主题图融合技术研究综述

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

摘 要 主题图在信息资源上层构建了一个结构化的语义网,提供了一个良好的语义模型,可以弥补Web 2.0在应用上存在的语义缺陷,而主题图融合作为主题图的重要研究内容,是将分布式环境下同一领域内分散的局部主题图合并为一个全局主题图,实现Web信息的有效组织和管理以及信息的集成与共享。归纳总结了主题图融合的处理过程,分析评价了主题图融合中的难点——主题图的相似性算法,并对融合的原则和算法进行了分析总结,明确了高相似度主题融合的过程。此外,在分析主题图融合冲突的基础上,提出了主题图融合冲突检测与消除的整体设计方案,并将主题图的动态更新进行了阶段划分,最后指出了主题图融合未来的研究方向。

关键词 主题图,融合,相似性算法,冲突检测
中图法分类号 TP391 **文献标识码** A

Survey of Topic Maps Merging Technology Research

LU Hui-min FENG Bo-qin ZHAO Ying-liang
(School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract Building a structured semantic Web on the layer of information resources, topic maps provide a good semantic model which can make up the existing semantic defects of Web 2.0. As one of the important researches, topic maps merging integrates the same field's partial topic maps into a new topic map. Through the overall topic map, the effective organization, management, integration and sharing of Web information can be easily realized. In this paper, the general steps of topic maps merging and their functions were summarized. With emphasized on the analysis and judgment of similarity algorithm which is the difficulty of topic maps merging, merging principles and strategies were further highlighted, and topic merging process of high similarity was clarified. Furthermore, the overall scheme of conflict detection and elimination was proposed. Topic maps dynamic updating was compartmentalized into several reasonable phases and topic maps merging research directions were also discussed.

Keywords Topic maps, Merging, Similarity algorithm, Conflict detection

1 引言

Web2.0本着自由、平等、开放、公正的原则,在为用户提供方便、简捷的创作方式的同时,也牺牲了语义关联;虽然Tag提供了简单的语义注解,但还远远不够[1]。主题图(Topic Maps)是一个由主题(Topic)、关联(Association)以及资源出处(Occurrences)组成的集合体(TAO)[2],可以定位于知识概念关联的资源的位置,也可以描述知识概念之间的语义联系。主题图继承了索引、词汇表、叙词表、本体、分类表等知识组织方式的特征,并吸取了人工智能领域的语义网思想,这使得它能够比较好地适应数字化环境中的知识组织[3]。如何将主题图融入Web2.0中,弥补其应用上存在的语义缺陷已经成为学者们比较感兴趣的问题。

主题图融合是主题图融入Web2.0中的一个关键技术,其目的是将分布式环境下离散的同一领域局部主题图合并为全局主题图,实现信息的组织、管理和共享,能够使用户以统一的模式查询和搜索异构数据源中的信息[4]。主题图标准ISO/IEC 13250 Topic Maps[5]中一般采用OASIS[6]制定的公共项目标识符[7](Published Subject Indicators, PSI)作为主题图设计模型[8]中主题的统一标识,PSI相同的主题均可融合,但PSI没有覆盖主题图的所有应用领域,仅在某些非常通用的领域(如语言、国家和地区名称)中定义了PSI,因此目前还无法实现基于PSI的主题图融合。在主题图标准中,主题图融合的前提是两个项目的描述必须一致,如果两个主题描述的是同一个项目就应当融合。但在分布式环境下,各主题图采用不同的词汇表导致对相同项目的描述很难一致而无法实现融合,因此基于相似主题的融合成为目前研究的热点,但大部分研究都集中在主题图相似性算法上,很少对主题图融合过程及其相关内容进行系统的研究。针对这一问题,归纳总结出相似主题图融合的处理流程及相关技术,为主题图融合

到稿日期:2008-05-15 本文受863国家高技术研究发展计划项目(2008AA01Z131)资助。
鲁慧民 女,博士生,CCF学生会员,主要研究方向为知识管理、知识融合等,E-mail:luhm@stu.xjtu.edu.cn;冯博琴 男,教授,博导,主要研究方向为数据挖掘、智能网络等;赵英良 男,副教授,博士,主要研究为最优化理论及应用、计算机网络等。

的系统研究提供借鉴。

2 主题图融合的处理流程

通过对主题图融合的分析与研究，将主题图融合的处理过程分为6个步骤，即XTM文件的解释、待比较主题对集的产生、主题相似度的计算、相似主题的融合、融合冲突的检测与修正、全局主题图的动态更新。处理框架如图1所示。

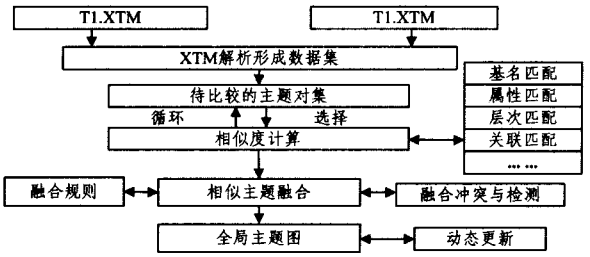

图1 主题图融合的处理框架

首先对XTM格式的主题图文件进行解析，抽取其中的元素（比如主题的基名、关联、资源等）形成所需要的数据集合，采取一定措施从中选择待比较的主题对集，然后从多个角度（比如基名、属性、资源、层次、关联等）综合计算主题的相似度，对相似度高的主题进行融合，在融合过程中对产生的冲突进行检测与消除，最后形成全局主题图。为了保证全局主题图与分布式环境下的局部主题图的一致性，需要对全局主题图进行动态更新。

3 主题图融合中的关键技术

主题图融合过程中主题图相似度计算是进行主题图融合的前提，也是主题图融合技术中的难点；相似主题的融合指的是对相似度高的主题进行融合，是整个融合技术的核心，对融合过程中产生的冲突进行检测并消除，提高全局主题图的质量。而全局主题图的动态更新可以保证全局主题图与局部主题图的一致性。

3.1 主题图的相似性算法

目前关于主题图相似度计算还处在理论研究阶段，主要采用统计方法从构成元素的字符组成上来考虑其相似性。比较典型的算法有SIM（Subject Identity Measure）算法、TOM（Topic & Occurrence-oriented Merging）算法和TM-MAP算法等。

3.1.1 SIM算法

Lutz Maicher和Hans Friedrich Witschel于2004年提出一种采用统计方法来实现相似主题判别的SIM算法[9]。首先计算主题名称和资源的相似度（主要通过比较两个字符串中的字符相似程度来计算），再通过主题名称和主题资源在相似性中所占比重来综合分析主题间的相似性，从而决定两个主题图是否能够合并。SIM算法采用式(1)计算主题间的综合相似度。

$$SIM=\lambda SIM.\ Names+(1-\lambda)SIM.\ Occs \quad (1)$$

其中 *SIM. Names* 是主题名相似度，*SIM. Occs* 是资源相似度。

if($SIM.\ Names<t_{Names}$ or $SIM.\ Occs<t_{Occs}$) 令 $SIM=0$

其中 t_{Names} 是设置的 *SIM. Names* 必须超过的阈值，避免在 *SIM. Names* 太小的情况下过分依赖 *SIM. Occs*；t_{Occs} 是设置的 *SIM. Occs* 必须超过的阈值，避免在 *SIM. Occs* 太小的情况下过分依赖 *SIM. Names*；λ 为 *SIM. Occs* 与 *SIM. Names* 在综合分析中所占比重。

SIM提供一种简单的、语言与结构独立的主题图相似度计算方法，克服了现有规则在分布式环境下的局限性，即使两个主题图采用不同的词汇也可以进行相似性计算，但SIM算法在进行相似度计算时只考虑主题名和资源的相似度，对主题图的其它元素没有考虑，因此它的准确性不够。

3.1.2 TOM算法

吴笑凡、周良等于2006年提出分布式主题图融合中的TOM算法[10]，通过判断不同主题图中的主题是否指示同一项目，检查它们所表示内容的相似性程度以实现融合，提高了主题相似性判别的准确性。

TOM算法的核心思想是内容的相似度越高，同一性的可能就越大。选择主题名称和主题事件作为检验对象，每个主题名称有且仅有一个基名，每个事件可以包括一个或数个属性，属性可以是事件数据，也可以是事件资源（通常以URI的形式出现），分别计算主题基名和事件的相似度，然后根据基名相似性和事件相似性的权重综合得出主题图的相似性。TOM算法判定主题相似性的计算公式如式(2)所示。

$$\text{if}(\delta_{N_i}>\psi_N \text{ and } \delta_{o_i}>\psi_o \text{ and } \lambda\delta_{N_i}+(1-\lambda)\delta_{o_i}>\psi_T) \quad \text{令 } \delta_{T_i}=1 \quad (2)$$

其中 δ_{N_i} 是基名的相似度，δ_{o_i} 是事件的相似度，δ_{T_i} 表示只有当 $\delta_{T_i}=1$ 的时候，两个主题才能合并。ψ_N 是基名阈值，ψ_o 是事件阈值，ψ_T 是整个相似性阈值，λ 是权值。

3.1.3 TM-MAP算法

SIM和TOM算法只是对主题图的主题名和资源进行相似性判别，没有考虑主题间的关联和结构上的相似性。Jung-Mn Kim，Hyopil Shin，Hyoung-Joo Kim于2006年提出一种多策略主题图相似度判别的TM-MAP算法[11]，利用名字、属性、层次和关联的相似度来综合计算主题图的相似度。TM-MAP算法采用的是平均值方法，公式如下：

$$SIM(T1,T2)=(SIM_{name}+SIM_{acc}+SIM_H+SIM_{assoc})/4 \quad (3)$$

其中 SIM_{name} 是主题名的相似度，SIM_{acc} 是属性相似度，SIM_H 是层次相似度，SIM_{assoc} 是关联相似度。设置一个阈值，如果 $SIM(T1,T2)$ 大于该阈值说明 $T1$ 和 $T2$ 具有很高的相似度，反之则不能融合。TM-MAP算法采用一种多策略方法来判别两个主题的相似度，比采用单一匹配的策略更能提高判别的准确性，该方法不需要扫描两个主题图的所有主题，提高了相似性判别的效率。但其采用基于名字、属性、层次和关联4种匹配策略来综合分析主题的相似性，增加了算法的复杂性。另外它采用平均值方法计算综合相似度，没有考虑各匹配策略在主题相似度中的重要程度。

现有的相似性算法还处在比较简单的初级阶段，为了提高相似性判断的效率和准确性，需要从以下几个方面进行研究。

(1)目前只是从构成元素的字符组成上来考虑，对其语义和上下文环境没有充分考虑，影响了相似性判别的精确性，在以后的研究中应充分考虑语义在相似性判别中的作用，提高判别的准确度。

(2)进一步研究主题图在层次结构上的相似性算法。不

仅从元素的组成和语义上来进行计算，还应充分考虑主题图结构上的相似性，提高判别的准确性。

(3)主题图用于信息的组织和管理，往往具有很大的数据量，进行相似性判别必须考虑执行的效率，采取一定的优化措施合理选择待比较的主题对，可以考虑采用聚类方法进行主题的划分，使比较在同类主题间进行。

(4)进行自适应相似性匹配算法的研究，针对不同主题图的特点合理选择匹配策略，提高算法的实际应用效果。

3.2 相似主题图的融合

通过相似性计算得到相似度高的主题，采用一定的规则与算法将其进行融合，得到全局主题图。

(1)相似主题图融合的原则

主题图融合的目的是为了减少主题图的冗余结构，主题图标准的一个很重要的原则是一个主题只能表示一个项目(One Topic for One Subject)，当主题表示同一项目时需要融合，满足以下条件之一的两个主题被认为是指向同一项目：

- 它们使用相同的资源作为项目；
- 它们使用相同的资源描述项目；
- 它们在同一个范围中使用相同字符串作为主题基名。

下面介绍主题融合、属性融合和关联融合的具体方法[11]。

主题融合：如果主题图 A 中的主题 a 与主题图 B 中的主题 b 一致或相似度高，那么就在合并后的主题图 C 中生成一个与 a 或 b 一致的新主题 c；如果主题图 A 中的主题 a 在主题图 B 中没有一致或相似度高的主题，则将其拷贝到主题图 C。主题图 B 中如果存在这样的现象，同样处理。

属性融合：如果主题图 A 中的主题 a 有属性 p，与 a 一致的主题图 B 中的主题 b 有属性 q，如果 p 与 q 一致或相似，合并后的主题 c 取 p 或 q；如果 p 和 q 不相似，则合并后的主题 c 必须包括所有的属性 p 和 q。

关联融合：如果联系属于 a，那么在与 b 合并后生成的 c 维持该关联；如果在主题图 A 中有关联 $R_a(a_1,a_2)$，在主题图 B 中有关联 $R_b(b_1,b_2)$，a_1 和 b_1 的合并主题 c_1 就有两个关联 $R_a(c_1,a_2)$和 $R_b(c_1,b_2)$。

(2)主题图融合的处理过程

以集合数据格式的融合措施为例，归纳出主题图融合算法的处理过程。首先将相似度高的主题存储在集合中，然后从中选择待融合的主题对，经适配器的融合处理添加到合并后的主题集中，循环执行直到所有的高相似度主题全部融合完毕，然后将局部主题图中低相似度的主题集与合并后的主题集进行并集运算，形成全局主题集，如图 2 所示。

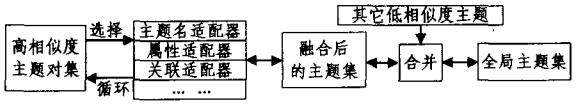

图 2 高相似主题融合示意图

融合的关键是适配器的设计，根据待融合的元素设计不同的融合算法，对相应的高相似度元素进行融合处理。目前关于主题图融合具体算法的研究还比较少，各种适配器融合算法是以后研究的重点。

3.3 主题图融合中的冲突检测与消除

局部主题图在融合成全局主题图时会产生各种类型的合并冲突。Jung-Mn Kim 等在对这些合并冲突进行分析的基础上，将其分为语义冲突、结构冲突和临时性冲突[11]，语义冲突主要包括名字冲突和属性冲突；名字冲突是指相同的主题有不同的名字或不同的主题有相同的名字，属性冲突又分为属性值和属性类型冲突，是指不同的属性类型有相同的属性值或不同的属性值表示相同属性类型。结构冲突主要是相同的主题在不同的信息源中采用不同的逻辑结构造成的，是指主题图的层次结构上产生的冲突，包括概括冲突和聚合冲突。概括冲突是指将多个主题汇总到单个主题引发的结构冲突，聚合冲突是指单个主题的属性或来自一个主题集合的属性或值。临时性冲突主要是指一些没有定义的实体造成的冲突和其它一些悬空实体、重复 ID 等造成的冲突。

主题图融合时要对这些冲突进行处理，主要包括两个方面——冲突的检测与冲突的消除，但目前该领域研究的文献很少。在本体领域已经对本体融合冲突展开了深入研究，文献[12]提出了一种基于本体的语义冲突处理方法，文献[13]在分析了计算机处理的信息中存在的语义不一致现象后，提出了语义冲突的处理模型等。借鉴以上的研究成果，设计出主题图融合的冲突检测与消除的整体解决方案。

冲突检测部分主要判断主题的语义相关程度，判断是否同一主题表示不同的事物或不同的主题表示相同的事物。如果存在上述情况，则说明相应的主题存在冲突，然后根据二者的相关性和特点，判断其冲突的类型。冲突消除部分主要根据冲突的类型设计不同的冲突处理器来进行处理，处理过程如图 3 所示。

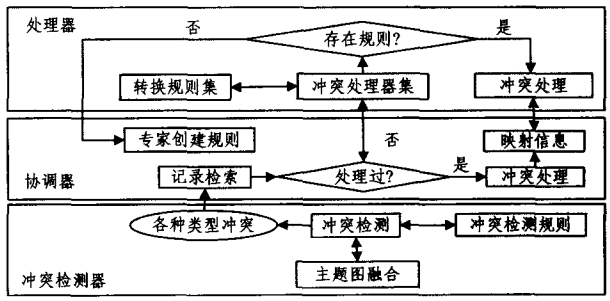

图 3 主题图融合冲突检测与消除处理过程

根据冲突规则对主题图进行冲突检测，判别主题图中是否存在冲突并判别其类型进行标注，对存在的冲突进行消除。根据冲突类型首先搜索已处理过的冲突记录，如果存在处理过的相同的冲突类型，则直接调用映射信息进行处理。如果不存在该记录，则从冲突处理器集中选择合适的冲突处理器进行处理，检索处理器的转换规则集，如果存在处理的规则即进行处理并将形成的映射信息进行存储；若没有相应的转换规则，则需要领域专家添加转换规则后再进行处理。通过以上对融合时产生的冲突进行及时的检测与消除，实现全局主题图融合的合理性，正确性和最优化，保证全局主题图的融合质量。

3.4 主题图的动态更新

主题图融合后，若局部主题图发生变化，全局主题图也应进行动态更新，另外还需要对局部的变化对全局产生的影响进行判别和处理，以保证全局主题图的一致性和融合质量。本文将主题图的动态更新分为 3 个阶段：更新的监听、更新信息的传输、全局主题图的更新。

(1)更新监听

建立更新监听器,监听局部智能主题地图的动态变化。

(2)更新信息的传输

监听器监听到局部主题图发生变化后,采用 SOAP 实现分布式系统中主题图同步数据的传输。SOAP(Simple Object Access Protocol),简单对象访问协议,是分布式环境中交换信息的简单的协议,是基于 XML 的协议,包括 4 个部分:

SOAP 封装(Envelop):定义一个描述消息中的内容。

SOAP 编码规则(Encoding Rules):表示应用程序需要使用的数据类型的实例。

SOAP RPC 表示(RPC Representation):表示远程过程调用和应答的协定。

SOAP 绑定(Binding):使用底层协议交换信息。

主题地图动态更新的数据传输如图 4 所示。

图 4 动态更新的数据传输示意图

SOAP 信封(SOAP Envelope)是一个包含 Header 和 Body 内容的 XML 文档。

信息结构如下所示:

```
〈soap-env:Envelope xmlns:soap-env="http://schemas.xmlsoap.org/soap/envelope/"〉
    〈soap-env:Header/〉
        〈soap-env:Body〉
            〈students:GetAllStudents xmlns:students="http://ctec.xjtu.edu.cn"〉
                〈students:name〉王飞〈/students:name〉
            〈/students:GetAllStudents〉
        〈/soap-env:Body〉
〈/soap-env:Envelope〉
```

(3) 全局主题图的更新

当接收到更新信息后,对全局主题图进行动态更新。通过描述源主题图的指定更新部分和目标主题图的拟更新部分,使用源主题图的 XML 描述以及转化的定义描述(convert.xslt)来产生更新以后的主题图,如图 5 所示。

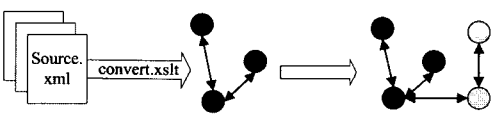

图 5 主题图更新示例

结束语 主题图融合是将主题图融入 Web 2.0 的关键技术之一,但主题图融合技术尚处于理论研究阶段,要将其实际应用到 Web,需要解决的问题还很多,如基于语义主题的相似性算法、分布式环境下主题图的逻辑融合、融合冲突的检测和消除以及主题图动态更新的协议和更新算法等等,有待我们去深入的探索和研究,主题图融合技术可以广泛应用到 Web 信息组织、信息管理、信息查询与搜索、信息集成以及构建信息导航地图等,有着极其广阔的应用前景。

参 考 文 献

[1] 朱良兵. Topic Maps:撬动 Web2.0 的语义杠杆[J]. 图书馆杂志,2007,26(5):48-51

[2] Pepper S. The TAO of Topic Maps [EB/OL]. http://www.gca.org/papers/xmleurope2000/,2000-06-12

[3] 朱良兵,纪希禹. 基于 Topic Maps 的叙词表再工程[J]. 现代图书情报技术,2006,141:81-84

[4] 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

[5] ISO/IEC 13250 Topic Maps:Information Technology Document Description and Processing Languages. 2nd ed. 2002[S]. http://www.y12.doe.gov/sgml/sc34/document/0322files/iso13250-2nd-ed-v2.pdf,2002-06-16

[6] Pepper S. OASIS Published Subjects:Introduction and Basic Requirements[EB/OI]. http//www.oasis-open.org/committees/geolang,2005-03-23

[7] Garshol L M, Moore G. Topic Maps-Data Model [EB/OL]. http://www.isotopicmaps.org/sam/sam-model/,2004-02-16

[8] Vatant B. Published subjects:from Confucius to topic maps and beyond [J]. Interchange,2002,39(1):25-32

[9] Maicher L, Witschel H F. Merging of Distributed Topic Maps based on the Subject Identity Measure (SIM) Approach. Department of Information Sciences, University of Leipzig, Chair of NLP, Augustusplatz 10-11, 04109 Leipzig, Germany, Sept. 2004

[10] 吴笑凡,周良,张磊,等. 分布式主题图合并中的 TOM 算法[J]. 武汉大学学报,2006,39(5):131-136

[11] Jung M K, Hyopil S, Hyoung J K. Schema and constraints-based matching and merging of Topic Maps[J]. Information Processing and Management,2007,43(4):930-945

[12] 吴克河,马应龙,林鹏程,等. 一种基于本体的语义冲突处理方法[J]. 计算机工程与应用,2007,43(13):182-185

[13] 聂志强. 语义冲突及冲突处理模型的设计[J]. Science & Technology Information,2007,2:491

(上接第 15 页)

[27] Musuvathi M, Park D Y, Chou A, et al. CMC: A pragmatic approach to model checking real code // Proc. OSDI'02. ACM, 2002:75-88

[28] Yang J, Twohey P, Engler D, et al. Using model checking to find serious file system errors // Proc. OSDI'04. San Francisco, 2004

[29] King J C. Symbolic execution and testing. Comm. of the ACM, 1976,19(7):385-394

[30] Khurshid S, Pasareanu C S, Visser W. Generalized symbolic execution for model checking and testing // Proc. TACAS'03. Springer, LNCS 2619, 2003:553-568

[31] Anand S, Pasareanu C, Visser W. Symbolic execution with abstract subsumption checking // Proc. SPIN'06. LNCS 3925, 2006:163-181

[32] Anand S, Pasareanu C S, Visser W. JPF-SE: A Symbolic Execution Extension to Java PathFinder // Proc. of TACAS'07. Springer-Verlag, LNCS 4424, 2007:134-138

[33] Schlich B, Kowalewski S. Model Checking C Source Code for Embedded Systems // IsoLA'05. Columbia, MD, USA, 2005