文章编号:1001 - 9081(2010)01 - 0240 - 03

基于主题图的本体信息检索模型研究

李清茂,杨兴江,周相兵,马洪江
(阿坝师范高等专科学校 计算机科学技术系,四川 郫县 611741)
(lqmhan@126.com)

摘　要:针对本体在定义领域概念时具有规范性、明确性和可共享性等特点,结合主题图对文档资源组织方式具有语义可导航性,提出了一种基于主题图的本体信息检索模型,并给出了模型的形式化定义。选择旅游领域作为研究对象,定义了旅游本体和旅游文档资源主题图,分析了在信息检索模型中利用本体来规范用户自然语言查询输入,识别用户检索意图和扩展查询语义方面的作用,并展示了主题图在语义导航和用户相关度排序方面的价值。最后通过实验表明基于主题图的本体信息检索模型较传统的检索系统有较好的性能。

关键词:本体;主题图;信息检索模型;语义扩展
中图分类号: TP319　　**文献标志码**:A

Research on topic maps-based ontology information retrieval model

LI Qing-mao, YANG Xing-jiang, ZHOU Xiang-bing, MA Hong-jiang
(*Department of Computer Science, Aba Teachers College, Pixian Sichuan* 611741, *China*)

Abstract: Ontology is normative, explicit and reusable when defining the domain concept, so it can be combined with topic maps to organize information resource for semantic navigation. An information retrieval model based on topic maps and ontology was proposed and defined formally. Firstly it specified a domain of tourism document. Secondly it defined the ontology and topic maps of tourism document in order to normalize query that user directly input in natural language, and identified the user's real meaning of search. Thus, it can expand user' semantic search. Therefore, analyzed the effect of the ontology was analyzed, and a valuable function of semantic navigation and sorting the retrieval result correlated with user's query was shown. Finally, the experimental result shows that the topic map-based ontology information retrieval model can perform better than the traditional model.

Key words: ontology; topic map; information retrieval model; semantic extension

0　引言

Ontology 作为具有明确规范化的概念模型语言,它通过提供类(Class)、关系(Relations)、函数(Functions)、公理(Axioms)和实例(Instance)等最基本的模型元素,可以对相关领域的共同概念和共同理解进行定义,达到知识共享的目的。由于 Ontology 具有良好的概念层次结构和知识表达能力,并能够根据一定的语义规则进行逻辑推理,因此非常适合用于构建基于概念和语义的信息检索模型,近几年来国内外的学者都在这方法做了大量的研究并取得了很多积极的研究成果[1-4]。然而很多研究人员却忽略了主题图在信息组织和模型构建方面的突出优点,主题图作为遵循 ISO/IEC13250[5] 规范的知识组织工具,可以通过定义主题、关联及其类别的方式来表达知识概念关系,从而形成领域的结构化知识模型。从知识模型表达的角度来看,本体和主题图都能通过自己的模型元素来定义概念,识别语义,都有共同的语法基础 XML 文档支持能力,不同之处是本体强调概念之间的逻辑推理能力,而主题图则更加重视知识模型对资源对象的导航能力,如果两者结合,将能显著改善基于关键词匹配的检索性能。因此本文提出基于主题图的本体信息检索模型,并选择旅游行业作为研究对象构建领域本体和主题图,在应用主题图与本体来改善检索性能方面作尝试性研究。

1　模型基本定义

定义 1　基于主题图的本体信息检索模型可抽象为五元组概念模型,表示为 $TMOIRM = \langle O, TM, M(tm,o), UI, R(VQ, SemSim(T_1,T_2))\rangle$。

定义 2[6]　领域本体 O 是五元组,即 $O = \langle C, A^c, H^c, R, I^c\rangle$,其中 $C = \{c_1, c_2, \cdots, c_n\}$ 表示概念集合,$|C|$ 表示概念的数目,$|C| = n$;A^c 表示概念 $c \in C$ 的属性集合,$A^c = \{a^c(i) \mid i = 1,2,\cdots, |A^c|\}$,$|A^c|$ 表示概念属性的数目;$H^c(Arc) \subseteq C \times C$ 是一个有向关系,$H^c(c_i \times c_j) \in H^c$, $c_i \in C, c_j \in C$,表示 c_j 是的 c_i 子概念,因此 $H^c(Arc)$ 为有向无环图;R 表示概念之间非层次关系的集合,$R = \{r(c_i, c_j) \mid c_i \in C, c_j \in C\}$,$r(\cdot)$ 表示 c_i 与 c_j 之间的二元连接关系;I^c 表示概念的实例集合,概念的实例集合记为 $I(c)$,$|I(c)|$ 表示概念的实例数目。

定义 3[7]　TM 为文档的主题模型,是关文档资源库的主题图,用七元组来表示,$TM = \langle T_{type}, T_{ass}, T_{occ}, T_{role}, T_{instance}, R_H, R_{acc}\rangle$,其中 T_{type} 表示主题类型集合;T_{ass} 表示关联类型集合;T_{occ} 表示资源指引(资源出处)类型集合;T_{role} 表示关联角色类型集合;$T_{instance}$ 表示主题类型的实例集合;R_H 表示主题层次关系集合;R_{acc} 表示除层次关系外的关联关系集合。

定义 4[8]　$M(tm,o)$ 表示主题图 TM 与本体 O 之间的映射,$M(tm,o) = \langle TM, O, P, Mr\rangle$。若用 $E(TM)$ 表示主题图 TM

收稿日期:2009 - 07 - 21;**修回日期**:2009 - 08 - 24。　**基金项目**:四川省教育厅自然科学基金资助项目(07ZC002)。

作者简介:李清茂(1973 -),男,四川米易人,副教授,硕士,主要研究方向:主题图技术、语义 Web;　杨兴江(1971 -),男,重庆人,副教授,硕士,主要研究方向:语义 Web;　周相兵(1980 -),男,四川仪陇人,讲师,主要研究方向:人工智能、语义网络、系统集成;　马洪江(1968 -),男,四川邛崃人,教授,主要研究方向:人工智能。

中的元素集合,$E(O)$ 表示本体 O 的元素集合,若 $\forall x \in E(TM), \exists y \in E(O), P = \{(p_i, P_{i+1})\}$ 是 TM 与 O 之间的映射参数约束集合,用队列 $Q(P)$ 表示,则映射关系 $Mr = ((E_1(TM) \times E_2(O)), Q(P)) \cup ((E_2(TM) \times E_1(O)), Q(P))$。

若 $Mr = \{mr_1, mr_2, \cdots, mr_i\}$ 是映射关系集合,且满足 $Mr(P) \rightarrow (TM, O)$,因此可以定义主题图 TM 与本体 O 的余弦相似度:

$$Sim(TM, O) = \frac{\sum_{par}[(|TM_{(p,k)} - E(TM)_i|)(|TM_{(q,k)} - E(O)_j|)]}{\sqrt{\sum_{par}(|TM_{(p,k)} - E(TM)_i|)^2 \cdot \sum_{par}(|TM_{(q,k)} - E(O)_j|)^2}} \quad (1)$$

其中:par 表示 $\sum$ 约束条件$(i, j \in Q(P); p \in E(TM); q \in E(O); k \in Mr(P))$,$i, j$代表映射的约束参数集合中的元素,且 $i \neq j$;k 是映射关系集合中的元素,(p,k)、(q,k) 表示主题图 TM 中元素 p、q 与映射关系 k 形成的序对。

定义 5　UI 表示用户接口,用二元组表示,$UI = \langle IN_{query}, Vq: IN_{query} \rightarrow RCset_{in} \rangle$。

IN_{query} 表示用户输入查询信息,表示未经语法和语义规范化处理的关键词、词组或是查询短句。Vq 是用户信息需求逻辑视图生成函数,将用户信息需求映射成符合领域本体 O 要求的相关概念集合,$Vq: IN_{query} \rightarrow RCset_{in}$。

定义 6　$R(VQ, SemSim(T_1, T_2))$ 是获得用户信息需求相关文档的算法,其中 VQ 为查询转换函数,即 $VQ: RCset_{in} \rightarrow Result$,$Result$ 表示检索结果集合,通过 VQ 函数把相关概念集合中元素通过主题图查询引擎转换成用户需求文档,$SemSim(T_1, T_2)$ 表示主题元素语义相似性计算函数,用于检索返回文档相关度排序依据。

2　模型实现

基于主题图的本体信息检索模型,如图 1 所示。

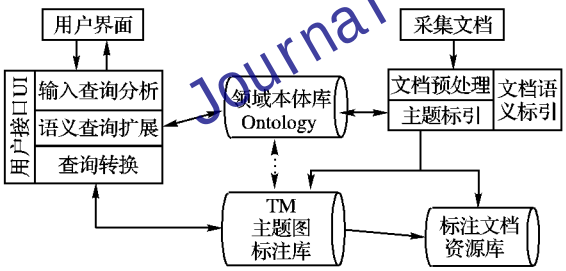

图 1　基于主题图本体信息检索模型

首先建立基于领域知识的本体库。领域本体库在模型中可以为主题图提供规范化的标注词汇和参照标准,用于对文档进行基于主题的语义标注。第二,根据领域本体中确定的基本概念及其关系,采用主题图技术标准对采集的文档进行语义标注,建立文档逻辑视图,并以 XTM 的形式保存在 TM 主题图标注库中,便于检索时建立概念语义与文档资源之间的链接关系。第三,对用户输入进行分析处理,形成符合本体 Ontology 要求的概念词汇。第四,根据本体库 Ontology 中定义概念及关系,对规范化的输入概念词进行语义扩展。第五,启动查询转换,从 TM 主题图标注库中提取相关检索结果呈现给用户。

2.1　构建领域本体库 Ontology

通常情况下,领域本体的构建方法有人工和自动两种方式[1]。本课题选用旅游领域作为本体建模对象,本体建模工具采用 Protege 4.1,手工方式输入基本的建模元素。旅游本体的目标是捕捉相关领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出词汇和词汇间相互关系的明确定义。采用通行的本体工程方法,创建测试用旅游本体片段如图 2 所示。

图 2　旅游本体中的类层次结构

2.2　文档语义标注

文档语义标注是实现语义检索的基础。模型中对采集到的旅游领域相关文档采用主题图标准进行标注,工具采用 OKS(Ontopia Knowledge Suite)。采用主题图标注的优点:1)主题图标准与本体标注兼容,本体中的模型元素概念、属性、关系、实例完全可以采用主题图的建模元素主题、资源指引、关联来实现。本模型中可以通过 $M(tm, o)$ 映射函数来实现主题图 TM 与本体 O 之间的映射,如果实际的工程实践中应该首先定义主题图与本体之间的映射机制,便于计算机自动转换,本模型测试中采用手工方式实现。关于主题图与本体之间的映射机制还应进一步深入的研究。2)主题图标注文档可以实现文档与标注文件分离,标注文件可以 XTM 格式存储在合适的数据库中形成主题图标注库。本模型测试用标注库为 XTM 文件。3)采用 XTM 格式的标准文档可以直接使用 TMQL 查询语义针对标注库进行基于语义的查询[9]。

2.3　输入查询分析

与传统的信息检索模型相区别,理想的检索系统除了提供关键词输入外,还要能提供基于自然语言输入的查询请求。要实现自然语言输入查询,就要对输入短句进行切词和分词。一般中文分词必须借助词典工具库才能实现,本模型中的查询输入是借助 Ontology 语义库来实现的。由于 Ontology 库中包含的领域概念不仅表达了领域共享的知识,还可以通过对本体库中的概念、属性、关系进行索引成为该领域的词典库,用于查询分析中的规范化检索请求。例如,在模型中输入“我想了解在卓克基土司官寨发生的历史大事?”,在没有领域本体支持下,普通分词的结果是“我/想/了解/在/卓/克/基/土司/官/寨/发生/的/历史/大事”。这个分词结果中仅有“土司”、“历史”、“大事”三个词具有检索意义,但与输入请求差别太大。在本体词典的支持下,重新处理的结果是“我/想/了解/在/卓克基/土司官寨/发生/的/历史大事”。去掉与检索无关的词汇,保留“卓克基/土司官寨/历史大事”,即为模型中 IN_{query} = {卓克基,土司官寨,历史大事}。

2.4　语义扩展

语义扩展是实现语义检索的关键,也是保证检索模型查全率和查准率的根本措施。模型中用用户逻辑视图生成函数 $Vq: IN_{query} \rightarrow RCset_{in}$ 来实现语义扩展。语义扩展可以从等价概念、相关关系概念和概念层次关系中的上下位关系三个方向进行扩展。为了控制扩展范围可以参照文献[10],定义语义集合 $SRC(c,r)$、$SRP(c,r)$ 和 $SRS(c,r)$ 分别表示子概念集合、上位概念集合和相邻关系概念集合,其中语义半径 $r(r \geqslant 1)$,

$c \in A$ 表示待扩展语义的概念。

因此逻辑视图生成函数 $Vq: IN_{query} \rightarrow RCset_{in}$ 的算法描述如下：

Vq 算法：逻辑视图生成函数即查询语义扩展算法。

```
输入：IN_query = {c1,c2,…,cn};
                    // 如 IN_query = {卓克基,土司官寨,历史大事}
输出：RCset_in;初始为空,即 RCset_in = ∅。
For Each x ∈ IN_query;C = x;
  For each C ∈ SRC(C,r)
    RCset_in = RCset_in ∪ C.getchild();      // 获取一个子节点
  EndFor
  For each C ∈ SRP(C,r)
    RCset_in = RCset_in ∪ C.getParent();     //获取一个父节点
  EndFor 置 Y = 0
  For each C ∈ SRP(C,r)
    RCset_in = RCset_in ∪ x.getSiblings(); // 获取一个兄弟节点
  EndFor
EndFor
```

2.5 查询转换

查询转换就是把经过语义扩展后形成的相关概念集合 $RCset_{in}$ 中查询元素，通过主题图引擎来访问主题图标注库中的主题图，检索出于相关概念集合 $RCset_{in}$ 中的主题及关联关系，并通过主题资源指引（Occurrence）把检索结果呈现给用户。主题图引擎是通过 TMAPI 的方式来实现，目前常用的 TMAPI 有 TM4J、tiny TiM 、XTM4XMLDB 和 OKS 等。本模型测试用查询转换采用 OKS 知识组件提供的 TMQL-tolog 查询语言直接访问主题图来实现，在编程接口方式下需要查询处理器（Query Processor），其核心组件是 QueryProcessorIF 接口，该组件位于 net. ontopia. topicmaps. query. core 包中[11]。

查询转换还有一项功能是按查询结果相关度排序。由于主题图标引词汇选择具有一定的灵活性，不一定完全采用领域 Ontology 中的规范术语，因此可能会出现主题图与本体表达语义一致而两者的术语词汇存在差异，影响查询排序结果，因而导致用户漏检，为此模型采用主题相似度作为最终排序依据。

关于主题图的相似度国内外学者都进行了相关研究，Lutz Maicher 等人 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[12]，$SIM = \lambda SIM_{\cdot Name} + (1+\lambda) SIM_{\cdot Occs}$；吴笑凡等人 2006 年提出分布式主题图融合中的 TOM 算法[13]；文献[7]中提出了一种多策略主题图相似度判别的 TM-MAP 算法，利用主题名字、主题属性、层次和关联的相似度来综合计算主题图相似度。

$$SIM(t1,t2) = (SIM_{\cdot Name} + SIM_{\cdot acc} + SIM_{\cdot H} + SIM_{\cdot assoc})/4 \quad (2)$$

其中：$SIM_{\cdot Name}$ 是主题名称相似度；$SIM_{\cdot acc}$ 是主题属性相似度；$SIM_{\cdot H}$ 主题层次相似度；$SIM_{\cdot assoc}$ 是主题关联相似度。由于文档是采用主题标引，可以用主题相似度来评价文档相关度，模型中采用 TM-MAP 算法来计算主题图相似度，用户需求文档的相关度由 $SIM(t1,t2)$ 的值来确定。

3 验证分析及比较

实验选用旅游主题图中 6 主题，从网上下载相关电子文档进行测试。分别对文档进行主题标引和关键词标引，采用信息检索领域广泛使用的查准率（Precision）和查全率（Recall）来评价实验结果。

查准率 = 检索到的相关文档数/检索到的全部文档数

查全率 = 检索到的相关文档数/系统全部相关文档数

表 1 关键法与基于主题图的本体信息检索模型比较

组号	主题范围	文档数量	关键词法		语义模型	
			查准率	查全率	查准率	查全率
1	民族文化	10	0.77	0.75	0.87	0.96
2	土司官寨	30	0.60	0.80	0.92	0.94
3	民俗节日	3	0.66	0.90	0.93	0.93
4	卓克基风景区	17	0.85	0.89	0.94	0.96
5	索官赢	3	0.92	0.92	0.92	0.90
6	西索民居	5	0.86	0.78	0.97	0.98

从表 1 提供的分析数据可以看出，由于模型中采用了本体作为概念词典，不仅规范用户查询输入，还能识别检索语义，从而提高了模型的准确率。在模型用户接口部分增加了语义扩展，因此检索模型的查全率比较高。从对比结果来看，该模型的查准率和查全率都比传统的关键词匹配要高。

该模型中还融入了主题图来组织和标引文档资源，不仅可以按照主题相关度排序输出，还可以利用主题工具实现可视化导航，如图 3 所示，检索“卓克基风景区”时系统所展示的主题关系导航图，用户可以沿着主题间展示的路径进行关联查询，不仅能增加查准率和查全率，还能增加系统交互性和易用性。

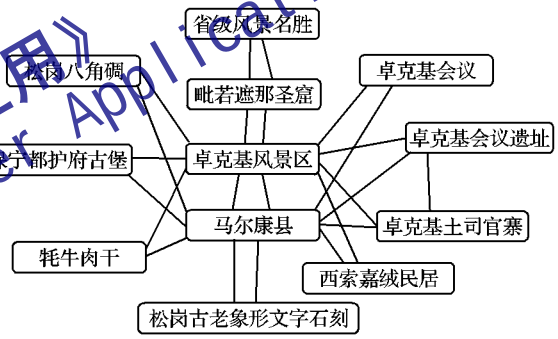

图 3 主题关系导航图

4 结语

将本体和主题图技术融合到信息检索系统中，首先可以利用领域本体中的概念词汇规范用户自然语言查询输入，便于提取有检索意义的概念词汇；其次可以利用本体中定义的概念关系来识别和扩展用户检索语义，从而提高检索模型查全率和查准率；同时也可根据领域本体定义的共享概念，并按主题图技术标准对文档资源进行主题标引，生成文档资源的知识结构图，实现标引库与资源库分离，便于用户根据自己的检索意图在语义层次上的匹配和获取文档资源。

参考文献：

[1] SHUN S B, MOTTA E, DOMINGUE J. Schol Onto: An ontology-based digital library server for research documents and discourse[J]. International Journal on Digital Libraries, 2000, 3(3): 237 – 248.

[2] SU XIAOMENG, GULLA J A. An information retrieval approach to ontology mapping[J]. Data and Knowledge Engineering, 2006, 58(1): 47 – 69.

[3] 程新荣, 杨仁刚, 康丽. 基于 Ontology 的 Web 信息检索方法[J]. 广西师范大学学报: 自然科学版, 2007, 25(2): 100 – 103.

[4] 卢林兰, 李明. 用户 ontology 的构建及其在个性化检索中的应用[J]. 计算机应用, 2006, 26(11): 2635 – 2638.

[5] ISO/IEC 13250: 2000. Topic maps: Information technology[EB/OL]. [2009 – 06 – 09]. http://www.y12.doe.gov/sgml/sc34/document/0129.pdf.

（下转第 248 页）