Digital Library Forum 数字图书馆论坛
http: // www.dlf.net.cn

基于本体的科技文献检索框架与技术实现*

□ 王莉 梁冰 白海燕 / 中国科学技术信息研究所 北京 100038

摘要：文章从提高科技文献检索质量的视角出发，提出基于本体的科技文献检索框架，就本体构建、文献语义空间、查询请求重构、检索过程等方面进行研究，并给出关键算法。指出本检索框架与现有研究相比，主要特征包括：基于规则自动生成文献资源的语义扩展模型；构造“特征词汇-文献-概念”三层子网结构的文献信息空间；引入用户兴趣模型，强调有关用户的这些知识将对新的检索策略的产生和发展产生影响。

关键词：本体检索框架，本体构建，语义空间，查询重构，语义检索
DOI：10.3772/j.issn.1673—2286.2012.07.006

1 引言

传统的科技文献检索主要通过查询请求与文献特征标识之间的简单匹配关系来获得查询结果。文献特征标识即索引词项，尽量采用经过规范化处理的词或代码，试图从形式和内容两个方面确切描述一篇文献。这些索引词项可以是从文献中直接提取出来的，也可以由人主观指定。不管其来源如何，将文献分解为一组词的集合，或多或少都将丢失部分语义信息。简单的字符串匹配方式则不可避免地存在表达一致性问题，一方面无法捕获用户的真实意图，另一方面也无法满足用户在其知识不完整的情景下的文献检索需求。

传统科技文献检索存在的诸多缺陷，使得人们开始期待“更加智能”的语义搜索技术，希望搜索结果与用户需求能够更精准地匹配，希望搜索系统能够具备人脑的逻辑与分析能力，发现用户的潜在需求，进而实现知识发现与分析预测评估。为了实现这一目标，人们在人工智能、机器学习、自然语言处理等领域展开大量研究工作，直到本体理论引入信息检索领域，为语义检索提供了重要的解决思路和实现途径。

信息检索领域对本体的研究[1-5]主要集中在本体构建、语义标注、相似度计算、概念查询扩展、查询匹配、查询推理等方面，普遍采用算法改进和数学建模的研究方法，旨在改善传统检索的效果。随着本体技术逐渐成熟，相关应用实践也越来越多，涌现出大量本体信息检索模型和原型系统[6-12]。归纳起来，本体在信息检索中的应用主要集中在改善检索效果与实现集成检索两个方面。改善检索效果是在单一检索系统中利用本体对文献资源进行语义标注，建立基于概念的文档索引；同时在用户检索过程中，借助本体分析用户输入的检索条件，通过人机交互的方式获得准确的信息需求。集成检索则是利用本体提供的共享概念模型，跨越多个异构文献资源实现语义集成。本文的研究属于第一个方向，从提高科技文献检索质量的视角出发，提出基于本体的科技文献检索框架，在该框架下探讨利用本体实现语义查询的技术和方法，通过挖掘科技文献的语义信息，利用本体重写检索表达式，提高查询质量要求，逐步迈向智能检索。

2 基于本体的科技文献检索框架

随着本体技术应用实践的深入，业界已经逐渐形

* 本研究得到国家十二五科技支撑计划课题“信息资源自动处理、智能检索与STKOS应用服务集成”（编号：2011BHA10B05）、中央级公益科研院所基本科研业务费专项资金中国科学技术信息研究所2011年度预研基金项目“信息与文献标准术语搜索关键技术研究”（编号：YY-201127）资助。

成一条清晰的研究思路，即：在本体的帮助下，采用两组概念分别表示“文献的语义”和“用户信息需求的语义”，查询过程自然地转换为具有语义关系的概念的匹配过程。从这一基本思路可以看到，对本体的利用主要是文献和用户信息需求的语义表示，即集中在文献预处理和查询扩展两个环节，复杂多样的词间关系并没有带到检索匹配过程中。预处理之后形成的信息空间，本质上并没有跳出图书馆传统的分类法和叙词表的框架，没有充分发挥本体在知识表示方面的强大功能。

针对这一问题，本文对该基本思路进行扩展，提出基于本体的科技文献检索框架，如图1所示。

2.1 本体构建

本文提出的科技文献检索模型主要用到了通用、领域和应用三类本体。其中，应用本体包括针对文献著录规范的语义扩展模型和用户兴趣模型，是本节重点论述的内容。通用本体是对最一般化概念及概念之间关系的描述，如空间、时间、事件、行为等，与具体的应用无关。文中采用国际上通用的WordNet词汇库，用以实现最基本的名称规范（见本文“2.3 查询请求重构”部分内容）。领域本体提供特定领域内的概念定义和概念关系，构建过程需要领域专家的参与，一般在确定领域范围后先考虑重用其他本体的可能性，继而在现有本体的基础上采用手工方法或本体学习技术生成。在科技文献检索模型中主要用领域本体对文献进行语义标注，实现词空间到概念空间的映射，完成文献语义空间的构建（见本文“2.2 文献语义空间”部分内容）。在本文中，假定领域本体已经存在，不讨论具体的构建问题。实际应用中，领域本体可以用相关领域内的词表代替。

（1）针对文献著录规范的语义扩展模型构建

针对文献著录规范的语义扩展模型利用基本书目信息自动构建，类似于一个书目本体，但又不是一个简单的书目信息描述框架，更多地面向实际应用，融合了数据结构和完整性约束等信息。自动构建的基本思路是建立数据模式到概念模式间的映射规则，基于规则对文献元数据进行抽取和转换。相关研究主要集中在数据库领域，并且已经提出了一些通用的、经过正确性证明的映射规则。例如，文献[13]抽象出ER模式与OWL本体间的概念对应关系。文献[14,15]给出本体元素与关系数据库模式各元素之间的对应关系表。文献[16]提出了关系数据库与本体之间的转换关系。文献[17]提出了一种直接通过数据库反向工程到一个本体知识库的正确性可保持转换算法。这些研究成果可以直接借用。考虑大多数文献检索系统在数据底层仍然采用关系数据库模式，本文给出从满足3NF的关系数据库结构抽取本体框架的两个主要步骤及通用转换规则[18]。（需要注意的是，这些转换规则在应用时是有先后顺序的。）

步骤一：从基础文献集合中抽取模式信息，主要包括关系名、属性名、主键、外键、完整性约束。

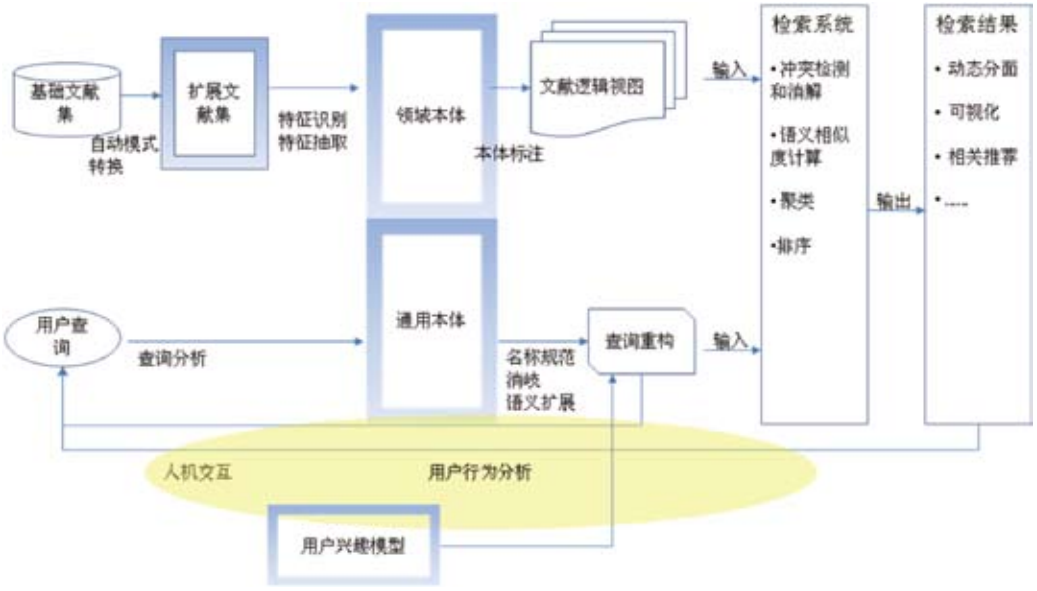

图1 基于本体的科技文献检索框架

步骤二：分析主键、外键、属性等信息，从关系、属性及实体完整性三个方面逐一完成本体的转换。

● 关系映射规则

关系映射是模式转换的第一步，主要遵循以下基本规则：

R1-当一个关系的主键没有引用其他关系的属性时，该关系是一个基本关系，创建一个概念。

R2-当同一个实体分散在多个关系中时，它们共享同一个概念。

R3-当一个关系包含依赖于另一个关系时，创建子概念及概念层次关系。

R4-那些只包含两个外键属性的关系，本身只是表示关系之间多对多的联系，在映射时不需要创建概念。

R5-缺省映射规则，即，当其他规则不适用时，创建一个新的概念。

● 属性映射规则

属性映射包括关系属性到本体属性的映射，以及属性的参照完整性约束映射。一般地，一个关系属性可以映射为本体概念的一个属性，但是由主键和外键体现的参照完整性描述了模式中关系之间的引用。这种表示引用的关系属性，可能并不需要映射为一个概念的属性，也可能需要映射为多个概念的属性。

通用的属性映射规则包括：

R1-当一个基本关系作为本体中的概念已经存在时，关系属性转换为该概念的一个属性。

R2-当一个基本关系包含依赖于其他关系时，关系属性转换为父概念的一个属性。

R3-当一个关系表示另外两个关系间的关联时，该关系的属性分别作为被引用关系的属性，并且两个属性保持引用关系。

R4-当关系A引用关系B，且关系A对应的概念包含于关系B对应的概念B’，则关系A中的属性作为概念B’的一个属性。

R5-当关系A引用关系B，且关系A对应的概念与关系B对应的概念无关时，每个关系中的属性作为其对应概念的一个属性，且两个属性保持引用关系。

● 实体完整性约束映射

实体完整性是基于主键的，一个主键由一个或多个关系属性组成。实体完整性约束要求主键中任一关系属性不能为空。除此之外，为了尽量保持关系模式中的语义信息，还需要完成DEFAULT、NOT NULL和UNIQUE等约束条件的映射。

概括地说，语义扩展模型的自动构建过程以实例数据的语义获取为目标，主要借助的是数据库模式与本体之间映射的相关研究成果，基于数据模式到本体框架之间的通用映射规则对文献元数据进行抽取和转换，一步完成本体建模和实例数据导入。事实上，在这一构建过程中，无需区分本体概念在哪里结束，实例数据从哪里开始。建成的扩展文献集合支持文献语义空间的构建。

(2) 用户兴趣模型构建

用户兴趣模型是信息检索领域的研究热点[19-23]，不仅仅是对用户个体的一般性描述，更是一种面向算法的，具有特定数据结构的、形式化的用户描述。用户建模是从用户静态信息和动态行为中归纳出可计算的用户模型的复杂过程。本文只关注用户的当前检索行为，将同一个用户（用IP标识）在一定时间的查询式集合作为一个会话，在一个会话过程中基于向量空间模型构建实时用户兴趣偏好，用于即时调整检索结果的排序策略。

模型构建的基本思路是：

● 采用用户输入的一组查询词表示用户兴趣，构造初始模型。初始状态下，检索结果直接反馈给用户。

● 用户在检索过程中会点击浏览感兴趣的文献，系统采用该文献的特征向量对初始用户模型进行修正。修正后的用户模型影响接下来的检索过程。

● 随着检索次数的增多，用户兴趣模型不断更新，越来越准确地反映用户兴趣。

以上方法构建的用户兴趣模型以实时调整排序策略为目标，关注的是用户的当前兴趣，只需要少量计算空间和物理开销即可实现。需要注意的是，用户输入的查询词可以反映用户的行为，但不能完整刻画用户的兴趣，更确切地说，采用这种方式构建的用户模型表示的是用户的目标，而非兴趣。因此这种构建方法只适用于本文设定的特定目标，并不适合单独用于用户建模。

2.2 文献语义空间[24-26]

经典检索模型用一组关键词（索引项/标引词）来描述每一篇文献，这些词被认为彼此之间相互独立。然而，在实际应用中关键词之间有时是相关的，借助本体的帮助不仅能对这些关键词进行规范，更重要的是揭示词间关系，进而构建多维的逻辑视图，形成文献语义

空间。

文献语义空间的构建思路如下：

（1）以文献题名、摘要、关键词为输入，采用自动语义标注工具，提取表征文献内容的词汇，形成文献和特征词汇两层子网。

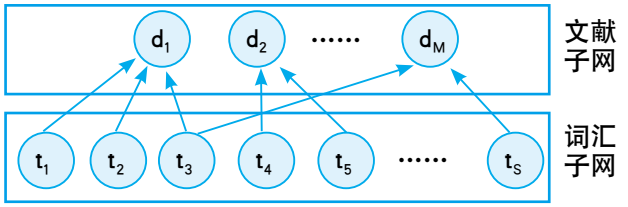

图2 词汇子网—文献子网简单拓扑结构

其中：

- 所有文献和特征词汇分别形成文献子网和词汇子网。
- 特征词汇和文献之间的连接反映了二者之间的依赖关系，可用词汇-文献相关度（term-document association）表示。
- 任意两篇文献之间没有连接。

（2）采用本体标注技术，将一篇文献归入一个或多个概念类目之下，图2的拓扑结构扩展为“特征词汇-文献-概念”三个层次，如图3所示。

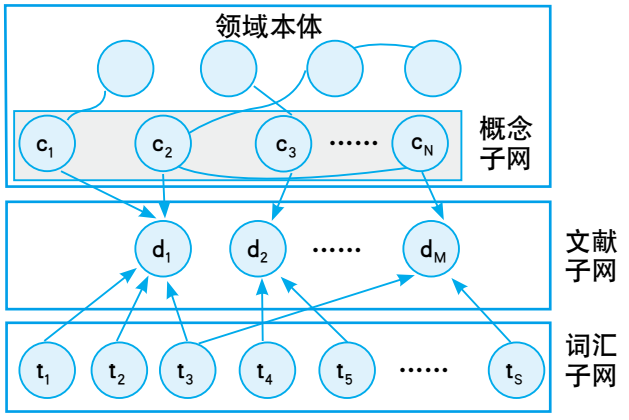

图3 概念子网—文献子网—词汇子网扩展拓扑结构

其中：

- 概念子网是领域本体的一个子集，概念之间保留了源本体中的语义关系，采用有向弧表示，可以表示为概念-概念相似度（concept-concept similarity）。
- 概念与文献之间的关系可用概念-文献相关度（concept-document association）表示。

（3）经过特征词汇抽取与本体标注之后，基础文献集转化为“词汇-文献-概念”三层子网。在文献层中文献与文献之间，在特征词汇层中词汇与词汇之间都是没有直接连接的，所有的关联汇聚在概念层，通过概念之间的语义关系建立起来。

借鉴文献[27]提出的K2CM（keyword-to-concept method）思想，词汇和概念之间可以从两个角度建立关联：一方面，考虑词汇通过文献和概念构成的所属关系，利用“概念-文献-词汇”所属程度描述词汇-概念相关度；另一方面，结合词汇-概念的共现、文本距离和分布特征等多种因素。这两个方面对词汇-概念相关度的贡献性差别忽略不计，分别计算两个相关度权值，再直接相乘，可以得到完整的词汇-概念相关度（term-concept association）。

以上步骤采用预处理方式完成，最终形成的文献语义空间分别生成概念-文献相关矩阵、词汇-文献相关矩阵、词汇-概念相关矩阵、概念-概念相关矩阵，建立相应的索引。

2.3 查询请求重构

查询扩展是公认的能够有效提高查全率的技术，其基本思想是利用与查询词相关的词语对查询进行修正和补充，以便能够找到更多的相关文档。借助本体实现的是查询词的概念扩展，包括语义相似与相关两个方面。这里使用“查询条件重构”强调这是一个查询请求优化过程，并不仅仅包括“扩展”，也可能根据需要“精简”。

查询请求重构的主要思路如下：

（1）采用分词技术将用户输入的查询请求切分为一组查询词。

（2）计算各查询词词义之间的语义相关性，根据“整体语义相关最大化”原则对查询词进行语义消歧。

参考文献[28]提出的方法，利用WordNet和WordNet Domains知识库，从结构相关性和领域相关性两个方面综合计算词义间的相关性。相关性最大的词义组合决定了各查询词的词义。如果得到的最大化词义组合数多于一个，则用词频信息作为第二选择标准。

由于科技文献检索中用户输入的查询请求多是短文本，查询词个数不会很多，所以计算所有查询词的各

种词义组合并不会产生组合爆炸的情况。

（3）查询词消岐之后，直接利用WordNet中的同义词关系进行第一步扩展。

（4）利用预处理生成的概念-词汇相关度矩阵，可以得到一个相应的查询词-概念相关度矩阵（keyword-concept correlation matrix），进而计算查询-概念相关度（query-concept relevance）。依据top-k策略进行筛选，将扩展概念和用户提交的查询词合并生成新的查询条件。

通过查询请求重写，一方面可以捕捉用户真正的信息需求，利用事先构建的语义索引；另一方面消除原始查询中的冗余条件和不必要的操作，真正实现查询的优化。

2.4 检索过程

文献检索是查询请求与文献集合的匹配过程，这一过程从用户提交查询请求时启动，查询重构应该看作检索过程的第一步，接下来的核心问题是查询词与文献的相关度（query-document relevance）计算。

（1）查询词与文献的相关度计算

在构建文献信息空间时，文档经过特征抽取和本体标注之后，可以分别用一组自由词和一组概念描述。概念之间具有语义相关性，而自由词是那些从文献中直接抽取的特征词汇，假定彼此之间相互独立。同样，查询请求也可以转化为一组自由词和一组概念。

按照经典向量模型匹配原理，查询与文献的匹配过程可以分解为概念向量和自由词向量两个向量模型的匹配。概念向量计算中引入概念之间的相似度权值，自由词向量计算则可以直接采用传统余弦法。

相关度计算之后形成有序的检索结果集。

（2）用户兴趣与文献相关度（interest-document relevance）计算

用户兴趣采用查询词和文献特征向量描述，同样可以使用向量空间模型匹配策略计算文献与用户兴趣之间的相关度，对检索结果集进行二次排序，检索结果输出。

捕获用户点击/浏览/下载的文献特征向量，修正用户兴趣模型。修正后的用户模型影响接下来的检索过程。

（3）浏览模型

查询具有很强的目标性，但是用户输入查询词的时候意图可能是模糊的，他的兴趣不在于提交一个简单的查询，而是浏览文献空间中感兴趣的内容。因此，建立一种护航式浏览模型非常重要，以检索结果集为基础动态分面，充分利用预构建的关联索引，通过自动聚类和关联展示，引导用户发现知识内容和知识点。相关内容不在本文研究范围内，可以参考文献[29-31]。

3 定义及关键算法

文献资源检索是从文献集中找出和用户信息需求相关内容的过程，不仅要求检出与用户信息需求相关的所有文献，同时还必须避免检出大量的无关文献。在这一过程中，相关度计算与排序非常重要，尤其在引入语义关系之后，计算的复杂度大大增加。以下给出本框架中采用的基本定义及关键算法。

3.1 基本定义

定义1. D表示文献集合，其中包含M篇文献，$d_j(j=1,\cdots,M)$表示集合D中第j篇文献。

定义2. T表示给定文献集合D的特征词汇集合，其中包含S个特征词汇，$t_i(i=1,\cdots,S)$表示集合T中第i个词汇。

定义3. C表示给定文献集合D的概念集合，其中包含N个概念，$c_i(i=1,\cdots,N)$表示集合C中第i个概念。

定义4. $Q=(q_1,q_2,\cdots,q_k)$表示给定查询，$q_k(k=1,\cdots,K)$表示查询请求中的一个查询词。

定义5. I表示描述用户兴趣的特征词汇集合，$i_k(k=1,\cdots,K)$表示用户兴趣中的一个特征词。

3.2 关键算法

（1）词汇-文献相关矩阵：TD-CM（term-document correlation matrix）

$$\text{DT-CM}=\begin{pmatrix} tda_{1,1} & \vdots & tda_{1,M} \\ \cdots & \cdots & \cdots \\ tda_{S,1} & \vdots & tda_{S,M} \end{pmatrix}$$

其中$tda_{i,j}$（term-document association）表示第i个词汇与第j篇文献的相关度，综合考虑特征词汇在文献中出现的位置、频率等因素，采用文献[1]中的TF-IDF改进算法计算。

（2）概念-文献相关矩阵：CD-CM（concept-document correlation matrix）

$$CD\text{-}CM=\begin{pmatrix} cda_{1,1} & \vdots & cda_{1,M} \\ \cdots & \cdots & \cdots \\ cda_{N,1} & \vdots & cda_{N,M} \end{pmatrix}$$

其中$cda_{i,j}$（concept-document association）表示第i个概念与第j篇文献的相关度，同样可以采用$tda_{i,j}$的计算方法得到。

（3）概念-概念相似度：cca（concept-concept similarity）

任意两个概念c_i和c_j之间的相似度记为$cca(c_i,c_j)$，具体的计算方法很多[33-39]，这里给出文献[39]中的方法。

$$cca(c_i,c_j)=1-\sqrt{\frac{1}{OL(c_i,c_j)}\times\frac{|Dep(c_i)+Dep(c_j)|+1}{Dep(c_i)+Dep(c_j)}\times\frac{1}{Int(c_i,c_j)}\times Dist(c_i,c_j)}$$

其中：

- $OL(c_i,c_j)$表示两个概念的语义重合度，通过概念c_i和c_j共有的上位概念集合计算。
- Dep(c)表示节点c的深度，根节点的深度设为1，非根节点的深度等于其父节点深度+1。
- $Int(c_i,c_j)$表示两个概念之间的强度，用概念节点的通路中最短路径所跨的边的强度之和来计算。
- $Dist(c_i,c_j)$表示两个概念的语义距离，用概念节点的通路中最短路径所跨的边的权重之和来计算。

注意，公式中边的强度和权重含义不同。边的强度引入本体库的统计特征，采用概念的信息量计算得到；边的权重则与概念宽度（用其直接子节点数目表示）相关。

（4）词汇-概念相关度：tca（term-concept association）

任意词汇t_k和概念c_i之间的相关度记为$tca_{k,i}$，采用文献[27]的方法定义如下：

$$tca_{k,i}=cw_{k,i}\bullet aw_{k,i}$$

其中：

- $cw_{k,i}$表示词汇t_k和概念c_i的共现程度权值（co-occurrence weight），基于有效窗口内词汇和概念的局部共现性计算得到。

$$cw_{k,i}=\frac{1.0}{|\Gamma_i|}\sum_{j=1,\cdots,|\Gamma_i|}=\frac{tpf_{k,ij}\bullet\log(\frac{m_{k,ij}}{M}+1.0)}{\log(avgdist_{k,ij}+1.0)}$$

– 设Γ_i为概念c_i的同义词集，$c'\in\Gamma_{i,j}=1,\cdots,|\Gamma_i|$，其中$c'_{ij}$是概念ci的同义词或入口词。

– $tpf_{k,ij}$表示(t_k, c'_{ij})在文献集合中出现的频率。

– $M_{k,ij}$表示(t_k, c'_{ij})在文献集合中出现的文献数目。

– $avgdist_{k,ij}$表示词汇-概念对(t_k, c'_{ij})在W大小窗口中位置距离的平均值。

– 有效窗口的确定参见文献[40]。

- $aw_{k,i}$表示词汇t_k对概念c_i所属程度的权值（attaching weight），基于词汇-文献-概念所属关系计算。

$$aw_{k,i}=\log(\frac{N}{n_k}+1.0)\bullet tf_{k,i}\bullet\log(\frac{l_{k,i}}{l_i}+1.0)$$

– n_k表示词汇t_k根据文献-概念关系映射到概念上的概念数目。

– $tf_{k,i}$表示词汇t_k通过文献映射到概念c_i的词频统计量。

– l_i表示概念c_i下的文献数，即概念c_i下的文献空间大小。

– $l_{k,i}$表示词汇t_k出现在概念c_i下文献空间中的文献数。

（5）查询-概念相关度：qcr（query-concept relevance）[27]

查询Q与任意概念c_i的相关度记为：$qcr(Q,c_i)$。

根据定义4，一个查询Q可以用一组查询词描述，查询词与概念的相关度可以直接采用词汇-概念相关度tca表示，得到一个查询词-概念相关矩阵KC-CM（keyword-concept correlation matrix）。

$$KC\text{-}CM=\begin{pmatrix} tca_{1,1} & \vdots & tca_{1,N} \\ \cdots & \cdots & \cdots \\ tca_{K,1} & \vdots & tca_{K,N} \end{pmatrix}$$

由各个查询词的概念向量的线性组合计算$qcr(Q,c_i)$，如公式所示：

$$qcr(Q,c_i)=\sum_{k=1\cdots K}w_k tca_{k,j}$$

其中，w_k表示查询词在整个查询请求中的权重，可以采用Rocchio相关反馈算法[32]设置。

（6）查询-文献相关度：qdr（query-document relevance）

查询Q与任意文献d_m的相关度记为$qdr(Q,d_m)$。

查询Q经过重构之后可以转换为N维概念向量和S维无对应概念的自由词表示，记作$Q=(c_1,c_2,\cdots,c_N,t_1,t_2,\cdots,t_s)$；同理，文档$d_m$也可以表示为$(c'_1,c'_2,\cdots,c'_N,t'_1,t'_2,\cdots,t'_s)$。相关度$Qdr(Q,d_m)$计算转化为N维概念向量和S维自由词向量的相似度之和，定义为：

$$qdr(Q,d_m)=qdrc(Q,d_m)+qdrt(Q,d_m)$$

其中：

- $qdrc(Q,d_m)$表示N维概念向量相似度，计算公式中引入了概念之间的相关度权值[27]。

$$qdrc(Q,d_m) = \sum_{i=1,\cdots,N} \sum_{j=1,\cdots,N} \gamma_i \eta_i cca(c_i,c'_j)$$

  - γ_i表示概念c_i在查询Q中的权重，即$qcr(Q,c_i)$。
  - η_j表示概念c'_j在文档d_m中的权重，即$cda_{j,m}$。

- $qdrt(Q,d_m)$表示S维自由词向量相似度，直接采用向量余弦夹角计算方法。

（7）用户兴趣-文献相关度：idr（intrest-document relevance）

用户兴趣与任意文献d_m的相关度记为$idr(I,d_m)$，可以直接采用余弦法计算。

4 结语

与现有研究相比，本文提出的检索框架具有以下特征：

（1）保留传统文献著录方法，在基础文献集之上，借用数据库模式与本体之间映射的相关研究成果，基于规则自动生成文献资源的语义扩展模型，基础文献集合转化为扩展文献集。文献资源的语义扩展模型类似于一个书目本体，由于基于数据库模式和实例构建，更多地考虑了数据结构和完整性约束等。

（2）在扩展文献集、通用本体、领域本体基础上构成的文献信息空间是一个知识库，既包含了来自领域本体/通用本体的概念及概念间的关系，也包含文献实例，形成一个包含“特征词汇-文献-概念”三层子网结构的信息空间。

（3）引入用户兴趣模型，关注对用户认知与行为问题的讨论，强调有关用户的这些知识将对新的检索策略的产生和发展产生影响。

文献语义空间的构建是本文的核心，第3节给出的主要相关度算法（词汇-文献、概念-文献、词汇-概念等）属于全局分析方法，可以最大限度地发掘词间关系，但是在文献集合较大时，性能是其主要瓶颈。虽然文中强调这是一个预处理过程，在建立索引之后可以获得较高的检索效率；但是在实际应用中，需要充分考虑时间和空间上的可行性，以及在文献集合改变之后的更新代价。

参考文献

[1] 陈欣,李晓菲.基于领域本体的专业文献信息检索研究[J].现代图书情报技术,2009(7/8):59-64.
[2] 宋华.本体向量文献检索模型研究[J].情报探索,2010,11:3-5.
[3] 徐静,孙坦,黄飞燕.近两年国外本体应用研究进展[J].图书馆建设,2008,8:84-90.
[4] 张娜,李宝敏.语义检索及其关键技术研究[J].计算机技术与发展,2006,16(11):22-25.
[5] 吴鸿汉.语义搜索若干关键问题研究[D].南京:东南大学计算机科学与工程学院,2010.
[6] CASTELLS P, FERNÁNDEZ M, VALLET D. An adaptation of the vector-space model for ontology-based information retrieval[J]. IEEE transactions on knowledge and data engineering, 2007,19(2):261-272.
[7] VALLET D, FERNÁNDEZ M, CASTELLS P. An ontology-based information retrieval model [EB/OL]. [2012-03-07]. http://ir.ii.uam.es/~search/publications/eswc05.pdf.
[8] SY M-F, RANWEZ S, MONTMAIN J, et al. User centered and ontology based information retrieval system for life sciences [J/OL]. BMCbioinformatics, 2012,13(Suppl 1):s4. [2012-03-07]. http://www.biomedcentral.com/1471-2105/13/S1/S4.
[9] MAHESWARI J U, KARPAGAM GR.A conceptual framework for ontology based information retrieval [J]. International Journal of Engineering Science and Technology, 2010,2(10):5679-5688.
[10] KRUMMENACHER R, TOMA I, FENSEL D, et al. Instrumenting and Monitoring the LarKCReasearch Infrastructure [EB/OL]. [2012-03-07]. http://www.larkc.org/wp-content/uploads/2011/03/IST-Africa2011_InstrumentingLarKC.pdf.
[11] BONINO D, CORNO F, FARINETTI L,et al. 2004 Ontology driven semantic search [J]. WSEAS Transaction on Information Science and Application 1, 1597-1605.
[12] BHAGDEV, RAVISH, CHAPMAN, et al. Hybrid Search: Effectively Combining Keywords and Semantic Searches [C]// European Semantic Web Conference 2008. 5th European Semantic Web Conference, 1-5 June 2008, Tenerife (Spain).
[13] XU Z-M, CAO X, DONG Y-S, et al. Formal approach and automated tool for translating ER schemata into OWL ontologies [J/OL]. Advances in Knowledge Discovery and Data Mining, 2004(3056):464-475.

[14] 许卓明,黄永菁.从OWL本体到关系数据库模式的转换[J].河海大学学报(自然科学版),2006,34(1):95-99.
[15] 朱姬凤,马宗民,吕艳辉.OWL本体到关系数据库模式的映射[J].计算机科学,2008,35(8):165-169,205.
[16] 周扬.基于关系数据库的本体映射方法研究[M].长春:吉林大学,2006:31-35.
[17] 彭劲松,徐德智.一种关系模式到OWL DL本体的翻译方法[J].计算机工程与应用,2008,44(12):166-169.
[18] 刘强.关系数据库语义集成关键技术研究[D].北京:中国科学院软件研究所,2008.
[19] HUANG C-K, CHIEN L-F, OYANG.Y-J.Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs [J].Journal of American Society for Information Science and Technology, 2003,54(7):638-649.
[20] 陈一峰,赵恒凯,余小清,等.基于本体的用户兴趣模型构建研究[J].计算机工程,2010,36(21):46-48,51.
[21] 李建廷.基于简化ODP的用户兴趣模型[J].计算机工程与科学,2010,32(5):121-123.
[22] 顾雅枫.基于用户兴趣模型的信息检索研究[D].兰州:兰州大学计算机应用技术专业,2009.
[23] 王昕光.基于关键词依赖的用户兴趣模型建模方法的研究[D].上海:上海交通大学计算机应用技术专业,2009.
[24] 杨断利,黄勇,王克俭,等.基于聚类BNR扩展模型的信息检索研究[J].计算机工程与应用,2008,44(13):137-140.
[25] ZAZO ÁF, FIGUEROLA CG, BERROCAL JLA, et al. Reformulation of queries using similarity thesauri [J]. Information Processing and Management, 2005,41(5):1163-1173.
[26] FRAENKEL AS, KLEIN ST. Information retrieval from annotated texts [J]. Journal of the American Society for Information Science, 1999, 50(10):845-854.
[27] 田萱,杜小勇,李海华.语义查询扩展中词语-概念相关度的计算[J].软件学报,2008,19(8):2043-2053.
[28] 王瑞琴,孔繁胜.基于查询扩展和词义消歧的语义检索[J].情报学报,2010,29(1):16-21.
[29] 赵琦,张智雄,孙坦.文本可视化及其主要技术方法研究[J].现代图书情报技术,2008(8):24-30.
[30] 周宁,文燕平.检索结果的可视化研究[J].中国图书馆学报,2002,6:48-50,53.
[31] SHNEIDERMAN B, FELDMAN D, ROSE A. Visualizing Digital Library Search Results with Categorical and Hierarchical Axes [C]// Proceedings of ACM Digital Libraries 2000, San Antonio, Texas, June 2-7, 2000.
[32] JOACHIMS T.A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization [C/OL].[2012-03-07]. International Conference on Machine Learning (ICML), 1997. http://www.cs.cornell.edu/people/tj/publications/joachims_97a.pdf.
[33] 黄果,周竹荣.基于领域本体的概念语义相似度计算研究[J].计算机工程与设计,2007,28(10):2460-2463.
[34] 高炜,梁立,张云港.基于图学习的本体概念相似度计算[J].西南师范大学学报(自然科学版),2011,36(4):64-67.
[35] 梅翔,孟祥武,陈俊亮,等.SSCM:一种语义相似度计算方法[J].高技术通讯,2007,17(5):458-463.
[36] GANESAN P, GARCIA-MOLINA H, WIDOM J. Exploiting hierarchical domain structure to compute similarity [J]. ACM Transactions on Information Systems. 2003,21(1):64-93.
[37] 文坤梅,李瑞轩,卢正鼎,等.一种语义搜索中的关联关系排序方法[C]//2009年中国计算机大会,915-925.
[38] 徐建民,田晋坤,付婷婷.基于共现分析法改进的PFIBF方法[J].情报杂志,2010,29(10):163-166.
[39] 张忠平,赵海亮,张志慧.基于本体的概念相似度计算[J].计算机工程,2009,35(7):17-19.
[40] LU S, BAI S. Quantitative analysis of context field in natural language Processing [J]. Chinese Journal of Computers, 2001,24(7): 742-747.
[41] 郁君甫,刘国华,唐军军,等.基于本体的关系数据库关键词语义查询扩展方法[J].燕山大学学报,2010,34(3):231-235,240.
[42] 王珊,张俊,彭朝晖,等.基于本体的关系数据库语义检索[J].计算机科学与探索,2007,1(1):59-78.

作者简介

王莉，硕士，中国科学技术信息研究所研究员，全国信息和文献标准化技术委员会技术协作分技术委员会委员，研究方向：数字图书馆。E-mail: wangli@istic.ac.cn
梁冰，博士，中国科学技术信息研究所高级工程师。E-mail: liangb@istic.ac.cn
白海燕，硕士，中国科学技术信息研究所研究员。E-mail: bhy@istic.ac.cn

Research on S&T Information Retrieval Framework Based on Ontologies

Wang Li, Liang Bing, Bai Haiyan / Institute of Scientific & Technical Information of China, Beijing, 100038

Abstract: In order to improve the quality of scientific and technical information retrieval, this paper proposes an ontology-based information retrieval framework, makes in-depth study on ontology construction, semantic space, query reformulation and retrieval process and gives the key algorithms. Compared with other studies, the main features of the framework include: to automatic generate the semantic expansion model based on literature's schema rules; to construct an information space with three-tier subnet structure(words-literature-concept); to work with a simple user interest model during retrieval process.

Keywords: Ontology based information retrieval, Ontology construction, Semantic space, Query reformulation, Semantic retrieval

（收稿日期：2012-03-07）