第45卷 第10期 西 安 交 通 大 学 学 报 Vol. 45 No. 10
2011年10月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Oct. 2011

扩展主题图本体融合策略与算法

薛咏[1,2]，冯博琴[1]，刘卫涛[1]
(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：针对分布式建立与存储的领域本体主题图在融合过程中的语义与结构重复问题以及冗余信息的判断与消除问题，提出了基于语义词典与语料库相结合的主题图融合算法(TMMC)，给出了概念相似度计算以及同义关系、整体部分关系等的处理方法. 对本体中概念进行基于 HowNet 语义词典及其他语义词典的多层次相似度计算，定义概念间不同语义关系的融合规则，针对专业领域本体中大量术语词典未收录的问题，提出基于语料库的概念相似度算法，并对计算机教育专业领域扩展主题图进行了融合实验. 实验结果表明，TMMC 提高了融合的准确率与查全率.

关键词：本体融合；相似度计算；语料库
中图分类号：TP391 **文献标志码**：A **文章编号**：0253-987X(2011)10-0013-06

Strategy and Algorithm for Merging Ontologies of Extend Topic Maps

XUE Yong[1,2], FENG Boqin[1], LIU Weitao[1]
(1. School of Electronics and Information Engineering, Xi′an Jiaotong University, Xi′an 710049, China;
2. Information Engineering College, Southwest University of Science and Technology, Mianyang, Sichuan 621010, China)

Abstract: Aiming at the distributed design and storage of topic map of domain ontology in abundance latent repetition, and redundancy in meanings and structures during merging process, a topic map merging algorithm(TMMC) based on common knowledge thesauri combined with corpus, is proposed, where the similarity of Chinese concepts in merging process is evaluated, then the hyponyms, part-of and other relations between concepts, are dealt with, and HowNet and the other thesauri are used to compose a multi-level similarity calculation during merging the concepts in ontology. The merging rules about the different semantic relationships among concepts are defined. For very large number of domain special terms are not included in the common knowledge thesauri, therefore a novel algorithm based corpus is presented for calculating similarities between domain special concepts. And the extended topic maps in computer specific education are tested to verify the improved *F*-measure value of TMMC.

Keywords: ontology merging; similarity calculating; corpus

本体奠定了网络中语义共享的基础. Neches 等将本体定义为：给出构成相关领域词汇的基本术语和关系，以及利用这些术语和关系构成的规定这些词汇外的规则. 本体具有分布式建立与存储的特点，当上层数据抽取与推理建立在多个本体上，就需要一种方法或机制来实现分布式本体之间的集成和融合. 其中，本体融合是计算机知识自动积累与增加的一种技术，用于整合领域内不同主体之间的概念与数据，并融合数据结构，考虑在新本体文件交互过程中，同义概念、同字异义概念、结构关系合并、冗余与冲突信息的检测处理等问题[1].

收稿日期：2011-01-26. 作者简介：薛咏(1978－)，男，博士生；冯博琴(联系人)，男，教授，博士生导师. 基金项目：国家高技术研究发展计划资助项目(2008AA01Z131).
网络出版时间：2011-07-29 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20110729.1609.002.html

1 扩展主题图本体融合

主题图在信息资源上层构建了一个结构化的语义网,是一个由主题、关联以及资源实体组成的集合体.主题图与资源描述框架都是基于可扩展标记语言的语义模型,是由国际标准组织标准化的元数据描述方法,一个领域主题图生成可以借用本体的多种应用,包括推理机.本体用某种语言生成,主要用来进行知识表示,而主题图是人工生成,用来组织互联网上的资源,可以把知识呈现与实现留给主题图,将概念化的指定工作交给本体[2].

国家高技术研究发展计划课题"面向领域的海量知识资源组织、管理与服务系统",突破了现有以资源、索引和元数据目录3要素为核心的资源管理和服务模式,建立了以扩展主题图为线索并综合考虑资源信息和用户兴趣的知识资源组织与管理新模式.扩展主题图在传统主题图概念关联资源模式中加入了知识元层,从而为用户提供多粒度、多层次知识服务体系.扩展主题图融合是将分布式环境下同一领域或不同领域的局部扩展主题图合并为全局扩展主题图,保证以统一的模式查询和搜索异构数据源中的信息,实现信息资源的组织、管理和共享.

扩展主题图本体融合算法可以描述为一个部分函数 $\psi:\chi\times\chi\times\varepsilon\rightarrow\chi$,其中 χ 为主题图文件,ε 为附加在融合过程中的环境约束.一般情况下环境约束不会对融合过程施加影响,这时融合定义为 $\psi:\chi\times\chi\rightarrow\chi$,即对于主题图 A 与 B,产生输出主题图 C,使得输入主题图与输出主题图语义一致.在此基础上主题图知识模型的本体知识融合可以分为两大步骤:①主题图元素相似度计算;②根据元素语义关系融合相应元素产生新主题图[2].

在主题图本体融合过程中,主要考虑元素间相似度与映射关系.根据文献[3],这种映射定义为一个5元组$\langle I_{id},e,e',n,R\rangle$,其中 I_{id} 为映射元素唯一标识符,e 和 e' 分别为第一、第二映射元素实体,n 为两元素关系之间的可信度,一般取值范围为[0,1],R 为两元素之间的关系(等于,包含,相交,不相交).本体融合需要首先对元素之间的语义距离也就是可信度 n 进行计算,并根据关系 R 决定最终生成主题图中元素间的逻辑关系.

2 研究现状

本体融合是模式/本体匹配技术关于信息整合方面的重要应用.现有的本体匹配算法可以分为元素级和结构级.元素级的本体匹配算法主要包括基于字符串、基于语言、基于约束、基于语言学(主要为公共领域词典,比如WordNet)、对齐、复用、上层形式化本体利用.结构级的本体匹配算法包括基于图技术、基于分类、基于结构知识库以及基于模型的方法.现有大部分算法对本体匹配进行综合考虑,不断丰富本体匹配算法库,同样面对本体概念相似度的问题,主题图融合需要进一步考虑主题图数据结构合并问题[3].

针对主题图融合,文献[4]提出了SIM算法.SIM算法是基于字符串的算法,对主题图中的主题基名进行两两间的字符串统计比较,判断两主题概念是否应该融合.文献[5]提出了TM-MAP算法,使用到主题图中的主题和主题间关系,将关系的计算结果以权重的形式综合到计算公式中,来判决概念的融合.这种算法在一定程度上涉及到主题图本身所蕴含的语义信息,但核心思想还是基于字符统计.文献[6]提出了基于主题和资源合并的TOM算法,基于字符统计,涉及到主题图3要素中的主题、关系、资源实体.文献[1]提出自动本体融合的知识积累方法,主要解决不同本体概念融合时的语义关系重写与添加问题,以及本体不一致性消除,并开发出原型系统.文献[7]提出了基于全信息的概念相似度判断与融合算法,在概念语法与语义相似度比较的基础上,加入了语义用相似度比较算法,考虑概念对中,与各概念有直接关系的概念之间的综合相似度,但没有考虑专业领域术语词典未登录的问题.

目前,主题图融合算法存在以下问题:①如何深入计算概念间的语义距离;②在融合过程中仅考虑到了概念同义关系,未考虑更复杂的关系,从而导致主题图融合时出现大量冲突;③专业领域主题图概念语义相似度计算.

3 主题图融合算法(TMMC)

针对目前主题图融合问题,提出分布式主题图融合过程中语法与语义相结合的多层次相似度计算与融合算法,进一步考虑元素间的相互关系,只有充分考虑待融合概念与主题图之间的相互关系才能保证最终主题图的语义一致性.针对专业本体中大量词汇公共领域词典未收录的问题,提出基于统计与分类结构相结合的相似度算法.算法框架见图1.

3.1 基于HowNet与同义词词林的相似度计算

目前,英文本体融合基于语言学的方法以WordNet为主,WordNet加入了丰富的词汇关系解

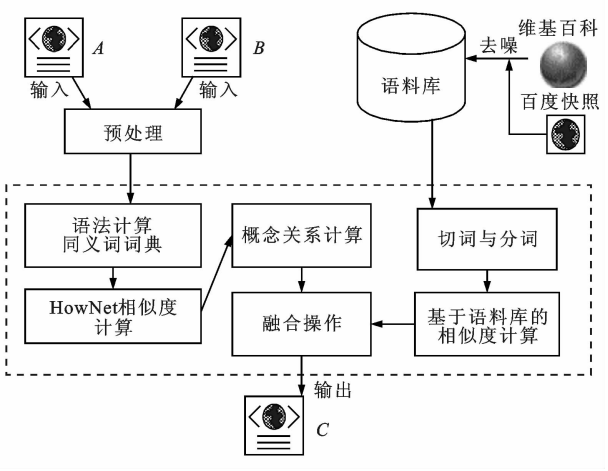

图 1　TMMC 框架

释. 中文语义词典以 HowNet 为基础，词典中每个词语可能有多个含义，每个含义对应 HowNet 中的一个义项，每个义项又由若干义元构成. 与 WordNet 类似，义元间存在上下位关系、同义关系、反义关系、整体-部分关系等. 现有算法无论是基于 WordNet 还是 HowNet，或者将 WordNet 翻译为中文 WordNet，对于词汇的相似度判断主要分为两类：一类是基于两个节点之间的路径长度，以刘群的算法[8]为代表；另一类是文献[9]提出的基于两个节点所含的共有信息计算相似度.

3.1.1　基于 HowNet 的词汇语义相似度计算　在中文词汇相似度计算方法中，刘群等[8]提出的基于 HowNet 的词汇相似度算法最具代表性：假设 W_1 有 n 个义元 $C_{11},C_{12},\cdots,C_{1n}$，$W_2$ 有 m 个义元 $C_{21},C_{22},\cdots,C_{2m}$，$W_1$ 和 W_2 的语义相似性是所有义元间相似性的最大值，可表示为

$$S(W_1,W_2)=\max_{i=1\cdots n,j=1\cdots m}S(C_{1i},C_{2j}) \tag{1}$$

义元依据上下位关系处在一个树形结构中，因此将义元在树中的语义距离转化为它们的语义相似性. 假设义元 C_1 和 C_2 在树中的距离为 $l(C_1,C_2)$，则两个义元的相似性

$$S(C_1,C_2)=\frac{\alpha}{l(C_1,C_2)+\alpha} \tag{2}$$

式中：α 为可调节参数，可解释为相似度为 0.5 时的路径长度.

文献[10]对义元相似度算法进行了合理的改进，在将概念相似度转换为两个义元向量距离计算的基础上，引入节点层次结构，相当于加入节点深度参数，相似度计算公式如下

$$S(C_1,C_2)=\frac{\alpha\min(d_{C_1},d_{C_2})}{\alpha\min(d_{C_1},d_{C_2})+l(C_1,C_2)} \tag{3}$$

式中：d_{C_i} 表示 C_i 在义元树中的层次深度. 这样在距离相等的情况下，层次越深的节点间有更深的相似度，因为层次越深意味着划分越详细，密度越大.

3.1.2　同义词词林相似度计算　为了进一步提高相似度计算准确性，本文在此基础上引入同义词词林相似度计算，结合 HowNet 多种义元关系计算，并在主题图融入最终本体时进一步考虑两概念之间的语义关系

$$S_t(W_1,W_2)=\begin{cases}0, & \text{if find}(W_1,W_2)=0 \text{ and find }(W_1,W_2)=0\\ \psi, & \text{if find}(W_1,W_2)=1 \text{ or find }(W_1,W_2)=1\end{cases} \tag{4}$$

式中：$\text{find}(W_1,W_2)=0$ 表示在词林 W_1 的条目中找不到 W_2 的编号；$\text{find}(W_1,W_2)=1$ 表示在词林 W_2 的条目中含有 W_1 的编号；ψ 表示两概念之间的近似值.

3.1.3　概念语义关系与主题图融合策略　本文进一步考虑义元间的其他非同义关系，使主题图融合时不仅仅是元素之间替换与简单写入，还能保证融合后主题图的结构一致性. TMMC 将词语间关系排序为同义关系、上下关系、整体-部分关系、反义关系及其他关系，从而在计算词语相似度时可根据词语相互关系决定主题图概念融合策略. 设待融合为主题图 A 和 B，存在概念 $W_1\in A,W_2\in B$，输出为融合后主题图 C，则融合策略如下.

(1)同义关系. 对于两待融合主题图，首先对待比较元素对进行基于语法与同义词词林的相似度计算，基于语法的计算是指基于字符串相似度的计算，此部分计算结果可用 $S_{syn}(W_1,W_2)$ 表示，该模块的最终相似度表示为

$$S_{ss}(W_1,W_2)=\begin{cases}1, & \text{if } S_{syn}(W_1,W_2)=1\\ S_t(W_1,W_2), & \text{其他}\end{cases}$$

将 S_{ss} 与式(1)计算结果作为词汇语义相似度，并为其设置阈值 $\sigma(0\leqslant\sigma\leqslant1)$，大于 σ 则认为两概念为同义关系，则融合策略为

$$T_C=T_C\cup W_1$$
$$R_C=R_C\cup R_A^{W_1}\cup R_B^{W_2}$$
$$W_2=W_1$$

式中：T_C 为概念和针对标准主题图扩展的知识元层元素集合；R_C 表示扩展主题图 C 的关系集合，包括主题图概念之间关系、扩展主题图中知识元之间关系、概念与知识元关系的集合；$R_A^{W_1}$ 为主题图 A 中关于 W_1 的所有关系. $W_2=W_1$ 表示替换所有 W_2 元素

值为 W_1. 其他子元素的融合可参见文献[5].

(2)上下位关系. 如果两概念义元中有上下位关系,则认为两概念为上下位关系. 设 $R_H(W_1,W_2)$ 表示 W_1 为 W_2 的上位关系,这种情况下融合策略为

$$T_C = T_C \cup W_1 \cup W_2$$

$$R_C = R_C \cup R_H(W_1,W_2) \cup R_A^{W_1} \cup R_B^{W_2}$$

特别要说明的是,当 $\exists R_H.w=W_2$,在 w 作为 W_1 的上下位关系融入主题图 C 时,删掉这个冗余关系 $R_H(W_1,w)$.

(3)整体-部分关系. 如果两概念义元中不存在上下位关系,而存在整体-部分关系,则认为两概念为整体-部分关系,策略与(2)类似.

(4)反义义元、对义义元. HowNet 中除了常用义元外,还有反义义元(如好对坏,大对小)、对义义元(如讲道理对不讲道理). 在计算义元相似性时,两义元有可能不属于同一个分类. 为了更好地利用这些语义关系,本文对于义元的相似度计算补充定义:①如果两义元不属于同一类,即 $d=1$,则 $S(W_1,W_2)=S(W_1,W_2)\tau_1$,其中 τ_1 为相似度惩罚系数;②如果两义元是反义义元或对义义元,则 $S(W_1,W_2)=S(W_1,W_2)\tau_2$,其中 τ_2 为相似度惩罚系数,并转入情况①中阈值判断部分,执行相应融合策略.

(5)其他关系. 针对主题图中自定义的其他概念语义关系,融合策略为

$$T_C = T_C \cup W_1 \cup W_2; R_C = R_C \cup R_A^{W_1} \cup R_B^{W_2}$$

通过以上策略融合后的主题图 C 会存在部分不一致或者关系冗余情况,比如可能因为概念插入已存在的上下位关系之间,导致出现冗余信息,所以主题图需要进一步转入一致性检查与消除模块,本文对此不做讨论.

3.2 利用语料库计算未登录词相似度

因为扩展主题图本体内包含未能在 HowNet 等词典中收录的大量计算机专业词汇,本文提出基于语料库概念的相似度计算方法,计算结果能形成一个可离线扩充的专业词汇相似度词典. 本方法建立的假设基础为:如果两个概念是同义词,那么它们在语料库中的上下文必然是相似的,从而将概念相似度比较转换为文本相似度比较.

向量空间模型是一个基础的文本相似度模型,它的基本原理是将语料库中的文本转换为向量 $\{(t_1,w_1),(t_2,w_2),\cdots,(t_n,w_n)\}$,$t_i$ 表示一个概念,w_i 表示概念 t_i 在文本中的权重,这个权重可以使用词频或者共信息法计算,进而通过计算两个文本向量的余弦夹角作为文本相似度.

本文算法是一种单纯基于词频统计的算法,完全忽略了文档的结构信息,将句子看作是半结构化文档的基本单元,这些文档来自于以待比较概念为关键词、在维基百科与谷歌中抓取搜索快照形成的语料库. 算法的基本过程描述如下.

(1)语料抓取:以主题图中概念的基名为关键字,抓取维基百科和谷歌快照.

(2)去噪纯化:删除网页片段中关于网页与搜索引擎本身的噪声数据、特殊字符与停用词,建立关于关键词的纯文本结构数据库.

(3)切分句子:将数据库中纯文本切分为一系列的句子,从而产生 2 个关于句子的集合

$$E_i = \{s_{i1},s_{i2},\cdots,s_{in} \mid s_{ij} \in D_i, i=1,2, n \in N\}$$

式中:E_i 为关于文档 D_i 的句子集合;D_i 表示第 i 个文档;n 为文档包含的句子数;s_{ij} 是词的集合,指文档中每个句子包含的词,本文采用最大正向匹配和最大逆向匹配相结合的分词方法,每个词集合可以表示为

$$s_{ij} = \{t_{j1},t_{j2},\cdots,t_{jk},\cdots,t_{jl} \mid k=1,2,\cdots,l, l \in N\}$$

t_{jk} 表示第 j 个句子中的第 k 个词.

(4)在以上定义的基础上,计算句子间的相似度

$$S_s(s_i,s_j) = \frac{2\theta(s_i,s_j)}{|s_i|+|s_j|} \tag{5}$$

式中:$\theta(s_i,s_j)$ 是用来计算词 s_i 与 s_j 中相同字的个数;$|s_i|$ 表示词 s_i 中的字数.

(5)取该句子与文档中所有句子相似度最高值作为的最终相似度,表示为

$$S_{sD}(s,D_i) = \max_{s_i \in S_i, j \in \{1,2,\cdots,n\}} (S_s(s,s_{ij})) \tag{6}$$

在式(5)、式(6)基础上,假设 $|D_1|>|D_2|$,$|D|$ 表示文档 D 包含的句子数,主题图概念语义相似度

$$S(D_1,D_2) = \sum_i^{|S_1|} (S_{sD}(s_{1i},D_2))/|D_1| \tag{7}$$

当 $S(D_1,D_2)$ 大于阈值 ρ,则视两概念同义,并转入 3.1 节中的同义关系融合规则.

基于语料库的概念相似度算法描述见图 2. 从算法可以看出,复杂度与文档中的句子数量、句子中的单词数量相关. 设文档的句子长度为 m,每个句子包含词的个数为 n,因为 $\theta(s_1,s_2)$ 函数的时间复杂度为 $O(m^2)$,因此基于语料库的概念相似度算法复杂度为 $O(n^2m^2)$. 其中,针对关键字的语料搜索、去噪和切分词都进行了集中运算,并将计算出的概念相似度做成可在线扩充和离线查询的相似度词典.

```
Text Corpus1=(craw1(W1)); //获得相关语料
Text Corpus2=(craw1(W2));
Document1=New ArrayList(term(pure(Corpus1)));
                    //去噪与分词
Document2=New ArrayList(term(pure(Corpus2)));
Result=0;           //存储句子与文档相似度
  For i=0 to Document1.length
    For j=0 to Document2.length
      Result=result+
          Max(sim(Document1.sentence[i],
          Document2.sentence[j]);
    S=result/|D1|;
Append the similarity value of topics to file;
IF S>ρ
    Then go to Merge(W1, W2)
```

图 2　基于语料库的概念相似度算法

4　实验与结果分析

对于相似性度量算法的评价，已有文献中均使用信息检索领域的 3 个评价标准：查准率 P、查全率 Q 和综合指标 F. 设通过人工对比认为两个词语相似性较高的对数为 N_P 个，算法判定符合认知的对数为 N_A 个，且 N_A 个元素对中正确的为 N_R 个，则 3 个标准定义为

$$P=\frac{N_R}{N_A};\quad Q=\frac{N_R}{N_P};\quad F=\frac{2PQ}{P+Q}$$

测试数据来源于“863”课题的部分计算机领域扩展主题图，语料库资源来自 Wikipedia 的 Google 搜索快照，共 1.5 GB 纯文本. 在测试过程中，参数 τ_1、τ_2 分别固定为 0.2 和 0.1，假设阈值 σ 与 ρ 不相关，通过实验找出两参数的最佳值，先固定参数 σ，当 ρ=0.65 时可取得最高 F. 图 3 给出 ρ=0.65 时扩展主题图融合结果与阈值 σ 的关系图，其中输入包括计算机教育类 7 大学科共 6 149 个术语概念.

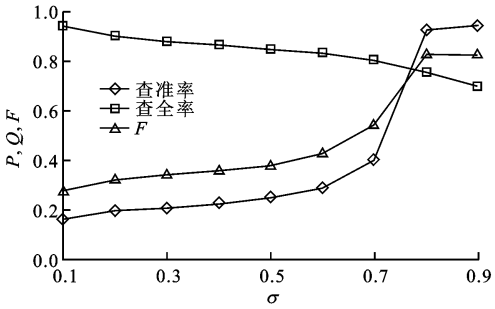

图 3　查准率、查全率、F 与阈值 σ 的关系

通过实验确定在阈值 σ=0.85 时，F 达到最大值 0.83，查准率为 0.91，查全率为 0.75. 在各参数值确定后，针对部分公共领域词语与专业术语，采用不同相似度算法的计算结果如表 1 所示. 其中，认知是通过问卷调查获得的相似度数据，ETMFC 为文献[7]提出的主题图融合算法.

表 1　不同算法相似度计算结果比较

词语对	相似度			
	认知	文献[8]	ETMFC	TMMC
把握 vs 领会	0.788	1.000	0.519	0.669
蘑菇 vs 大豆	0.102	1.000	0.916	0.134
篮球 vs 运动员	0.591	0.168	0.721	0.183
篮球 vs 宗教	0.054	0.189	0.234	0.106
中国 vs 联合国	0.411	0.136	0.235	0.281
宗教 vs 历史	0.600	0.171	0.138	0.076
宗教 vs 电脑	0.058	0.111	0.103	0.106
军备 vs 武器	1.000	0.600	0.643	0.553
外部 vs 内部	0.400	0.861	0.354	0.954
安康 vs 抱恙	0.255	0.211	0.560	0.000
局域网 vs 广域网	0.548	0.000	0.745	0.897
局域网 vs 园区网	0.701	0.000	0.641	0.784
计算机网络 vs 以太网	0.627	0.000	0.213	0.706
TCP/IP 协议 vs IP 协议	0.517	0.000	0.562	0.713
JAVA vs C++	0.726	0.000	0.123	0.564

为了测试算法的综合能力，选取 180 个日常生活用词组成 90 对测试概念对，选取 20 个计算机专业领域词汇组成 10 对测试概念对. 采用 TF-IDF 算法进行基于语料库的文档比较计算概念相似度，文献[7]算法和本文算法的比较结果如图 4 所示.

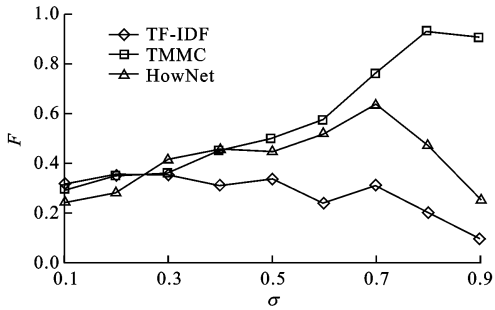

图 4　各算法的性能参数 F 对比图

由于利用知网的相似度算法没有对专业领域词汇的判断能力，导致查全率偏低，而 TF-IDF 算法在计算过程中没有充分利用文档的结构信息，且对于常用词的准确率远不如人工标注的字典，导致其查全率与差准率都偏低. 本文提出的 TMMC 不仅充分利用了经过人工标注的语义词典，同时利用在线搜索语料库，比较过程中保留了文档的结构信息，从而获得较高的综合性能，F 最高为 0.93，HowNet 算法与 TF-IDF 的 F 最高分别为 0.35 和 0.64. TMMC 在针对主题图中含有大量专业领域词汇时

有更好的优势，但是同一来源的主题图概念命名比较统一，同义词与近义词对较少，算法在多数据来源情况下会体现出更好的适用性.

从表1可以看出，HowNet算法对词典未收录词不能计算，ETMFC因为考虑了概念在本体周围节点的关系，导致相关度高的节点相似度很高，比如“蘑菇”与“大豆”，另外共有字符串长度较长的词对具有比较高的相似度，比如“局域网”与“广域网”，而“JAVA”与“C++”说明了相反的情况，可以看出TMMC与人的认知问卷调查结果最接近.

TMMC与已有主题图、扩展主题图融合算法查准率、查全率以及F的比较如表2所示.

表2 不同算法性能指标比较

算法	P	Q	F
SIM	0.88	0.30	0.45
TM-MAP	0.89	0.55	0.68
ETMFC	0.82	0.69	0.75
TMMC	0.91	0.75	0.83

当选取能使F达到最高的相似度阈值时，公共领域的测试结果稍高于专业领域，原因在于公共领域的相似度计算既能充分利用语义词典的语义信息，又能利用语料库的语义信息，而专业领域的词语很少能出现在语义词典中，可利用的语义信息较少.

通过对比在不同领域测试数据集下确定的阈值，可得出阀值与数据集有关. 在确定最终的阈值之前必须进行不同阀值下的对比测试，在F最好的情况下选取阈值.

5 结 论

针对扩展主题图本体融合问题，提出了基于HowNet中文语义词典的算法，并且考虑词典中义元之间的多种关系在融合过程中的含义，制定了相应的融合规则. 对于词典未收录的大量专业领域术语，提出了基于语料库向量空间相似度的计算方法，使相似度计算能够自动化并扩展到任意现有领域本体. 对于扩展主题图本体融合，如何能够提高语料库词语相似度算法的效率，挖掘词汇之间的潜在语义关系，将是下一步需要做的工作.

参考文献：

[1] ADOLFO G A, CUEVAS A D. Knowledge accumulation through automatic merging of ontologies [J]. Expert Systems with Applications, 2010, 37(3): 1991-2005.

[2] BIEZUNSKI M, BRYAN M, NEWCOMB S. A mathematical formalism for the topic maps reference model [EB/OL]. [2011-01-22]. http://www.isotopicmaps.org/tmrm/.

[3] SHVAIKO P, EUZENAT J. A survey of schema-based matching approaches [M] // Lecture Notes in Computer Science: Vol 3730. Berlin, Germany: Springer, 2005: 146-171.

[4] MALCHER L, WITSCHEL H F. Merging of distributed topic maps based on the subject identity measure (SIM) approach [M]. Leipzig, Germany: LIT, 2004: 1-11.

[5] 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.

[6] 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.

[7] 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.

[8] 刘群，李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学期刊，2002，7(2)：59-76.
LIU Qun, LI Sujian, Word similarity computing based on HowNet [J]. Computational Linguistics and Chinese Language Processing, 2002, 17(2): 59-76.

[9] LIN D K. An information theoretic definition of similarity semantic distance in WordNet[C] // Proceedings of the 15th International Conference on Machine Learning. Madison, USA: [s. n.], 1998: 296-304.

[10] 吴健，吴朝晖，李莹. 基于本体论的词汇语义相似度的Web服务发现[J]. 计算机学报，2005，28 (4)：595-602.
WU Jian, WU Zhaohui, LI Ying. Web service discovery based on ontology and similarity of words[J]. Chinese Journal of Computers, 2005, 28 (4): 595-602.

（编辑 杜秀杰 武红江）