是概念集;

b) $P^x=\{P\rightarrow\{C^x\}\}$、$I^x=\{I\rightarrow\{C^x\}\}$是属性集、实例集;

c) $Hc^x=\{Hc\rightarrow\{C^x\}\}$是 O^x 的概念间层次关系集;

d) $R^x=\{R\rightarrow\{C^x\}\}$是 O^x 的概念间其他关系集。

推论 1　给定本体 O 及其锚点概念 x 导出的本体 O^x,若有向图 $G(O)$、$G(O^x)$分别为其对应的图结构表示,则有

$$G(O^x)\subseteq G(O) \tag{6}$$

证明　由定义 1、2、7 可知推论 1 成立。

推论 2　对于给定的本体 O 及其锚点概念 x 导出的本体 O^x,对其有向图表示 $G(O)$、$G(O^x)$来说,有:

a)若 x 为 $G(A)$的根节点,则 $G(O^x)=G(O)$;

b)x 为 $G(A)$的非根节点,则 $G(O^x)\subset G(O)$。

特别地,当 x 为 $G(A)$的叶节点时,$G(O^x)$退化为 $G(O)$中的一个节点。

证明　由上文对锚点概念在本体层次结构中所处位置的分析及推论 1 可知,推论 2 成立。

3.2.3 本体有向图的结构相似性计算

对于待匹配本体 A、B 及其有向图表示 $G(A)$、$G(B)$，$G(A)$ 与 $G(B)$ 相似性计算由四部分构成：

a）节点编辑距离相似度计算。由节点所代表的概念相似度、属性相似度综合计算得到。具体方法如下：设 x 与 y 分别为 $G(A)$、$G(B)$ 中的两个节点，$S_e^c(x,y)$ 为 x 与 y 的概念编辑距离，$S_e^p(x,y)=\dfrac{2|p|_{p\in P^A\cap P^B}}{|P^A|+|P^B|}\sum_p S(p(x),p(x))$ 为 x 与 y 的共有属性编辑距离，应用式(1)计算，则节点 x 与 y 间相似度计算公式为

$$S_e(x,y)=\alpha\cdot S_e^c(x,y)+\beta\cdot S_e^p(x,y) \tag{7}$$

其中：α、β 为权重调整参数，且 $0<\alpha,\beta\leqslant 1\wedge\alpha+\beta=1$。

b）节点间层次关系相似度。设 $G(A)$ 中 x 的层次关系入度集为 $x_{in}=\{x_i\in V(A)\mid\exists x_iHcx\}$，层次关系出度集为 $x_{out}=\{x_j\in V(A)\mid\exists xHcx_j\}$，类似 y 在 $G(B)$ 中的层次关系入度集、层次关系出度集分别为 y_{in}、y_{out}，则 x 与 y 间层次关系相似性计算公式为

$$S_{Hc}(x,y)=\frac{|x_{in}\Leftrightarrow y_{in}|+|x_{out}\Leftrightarrow y_{out}|}{|x_{in}|+|x_{out}|+|y_{in}|+|y_{out}|-|x_{in}\cap x_{out}|-|y_{in}\cap y_{out}|} \tag{8}$$

$x_{in}\Leftrightarrow y_{in}=\{x\mid x\in x_{in},y\in y_{in}:S_e(x,y)\in OM(x,y)\}$ 表示与 x、y 有层次关系的父节点中可匹配的节点集；

$x_{out}\Leftrightarrow y_{out}=\{x\mid x\in x_{out},y\in y_{out}:S_e(x,y)\in OM(x,y)\}$ 表示与 x、y 有层次关系的子节点中可匹配的节点集。

c）节点间其他关系相似性。记与 x、y 存在关系 R 的节点集分别为

$$x^R=\{x'\in V(A)\mid r\in R^A:\exists(x'rx)\vee(xrx')\}$$

$$y^R=\{y'\in V(B)\mid r\in R^B:\exists(y'ry)\vee(yry')\}$$

若 $\exists r_1\in R^A,r_2\in R^B$，有

$$((x'r_1x)\wedge(y'r_2y)\wedge OM(x',y'))\vee((xr_1x')\wedge(yr_2y')\wedge OM(x',y')) \tag{9}$$

把满足式(9)的节点集记为 $x^R\Leftrightarrow y^R$，辅以权重调整系数，则节点间其他关系相似度计算公式可表示为

$$S_R(x,y)=\sum_{r\in R^A}\gamma_i\frac{|x^r\Leftrightarrow y^r|}{|x^r|+|y^r|}+\sum_{r\in R^B}\lambda_i\frac{|x^r\Leftrightarrow y^r|}{|x^r|+|y^r|}-\sum_{r\in R^A\cap R^B}(\gamma_i+\lambda_i)\frac{|x^r\Leftrightarrow y^r|}{|x^r|+|y^r|} \tag{10}$$

权重调整参数 γ_i、λ_i 满足 $0<\gamma_i,\lambda_i\leqslant 1\wedge\sum_i r_i=1\wedge\sum_i\lambda_i=1$。

d）图的结构相似性。待匹配本体 A、B 及其有向图 $G(A)$、$G(B)$，$\langle x,y\rangle$ 为 A、B 的一对锚点，则 x 与 y 导出本体的有向图 $G(x)$、$G(y)$ 间相似性计算公式可表示为

$$S(G(x),G(y))=\eta\cdot S_e(x,y)+\varphi\cdot S_{Hc}(x,y)+\xi\cdot S_R(x,y) \tag{11}$$

其中：$\eta+\varphi+\xi=1$ 为权重调整系数。

3.2.4 基于子图近似同构的本体匹配算法

定义 6 如果有向图 G 与 G' 的点之间保持一一对应，边之间保持一一对应，而且对应点与对应边之间保持相同的关联关系，则称 G 与 G' 同构，记为 $G\cong G'$。

由于本体匹配中一般难以达到严格的一一对应，只要本体间的相似程度满足预设阈值，即可判定为匹配，因此本文提出本体图结构的近似同构概念。

定义 7 给定目标本体 A 及待匹配本体 B，其有向图表示为 $G(A)$、$G(B)$，若：a）对于 $G(A)$ 的根节点 a，存在 $G(B)$ 中的一个节点 b，有 $\langle a,b\rangle$ 为 A、B 的一对锚点；b）对于 $G(A)$ 与 b 导出本体 B^b 的有向图 $G(B^b)$，有

$$V(A)\subseteq V(B^b),E(A)\subseteq E(B^b)$$

$$\forall x\in V(A):\exists y\in V(B^b):OM(x,y)$$

$$\forall e\in E(A):\exists e'\in E(B^b):OM(e,e')$$

对于设定的匹配阈值 θ，有 $S(G(A),G(B^b))\geqslant\theta$，则称本体 A 与 B 是图近似同构的，记为 $G(A)\approx G(B)$。

基于图近似同构，SIOM 算法的主要思想是：按图的广度优先遍历次序，在确定了匹配的锚点节点后，基于节点出入度实现图节点与边的交替匹配，从而在待匹配本体 B 的图 $G(B)$ 中找到一个与 $G(A)$ 近似同构的子图。关键步骤主要有：首先，确定 $G(B)$ 中与 $G(A)$ 根节点 a 对应的锚点节点 b；接着，生成 B 的锚点导出本体 B^b 及其有向图表示 $G(B^b)$；再次，在 $G(A)$ 与 $G(B^b)$ 间进行图的近似同构判定。若两者满足近似同构关系，则称 A、B 是可匹配的，否则，迭代上述过程直至达到收敛要求。下面给出算法主要操作的伪码描述。

```
Algorithm OM(A,B)
Input: A、B、G(B)、G(A)、a、θ
Output: Y or N
1  for each node n-anchor(a,b)
2  generate B^b;
3  get G(B^b) from G(B)
4  node-add(N^a,N^Bb);arc-add(E^a,E^Bb)
5  while N^a≠∅ do
6     x∈N^a: select y∈N^Bb s.t. S_e(x,y)≥θ_e
7     for each arc e∈E^a related to node x
8.       for arc e∈E^b related to node y in E(B^b)
9.          map(x→y);
10         Calculate S_Hc(x,y),S_R(x,y)
11       Calculate S(G(x),G(y))
12    Generate subgraph G^x(A),G^y(B^b)
13    Test = DAI(G^x(A),G^y(B^b)
14    if N^a=∅ then OM(A,B)=T else OM(A,B)=F
15 end
```

4 学习资源本体匹配演示与分析

4.1 e-Learning 学习资源本体

以课程本体构建为例，说明 e-Learning 学习资源本体的构成要素。通常，一门课程包含知识点、习题、案例、答疑等多要素，其中知识点是指根据课程教学大纲对课程分解的、构成学习资源的逻辑独立的基本单元。依据教学实践经验和学习规律，知识点间主要有以下几类关系：

a）前驱/后继关系（pre/suc-of）。若学习知识点 B 前必须学习知识点 A，则 A 为 B 的前驱，B 为 A 的后继。

b）包含关系（include-of）。若知识点 A 由更小粒度的知识点 A_1、A_2…构成，且 A_1、A_2…本身也是可以独立使用的逻辑单元，则 A 与 A_1、A_2…间存在包含关系。

对一个知识点 A 来说，若 A 包含其他知识点，则 A 称为复合知识点；若 A 不再包含更小粒度的知识点，则 A 称为元知识点。特别地，若 A、B 有完全相同的前驱/后继知识点，且包含内容也完全一致时，则称 A 与 B 等价。

c）相关关系（related to）。若知识点 A、B 同时包含知识点 C，则 A 与 B 具有相关关系。引用关系（quoted-of），若知识点 A 的内容涉及到知识点 B 的内容，但 A 与 B 并不属于同一领域范畴，则称 A 与 B 间有引用关系。

上述关系中，前驱/后继关系、包含关系、引用关系有传递性，相关关系具有对称性和自反性。除此之外，知识点 ontology

间还延用了传统的实例关系(instance-of)、属性关系(attribution-of)。

依据上述分析,给出一个知识点本体的定义如下:

定义 8　一个知识点本体(knowledgy ontology, KO)可由一个七元组表示:

$$KO(name)=\langle id,name,define,function,content,includedKO,R_{KO}\rangle \quad (12)$$

其中:id、name、define、function、content、includedKO、R_{KO}分别是知识点本体 KO 的编号、名字、定义、功能、内容描述、KO 关系集。

按照定义 8,选择网络管理课程为例,针对崔北亮等人著的《网络管理—从入门到精通》第五章"配置常用服务器",以 DNS 服务器配置知识点为例,构建其对应的知识点本体,如图 5 所示。

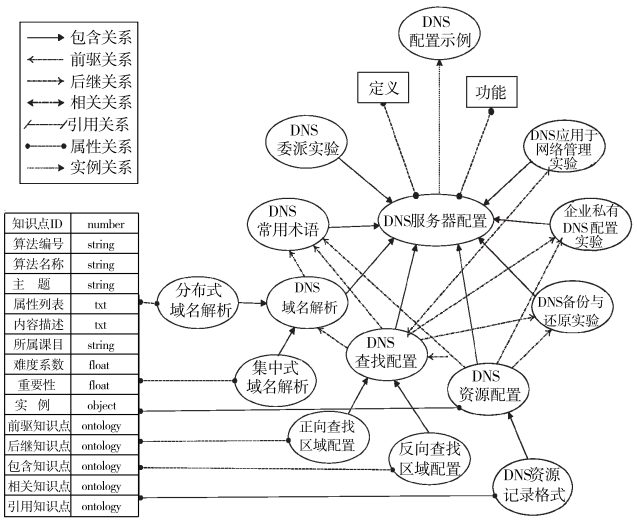


图 5　DNS 服务器配置知识点 ontology

若假设该章所含知识点本体已构建完好的基础上,图 6 示意了该章所对应的学习资源本体层框架。

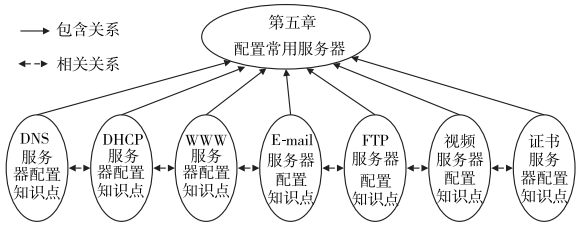
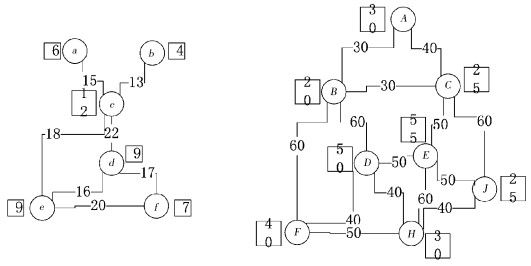


图 6　配置常用服务器的学习资源 ontology 模型

4.2　知识点本体匹配过程演示

为演示方便,对图 5 所示的知识点本体简化,抽象成如图 7(a)所示的有向图,作为目标本体 Q。其中节点上标注数字代表属性,有向弧上标注数字代表节点间的其他关系要求。给出一个待匹配本体 Q'如图 7(b)所示。



(a)目标本体 Q 的表示图　　(b)目标本体 B 对应的图

图 7　有向图

首先选定并匹配出一对锚点概念节点$\langle c,B\rangle$,如图 8 所示。

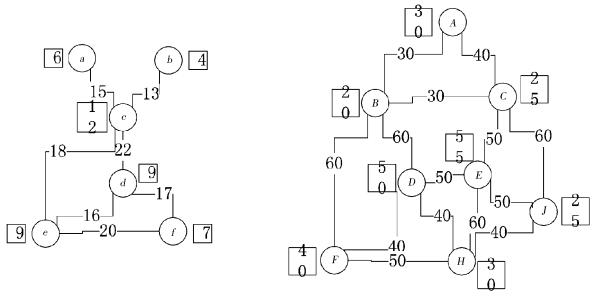


图 8　Q 与 Q'的第一对锚点概念匹配

由锚点概念节点开始,按顺序生成 Q 与 Q'的第一步子图,如图 9 所示,依此计算和判定生成子图的节点匹配、边匹配和结构匹配,实现 Q 与 Q'的第一步匹配。

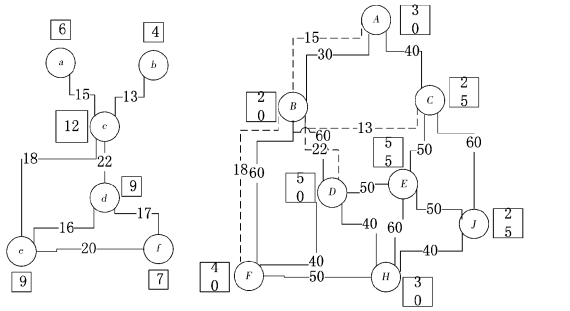


图 9　基于锚点$\langle c,B\rangle$的 Q 与 Q'生成子图匹配

在已匹配的本体子图中,匹配一对锚点概念节点,重复上述过程,直至实现本体 Q 的全部图匹配或无法再匹配,算法终止,如图 10 所示。

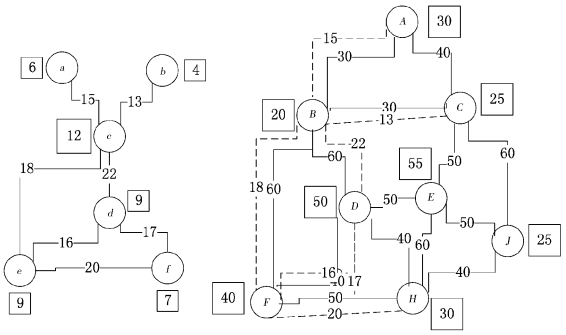


图 10　Q 与 Q'的完全匹配

上述匹配过程中,本体 Q 的表示图在本体 Q'的表示图中实现近似同构匹配,从而认为本体 Q'是可以匹配本体 Q 的。

4.3　算法时间复杂性分析

前文给出的算法伪码描述中,三层循环嵌套的子图近似同构运算量规模决定算法的时间复杂度。

设本体 Q 的表示图 $G(Q)$中 $|V(Q)|=n$、$|E(Q)|=m$,本体 Q'的表示图 $G(Q')$中 $|V(Q')|=N$、$|E(Q')|=M$,则为完成本体 Q'与 Q 的匹配,算法主要运算所需单位时间分别为:

a)第一对锚点节点的匹配所需时间量为 $n\times N$;

b)对于节点$\langle x,y\rangle$,边匹配所需时间量为 $E(x)\times E(y)$;

c)对于子图 $G(x)$与 $G(y)$的同构判定,所需时间量为

$$C^2_{|V(G(x))|}\times C^2_{|V(G(Y))|}\times E(x)\times E(y)$$

当节点、边的数目分别取为 n、N、m、M 时,算法主要操作运算所需时间量规模为

$$\frac{n(n-1)}{2}\times\frac{N(N-1)}{2}\times m\times M$$

因此,算法时间复杂度是 $O(n^6)$级的,是一个有效算法。

(下转第 434 页)

样就大大降低了系统的 I/O 开销,平衡了计算资源,提高了系统整体计算性能,从而缩短了迭代计算时间,提高了算法的效率。同时,由图 3 所示,当数据块不断增大时,NumPR 算法迭代执行时间上升的增幅也较大,由此可见,SGPR 算法在性能上有一定的提升。

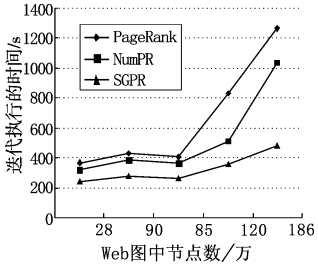


图 3 三种算法在不同数据集下迭代执行时间

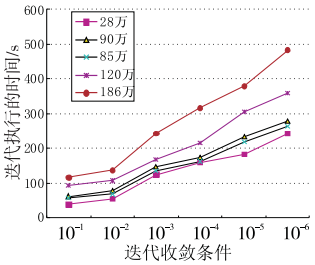


图 4 SGPR 算法在不同收敛条件下迭代执行时间

对比改进后 SGPR 算法在 10^{-1}、10^{-2}、10^{-3}、10^{-4}、10^{-5}、10^{-6}六种情况下执行迭代计算时间,以验证改进后的 SGPR 算法性能是否平稳;由图 4 所示,随着 Web 图中节点数的增大,迭代执行时间的增幅不大,而传统的 PageRank 算法在 Web 图中节点越来越大时,执行时间则急剧增大。由此可见,改进后的 SGPR 算法在迭代计算的执行时间上增长相对较为平稳,迭代收敛的速度较高。

4 结束语

本文主要研究在开源 Hadoop 平台下,结合 MapReduce 模型,实现 PageRank 算法并行化,提出了基于块结构划分的方法,将网页之间的链接关系转换成网络块间的关系,减少了 map 和 reduce 操作的调用次数,降低了 I/O 传输造成的开销,提高了计算的效率。实验证明,本文方法具有一定的优越性。下一步的工作是对 PageRank 算法过程中每次迭代加载不变数据,造成 CPU 资源浪费以及通信开销,以及对于每次迭代结果检测频繁调用 MapReduce 任务的问题继续研究,以期进一步提高 PageRank 算法的并行计算效率。

参考文献:

[1] PASQUINELLI M. Google's PageRank algorithm: a diagram of cognitive capitalism and the rentier of the common intellect[EB/OL]. [2009]. http://matteopasquinelli.com/docs/Pasquinelli_PageRank.pdf.

[2] RIDINGS C, SHISHIGIN M. PageRank uncovered[EB/OL]. [2009-06-18]. http://www.voelspriet2.nl/PageRank.pdf.

[3] 王俊生,施运梅,张仰森. 基于 Hadoop 的分布式搜索引擎关键技术[J]. 北京信息科技大学学报,2011,26(4):53-54.

[4] 李建江,崔健,王聃,等. MapReduce 并行编程模型研究综述[J]. 电子学报,2011,39(11):2635-2642.

[5] Apache MapReduce architecture[EB/OL]. [2012-05-28]. http://hadoop.apache.org./mapreduce/.

[6] JEFFREY D, SANJAY G. MapReduce: a flexible data processing tool [J]. Communications of the ACM, 2010, 53(1): 72-77.

[7] 梁秋实,吴一雷,封磊. 基于 MapReduce 的微博用户搜索排名算法[J]. 计算机应用,2012,32(11):2989-2993.

[8] 李远方,邓世昆,闻玉彪,等. MapReduce 下的 PageRank 矩阵分块算法[J]. 计算机技术与发展,2011,21(8):6-9.

[9] BU Y, HOWE B, BALAZINSKA M, *et al*. The HaLoop approach to large-scale iterative data analysis[J]. International Journal on Very Large Data Bases, 2012, 21(2): 169-190.

[10] WANG Yuan, DEWITT D J. Computing PageRank in a distributed Internet search system[C]//Proc of the 30th International Conference on Very large Data Bases. 2004: 420-431.

[11] BAHMANI B, KUMAR R, MAHDIAN M. PageRank on an evolving graph[C]//Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2012: 24-32.

[12] GAO Bin, LIU Tie-yan, WEI Wei, *et al*. Semi-supervised ranking on very large graphs with rich metadata[C]//Proc of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2011: 96-104.

[13] ZHONG Cai-ming, MIAO Duo-qian, WANG Rui-zhi. A graph-theoretical clustering method based on two rounds of minimum spanning trees [J]. Pattern Recognition, 2010, 43(3): 752-766.

[14] SANKARALINGAM K, YALAMANCHI M, SETHUMADHAVAN S, *et al*. PageRank computation and keyword search on distributed systems and P2P networks[J]. Journal of Grid Computing, 2003, 1(3): 291-307.

(上接第 421 页)

5 结束语

e-Learning 学习资源本体常常基于不同规范构建,在资源发现时难以通过语言学本体匹配方法实现,现有结构匹配方法又未能较好解决结构的相似性计算问题,因此本文提出一种基于图同构的本体匹配方法,在计算图结构整体相似性的基础上对本体表示有向图中的点、边进行交替匹配,从而以子图同构判定来实现本体匹配。该方法旨在发现高效的相似子图,提高本体匹配的精准性和效率。

参考文献:

[1] 美国教育部教育技术白皮书[K]. 上海:上海市教科院智力开发研究所,2001.

[2] STUDER R. Knowledge engineering: principles and methods[J]. Data & Knowledge Engineering, 1998, 25(1-2): 199-220.

[3] GANGEMI A, PISANELLI D M, STEVE G. An overview of the ONIONS project: applying ontologies to integration of medical terminologies[J]. Data & Knowledge Engineering, 1999, 31(2): 183-220.

[4] DOAN A H, MADHAVAN J, DOMINGOS P, *et al*. Learning to map between ontologies on the semantic Web[C]//Proc of the 11th International Conference on WWW. New Yrok: ACM Press, 2002: 662-673.

[5] JI Qiu, HAASES P, QI Gui-lin. Combination of similarity measures in ontology matching using the OWA operator[M]//Recent Developments in Orderad Weighted Averaging Operators: Theory and Practice. [S.l.]: Springer, 2011: 281-295.

[6] EUZENAT J, SHVAIKO P. Ontology matching[M]. Berlin: Springer-Verlag, 2007.

[7] COHEN W W, RAVIKUMAR P, FIENBERG S E. A comparison of string distance metrics for name-matching tasks[C]//Proc of IJCAI. 2003: 73-78.

[8] MELNIK S, GARCIA-MOLINA H, RAHM E. Similarity flooding: a versatile graph matching algorithm and its application to schema matching[C]//Proc of the 18th International Conference on Data Engineering. 2002: 117-128.

[9] MADHAVAN J, BERNSTEIN P A, RAHM E. Generic schema matching with CuPid[C]//Proc of the 27th International Conference on Very Large Data Bases. San Francisco: Morgan Kaufmann Publishers, 2001: 49-58.

[10] NOY N, MUSEN M. The PROMPT suite: interactive tools for ontology merging and mapping[J]. International Journal of Human-Computer Studies, 2003, 59(6): 983-1024.

[11] GIUNEHIGLIA F, SHVAIKO P, YATSKEVIEH M. S-Match: an algorithm and an implementation of semantic matching[C]//Proc of ESWC. 2004: 61-75.

[12] 唐杰,梁邦勇. 语义 Web 中的本体自动映射[J]. 计算机学报,2006,29(11):1957-1974.