第44卷 第2期 西 安 交 通 大 学 学 报 Vol.44 №2
2010年2月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Feb. 2010

面向多源知识融合的扩展主题图相似性算法

鲁慧民， 冯博琴，李旭
（西安交通大学电子与信息工程学院，710049，西安）

摘要：针对基于元数据或传统主题图的知识组织模式没有实现知识的多层次多粒度表示，以及知识融合过程中相似性算法准确性不高而影响融合质量的问题，结合全信息理论与扩展主题图结构特点及语义信息，提出了面向多源知识融合的扩展主题图相似性算法(ETMSC)和阈值选取的相关性、层次对应和实验确定三原则. 该算法综合了语法、语义和语用的相似性，扩展了主题图元素间组成结构上的相似性，同时充分考虑了涵义及所处语境的相似性. 主题图相似性的判别准则与阈值有关，阈值的确定与数据集相关. 实验结果表明，ETMSC算法与单纯基于语法或语义的相似性算法相比，准确性提高了9.2%～11.1%.

关键词：知识融合；主题图；相似性算法
中图分类号：TP391 **文献标志码**：A **文章编号**：0253-987X(2010)02-0020-05

Novel Similarity Algorithm of Extended Topic Maps for Multi-Resource Knowledge Fusion

LU Huimin, FENG Boqin, LI Xu
(School of Electronics and Information Engineering, Xi′an Jiaotong University, Xi′an 710049, China)

Abstract: A novel similarity algorithm of extended topic map called ETMSC for multi-resource knowledge fusion is proposed to improve the drawbacks that the knowledge organization model based on metadata or traditional topic map can not represent knowledge multi-level and multi-granularity, and the low accuracy of existing similarity algorithms. Three principles of the correlation, levels corresponding, and the experimental determination in selecting threshold are presented. The algorithm combines the comprehensive information theory with the structure and semantic information of extended topic map. The syntactic matching, semantic matching, and pragmatic matching are comprehensively considered, in which not only the structural similarity of topic map elements are extended, but also the meaning and relevance in linguistic contexts are thoroughly taken into account. Topic map similarity criterions are related to a threshold, and the determination of the threshold is associated with the data sets. Experimental results and comparisons with the traditional algorithms that are purely based on the syntactic or semantic similarity show that the F-measure of ETMSC is improved by 9.2%-11.1%.

Keywords: knowledge fusion; topic map; similarity algorithm

知识融合可描述分布、异构的多源知识集成过程，成为当今知识管理领域研究的热点. 目前，有关知识融合的算法主要有基于本体[1]的融合算法、构建语义网的融合算法、构建知识网格的融合算法以及基于主题图[2]的融合算法等. 主题图融合技术是将多个具有自治管理的局部主题图进行融合，构建一个独立于知识资源的全局主题图，从而实现多源知识的逻辑集成. 主题图相似性计算是主题图融合

收稿日期：2009-07-15. 作者简介：鲁慧民(1972—)，女，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)；国家自然科学基金资助项目(60803162).

的前提和基础，直接关系到主题图融合的质量，代表性算法有 SIM[3]、TOM[4] 以及 TM-MAP[5] 等，但大多数算法仅在语法层面从概念的组成结构上来考虑相似性，却未充分涉及概念的语义和语用相似性.

本文基于扩展主题图实现了分布式、异构的多源知识融合，提出了一种基于全信息理论的主题图相似性算法，综合了概念的语法相似度、语义相似度和语用相似度，在考虑了概念组成结构上的相似性的同时，还充分考虑了涵义和所处语境的相似性，从而提高了相似性算法的准确性.

1　扩展主题图融合

1.1　扩展主题图

主题图可以定位那些与知识概念关联的资源位置，描述知识概念之间的语义联系，是一个由主题(Topic)、关联(Association)以及资源出处(Occurrence)组成的集合体(TAO)[6]. 本文对传统主题图进行了扩展，在主题层与资源层之间引入了知识元及其关联关系，并建立了主题-知识元-资源实体三者间的联系，从而体现出知识的多层次、多粒度，为用户提供更为详细的知识细节，实现了知识元的关联导航[7]. 扩展主题图的知识逻辑组织结构如图 1 所示.

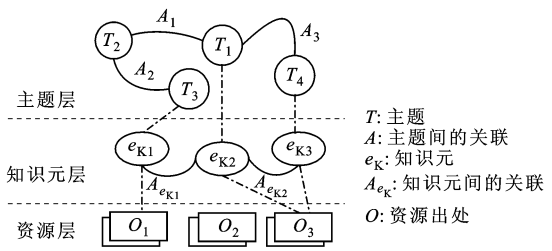

图 1　扩展主题图的知识逻辑组织结构

1.2　扩展主题图融合

扩展主题图融合是将多个局部扩展主题图融合为一个新的全局扩展主题图. 假设扩展主题图分别为 G_a、G_b 和 G_c，则扩展主题图融合可定义为

$$G:(G_a \times G_b) \rightarrow G_c \tag{1}$$

该融合主要包括 2 个步骤：①计算局部扩展主题图元素间的相似度；②依据融合规则对局部扩展主题图进行融合，生成全局扩展主题图.

2　全信息相似性算法

2.1　算法流程

依据全信息自然语言理解方法论[8]，本文提出了一种基于全信息的扩展主题图相似性算法(ET-MSC)，主要包括数据预处理、全信息相似性计算和元素对筛选 3 个方面. 数据预处理是对扩展主题图文档进行解析，并转换为主题图对象模式，形成待比较的元素对；全信息相似性计算是算法的核心，它从语法、语义和语用 3 个层面计算元素的相似性；元素对筛选通过阈值 t 比较来实现.

2.2　语法相似性计算

语法相似性计算主要用于统计相同字符的比例，其描述如图 2 所示.

输入：元素 E_a 和 E_b
输出：语法相似度 $S_{\text{syntactic}}(E_a,E_b)$
(1) 中文分词.元素 E_a 和 E_b 分词得到的词块数分别构成相应的字符串集合 S_a 和 S_b.
(2) $\forall S \in S_a \vee S_b$ 删除代词、介词、数词等没有特别含义的单词.
(3) 根据式

$$S_{\text{syntactic}}(E_a,E_b)=\sum_{m=1}^{|S_a|}\sum_{n=1}^{|S_b|} S(s'_m,s'_n)/|S_a||S_b|,$$

$s' \in S_a, s' \in S_b, S(s'_m,s'_n)$ 计算字符串 s'_m 和 s'_n 的相似度 $S(s'_m,s'_n)=2C/(|s'_m|+|s'_n|)$

$|S_a|$、$|S_b|$ 分别表示集合 S_a、S_b 中字符串的个数；s'_m、s'_n 表示字符串；C 表示 2 个字符串的最长公共字符串的字符数；$|s'_m|$、$|s'_n|$ 表示字符串 s'_m、s'_n 包含的字符数

图 2　语法相似性算法

2.3　语义相似性计算

经过中文分词之后，扩展主题图中元素值的最小单位均为词语，这样文中存在大量的同义词，它们同义不同形. 本文提出基于 HowNet 的、采用分区间的词语语义相似性算法，即根据 HowNet 的词语构成方式，按照词语粒度从细到粗的顺序排序. 该算法包括了义原相似性、义项相似性、词语相似性计算.

义原相似性计算主要利用义原层次体系计算义原的语义距离，计算式为

$$S_P(p_1,p_2)=\frac{\partial D}{\partial D+d} \tag{2}$$

式中：∂ 为可调参数；d 为 2 个义原间的距离；D 为 2 个义原最近公共父节点在树中的深度. 可以看出，D 越大则 S_P 越大，而 d 越大则 S_P 越小.

义项由多个义原按结构组合而成，将各部分的相似性度量值组合起来即可得到义项的相似度. 义项中的义原分为 4 个部分：主要义原、其他义原、符号义原和关系义原. 对这 4 个部分的义原相似度加权求和，可得到义项相似度.

主要义原为义项的基本特征，主要义原相似度

可由式(2)获得,即 $S_{MP}=S_P$.

其他义原可能有若干个,将2个义项的其他义原分别组成集合 P_1 和 P_2,对2个集合间的元素按图3所示算法计算,可以得到其他义原相似度 $S_{OP}(P_1, P_2)$.

输入:集合P_1和P_2
输出:$S_{OP}(P_1, P_2)$
假设对象SimPair有3个属性,即ξ_1、ξ_2及其相似度值v;$S=0$.
(1) $\forall a\in P_1, \forall b\in P_2$, List.add (new SimPair($a$, b,$S_{P(a,b)}$))
(2) 将List中的元素按v值降序排列
(3) for $i=0$ to List.length
//s_p是List中第i个元素
if HashSet.contains($s_p.\xi_1$) = False &&
HashSet.contains($s_p.\xi_2$) = False
then
{
ξ_1与ξ_2配对;
$S=S+v$;
HashSet.add ($s_p.\xi_1$) and HashSet. add ($s_p.\xi_2$);
}
end if
end for
(4) 若P_1和P_2集合中元素个数不同,则多余的元素和null配对
(5) 计算 $S_{OP}(P_1, P_2)=S/\min(|P_1|, |P_2|)$

图3 其他义原相似性算法

同一种结构的关系义原相似度为

$$S_{RP}=(S_P(c_1, c_2)S_P(c_1', c_2'))^{1/2} \tag{3}$$

式中:c_1、c_2 为特征;c_1'、c_2' 为特征值.

符号义原按结构特征分类,每类为一个集合.分别计算相同类别间的相似度,最后对所有分类的相似度求平均,从而获得符号义原的相似度 S_{SP}.

词语 w_1 和 w_2 的义项相似度集合为 $S_w=\{s_{w_1}, s_{w_2}, \cdots, s_{w_{m\times n}}\}$.将 S 中的值划分成4个区间,即 α:[0.0,0.1)、β:[0.1,0.2)、γ:[0.2,0.8)、δ:[0.8,1.0],这样可分析4个区间对于词语相似性认知的模糊性和确定性的贡献.

(1)α 区间的值属于 S_α,义项间的语义相似性很低,词语的认知会产生模糊和不确定性.

(2)β 区间的值属于 S_β,与 α 区间的相近,但贡献比 α 区间稍差.

(3)γ 区间的值属于 S_γ,由于该区间的值对认知的模糊性和确定性都有贡献,可将此区间舍去.

(4)δ 区间的值属于 S_δ,该区间的值对词语的认知产生确定性.

词语语义相似性计算式为

$$S_{semantic}=\max(S_\delta)-\frac{(0.2\,|\,S_\alpha\,|-\sum S_\alpha)+\sum S_\beta}{|\,S_\alpha\,|+|\,S_\beta\,|} \tag{4}$$

式中:$|\,S_\alpha\,|$ 为 S_α 中的元素个数;$|S_\beta|$ 为 S_β 中的元素个数.当 $S_\alpha=S_\beta=S_\delta=\varnothing$ 时

$$S_{semantic}=\max(S_\gamma) \tag{5}$$

2.4 语用相似性计算

语法和语义相似性都是孤立地考虑元素本身,因此存在歧义,无法分辨具体的意义指向.对于有多种含义的词"深",无法确定具体的意思,但是放在语境中,如"内容太深,很难理解"和"颜色太深,不好看",则其含义就非常清楚.本文依据语用相似性来区分可能存在的歧义.与目标主题(知识元)有关的,在一定知识半径范围内的其他主题和知识元将构成该主题(知识元)的语境.语用相似性计算式为

$$S_{pragmatic}(E_1, E_2)=\omega S(T_1, T_2)+(1-\omega)S(e_{KS1}, e_{KS2}) \tag{6}$$

式中:$S(T_1, T_2)$ 是2个主题集合间的相似度;$S(e_{K1}, e_{K2})$ 是知识元集合间的相似度;$\omega\in[0,1]$,是针对主题集合和知识元集合重要性定义的权重.

2.5 ETMSC

ETMSC从语法、语义和语用3个层面上综合衡量了概念的相似性.首先计算元素对的语法相似性,若元素对的语法相似性较高,说明元素对不存在同义不同形的情况,不必进行语义相似性计算,直接考虑其语境,进行语用相似性计算.若元素对的语法相似性较低,则进行语义和语用相似性综合计算.算法描述如图4所示.

输入:待比较元素对E_1, E_2
输出:$S(E_1, E_2)$
tag=0; //ω、φ_1、φ_2、φ_3是权重
Csim1: $S_{syntactic}(E_1, E_2)$
if $S_{syntactic}\geqslant t$(阈值)
tag=1;
goto Csim3;
Csim2: $S_{syntactic}(E_1, E_2)$
Csim3:$S_{pragmatic}(E_1, E_2)$
if tag=1
$S(E_1, E_2)=S_{syntactic}(E_1, E_2)+(1-\omega)S_{pragmatic}(E_1, E_2)$
else
$S(E_1, E_2)=\varphi_1 S_{syntactic}(E_1, E_2)+\varphi_2 S_{syntactic}(E_1, E_2)+\varphi_3 S_{pragmatic}(E_1, E_2)$
return $S(E_1, E_2)$;

图4 ETMSC

3 实验及结果分析

3.1 实验配置

实验配置:Pentium Ⅳ 2.0 GHz,1 GB内存,Windows XP操作系统.采用JAVA编写程序.实验数据:"863计划"项目中有关计算机网络领域的部

分主题图的 XTM 文件.

3.2　实验结果及分析

采用查准率 R_P、查全率 R_F 和 F 值 3 个评价指标来衡量算法，相应指标定义如下

$$R_P=\frac{N'_A}{N_A};\quad R_F=\frac{N'_A}{N};\quad F=2\frac{R_PR_F}{R_P+R_F}\quad(7)$$

式中：N 为人工比对认为应该融合的元素对数；N_A 为算法判定应该融合的元素对数；N'_A 为 N_A 个元素对中正确的元素对数.

利用本文的语法、语义相似性算法以及综合语法、语义和语用相似性算法，对计算机网络领域扩展主题图 XTM 文件进行两两间计算，并对 3 个指标求平均，结果如图 5～图 7 所示.

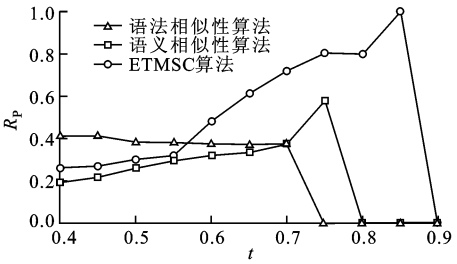

图 5　语法、语义和 ETMSC 相似性算法的查准率

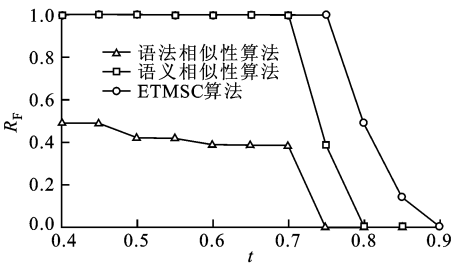

图 6　语法、语义和 ETMSC 相似性算法的查全率

从图 5 可以看出，ETMSC 算法的查准率在整体上比较高，当阈值 $t<0.55$ 时，语义和 ETMSC 算法的查准率比语法的低. 在语义和 ETMSC 算法的实验结果中，待融合的元素对相似度值偏大的较多，而在语法相似性的实验结果中，待融合的元素对的相似度值分布均匀，所以当 t 逐渐减小时，语法相似性算法的查准率没有明显变化，而语义和 ETMSC 相似性算法的查准率则迅速降低.

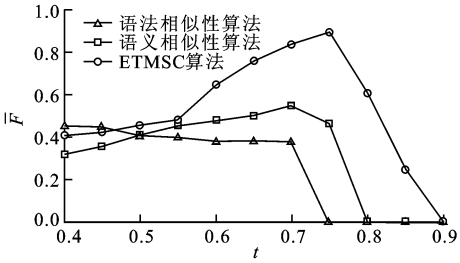

图 7　语法、语义和 ETMSC 相似性算法的 F 均值

从图 6 可以看出，语义和 ETMSC 算法的查全率在一定的阈值取值区间内均达到 1.0，这是因为数据是专业领域内的扩展主题图数据，这些数据之间的相似度存在明显的界限，分布不均匀，应该融合的元素对相似度比较高，而不应该融合的元素对之间相似度普遍偏低，所以在一定的阈值取值区间内，查全率可以达到 1.0. 语义和 ETMSC 相似性算法的查全率较高，当二者查全率达到峰值时，t 仍然比较大.

从图 7 可以看出，当 $t>0.55$ 时，ETMSC 算法的 F 均值高于语法和语义相似性算法，但 F 均值达到某个峰值后，随着 t 的增大而下降. ETMSC 算法比语法和语义相似性算法的 F 均值平均提高了 13.5%～17.1%.

将 ETMSC 算法的 F 值与已有的相似性算法（SIM 算法和 TM-MAP 算法）进行了比较测试，结果如表 1 所示. 从表中可以看出：t 在 0.4 左右，SIM 算法的 F 达到最大值为 0.464；$t=0.4$ 时，TM-MAP 算法的 F 达到最大值为 0.697；$t=0.75$ 时，ETMSC 算法的 F 达到最大值为 0.891. SIM、TM-MAP 和 ETMSC 算法的 F 均值（$\overline{F}$）分别为 0.428、0.409 5 和 0.520 6. 3 种算法的 F 值比较如图 8 所示. 从图 8 中可以看出，ETMSC 算法在阈值区间 [0.15，0.75] 的 F 值呈逐渐上升的趋势，整体性能比 SIM 和 TM-MAP 算法有所提高，F 均值提高了 9.2%～11.1%.

表 1　不同阈值 t 下算法的 F 值比较

算法名称	F															
	$t=0.15$	$t=0.20$	$t=0.25$	$t=0.30$	$t=0.35$	$t=0.40$	$t=0.45$	$t=0.50$	$t=0.55$	$t=0.60$	$t=0.65$	$t=0.70$	$t=0.75$	$t=0.80$	$t=0.85$	$t=0.90$
SIM	0.456	0.444	0.450	0.464	0.464	0.464	0.464	0.400	0.400	0.379	0.379	0.379	0.000	0.000	0.000	0.000
TM-MAP	0.056	0.057	0.057	0.531	0.674	0.697	0.619	0.585	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ETMSC	0.410	0.410	0.410	0.410	0.410	0.410	0.424	0.456	0.479	0.648	0.760	0.838	0.891	0.609	0.245	0.000

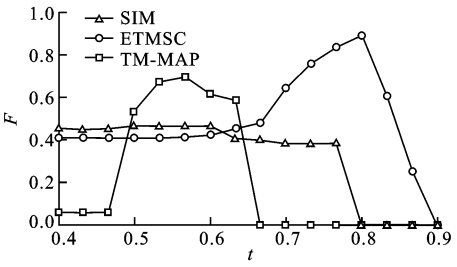

图 8 SIM、TM-MAP 和 ETMSC 算法的 F 值比较

ETMSC 算法在主题图相似性判别的准确性上也有所提高，但与 t 的选取有较大关系. t 与数据集有关，t 的选取遵循以下规则.

(1)相关性规则：与数据集的专业性有关. 专业性强的主题图数据阈值偏小，因为专业领域内主题图数据中很少出现歧义性的事物描述，而公共领域主题图数据阈值偏大.

(2)层次对应原则：与相似性重要程度、融合需求的不同层次相对应，要充分考虑达到预期目标和效果所需要的时间和精力.

(3)实验确定原则：在确定 t 之前必须用目标数据集中的数据进行测试，应在 F 均值最好的情况下选取阈值.

4 结 论

本文提出了一种面向多源异构知识融合的扩展主题图相似性算法，基于全信息理论，从语法、语义和语用 3 个层面进行综合性相似性计算. 经国家“863 计划”项目主题图融合的实验结果表明，全信息相似性的算法在查全率、查准率和 F 值方面均优于单纯采用语法或语义的相似性算法. 与目前已有的相似性算法相比，所提算法的相似性计算的准确性有一定程度的提高. 下一步工作是，对生成的待比较元素对进行归类处理，减少比较元素对的数目，进一步提高算法的效率.

参考文献：

[1] 田磊，覃征，衡星辰，等. 基于本体的多源异构 XML 数据近似查询方法 [J]. 西安交通大学学报，2007，41(6)：702-706.
TIAN Lei，QIN Zheng，HENG Xingchen，et al. Approximate query approach based on ontology for multi-source and heterogeneous XML data [J]. Journal of Xi′an Jiaotong University，2007，41(6)：702-706.

[2] 鲁慧民，冯博琴，赵英良，等. 基于扩展主题图的分布式知识融合 [J]. 吉林大学学报(理学版)，2009，47(3)：543-547.
LU Huimin，FENG Boqin，ZHAO Yingliang，et al. Distributed knowledge fusion based on extended topic maps [J]. Journal of Jilin University (Science Edition)，2009，47(3)：543-547.

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

[4] 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.

[5] 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.

[6] PEPPER S. The TAO of topic maps [EB/OL]. [2008-12-20]. http://www.gca.org/papers/xmleurope2000/.

[7] LU Huimin，FENG Boqin，ZHAO Yingliang，et al. A new model for distributed knowledge organization management [C]// Proceedings of the 7th International Conference on Grid and Cooperative Computing. Los Alamitos，CA，USA：IEEE Computer Society，2008：261-265.

[8] 钟义信. 自然语言理解的全信息方法论 [J]. 北京邮电大学学报，2004，27(4)：1-12.
ZHONG Yixin. A methodology of natural language understanding based on comprehensive information [J]. Journal of Beijing University of Posts and Telecommunications，2004，27(4)：1-12.

(编辑 苗凌)