基于本体的林业领域语义查询扩展模型①

张乃静, 鞠洪波, 纪 平

(中国林业科学研究院 资源信息研究所, 北京 100091)

摘 要: 传统信息检索模型仅考虑考虑关键词本身的匹配程度, 在林业领域信息检索时得到的检索结果不全面或不准确. 为了改善检索质量, 提出了一种基于本体的林业领域语义查询扩展模型. 该模型利用了本体的语义推理的能力和语义结构对关键词进行语义查询扩展, 最终实现提高检索查全率和查准率的目的, 是对传统基于关键词匹配的信息检索模型的语义补充. 结果表明该模型在一定程度上改善了林业领域信息检索结果的查准率和查全率.

关键词: 本体; 查询扩展; 语义推理; 语义相似度; 林业

Modeling Semantic Query Expansion of the Foresty Domain Based on Ontology

ZHANG Nai-Jing, JU Hong-Bo, JI Ping

(Research Institute of Forestry Information Techniques, Chinese Academy of Forestry, Beijing 100091, China)

Abstract: The traditional information retrieval model only considers the matches of key words in the corpus, so it cannot recall more comprehensive and accurate results in the forestry domain. Aiming to solve the disadvantage above, the model of semantic query expansion based on ontology is presented for forestry domain. The semantic query expansion is modeled using the semantic reasoning and structure of ontology, and improves the recall and precise ratio in information retrieval finally. This model is a supplementary semantic function for the traditional information retrieval model. The experiment proves that this semantic query expansion model can increase the recall and precise ratio in information retrieval of forestry domain.

Key words: ontology; query expansion; semantic reasoning; semantic similarity; forest

随着林业领域相关科学研究的不断深入, 林业科学数据量迅速增加, 海量数据在为科研和决策提供便利的同时, 也不断对林业领域数据和成果的共享技术提出新的要求, 如何从海量数据中准确全面的检索出需要的内容, 是林业科学数据共享中亟待解决的问题. 一直以来, 基于字符串匹配原理的传统的信息检索技术在信息检索中应用较为广泛, 由于该技术利用关键词在语料库中全部或部分字符的匹配来进行检索, 不会得出与关键词同义或近义的检索结果, 从而影响检索结果的客观性. 自 Gruber 首次将本体引入信息科学后, 以本体作为知识模型的语义信息检索技术成为了研究热点[1]. 文献[2]结合兴趣都和本体提出一种语义搜索算法; 文献[3]利用本体对语料库进行语义标注, 提高检索性能; 文献[4]基于本体改进了文档特征权重模型; 文献[5]对信息检索中的语义相似度进行了相关研究. 但针对语义查询扩展的研究相对较少.

本文利用林业领域本体中概念之间的语义联系和结构差异, 结合语义推理和语义相似度, 提出一种基于本体的林业领域语义查询扩展模型, 为林业领域语义信息检索和数据挖掘提供理论基础和新的途径.

1 林业领域本体

领域本体是某一领域知识的客观描述, 包含了领域内的概念、概念的属性、概念之间的关系以及属性

① 基金项目:中国林业科学研究院资源信息研究所中央级公益性科研院所基本科研业务费专项资金(IFRIT201304)
收稿时间:2015-06-17;收到修改稿时间:2015-07-24

和关系的约束等[6]. 一般来说, 一个本体由概念、关系、函数、公理和实例 5 个基本元素构成. 林业领域本体的构建步骤如图所示: (1)确定林业领域知识范围; (2)抽取领域内的关键词或术语词集合; (3)根据关键词和术语词建立本体的核心概念集合和概念之间的关系集合; (4)根据关键词和术语词建立属性和属性之间的关系集合; (5)领域专家审核, 判断集合客观性; (6)创建本体实例, 利用本体工具以本体描述语言 OWL(Ontology Web Language)进行形式化编码. 通过以上步骤, 建立基于林业领域标准规范、网络和专业叙词表的林业领域本体, 如图 1 所示, “Thing”表示万事万物, 本体中所有概念和关系都是“Thing”成员.

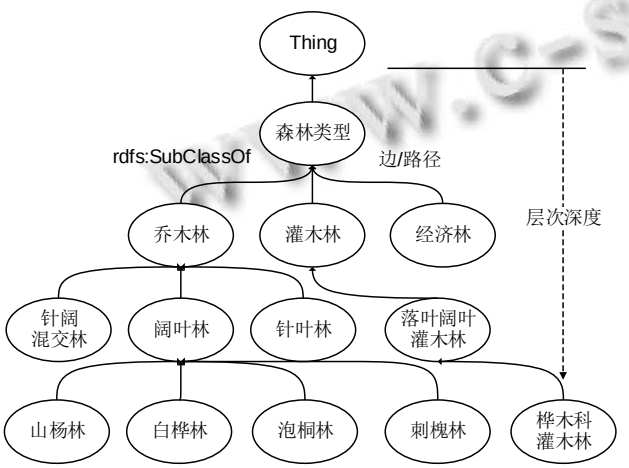

图 1　林业领域本体描述片段

本体中的信息被表示为一个客观事实的集合, 称为陈述, 表示为主谓宾三元组—— {主语, 谓语, 宾语}({Subject, Predicate, Object }), 主语和宾语对应着本体中概念和实例, 谓语对应本体中概念或实例之间的属性关系[7, 8]. 从本质上说, 本体是一个客观事实的集合, 而这些集合是实现语义信息检索的基础.

本体语义查询使用 W3C 提出的标准语言 SPARQL, 其核心实际上是一个图模式匹配的过程[9]. 本体中最简单的图就是包括主谓宾的三元组陈述, 例如 SPARQL 语言片段{ ?forest fd:拉丁学名 “Pinus koraiensis”}表示查询“拉丁学名”为“Pinus koraiensis”的树种是什么(红松).

2　语义查询扩展

查询扩展的基本方法是利用计算机语言学和信息学等相关技术将与原查询关键词相关的词或词组扩展到原查询中, 从而改善检索的查全率和查准率[10]. 基于本体的林业领域语义查询扩展模型的流程如图 2 所示:用户发出查询请求, 首先经过分词系统进行分词, 得到的关键词进入本体分别进行同义词扩展, 当多关键词时进行语义推理扩展, 如果语义推理扩展结果为空则进行语义相似度扩展, 单关键词直接进行语义相似度扩展, 最后将语义查询扩展结果出入到信息检索系统.

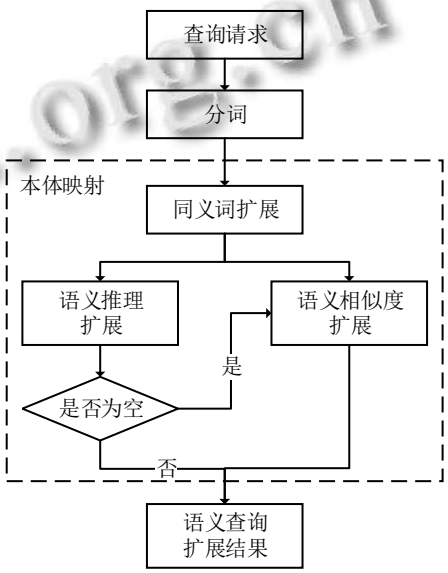

图 2　语义查询扩展流程图

2.1 语义推理查询扩展

本体中主谓宾三元组陈述是查询的语义推理的基础, 当使用语句作为查询词时, 通过分词, 将语句分解为多关键词, 一般可以得到主谓宾结构, 将该主谓宾结构在本体中映射, 可推理出用户查询请求的目的.

林业领域内存在大量的同义专业词汇, 用户在信息检索过程中不可能列举某一概念的所有同义词汇, 因此会影响信息检索结果的准确性, 例如用户查询“红松”, 如果直接进行关键词匹配, 许多包含“海松”和“籽松”的文档无法被检索出, 而“红松”、“海松”和“籽松”的概念是一致的. 为了解决同义词查询扩展的问题, 在构建本体中, 创建资源时以某一特定属性标记同义词, 例如使用“owl:sameAs”标记 “红松”、“海松”和“籽松”之间的属性关系, 在本体中的表示如图 3 所示, 不论用户查询请求是三个同义词中的任何一个, 使用 SPARQL 语言对本体中与之关联的隐含三元组陈述进行查询后可扩展出该词汇的其它同义词.

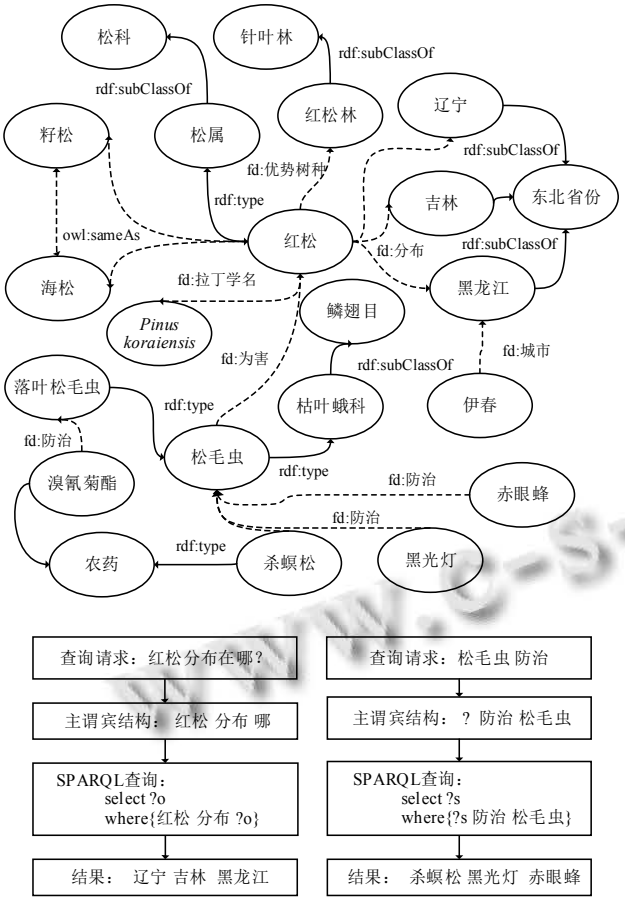

图 3 语义推理查询扩展

语义推理过程中, 以疑问句查询需要作为特例处理, 因为从疑问句中可分解出的主谓宾结构, 疑问代词指代的内容就是用户查询实际的需求, 利用 SPARQL 语言在领域本体中推理出疑问代词所指代的概念或属性, 并将这些资源扩展到查询词, 或直接返回给用户. 如图 3 所示, 用户发出“红松分布在哪里？”的查询请求, 经过分词和词性判断处理, 可得到“红松 分布 哪里”的主谓宾结构, 对该结构进行本体映射, 推理出“红松 分布 <辽宁 吉林 黑龙江>”, 推理的同时需要考虑下位概念的影响, 例如本体中“伊春”是“黑龙江”的下位概念, 但疑问句查询中, 用户目的比较明确, 考虑概念的下位概念如果过多, 会导致查询词的过度扩展而影响查准率, 所以疑问句查询时不使用下位概念扩展.

当查询词为关键词集合时, 由于随机性大, 建立准确的三元组集合较为困难, 需要进行一些词性判断和本体查询操作, 再进行本体映射.

首先判断关键词集合中是否存在动词, 如果有动词, 那么把该关键词作为谓语, 其他名词作为主语或宾语, 进入本体中进行推理, 得出宾语或主语. 例如查询“防治 松毛虫”, 经过推理, 结果为“杀螟松 黑光灯 赤眼蜂”(图 3). 如果没有动词, 则判断关键词集合中是否存在本体中的属性或关系, 如果是属性或关系, 该关键词做谓语, 其他名词作为主语或宾语, 进入本体进行推理宾语或主语, 例如查询“红松 拉丁学名”(图 3), 推理结果为“*Pinus koraiensis*”. 如果关键词集合均不是本体中的属性或关系, 则需判断关键词集合是否是本体中的概念, 如果是则提取这些概念的直接下位概念(具有 SubClassOf 或 type 属性), 作为推理结果. 例如用户查询“东北省份 红松”(图 3), 本体中“东北省份”的下位概念“辽宁 吉林 黑龙江”可以扩展到查询中. 如果关键词与本体之间没有找到直接联系, 则该关键词集合不适用语义推理, 需要对关键词分别进行语义相似度扩展.

2.2 语义相似度查询扩展

2.2.1 语义相似度

领域本体中概念之间的语义相似度是语义信息检索研究的重点内容, 其理论基础来自于离散数学中的图和树的匹配技术[11]. 本体中概念的父类(Class)、子类(SubClass)及其实例(Type)的上下级结构可抽象为树形结构, 如图 1 所示, 本体树形结构中的概念被表示为节点, 例如“森林类型”和“经济林”等均表示为树节点, 概念之间的关系表示为各节点之间的路径, 例如“SubClassOf”表示了概念的上下级关系, 本体树形结构中所有的节点都利用路径进行连接, 表示本体中任何两个概念相互都存在着联系, 这些联系可以使用语义相似度来衡量, 根据树形结构特点和前人研究经验的总结, 概念之间的语义相似度与它们的语义距离、语义重合度以及在本体中的层次深度等因素密切相关. 为方便使用, 研究中构建的语义相似度计算模型取值范围为(0, 1]: 两个概念重合时, 语义相似度为 1; 两个概念联系较小时, 语义相似度趋近于 0.

定义 1. 设 X、Y 是本体中的任意两个概念(或节点), X 与 Y 的最短路径称为它们之间的语义距离, 表示为 Dis(X, Y).

语义距离衡量本体中两个概念在语义上的近义程度. 语义距离与语义相似度呈反比. 当两个概念的语义距离为 0 时, 二者为同一概念, 语义相似度最大; 当两个概念之间路径逐渐增加时, 语义距离随之增大,

语义相似度随之减小. 例如根据图 1 中的本体树形结构, 可以计算 Dis(山杨林, 阔叶林)=1, Dis(山杨林, 桦木科灌木林)=6, 结果与客观事实相符, “山杨林”与“阔叶林”的语义相似度大, 因为“山杨林”是“阔叶林”的一种; “山杨林”与“桦木科灌木林”的语义距离较大, 所以语义相似度与前者相比较小, 因为二者不是同一种森林类型.

定义 2. 设 X、Y 是本体中的任意两个概念(或节点), X 和 Y 与根结点“Thing”的最短路径中所包含的节点数分别表示为 N(X)和 N(Y), 则 X 和 Y 的语义重合度表示 X 和 Y 与根节点最短路径公共节点的个数与总节点个数之和的比值, 记为 Con(X, Y)=(|N(X)∩N(Y)|)/(|N(X)∪N(Y)|).

语义重合度衡量本体中两个概念在语义上的相互重合的程度. 语义重合度与语义相似度呈正比, 如果两个节点与根节点最短路径所含的公共节点数量较多, 那么两个节点表示的概念之间的语义重合程度就越高, 相应的语义相似度越大. 以图 1 为例, Con(山杨林, 白桦林)=0.5, Con(山杨林, 桦木科灌木林)=0.125, 通过比较数据很容易看出“山杨林”和“白桦林”之间的语义重合度相对较大.

定义 3. 设 X、Y 是本体中的任意两个概念(或节点), L(X)和 L(Y)分别是概念 X 和 Y 所处在本体树形结构中的层次, 则 X 和 Y 的层次差表示为 Dif(X, Y)=|L(X)-L(Y)|.

本体中处于同一层次的两个概念所包含的信息量相似; 当两个概念之间的层次差逐渐增大时, 语义相似度就相应减小, 即层次差与语义相似度呈反比关系. 从图 1 可以看出“山杨林”和“白桦林”两个概念处于本体树的层次是相同的, 所以 Dif(山杨林, 白桦林)=0, “山杨林”和“针叶林”处于不同层次, Dif(山杨林, 针叶林)=1, 从语义角度上来说, “山杨林”和“白桦林”是“乔木林”的两个实例化对象, 所包含的信息量相同, 语义相似度相对较大; “针叶林”是“乔木林”的一个子类, 所以“山杨林”和“针叶林”包含的信息量是不同的, 语义相似度相对较小.

定义 4. 设 X、Y 是本体树中的两个概念(或节点), 结合定义 1~3, X 和 Y 之间的语义相似度计算公式如下:

$$Sim(X,Y)=\frac{|N(X)\cap N(Y)|}{\alpha\cdot Dis(X,Y)+\beta\cdot Dif(X,Y)+|N(X)\cup N(Y)|} \quad (1)$$

式中 Dis(X, Y)表示概念之间的语义距离; (|N(X)∩N(Y)|)/(|N(X)∪N(Y)|)表示概念之间的语义重合度; |L(X)-L(Y)|表示概念之间的层次差, α、β 分别是调节参数, 取值范围均为(0,1], 该研究中 α=1, β=1.

2.2.2 语义相似度查询扩展模型

在信息检索和文本数据挖掘研究中, TF-IDF 模型是一种常用的文档特征权重模型, 用来表示文档与用户查询之间的相关程度[12]. 对于文档集合 D={d1, d2…di}来说, 其 TF-IDF 模型表示为:

$$TF\text{-}IDF=TF_{ij}\times IDF_{ij}=\frac{f_{ij}}{\sum_i f_{ij}}\times\log\frac{N}{n_i+1} \quad (2)$$

其中 TF_{ij} 是文档 d_j 中词汇 t_i 的词汇频率, IDF_{ij} 表示 t_i 在文档集合 D 中的逆文档词频, f_{ij} 表示词汇 t_i 在文档 d_j 中出现的次数, N 表示文档数据集中文档的数量, n_i 表示词汇 t_i 出现过的文档数量.

当单纯使用 TF-IDF 模型进行信息检索时, 检索结果反映了关键词在文档集合中的匹配程度, 与关键词语义相关的词汇并没有考虑, 例如检索“阔叶林”时, 关键词在某一文档中的 TF-IDF 权重为 0, 但如果该文档中包括“白桦林”和“山杨林”等在语义上与“阔叶林”十分相近的内容时, 则 “阔叶林”在该文档中的权重应体现出这些相关词汇的存在, 使该文档能够被检索. 所以, 应结合关键词与文档中词语之间的语义相似度进行查询扩展. 结合语义相似度与 TF-IDF 模型, 语义相似度查询扩展后关键词在文档中的特征权重公式如下:

$$TF-IDF_{sim}(d_j,t_i)=TF-IDF(d_j,t_i)+\sum_k[TF-IDF(t_k)\times Sim(t_i,t_k)] \quad (3)$$

即在查询扩展后的文档召回时, 重新计算关键词在文档中的特征权重, 将词语 t_i 与文档 d_j 中其它相关词语 t_j 之间在本体中的语义相似度 $Sim(t_i,t_k)$作为词语 t_k 对词语 t_i 的文档词汇特征权重的贡献度, 并重新对检索结果进行排序. 但如果将领域本体中所有 $Sim(t_i,t_k)>0$ 的概念全部作为查询扩展必然会导致查准率较低, 所以在实际应用中, 需要对语义相似度设置适合的阈值来控制关键词查询扩展的幅度.

3 实验分析

3.1 实验设计

为了对基于本体的林业领域语义查询扩展模型进行实验和评价, 从国家科技基础条件平台中心林业科

学数据平台(www.forestdata.cn)获取林业领域相关的文档, 共保存为 2066 个文本文件(txt), 以这些文件作为实验语料库. 实验内容包含两个方面: 首先使用不同关键词进行语义相似度查询扩展, 查询过程中, 对语义相似度设置不同的阈值, 分析在不同语义相似度阈值条件下查询扩展模型的表现, 以确定最佳阈值; 然后使用不同关键词分别进行使用和不使用语义查询扩展两种方法对语料库进行检索, 以比较两种检索结果的性能差异.

实验采用信息检索中常用的查全率、查准率和 F-Score 作为评价指标, 查全率衡量检索结果中正确文档在语料库所有正确文档中的百分比, 查准率衡量检索结果中正确文档在检索结果中的百分比, F-Score 是综合评价指标, 考虑了查准率和查全率的平衡性, F-Score 值得大小与检索系统性能值呈正比关系[13].

实验运行环境如下:

硬件平台: Intel I5 2500K, 8GB 内存;

操作系统: Windows 7 SP1 x64;

本体构建工具: Protégé 4.1;

语义 Web 框架: Jena-API 2.10;

开发环境: JDK 1.7;

开发工具: Eclipse 4.2;

搜索框架: Lucene 4.2.

3.2 实验结果

不同语义相似度阈值下查询扩展的检索结果(表 1)表明, 语义相似度的阈值与检索系统的查全率呈反比趋势, 与查准率呈正比趋势, F-Score 值呈单峰形变化趋势, 在语义相似度为 0.5 时 F-Score 值最大, 说明当语义相似度设置为 0.5 时, 语义相似度查询扩展表现最好.

表 1 不同语义相似度阈值下查询扩展检索结果

阈值	查全率	查准率	F-Score
0.0	100.0	12.2	0.22
0.1	100.0	13.6	0.24
0.2	98.1	18.4	0.31
0.3	95.5	24.5	0.39
0.4	90.2	40.1	0.56
0.5	85.3	55.7	0.67
0.6	75.1	54.9	0.63
0.7	50.5	54.2	0.52
0.8	44.6	58.3	0.50
0.9	37.1	60.2	0.46
1.0	36.8	62.1	0.46

语义查询扩展实验结果如表 2 所示, 当使用基于本体的林业领域语义查询扩展模型时, 信息检索系统的查全率为 100.0%; 查准率为 58.3%; F-Score 值为 0.737, 均优于未使用语义查询扩展的信息检索系统.

表 2 语义查询扩展实验结果

模型	查全率(%)	查准率(%)	F-Score
语义查询扩展模型	100.0	58.3	0.737
未使用查询扩展模型	79.6	48.1	0.600

从查全率-查准率曲线(图 4)中同样可以看出, 未使用语义查询扩展的传统检索模型无法检索到关键词在语料库中语义相关的内容, 检索结果的查全率较低, 查准率也并不理想. 相比之下, 使用了基于本体的林业领域语义查询扩展模型的信息检索系统在查全率和查准率方面比的传统检索模型有明显的提高.

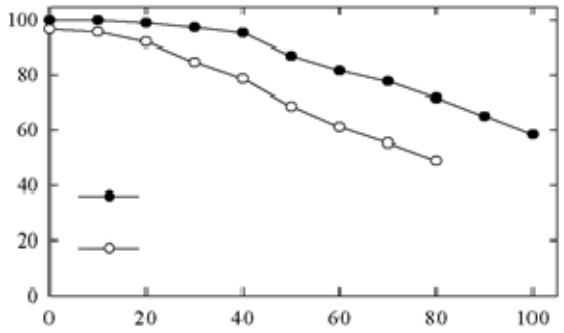

图 4 查全率-查准率曲线

4 结语

利用本体的语义推理的能力和语义结构, 提出了一种基于本体的林业领域语义查询扩展模型, 该模型可以实现关键词的语义查询扩展, 最终实现提高检索查全率和查准率的目的, 是对传统基于关键词匹配的信息检索模型的语义补充. 但该模型还有以下缺点需要改进: 该模型的核心是领域本体, 而构建领域本体是一项十分复杂的工作, 由于研究者对领域知识分类的理解存在差异, 构建的领域本体结构也各不相同, 在一定程度上导致了本体的异构型, 该模型很大程度上依赖于本体的语义结构, 也为本体的复用带来一定的困难, 所以该模型的使用可能具有一定的局限性, 如何制定领域内本体的构建标准将是以后本体研究的重点. 语义查询扩展过程中, 因为该模型未对算法进行优化, 如果使用了大量的关键词进行查询检索, 语义推理和语义相似度将执行大量的穷举计算, 导致检

索效率的降低, 如何利用关键词的词性制定有效的规则和算法来实现高效的语义推理也是一个有意义的研究方向.

参考文献

1 Thomas RG. Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, 1995, 43(5): 907–928.

2 孙萍萍.基于兴趣度和本体自适应学习的语义搜索算法研究.计算机应用与软件,2013,30(5):137–139.

3 吴振忠,王曼,宋婧文,蒋运承.一种基于领域本体的论文检索方法的研究与应用.计算机应用与软件,2013,30(10):177–180.

4 张乃静,鞠洪波,纪平.基于本体的林业领域文档特征权重模型.计算机工程与应用,2013,49(18):20–23.

5 王旭阳,万里.信息检索中语义相似度算法研究.计算机工程与应用,2014,50(10):124–127.

6 张乃静,鞠洪波,纪平.本体构建理论在林业科学数据共享中的应用研究.西北林学院学报,2013,28(6):151–156.

7 Alexander S, Paul B. RelExt: A tool for relation extraction from text in ontology extension. In: Yolanda G, Enrico M, Benjamins VR, Mark AM, eds. The Semantic Web-ISWC 2005. Berlin. Springer-Verlag Berlin Heidelberg. 2005. 593–606.

8 程显毅,施佺,沈学华,田宇贺.属性和属性值组合的概念模板.北京大学学报(自然科学版),2013,49(1):15–19.

9 Ilianna K, Birte G. Optimizing sparql query answering over owl ontologies. Journal of Artificial Intelligence Research, 2013, 48(1): 253–303.

10 王忠民,霍艺伟,邓万宇.基于环境信息的移动搜索个性化查询扩展.计算机科学,2013,40(9):182–184.

11 刘宏哲,须德.基于本体的语义相似度和相关度计算研究综述.计算机科学,2012,39(2):8–13.

12 Gerard S, Edward AF, Harry W. Extended boolean information retrieval. Communications of the ACM, 1983, 26(11): 1022–1036.

13 Bing L. Web 数据挖掘.第 2 版.北京:清华大学出版社,2009.