

# 基于本体的语义检索方法

王旭阳, 尉醒醒

(兰州理工大学 计算机与通信学院, 甘肃 兰州 730050)

**摘要:** 为有效解决本体查询扩展的语义边界问题, 提出一种基于本体和局部共现的查询扩展方法。计算概念在本体中的语义相似度和局部文档中的相关度, 构成概念之间总的关联度来筛选扩展词; 在此基础上, 设计一个扩展词权重的计算方法, 考虑原查询词的重要程度以及扩展词与原查询词的关系, 改进查询与文档相关度的计算方法。实验结果表明, 在设计的语义检索系统中, 该方法在保证查全率的基础上能有效提高查准率。

**关键词:** 本体; 查询扩展; 语义相似度; 查全率; 查准率

中图分类号: TP391 文献标识码: A 文章编号: 1000-7024 (2016) 09-2538-05

doi: 10.16208/j.issn1000-7024.2016.09.046

## Semantic retrieval method based on ontology

WANG Xu-yang, WEI Xing-xing

(College of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China)

**Abstract:** To solve the semantic boundary problem of ontology query expansion effectively, the method of query expansion based on ontology and local co-occurrence was proposed. The semantic similarity in the ontology and the relevance in the local documents were computed for concepts. The overall correlation between concepts was composed to select expansion words. Based on that, the calculation method of expansion words weight was presented. The importance of the original query words and the relationship between the expansion words and the original query words were considered. And the calculation method of relevance between query and document was improved. Experimental results show that, in the design of semantic retrieval system, the precision is effectively improved using the method based on guaranteeing the recall.

**Key words:** ontology; query expansion; semantic similarity; recall; precision

### 0 引言

基于本体的语义检索<sup>[1-3]</sup>与基于关键词的传统检索相比, 检索性能有所提高, 其中查询扩展技术作为解决“词不匹配”、“信息迷向”等问题的有效手段备受学者的关注和研究。基于关键词的查询扩展方法<sup>[4-6]</sup>, 不仅存在自身不可避免的缺点, 而且对查询请求的扩展也缺少语义上的理解。基于本体的查询扩展<sup>[7-9]</sup>虽然能解决关键词查询扩展存在的缺陷, 但是也有人质疑扩展脱离了待检索的语料集以及存在语义边界确定困难的问题。此外, 由于用户关心的往往只是排在前面的检索结果, 如果排序靠前的文档与用户的查询相符则认为是检索成功, 否则对用户来说检索是不理想的, 所以设计一个排序上相对准确的算法显得尤为重要。针对上述问题, 改进了基于本体的查询扩展算法, 同时为了满足用户的需求改进了查询-文档相关度的计算。

### 1 基于本体和局部共现的查询扩展方法

基于本体的查询扩展是根据构建的领域本体为用户的查询请求筛选相关的概念加入原查询, 从而达到扩展原查询的目的。然而这种方法仅仅是依靠建立好的本体来扩展概念, 并没有考虑概念在要检索的文档中的有关信息, 因此扩展时加入了概念之间在文档中的相关度, 以便更准确的衡量概念之间的关系。

#### 1.1 基本思想

设置初始查询各概念的权重都为1, 提交检索系统后得到与查询相关的前 $n$ 篇局部文档集 $S$ 。利用概念在本体中的相似度和在局部文档中的相关度, 构造概念之间的关联度矩阵, 并根据关联度大小选出各初始查询概念的候选扩展概念, 之后计算候选扩展词与整个查询的相关度进一步对候选扩展概念进行筛选。

收稿日期: 2015-09-22; 修订日期: 2015-11-27

作者简介: 王旭阳 (1974-), 女, 甘肃定西人, 硕士, 副教授, 研究方向为智能信息处理、数据挖掘、模式识别; 尉醒醒 (1989-), 女, 安徽宿州人, 硕士研究生, 研究方向为智能信息处理。E-mail: wei163xingxing@163.com

1.2 语义相似度

在描述的本体层次结构中, 下层概念是对上层概念的细化, 当概念之间最短路径上的有向边个数相等时, 位于下层概念之间的相似度要大于上层的。如概念对 (内存、外存) 和概念对 (RAM、ROM), 路径距离都是 2, 显然后一对概念的相似度较大。

定义 1 给定有向边  $F \rightarrow S$ , 它的深度为节点  $F$  的深度, 则该有向边所表示的语义距离  $Dist(S, F)$  为

$$Dist(S, F) = 1/W(S, F) \quad (1)$$

式中:  $W(S, F) = \frac{1}{2^{Deep(F)}} + \frac{1}{2^{Deep(F)-1}} + \dots + \frac{1}{2}$  为该有向边的权重。

定义 2 概念  $C_1$  和  $C_2$  之间总的语义距离  $Dist(C_1, C_2)$

$$Dist(C_1, C_2) = Link[C_1, Nc(C_1, C_2)] + Link[C_2, Nc(C_1, C_2)] \quad (2)$$

$$Link[C_1, Nc(C_1, C_2)] = \sum_{n \in path(C_1, Nc(C_1, C_2))} Dist(n, parent(n)) \quad (3)$$

其中,  $Nc(C_1, C_2)$  表示节点  $C_1$  和  $C_2$  的最近公共祖先节点,  $path(C_1, C_2)$  表示在本体层次结构中  $C_1$  到  $C_2$  的最短路径上所有有向边的集合,  $Link[\cdot, Nc(C_1, C_2)]$  表示节点与最近公共祖先节点的语义距离。

我们用下面的公式将语义距离转化成语义距离相似度

$$Sim(C_1, C_2) = 1 - \frac{Dist(C_1, C_2)}{2(Maxdep - 1)} \quad (4)$$

式中:  $Maxdep$  表示本体层次结构的最大深度。

1.3 语义相关度

局部分析法是用词项在局部文档中的词频信息来扩展初始查询。如果选出的局部文档与查询的相关性很大, 那么该方法就能很好的扩展查询, 相反, 会得到大量与原查询不相关的词语。

概念在一定的文本窗口范围内的共现性分析<sup>[10]</sup>从某种程度上反映了它们之间的相关度, 共现度越大, 说明相关度也越大。本文以一篇文档为有效窗口, 则在有效窗口范围内概念  $C_1$  和  $C_2$  的共现频度为

$$cf(C_1, C_2 | d) = \log(tf(C_1 | d) + 1.0) \times \log(tf(C_2 | d) + 1.0) \quad (5)$$

式中:  $tf(\cdot | d)$  指在概念在文档  $d$  中出现的次数。

由于我们要在整个局部文档集  $S$  中分析概念  $C_1$  和  $C_2$  的关系, 所以在  $S$  中的平均共现频度为

$$acf(C_1, C_2 | S) = \frac{\sum_{d \in S} cf(C_1, C_2 | d)}{N} \quad (6)$$

式中:  $N$  为共同出现概念  $C_1$  和  $C_2$  的文档数。

概念之间的相关度与它们在文档中的距离、在文档中共同出现的次数以及文档集中出现两者的文档数有关<sup>[11]</sup>。因此用下面给出的公式计算概念之间的共现度权值<sup>[12]</sup>

$$cw(C_1, C_2) = \frac{acf(C_1, C_2 | S) \times \log(f_{C_1, C_2} + 1.0)}{\log(avs(C_1, C_2) + 2.0)} \quad (7)$$

式中:  $f_{C_1, C_2}$  为出现概念  $C_1$  和  $C_2$  的文档数占局部文档集总数的比率,  $avs(C_1, C_2)$  为两者在局部文档集中的平均距离。

1.4 扩展概念的选择

得到概念之间在本体中的语义相似度和在局部文档中的相关度后, 两者共同构成概念之间的关联度  $Rel(C_1, C_2)$

$$Rel(C_1, C_2) = Sim(C_1, C_2) \cdot cw(C_1, C_2) \quad (8)$$

同一个概念之间关联度我们定义为 1, 所以由以上分析和计算可以得到本体中概念之间的关联度矩阵  $RelM$

$$RelM = \begin{bmatrix} 1 & Rel(C_1, C_2) & \dots & Rel(C_1, C_n) \\ Rel(C_2, C_1) & 1 & \dots & Rel(C_2, C_n) \\ \vdots & \vdots & \ddots & \vdots \\ Rel(C_n, C_1) & Rel(C_n, C_2) & \dots & 1 \end{bmatrix}$$

基于本体和局部共现的查询扩展结合了局部查询扩展的优点和本体语义查询扩展的优点, 使得概念之间关联度的衡量更加准确。

对于初始查询经预处理映射成本体中的概念, 表示成  $Q(C_1, C_2, \dots, C_k)$ 。查找关联度矩阵, 当某个概念与初始查询概念的关联度  $Rel(C_1, C_2) > \lambda_1$  时, 该概念则被选为候选扩展概念。然而单个扩展词并不能完全真实的反映查询意图, 所以我们根据概念与整个查询的关系对候选扩展概念进行再一次的筛选。其计算方法如下

$$Rel(Q, C_i) = \sum_{C_k \in Q} Rel(C_i, C_k) \quad (9)$$

当概念  $C_i$  与初始查询  $Q(C_1, C_2, \dots, C_k)$  的相关度  $Rel(Q, C_i) > \lambda_2$  时则将概念  $C_i$  作为最终的扩展概念。查询扩展算法描述如下:

- (1) 对用户提交的查询请求  $Q$ , 结合本体中的概念经分词和概念的提取后, 得到初始查询概念集  $Q(C_1, C_2, \dots, C_k)$ 。
- (2) 初始查询概念集中的各概念的权重设为 1, 送交检索系统检索, 选取检索结果的前  $n$  篇文档作为局部文档集  $S$ 。
- (3) 在本体通过式 (4) 计算语义相似度, 利用式 (7) 在局部文档中计算相关度, 构造概念之间的相关度矩阵  $RelM$ 。
- (4) 查询概念关联度矩阵, 根据参数  $\lambda_1$  为每个初始查询概念  $C_i$  选取候选扩展概念集  $C_i(C_{i1}, C_{i2}, \dots)$ 。
- (5) 用式 (9) 计算查询和候选扩展概念之间的相关度, 将相关度  $> \lambda_2$  的扩展词作为最终的扩展词。

2 查询-文档相关度计算

检索系统将检索结果返回给用户后, 用户只关心排在前面的检索结果, 所以计算查询和文档的相关度关系到检索系统的性能和用户对检索系统的评价。在计算方法上既要考虑到复杂度又要兼顾计算的准确性, 计算太复杂会影响

响响应时间, 太简单又不能准确的反应查询-文档之间的相关度大小。

### 2.1 权重计算

对于查询和文档来说, 不同的词语或者概念具有不同的重要程度。查询中词语或概念权重的分配是计算查询-文档相关度的基础。大多数查询扩展方法选择, 传统的 Rocchio 方法计算扩展词的权重

$$W(q|Q_{new}) = \alpha \cdot W(q|Q) + \beta \cdot \frac{\sum_{d \in S} W(q|d)}{n} \quad (10)$$

式中:  $W(q|Q)$  为词语  $q$  在初始查询  $Q$  中的权重, 一般由  $q$  在  $Q$  中的频度得到;  $W(q|d)$  表示  $q$  在文档  $d$  中的权重, 通常可由 TF-IDF、BM250 等方法得到;  $n$  为局部文档集  $S$  中的文档数。

然而上述方法存在以下问题: ①原始查询经分词得到查询关键词, 这部分词用上述公式计算时若按词频表示权重, 则它们的  $W(q|Q)$  都是相等的; 而对于扩展词来说不在原查询中出现, 它们的  $W(q|d)$  都为 0。即公式前一部分的计算在上述两种情况下并不能细致区分他们的权重。②分配权重时, 并没有考虑各原查询词的重要程度以及扩展词与原查询的关系。

基于对上述问题的考虑, 本文首先通过局部文档得到概念的权重  $W(C_i|d)$

$$W(C_i|d) = \frac{\sum_{d \in S} \log(tf(C_i|d) + 1.0)}{n \cdot \text{Max}\{W(C_i|d)\}} \quad (11)$$

式中:  $tf(C_i|d)$  为概念  $C_i$  在文档  $d$  中出现的次数, 除以  $\text{Max}\{W(C_i|d)\}$  是为了将权重归一化。

初始查询概念集  $Q(C_1, C_2, \dots, C_k)$  中的每个概念按式 (11) 计算得到权重, 而对于他们的候选概念  $C_i(C_{i1}, C_{i2}, \dots)$ , 概念  $C_{ij}$  的权重既要考虑到原查询概念  $C_i$  的重要程度, 又要考虑到与概念  $C_i$  的关联程度, 以及扩展词本身在局部文档中的重要程度, 因此本文给出一个扩展词权重  $W(C_{ij})$  的计算方法

$$W(C_{ij}) = \alpha W(C_i|d) + \beta \text{Rel}(C_i, C_{ij}) + \gamma W(C_{ij}|d) \quad (12)$$

式中:  $\alpha, \beta, \gamma$  为调节参数, 且  $\alpha + \beta + \gamma = 1$ ,  $W(C_{ij}|d)$  按式 (11) 计算得到。

### 2.2 相关度计算

用户最终的查询向量由初始查询概念和扩展概念组成, 原查询概念和各扩展概念的权重分别由式 (11) 和式 (12) 计算得到, 如果一个概念存在于多个原查询概念的扩展概念集中, 取较大的权重作为该概念在查询向量中的权重。则扩展后的  $N$  维用户查询向量为  $QE(C_1, C_2, \dots, C_N)$ 。

待检索的文档集  $D$  中的每一篇文档  $d$ , 经过预处理后, 对  $d$  构造  $N$  维文档向量  $Td(t_1, t_2, \dots, t_N), t_i \in C$ 。对于文档中的概念我们用方 TF-IDF 法计算其权重

$$\text{Rel}(QE, d) = \sum_{i=1, \dots, N} \sum_{j=1, \dots, N} W_{QE}(C_i) W_d(C_j), \quad \text{Rel}(C_i, C_j) \text{ 且 } \text{Rel}(C_i, C_j) > \lambda_1 \quad (13)$$

式中:  $W_{QE}(C_i)$  为查询向量中概念的权重,  $W_d(C_j)$  表示文档中概念的权重。在扩展时, 原查询概念与某个概念  $C_i$  的关联度大于  $\lambda_1$  时被选为该概念的候选扩展概念, 所以为了避免相关度小的概念降低整个文档的相关度, 本文加了一个  $\text{Rel}(C_i, C_j) > \lambda_1$  的限制条件。

## 3 实验及结果

### 3.1 实验设计

为验证本文方法的有效性, 设计了一个小型检索系统, 其结构如图 1 所示, 主要包括本体的构建及推理、查询和文档预处理、查询扩展和检索模块。

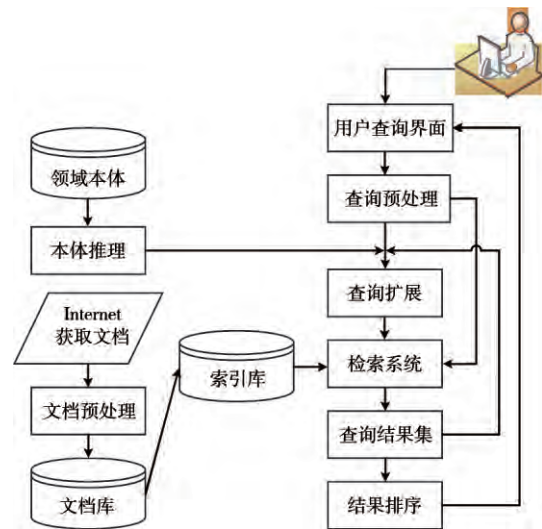


图 1 检索系统结构

构建领域本体时, 我们选择计算机领域, 借助领域术语和相关计算机专家利用 Protégé 工具构建的一个简单的由 OWL 语言描述的计算机领域本体, 其中概念之间的主要关系为 is\_a 和 part\_of 关系。对构建好的本体利用开源工具 Jena 对此领域本体进行相应的推理操作。对构建好的本体利用开源工具 Jena 对此领域本体进行相应的推理操作。

查询扩展是利用基于本体和局部共现性分析的算法对初始查询概念进行扩展, 使扩展后的查询更能反映用户的查询请求。

查询预处理主要过程如下:

- (1) 导入计算机领域本体相关概念词典  $C$ , 该词典由 Jena 推理得到。
- (2) 使用 ICTCLAS 自带的方法将概念词典  $C$  写入用户词典, 并更新词库。
- (3) 利用 ICTCLAS 对用户的查询请求进行分词和去停用词处理。

(4) 分词结果与本体概念词汇集进行比较, 提取出查询概念及该概念的同义概念, 供查询扩展使用。

文档预处理的最目的是建立索引。由于标准的测试文本集 (如 TREC) 包含的领域范围较广, 很难建立一个与之对应的本体来验证该检索方法的有效性, 所以实验中利用网络爬虫工具从 Web 上得到 2800 篇有关计算机领域的文档作为测试集, 其处理过程如下:

(1) 通过与 HTML 网页中标记符 (如 <title>、<body> 等) 匹配的方法来抽取 Web 文档中的文本信息并进行存储。

(2) 对提取出的文本信息进行分词, 类似于查询预处理过程的分词同样需要借助本体概念集 C 完成。

(3) 分词结果与本体中的概念进行比较, 提取出文档概念集, 比较时如果有同义概念则添加到文档概念集中。

(4) 利用 TF-IDF 算法计算概念的权重, 满足一定阈值的概念作为文档的语义向量即  $Td(t_1, t_2, \dots, t_N), t_i \in C$ 。

(5) 构建倒排索引, 结构如图 2 所示, 其中概念列表存储在内存, 而倒排列表和文档集存在磁盘。

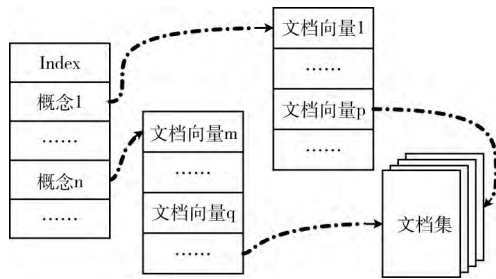


图 2 文档概念索引结构

### 3.2 实验结果与分析

以标准的查全率 Recall、查准率 Precision 和 F 测试值为主要的评价指标<sup>[13]</sup>, 所以本文查询扩展算法仍采用该方法评价系统性能, 以  $Pr@n$  为辅助评价指标。并用文献 [12] 中根据标准指标 MAP (mean average precision) 定义的  $AP@m$  指标评价查询-文档相关度算法

$$Recall = \frac{\text{检索出的相关文档数}}{\text{文档库中全部相关文档数}} \quad (14)$$

$$Precision = \frac{\text{检索出的相关文档数}}{\text{检索出的文档总数}} \quad (15)$$

F-测试值综合反映了查全率和查准率两个方面, 能够较好的反应查询性能

$$F\text{-测试值} = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (16)$$

$Pr@n$  表示的是在检索出的前 n 篇文档中的查准率, 采用该评价指标主要是由于用户通常只关心靠前的检索结果, 如检索结果的第一页或者前两页

$$Pr@n = \frac{\text{检索结果的前 } n \text{ 篇中相关文档数}}{n} \quad (17)$$

$AP@m$  描述前 m 个检索结果中相关文档的准确率平均值

$$AP@m = \frac{1}{k} \sum_{rank_i \leq m} \frac{i}{rank_i} \quad (18)$$

式中: k 表示前 m 个结果文档中相关文档的个数, i 表示前 m 个结果文档中第 i 个相关文档,  $rank_i$  表示第 i 个相关文档在结果文档中的排序。

实验中将无查询扩展 (即初始检索结果)、基于本体的查询扩展 (即只根据本体中的语义相似度进行查询扩展) 和本文查询扩展方法相比较, 在比较中文档的排序采用传统的向量空间模型。实验中利用不同的查询关键字分别进行了 30 次查询, 其平均查询性能的比较见表 1。

表 1 3 种扩展方法的查询性能比较

| 查询扩展方法    | Recall | Precision | F-measure | Pr@10 | Pr@20 |
|-----------|--------|-----------|-----------|-------|-------|
| 无查询扩展     | 0.432  | 0.354     | 0.389     | 0.411 | 0.382 |
| 基于本体的查询扩展 | 0.578  | 0.602     | 0.590     | 0.664 | 0.643 |
| 本文查询扩展方法  | 0.593  | 0.685     | 0.635     | 0.732 | 0.706 |

从表中可以看出, 与未扩展的检索相比, 基于本体的扩展和本文查询扩展方法在查询性能上均有所提高。而本文查询扩展方法与基于本体的查询扩展方法相比, 在保证查全率的基础上查准率有所提高。其主要原因是, 在查询扩展时, 原查询得到候选概念后, 又根据查询-概念的相关度对候选扩展概念进行了第二次筛选, 保证了扩展词的质量。

本文对查询-文档相关度排序算法进行了改进, 所以为了验证该算法的有效性, 实验中用  $AP@k$  为评价指标, 分别将传统向量空间模型排序和本文查询-文档相关度排序进行比较, 其扩展算法均采用本文查询扩展算法。由于用户往往最多只关心前 20 个检索结果, 所以在实验中取  $k=20$ 。选取 30 次查询中的 10 次进行  $AP@20$  的比较, 其中相关性的判断采用专家评判的方法, 结果如图 3 所示。

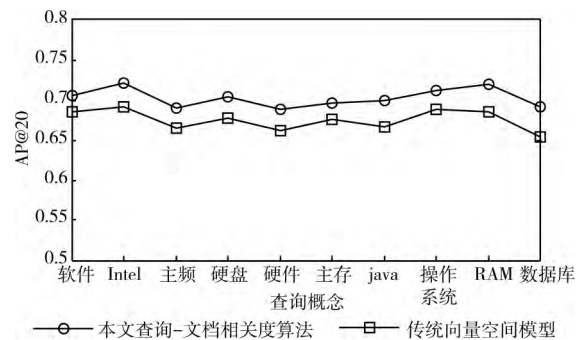


图 3 查询-文档相关度算法比较

在查询扩展中, 扩展规模也是影响查询性能的重要因素。扩展词过少不能充分表达用户的查询需求, 而扩展词

过多又会使查询结果精度降低。图 4 通过比较查准值，反映了扩展规模对查询性能的影响。

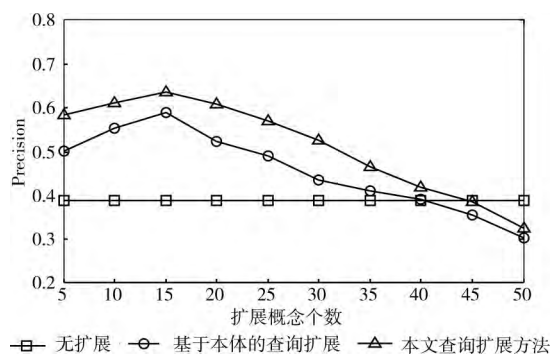


图 4 扩展规模对查询扩展方法的影响

当扩展词个数在 15 个左右时，基于本体的查询扩展和本文方法可获得较好的查询性能，当扩展词个数超过 40 个时其查询性能甚至低于未扩展的检索性能。

#### 4 结束语

本文将语义查询扩展和统计查询扩展相结合，更准确的反应了概念之间的关系，扩展概念更接近用户的查询意图。扩展词权重计算的改进和查询-文档相关度计算的改进，使得检索结果的排序更满足用户的查询需求。实验结果表明该语义检索方法在保证查全率的基础上能有效提高查准率。然而以上方法都是建立在对本体中所有概念建立了关联度的基础之上的，其计算量大。所以下一步的研究重点是如何只分析部分相关概念之间的关联度，减少关联度的计算量。

#### 参考文献：

[1] MENG Hongwei, ZHANG Zhiping, ZHANG Xiaodan. Research on intelligent information retrieval model based on domain ontology [J]. Journal of Intelligence, 2013, 32 (9): 180-184 (in Chinese). [孟红伟, 张志平, 张晓丹. 基于领域本体的文献智能检索模型研究 [J]. 情报杂志, 2013, 32 (9): 180-184.]

[2] WANG Lu, WANG Guochun, GUI Jinhua, et al. A semantic retrieval system based on ontology [J]. Journal of Changchun University of Technology (Natural Science Edition), 2013, 34 (6): 726-730 (in Chinese). [王璐, 王国春, 桂金花, 等. 本体语义检索系统 [J]. 长春工业大学学报 (自然科学版), 2013, 34 (6): 726-730.]

[3] ZHANG Sheng. A model of semantic retrieval based on domain ontology [J]. Software Guide, 2014, 13 (3): 18-20 (in Chinese). [张胜. 一种基于领域本体的语义检索模型 [J]. 软件导刊, 2014, 13 (3): 18-20.]

[4] Latiri C, Haddad H, Hamrouni T. Towards an effective automatic query expansion process using an association rule mining approach [J]. Journal of Intelligent Information Systems, 2012, 39 (1): 209-247.

[5] LIU Caihong, QI Ruihua, LIU Qiang. Efficient query expansion based on positive and negative association rules [J]. China Sciencepaper, 2013, 8 (1): 51-56 (in Chinese). [刘彩虹, 祁瑞华, 刘强. 一种正负关联规则的快速查询扩展算法 [J]. 中国科技论文, 2013, 8 (1): 51-56.]

[6] Tannebaum W, Rauber A. Using query logs of USPTO patent examiners for automatic query expansion in patent searching [J]. Information Retrieval, 2014, 17 (5-6): 452-470.

[7] Chauhan R, Goudar R, Sharma R, et al. Domain ontology based semantic search for efficient information retrieval through automatic query expansion [C] //International Conference on Intelligent Systems and Signal Processing, 2013: 397-402.

[8] LI Xingchun. The research of query expansion based on ontology in information retrieval technology [J]. Journal of Chongqing Normal University (Natural Science Edition), 2013, 30 (4): 115-118 (in Chinese). [李兴春. 信息检索中基于语义的查询扩展研究 [J]. 重庆师范大学学报 (自然科学版), 2013, 30 (4): 115-118.]

[9] Wang H, Fan H, Li J, et al. Research on the method of semantic query expansion in civil aviation emergency domain ontology [J]. International Journal of Digital Content Technology and its Applications, 2014, 8 (5): 128-135.

[10] DING Guodong, BAI Shuo, WANG Bin. Local co-occurrence based query expansion for information retrieval [J]. Journal of Chinese Information Processing, 2013, 20 (3): 84-91 (in Chinese). [丁国栋, 白硕, 王斌. 一种基于局部共现的查询扩展研究 [J]. 中文信息学报, 2013, 20 (3): 84-91.]

[11] TIAN Xuan, DU Xiaoyong, LI Haihua. Computing term-concept association in semantic-based query expansion [J]. Journal of Software, 2008, 19 (8): 2043-2053 (in Chinese). [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.]

[12] WANG Xuyang, XIAO Bo. Query expansion method based on ontology and local context analysis [J]. Computer Engineering, 2012, 38 (7): 57-60 (in Chinese). [王旭阳, 萧波. 基于本体和局部上下文的查询扩展研究 [J]. 计算机工程, 2012, 38 (7): 57-60.]

[13] QIN Yuping, MENG Xiangna, WANG Xiukun, et al. Research on QE based on local co-occurrence analysis and concept semantic [J]. Microcomputer Applications, 2010, 31 (6): 8-13 (in Chinese). [秦玉平, 孟祥娜, 王秀坤, 等. 基于局部共现分析和概念语义的查询扩展研究 [J]. 微计算机应用, 2010, 31 (6): 8-13.]