

DOI: 10.7652/xjtuxb201509013

ABox 推理计算实体相似度

薛咏^{1,2}, 冯博琴¹, 武艳芳²

(1. 西安交通大学电子与信息工程学院, 710049, 西安; 2. 西南科技大学信息工程学院, 621010, 四川绵阳)

摘要: 针对主题图本体映射中主题上下文语义不易确定, 缺少分类信息, 但有明确的结构语义特征的特点, 提出了 A-Sim 主题图本体映射算法。首先对主题图中的主题进行分类, 利用描述逻辑语言 ALCIR₊ 的表达能力, 根据主题属性及关联关系建立各种能表达上下文语义信息的概念术语集, 将主题图本体转换为断言集合; 其次在断言集合上进行多项式复杂度实例检测, 记录模型构造过程中的个体, 获得主题的上下文语义; 最后将上下文语义结合元素级概念相似度算法计算实体综合相似度, 获得异构主题图实体间的实体映射关系。通过对主题图进行相似度计算实验, 结果证明新算法对主题图实体相似度计算综合性能至少提升了 14%。

关键词: 本体映射; 相似度计算; 描述逻辑; 实例检测

中图分类号: TP391 **文献标志码:** A **文章编号:** 0253-987X(2015)09-0070-07

Measuring Similarity of Entities Based on ABox Reasoning

XUE Yong^{1,2}, FENG Boqin¹, WU Yanfang²

(1. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China; 2. School of Information Engineering College, Southwest University of Science and Technology, Mianyang, Sichuan 621010, China)

Abstract: Aiming at the problems that the context semantics of concepts are hard to make sure in the ontology mapping of topic map, and that the topics are lack of classification information but have well-defined constructive semantic characters, an A-Sim algorithm for ontology mapping of topic maps is presented in this paper. At first, the topics in topic maps are classified using expression ability of description logic language ALCIR₊; then the concept terminology box expressing various context semantics of concepts is built according to the attributes and associations of topics, and the topic map ontologies are transformed into an assertions box. Second, a polynomial complexity algorithm for instance detection is presented and implemented on the assertions box, to obtain the context semantics of concepts by storing individuals in the process of constructing models. At last, the semantic similarity which combines the syntax-based and semantic similarity measurements is measured to calculate the synthetic similarity of entities, and then the mapping relationship between entities of heterogeneous topic maps is obtained. The experiments of similarity measurement of topic maps demonstrated that the novel method has achieved better performance and improved the comprehensive performance value at least 14% than other methods.

Keywords: ontology mapping; similarity measurement; description logic; instance detection

实体相似度计算是本体映射的基础,用以解决异构本体语义一致性与本体复用问题,在知识地图中,需要不断更新融合新数据,这些数据中存在术语异构等问题。主题图是一种应用广泛的知识地图结构,主题、关联(association)、资源出处(occurrences)组成的集合体(TAO)^[1]通过主题自身特征及与其他概念的关联关系确定概念语义。

主题图本体的融合算法,主要有文献[2]提出的SIM算法,属于元素级算法;文献[3]提出的TM-MAP算法,计算主题图中主题和主题间关系,将关系的计算结果以权重的形式综合到计算公式来判决概念的融合,考虑了数种固定的属性关系;文献[4]提出的TOM算法,与SIM算法类似,基于字符串统计,涉及到主题图三要素中的主题、关系、资源实体;文献[5]利用关键词聚类图与文本聚类图方法确定主题关系;文献[6]提出的扩展主题图融合ETM-SC算法,该算法对词典未收录词汇的相似度来自于两个方面,一是进行切词然后进行基于字符串的统计,二是来自于与之有直接关系的概念相似度;文献[7]对专用领域概念利用语料库技术形成专业领域词典,提高对领域本体元素级相似度计算性能;文献[8]提出的一种将语义进行命题逻辑描述的方法,对一般概念相似度计算,语义引入合取、析取与否定构子,表达能力差,利用WordNet词典进行概念可满足性判断。

针对主题图中数据量大、结构化良好的特点,本文提出一种混合式相似度A-Sim算法,利用ALCIR₊描述逻辑语言实现对概念上下文语义的灵活描述。根据主题图元素之间的关系定义概念术语集(TBox),将主题的结构语义描述为多种逻辑概念,将待匹配数据转化为断言集(ABox),设计了一种多项式复杂度的概念实例检测推理算法,每一种可满足的概念都是一个主题的结构语义,将结构语义与元素级语义相似度相结合,进行综合实体相似度计算。

1 背景知识与问题描述

标准主题为主题、资源两层结构,扩展主题图在标准知识地图-主题图上增加知识元层,对主题知识结构进行细化,增加知识地图导航能力,是一系列E-Learning知识获取服务国家级项目的知识组织结构,主题、知识元、资源的教育类资源知识组织方式已有很多实际应用^[6]。图1是两个异构扩展主题图片段,图中节点为主题图实体,边标注为关系类型,从图中可以看出主题图间“数据通信”与“服务”、

“面向无连接”与“无连接服务”、“IP地址”与“网络协议地址”都属于异构术语,是相似主题。此外,有同形异意异构等,相似度计算的目即对不同本体内容实体进行相似度计算,同义术语赋予高相似度。

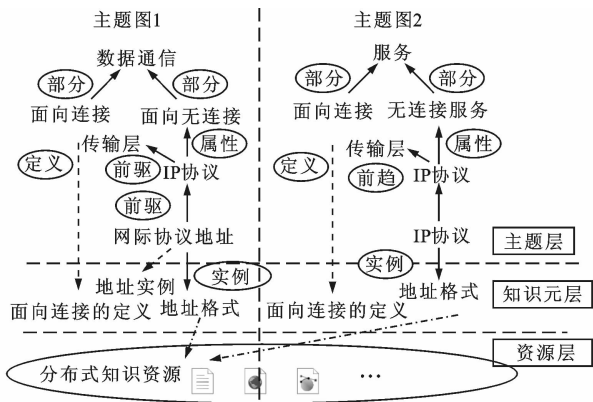


图1 异构的扩展主题图

本文中实体相似度指主题图中主题的相似度,可以定义为一个4元组 $\langle e_1, e_2, k, r \rangle$, e_1, e_2 分别为第一、第二映射元素实体, k 为两元素关系之间的相似度,一般取值范围为 $[0, 1]$, r 为两元素之间的关系,包括对称、相交、包含、不相交等。

2 主题图的描述逻辑表示方法

2.1 ALCIR₊描述逻辑

描述逻辑是语义网OWL语言的理论基础,描述逻辑语言建立的本体知识库可表示为 $KB = \langle TBox, ABox \rangle$, TBox是有关概念和角色的蕴含公理集合,声明概念或角色间的包含关系。ABox是有关个体的实例断言集合,断言包括声明个体是某概念的实例,以及个体之间的二元关系。描述逻辑语言包含逻辑构子越多表达能力越强,推理复杂度越高,在基本的AL语言上加入概念析取、完全存在量词、概念否定后为ALC语言,加入逆角色、允许角色传递后构成ALCIR₊逻辑描述语言。根据扩展主题图实体关系特点,以ALCIR₊作为TBox描述语言。

描述逻辑中Tbox主要包括蕴含公理与等价公理,公理的语义由建立者决定,如果解释 I 满足 $C' \sqsubset D'$ 则 $C \sqsubseteq D$,同理有 $C \equiv D$ 等。TBox中出现的概念分为:基本概念(primitive concept),只出现在公理的右边(RHS);定义概念(defined concept),出现在定义左边的概念(LHS)。

2.2 构建TBox

扩展主题图的基本概念有主题、知识元包含的

子类以及资源包含的子类,其中主题间关系有前续和后继、整体和部分、同义、类似、对比、不相关等,根据需要还可以定义并列、用途和环境、因果等任意关系。主题与知识元、资源的结合中,不同的结构有明确的自然语义信息,比如主题是知识点,在学习过程中有先后顺序,不考虑知识结构文件来源,与知识元关系构成学习单元,这些学习单元的结构是相对稳定的,因此以知识类别、知识单元进行结构语义匹配,比一般模式匹配更加准确,利用描述逻辑进行精确形式化结构描述是非常有意义的^[8-9]。

根据主题图实体结构关系进行逻辑语法描述,构建 TBox = { N_C, N_R }, 其中 N_C 表示概念集合, N_R 表示角色集合, 又分为一般角色、传递性角色 N_{R+} 。将扩展主题图直接定义出的类作为基本概念, 根据主题图知识导航特点给出主题图的定义概念, 部分 TBox 公理集表示为: ①前驱主题类 $\sqsubseteq$ 主题; ② $\forall$ 后继. 主题 $\sqsubseteq$ 前驱主题类; ③后继主题类 $\sqsubseteq$ 主题; ④ $\forall$ 前驱. 主题 $\sqsubseteq$ 后继主题类; ⑤结构知识 $\equiv \exists$ 知识元关系. $\exists$ 资源关系 $\sqsupset$; ⑥相近知识 $\equiv \exists$ 相同. 主题 $\sqsubseteq \exists$ 类似. 主题; ⑦有后继 $\sqsupset$ 有前驱; ⑧知识元 $\equiv$ 实例 $\sqsubseteq$ 定义 $\sqsubseteq$ 属性 $\sqsubseteq \dots \sqsubseteq$ 功能; ⑨资源出处 $\equiv$ 目标指示符 $\sqsubseteq$ 资源 $\sqsubseteq$ 数据类资源。

主题为基本概念, 前驱主题类为原子定义概念, $\forall$ 后继. 主题为复杂定义概念。①~⑥为对主题结构概念的定义, 同理可定义出整体概念类与部分概念类公理, 区别知识等; ⑧~⑨为知识元与资源出处概念, 其中⑦为角色公理, 表示两角色为互逆角色, 定义 $N_{R+} = \{\text{有后继, 有前驱, 整体, 部分}\}$ 。实体基本类概念、知识元结构概念以及角色关系与性质用类似方法定义。通过定义的 TBox 可以看出, 图 1 中的“IP 协议”属于前驱主题类、后继主题类、属性类的交集, 而“IP 地址”属于知识单元。利用公用推理机可验证构建的 TBox 是非循环一致的 ALCIR₊ 描述公理集合。

2.3 构建断言

Abox 包含两类断言, $C(a)$ 形式的概念断言, 以及 $R(a, b)$ 形式的角色断言, 扩展主题图格式文件 XTM (基于 XML 语法的 TM 文件) 以标准主题图语法为基础, 从数据库中抽取个体声明, 个体之间的关系声明, 形成 XTM 文件。根据 XTM 文件建立 ABox 的过程, 即为抽取 TBox 定义的相关概念实例与角色实例的过程, 利用 XStream 技术实现 XTM 与 Java 对象之间的转换, 去除与结构信息无关的范围

约束等元素。将 XTM 的个体与角色信息读入定义的数据结构中, 形成需要的 ABox, 因为大规模主题图本体标注困难, 只能给出个体的基本概念类断言。

3 结构语义信息计算

3.1 可满足性推理

ABox 的推理都可以转换为一致性推理, 目前有两种方法, 一种将 ABox 展开为预完整形式^[10], 将 ABox 中的断言根据 TBox 不断扩充, 得到一系列等价的 ABox 集合, 如果没有出现空概念或者冲突, 表明 ABox 一致, 算法复杂度为 4 倍指数级。另一种将 ABox 推理规约到 TBox 推理中^[11]。

定义 1 一个 ABox 相对于 TBox 是可满足的, 是指存在一个模型 $I = \{\Delta^I, \cdot^I\}$ 满足 TBox, 并且对于 ABox 有: 如果 $C(a) \in \text{ABox}$, 那么 $a^I \in C^I$; 如果 $(a, b) \in \text{ABox}$, 那么 $(a^I, b^I) \in R^I$; 如果 $a = b \in \text{ABox}$, 那么 $a^I = b^I$; 如果 $a \neq b \in \text{ABox}$, 那么 $a^I \neq b^I$ 。

在定义 TBox 的各种结构语义概念后, 需要进行可满足性推理, 才能得到个体与语义概念的关系。对于知识库 KB, 假设 $\text{TBox} = \{A \sqsubseteq C, B \sqsubseteq C, \exists R \sqsubseteq B, \exists R^- \sqsubseteq D, R_+\}$, 对应的 ABox 为 $A_1 = \{A(a)\}$, $A_2 = \{A(a), B(a), C(a), R(a, b)\}$, 如果对这两个断言集进行推理, 都是一致的, 但是根据 A_2 的预完整表可以得出 a 也是 D 的一个实例, 对 $\tilde{A}_1 = A_1 \cup \{\neg D(a)\}$, $\tilde{A}_2 = A_2 \cup \{\neg D(a)\}$ 进行可满足性推理, 可以得到 $\tilde{A}_1$ 是一致的, 而 $\tilde{A}_2$ 是不一致的。也就是说对于不直接出现在 ABox 中的概念, 只有通过明确的概念可满足性推理, 对 ABox 进行实例检测, 才能得出实例与概念的关系, 比如 TBox 中的定义概念 $\exists R, \exists R^-$, 因此要获得概念语义信息必须对个体进行关于概念的实例检测。

3.2 结构语义计算

定义 2 同知识单元个体集合。假设个体 a 属于知识单元 C , 以 a 为根节点, 构造概念 C 完备语义树过程中经过个体的集合。

对于个体主题 a , 同知识单元个体集合就包含了 a 关于 C 的结构语义信息, 而所有知识单元概念都是定义概念, 所以结构信息计算中只对定义概念进行实例检测。为了记录同知识单元集合, 设计一个储存结构语义的数据结构

$$M_{\text{sem}} = \{C_1\{a_1, a_2, \dots, a_{m_1}\}, \dots, C_i\{b_1, b_2, \dots, b_{m_i}\}, \dots, C_n\{c_1, c_2, \dots, c_{m_n}\}\} \\ n, m_1, \dots, m_i, \dots, m_n \in \mathbb{N}$$

其中 $C_i\{b_1, b_2, \dots, b_{m_i}\}$ 表示个体 a 所属概念 C_i 中包含的所有个体。每个个体 a 都有一个对应的 M_{sem} 集合,记录 ABox 中实例所属概念语义信息。

设计实例检测函数 instanceDetection 返回实例检测 $a; C_n$ 关于 TopicMap = <Tbox, Abox> 的结果,同时生成一个以 L 为对象的列表,存储 $C_i\{b_1, b_2, \dots, b_{m_i}\}$,记录实例模型构建过程中经过的节点,当实例检测结果为真时,将该列表与其对应的概念加入 M_{sem} 集合。

本文中的 TBox 是一个可展开的 ALCIR₊ 逻辑系统,因此首先需要将知识库 TopicMap 中的 TBox 展开并保持语义不变,展开后的术语集右边只包含基础概念,左边为定义概念,这样定义概念可以内化为一个标准形式

$$C = D \sqcap C \sqcap \forall U. C \quad (1)$$

式中: D 指基本概念的组合与吸取; U 为所有角色的超角色,这样所有的实例检测都可以表示为对迭代表达式的概念可满足性检测。经典的概念可满足性检测采用对个体断言集合加入否定概念断言,再利用 Tableaux 算法不断对个体断言集进行扩展,如果扩展树出现矛盾,则表明概念是可满足的。基于语义的实例检测与经典的方法有如下两点不同。

(1)经典的实例算法通过加入否定概念,运用 tableaux 算法检测 ABox 关于 TBox 是否一致,会因为存在量词和数量限定而产生新节点,随着 ABox 扩展可能会在新节点中发现冲突,从而判断不一致性。比如对于 $\text{ABox} = \{a; \text{主题}, a; \exists R. C\}$,一致性检测会产生一个 R 后继节点 $x = \{C\}$,若有 a 的邻居再次扩展 x 节点,冲突是有可能的,当没有其他节点对 a 节点进行扩展时它是一致的,但是增加节点会改变主题图原有结构,而在主题图中这样的语义结构是不存在的。

(2)现有的实例检测算法,只给出是否满足的结果,而不能记录同知识单元个体集合。

针对上述问题,本文提出一种不加入否定概念,而直接在 ABox 上构建 Tableau 表的方法,寻找概念客观存在的语义模型,同时记录同知识单元个体集合,该方法查找可满足概念集合的一个子集,不包含需要进行增加新节点的可满足性概念。在构建前需要进行预处理:①将 TBox 中的概念表达式通过 De Morgan 定律转化为否定范式 NNF;②对所有包含断言转化为析取式,比如 $C \sqsubseteq D$ 转化为 $\neg C \sqcup D$,与要检测概念文字做合取,作为概念实例检测的

参数。以下给出 instanceDetection 函数的算法伪代码,将其简写为 iD。

Input: $a, C_n, \text{TBox}, \text{ABox}; C_n = D \sqcap C \sqcap \forall U. C$

Output: L

begin

```

1. 初始化  $L$  包含  $C_n$  的全局变量
2. for all subconcept  $C_i$  in  $C_n$ 
3.   if  $a; C_i$ 
4.      $L.add(a)$ ; return true;
5.   else
6.     switch( $C_i$ )
7.       case  $D_1 \sqcap D_2$  if  $iD(a; D_1)$  and  $iD(a; D_2)$ 
8.         return true;
9.       else return false;
10.      case  $D_1 \sqcup D_2$  if  $iD(a; D_1)$  or  $iD(a; D_2)$ 
11.        return true;
12.      else return false;
13.      case  $\exists R. D$  if exist R-successor  $b_i$ 
14.        if  $iD(b_i; D)$ ;
15.        return true
16.      else return false;
17.      case  $\forall R. D$  for all R-successor of  $b_i$ 
18.        if all  $iD(b_i; D)$ 
19.          return true;
20.        else return false;
21.      case  $\forall R_+. D$  for all R-successor of  $b_i$ 
22.        if all  $iD(b_i; D)$ 
23.          if  $b_i$  has R-successor  $c_i$ 
24.            if all  $iD(c_i; D)$ 
25.              return true;
26.            else return false;
27.          else return true;
28.        else return false;
29.      default return false;
30. end for
end

```

如果 instanceDetection 返回为 false 值,将 L 中的 C_i 删除,这样 L 只包含 a 可满足概念。算法补充说明:①对于所有的涉及 R^- 角色,如果存在 $R^- . D$ 则将 R^- 看作一般角色展开,否则检测是否有 R 角色的前驱节点 b 存在,则对 $b; D$ 进行实例检测,关于前驱、后继、邻居的概念可参考文献[10];② $\forall R_+$ 可能引起循环检测,所以需要对其经过的个体节点

进行记录并检测,进行阻塞,使算法能够中止,也就是进入第一次迭代后,以下所有迭代都不能再对根节点进行实例检测,这说明存在一个传递关系组成的环路径;③所有需要对邻居进行实例检测时,如果不存在满足要求的邻居节点,都返回 false 值;④因为析取与存在规则,个体对概念的可满足模型可能有多个,所以需要通过对加标记的方法,对所有分支都进行检测。

因为本算法并不产生新的断言节点,而且除过逆角色外所有以个体为根节点向后继角色单向扩展的,只有传递性角色可能产生循环检测,通过阻塞后,算法中的5条规则必然会在最多遍历整个主题图节点后中止。

证明 一个断言 $a:C$ 关于 TBox 是可满足的,根据定义1,即证明存在一个解释 $a^I \in C^I$,设根据 instanceDetection 函数中的5个规则遍历后树的节点集合为 $\tilde{A}$,对于任意个体 $a \in \Delta^I$,如果 $a:C \in \tilde{A}$,都有 $a^I \in C^I$,这里只证明最复杂的 $\forall R_+$ 关系,其他规则方法类似。

假设 $C \equiv \forall R.D$, $R \in R_+$,这里 D 是一个形如式(1)的复杂概念,实例检测为真,即有 $a:C \in \tilde{A}$,则个体 a 至少有一个 R 后继,同时如果:① D 为基本概念,则有 R 后继集合 $\{b_1, b_2, \dots, b_n\} \in D$,且 $b_1, b_2, \dots, b_n$ 都没有后继,由 $\langle a^I, b_1^I \rangle, \langle a^I, b_2^I \rangle \dots \langle a^I, b_n^I \rangle \in R^I$ 可以看出所有 R^I 值域的都属于 D^I ,定义域都属于 C^I ,即 $a^I \in C^I$;② D 为基本概念,假设 b_i 有后继 c_i ,因为 R 有传递性,则必有 $\langle a^I, c_i^I \rangle \in R^I$,也就是要进一步检测实例 $c_i:D$,因为包含检测为真,所以有 $a^I \in C^I$;③ D 为一般概念,则概念 D 的文字必然为5种规则的一种,如果适用于 $\forall R_+$ 规则则进入①。

证明完毕。

3.3 相似度计算

设待匹配扩展主题图中的主题集合分别为 $E_1 = \{e_1, e_2, \dots, e_n\}$ 、 $E_2 = \{e_1, e_2, \dots, e_m\}$,任意两个主题之间的相似度计算由语法相似度 S_G 、语义相似度 S_S 及结构相似度 S_T 三部分组成,通过实例检测得到个体的 M_{sem} 集合包含完整的上下文结构信息。

对于语法、语义相似度的计算方法参见文献[7],其中语义相似度使用 WordNet 词典、HowNet 词典、同义词词林进行相似度计算,针对字典中没有的专业领域主题相似度计算采用了基于语料库的方法,获得了较好的实验结果,因此本文只阐述 S_T 的计算算法。设两个待比较主题对为 (e_i, e_j) ,对应的

题可满足的概念集合分别为: $M_{semi} = \{C_1, C_2, \dots, C_i, \dots, C_n\}$, $M_{semj} = \{C_1, C_2, \dots, C_j, \dots, C_m\}$,其中 C_i, C_j 为概念包含个体的集合,设 M_{semi} 与 M_{semj} 中概念名的交集为 G ,并集为 E ,对所有概念名相同的集合做 Jaccard 相似度计算,得到 $|G|$ 个集合相似度

$$C_{sim}(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \quad (2)$$

式(2)中以词法语法相似度作为集合交集运算的基础。设置阈值 t_1 ,若 $C_{sim} > t_1$ 的概念名个数为 p ,则主题间的相似度可表示为 $S_T(e_i, e_j) = (|G| - p) / |E|$,在相似性判断中通过实验设置阈值 t_2 ,如果 $S_T > t_2$,则认为两主题相似。

3.4 算法复杂度分析

设 TBox 左侧概念数为 n ,根据展开后的 TBox 计算 LHS 概念间包含关系,共需要进行 n^2 次计算,本部分复杂度为平方复杂度。

设 ABox 中包含的个体数为 n ,需要检测的概念数为 m ,每个概念的子概念数最多为 p ,因为算法 $\forall R_+$ D 规则复杂度最高,传递性检测最多可以设计 $n-1$ 个节点,假设子概念约束都是全称量词,角色深度为 q ,则每个子概念需要检测的次数为 $(n-1)^q$,则算法复杂度为 $O(mpn^{q+1})$ 。

4 实验与结果分析

相似性度量算法的评价,使用信息检索领域通用的查准率(Precise)、查全率(Recall)和 F 值(F -measure) 3 个评价标准。

扩展主题图 TBox 共 15 个基本概念,18 个定义概念,23 个角色,其中传递角色 4 个,逆角色 2 个。ABox 测试数据来源于“863 课题”的计算机领域扩展主题图。将 3 组不同标注来源的《计算机网络》扩展主题图分别进行相似度计算,其中节点最大连通图分别为 756、452、563,融合时删除了所有孤立节点。应用开源推理工具 pellet 调试一致并分类后,解析成专用 TBox 数据结构。实验平台 Inter i5 4 200MB CPU,内存 4 GB,Java 虚拟机内存 1 GB,3 组比较最大用时 82 s。

阈值 t_1, t_2 使用实验的方法确定,以计算结果的综合性能为依据,假定两阈值对结果的影响是相互独立的,采用经验给定 t_1 值后,取使得 F 值最高的 t_2 为最终阈值,再对 t_1 进行调整。图 2 为 $t_1 = 0.4$ 时不同 t_2 值对各项性能指标的影响,在进行集合相似度计算时,仅使用了字符串比对的方法来提高算法时间性能。准确性由计算结果与人工标注结果进

行对比得出,除去完全相同的主题,人工标注给出了 50 对相似度最高的主题,根据实验结果取 $t_1=0.4$ 、 $t_2=0.6$ 作为最终阈值。

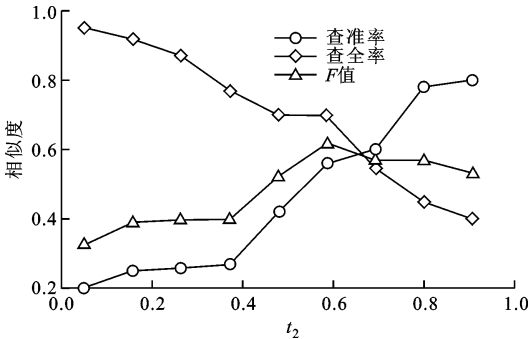


图 2 不同阈值 t_2 与算法性能关系

在阈值确定过程中,为了提高速度与结果检查效率,未计入语义相似度 S_s 。以获得性能最好时的阈值为目的,图 3 给出了结构语义相似度对匹配结果的影响,在语法相似度计算的基础上,加入 S_T 的计算结果,查全率有 11% 的提高,同时对异构术语相似度的准确性也得到了提高, F 值提高了 6%。

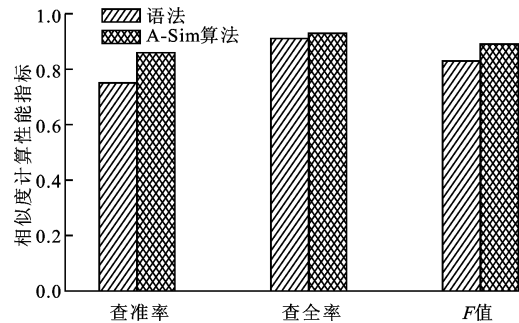


图 3 语法相似度与总相似度的比较

经过语法相似度筛选以后,共得到 117 936 个对比结果,其中 90% 以上的值为 0,其他值主要分布在 0.3~0.7 之间,实验结果显示查全率随着阈值 t_2 的升高下降很快。这是由于一些关系复杂的相似主题,在另一个主题图中却关系简单,导致相似度下降,主要中心节点比如“协议”、“网络协议”关联的实体可能很多。一些相似度很高的主题,可能完全不相似,比如“子网掩码”与“网络网关”都是“TCP/IP”参数的后继,但是又都没有其他个体关系。这种情况比较多,所以设置策略将大于阈值,但只有一个邻居的个体对相似度设为 0。计算机网络主题图中的主题以基本概念为主,产生异构的来源主要是人工标注过程中的随意性,另外一部分来自网络抽取的数据,但是这些基本概念一般比较固定,一个异构实体的关联实体都异构的情况很少,而且主题关系的

关联以前后序、整体部分关系较多,很适合本文算法的特点,但是同样存在一些不能识别的实体,比如“网关”与“协议转换器”,一个关联着“网络层设备”,一个关联着“网络层”,这类异构主题降低了相似度算法的性能,需要依赖更好的领域知识词典配合。

将 A-Sim 算法与 SIM、TM-MAP、ETMSC 算法进行了比较,结果如表 1 所示,因为 SIM 算法的语法相似度计算来自于通用本体词典,因此性能较差。TM-MAP、ETMSC 的相似度计算包括了主题上下文语义,因此对查全率有较大的提升作用,但是以相邻实体作为上下文语义不够精确。本文算法在各方面都体现出较好的性能结果,综合相似最少提高了 14%。

表 1 当前主题体相似度算法性能的比较

算法	查准率	查全率	F 值
SIM	0.88	0.3	0.45
TM-MAP	0.89	0.55	0.68
ETMSC	0.82	0.69	0.75
A-Sim	0.93	0.86	0.89

表 2 为不同算法对部分异构主题相似度判断结果的描述,括号内为 1 表示认为相似,0 表示认为不相似。因为本文算法的概念是以相对固定的知识结构定义的,因此对于关联节点不但涉及到相邻实体,还涉及到多级相连实体,尤其对前后续、整体部分中的异构术语能计算出较准确的相似度。例如图 1 中的“网际协议地址”与“IP 地址”的结构相似度可以追溯到“传输层”主题,向下可以达到资源层,对于主题对“数据通信”和“服务”,因为 A-Sim 可以将两个 part-of 关系主体合并为一个“部分概念”,从而减小了主题关系数量不同产生的干扰。在最坏情况下,

表 2 部分异构相似主题的匹配

主题对	SIM	TM-MAP	ETMSC	A-Sim
软件与计算机软件(1)	×	✓	✓	✓
广域网与局域网(0)	×	✓	✓	✓
信息量与信息(0)	×	×	✓	✓
vlan 与虚拟网(1)	×	×	×	✓
点到点与 Point to point(1)	×	×	✓	✓
网关与默认网关(0)	×	✓	×	✓
IP 地址与网际协议地址(1)	×	✓	✓	✓
数据通信与服务(1)	×	×	×	✓

注:✓表示正确

算法将退化为一般结构相似度算法,比如 ETMSC 的关联结构算法,即主题不属于 TBox 中定义的复杂结构概念类。

实验表明,A-Sim 算法对本体中主题上下文环境的确定更加全面、准确,提升了实体相似度计算的查全率与查准率。实验同时显示,进一步研究实体结构语义与语法语义相似度有效结合的方法,比如自适应的参数调整,有利于提高对同形异异构实体的相似度计算性能。

参考文献:

- [1] GARRIDO A, ILARRI S. TMR: a semantic recommender system using topic maps on the items' descriptions [C]// The Semantic Web: ESWC 2014 Satellite Events. Berlin, Germany: Springer, 2014: 213-217.
- [2] MALCHER L, WITSCHER H F. Merging of distributed topic maps based on the subject identity measure (SIM) approach [M]. Leipzig, Germany: LIT, 2004: 1-11.
- [3] KIM J M, SHIN H, KIM H J. Schema and constraints-based matching and merging of topic maps [J]. Information Processing and Management, 2007, 43(4): 930-945.
- [4] 吴笑凡,周良,张磊,等. 分布式主题地图合并中的 TOM 算法 [J]. 武汉大学学报: 工学版, 2006, 39(5): 131-136.
WU Xiaofan, ZHOU Liang, ZHANG Lei, et al. TOM algorithm in distributed topic maps merging [J]. Journal of Wuhan University: Engineering Edition, 2006, 39(5): 131-136.
- [5] DING Y, FU X. The research of text mining based on self-organizing maps [J]. Procedia Engineering, 2012,

29(4): 537-541.

- [6] 鲁慧民,冯博琴,李旭. 面向多源知识融合的扩展主题图相似性算法 [J]. 西安交通大学学报, 2010, 44(2): 20-25.
LU Huimin, FENG Boqin, LI Xu. Novel similarity algorithm of extended topic maps for multi-resource knowledge fusion [J]. Journal of Xi'an Jiaotong University, 2010, 44(2): 20-25.
- [7] 薛咏,冯博琴,刘伟涛. 扩展主题图本体融合策略与算法 [J]. 西安交通大学学报, 2011, 45(10): 13-18.
XUE Yong, FENG Boqin, LIU Weitao. Strategy and algorithm for merging ontologies of extend topic maps [J]. Journal of Xi'an Jiaotong University, 2011, 45(10): 13-18.
- [8] GIUNCHIGLIA F, SHVAIKO P, YATSKEVICH M. S-Match: an algorithm and an implementation of semantic matching [C]// ESWS. Berlin, Germany: Springer, 2004: 61-75.
- [9] CHIU D Y, PAN Y C. Topic knowledge map and knowledge structure constructions with genetic algorithm, information retrieval and multi-dimension scaling method [J]. Knowledge-Based Systems, 2014, 67(9): 412-428.
- [10] FOKOUE A, KERSHENBAUM A, MA L, et al. The summary abox: cutting ontologies down to size [M]. Berlin, germany: Springer, 2006: 343-356.
- [11] HAARSLEV V, MÖLLER R. Expressive Abox reasoning with number restrictions, role hierarchies, and transitively closed roles [C]// International Conference on Principles of Knowledge Representation and Reasoning. Hamburg, Germany: Universität Hamburg, 2000: 273-284.

(编辑 赵炜)

(上接第 57 页)

- ASME Turbo Expo 2014. New York, USA: ASME, 2014: 25408.
- [6] GUO T, WANG T, GADDIS J L. Mist/steam cooling in a 180-degree tube bend [J]. ASME Journal of Heat Transfer, 2000, 122(4): 749-756.
- [7] DHANASEKARAN T S, WANG T. Numerical model validation and prediction of mist/steam cooling in a 180-degree bend tube [J]. International Journal of Heat and Mass Transfer, 2012, 55(13): 3818-3828.
- [8] 姜澎,黄洪雁,冯国泰,等. 水雾/蒸汽相变冲击冷却的数值模拟 [J]. 哈尔滨工程大学学报, 2009, 30(10): 1097-1101.
JIANG Peng, HUANG Hongyan, FENG Guotai, et

al. Numerically simulating mist/steam impingement cooling with phase change [J]. Journal of Harbin Engineering University, 2009, 30(10): 1097-1101.

- [9] FU W L, WRIGHT L M, HAN J C. Rotational buoyancy effects on heat transfer in five different aspect-ratio rectangular channels with smooth walls and 45degree ribbed walls [J]. ASME Journal of Heat Transfer, 2006, 128(11): 1130-1141.
- [10] DHANASEKARAN T S, WANG T. Validation of mist/steam cooling CFD model in a horizontal tube [C]// ASME 2008 Heat Transfer Summer Conference. New York, USA: ASME, 2008: 611-624.

(编辑 苗凌)