多关键字查询中LCA剪枝概念树的查询扩展技术研究

王昭龙[1] 李 霞[2] 许瑞芳[3]
(中国航空计算技术研究所 西安 710068)[1] (西北工业大学计算机学院 西安 710072)[2]
(中国电子科技集团公司第二十八研究所 南京 210007)[3]

摘 要 语义查询扩展中,关键一步是扩展词的选择方法和扩展词权重的计算。提出一种改进的LCA(局部上下文分析法):OLCA(Optimize Local Context Analysis)。OLCA应用于分权重的多关键字查询中,结合WordNet概念树,从语义和实际查询语料两方面对初始查询词进行扩展,并根据初始查询词中多个关键词的位置,结合扩展候选集中词间关系计算修正各扩展词的权重。实验证明,与单独基于统计或基于语义的查询扩展方法相比,其查准率和查全率均有较大提高。

关键词 多关键字,查询扩展,概念树,局部上下文分析法
中图法分类号 TP311 **文献标识码** A

Multi-keywords Query Expansion with OLCA Based Concept Tree Pruning

WANG Zhao-long[1] LI Xia[2] XU Rui-fang[3]
(China Aeronautics Computing Technique Research Institute, Xi'an 710068, China)[1]
(School of Computer Science, NorthWestern Polytechnical University, Xi'an 710072, China)[2]
(No. 28 Institute of China Electronic Technology Corporation, Nanjing 210007, China)[3]

Abstract In semantic-based query expansion, computing expansion words and its weight is a key step to describe the needed query. We proposed a method called OLCA (keyword to concept method), the idea comes from LCA(Local context analysis). We made some improvement and applied it to multi-keywords query with different weight according to their attribute to the query. Combined with concept tree based on WordNet, we made the query expansion performed from both semantic and the real query documents aspects, and calculated the weight of expansion term based on this technique. Compared with the one based on semantic or the traditional expansion which merely is based on statistic, the experiments reveal that this method can achieve a better query quality.

Keywords Multi-keywords, Query expansion, Concept tree, Local context analysis

1 相关工作

在信息检索领域,查询扩展(query expansion,简称QE)早在20世纪60年代就被提出[7],是公认的能够有效提高查全率的技术之一。其基本思想是利用与查询关键词相关的词语对查询进行修正,以找到更多相关文档,提高查全率。然而,基于关键词的传统查询扩展方式常常会带来许多语义理解错误,文献[8]中称其为词语问题(vocabulary problems),如同义词问题(synonyms)、歧义问题(polysemy)、异体问题(lemmas)、准同义问题(quasi-synonyms)等,在提高查全率的同时难以保证查准率[15]。

词语问题的根本原因在于,人们描述同样的对象或事件时,用词存在多样性,例如,"教材"和"教科书"都是对"课本"这一概念的称谓。为解决该问题,人们提出了基于概念的语义查询扩展(semantic-based QE),用概念来描述查询主旨,找到与查询语义相关的概念对查询进行扩展[9-11,15]。

查询扩展技术是利用计算机语言学、信息学等多种技术,将与初始查询词词义相关的词添加到初始查询词中,得到比初始查询词更丰富的查询词,避免同义词、近义词、缺词等问题,同时,剪去不相关的扩展词词义,从而提高检索的查全率和查准率,解决困扰信息检索领域的词不匹配问题。查询扩展技术的核心问题是扩展词的选择(selection)和权重(weight)[12-14]的分配。

文献[12]中介绍了一种LCA(Local Context Analysis)剪枝概念树的方法。用LCA方法初次检索出与初始查询词最相关词,将其加入到备选扩展词中,用以剪枝构造语义词典的概念树,并补充概念树上不存在的新词,在单关键字的查询中获得了较好的查询质量。但在实际运用中,单个查询词的应用场景几乎不存在,绝大多数的查询都是基于2~3个关键字的,所以文献[12]最大的缺陷是缺少实际使用价值,本文针

到稿日期:2009-10-09 返修日期:2009-12-10 本文受国家自然科学基金(No. 60803043),国家高技术研究发展计划(863)(No. 2009AA1Z134)资助。

王昭龙(1976-),男,工程师,主要研究领域为计算机电路板设计评测、软件设计等,E-mail:13351745@qq.com;李 霞(1977-),女,博士生,讲师,主要研究领域为信息检索、数据管理、软件工程等;许瑞芳(1976-),女,工程师,主要研究领域为地面情报、软件工程等。

对该问题，研究基于上下文分析法和剪枝概念树的查询扩展技术在多关键字查询中的应用方法，并实验证明之。

在多关键字的查询中，由于多个查询词间语义相关，降低了产生歧义的可能性，查准率会有所提升。但由于词义的多义性，Furnas 提出的“词典问题”(dictionary problem)依然存在。例如用户提供的初始查询词为：“熊猫 电视”，查询结果可能为“熊猫牌电视机”，也可能为“播放熊猫的电视节目”，通过查询扩展技术，可以将“电视 熊猫”扩展为“熊猫 电视 彩电”，从而避免初始查询词造成的歧义。

本文拟通过概念树对各查询词进行语义扩展，并提出一种修正的 LCA 方法，称之为 OLCA(Optimize Local Context Analysis)方法，计算出同一语料库中与各查询词语义最相关的词，作为扩展词加入到原始查询词中，用以弥补初始查询词信息不足的缺陷。由于在多关键字查询中，各查询词作用不同，权重也应不同，例如查询词“鲁迅 作品”，“鲁迅”的权重应该比“作品”高。本文将通过对查询词扩展结果集的 Top-k 排序，结合原始查询词的位置和扩展词相互关联度来计算核心关键词和非核心关键词的权重，让参与查询的扩展词更接近描述用户的查询意图，从而提高查准率。

2 算法思想

第一步，使用 LCA 方法构造统计候选词集，这里的 LCA 方法是在 Xu J[1]的 LCA 方法基础上，针对分权重的多关键字查询而改进的方法，称之为 OLCA(Optimize Local Context Analysis)。OLCA 方法修改了计算潜在扩展词和原始查询词的关联程度公式[1]，引入不同查询词对选择扩展词的影响权重的不同，体现出不同重要程度的初始查询词对扩展词选择的影响；第二步，利用 WordNet[3,4]对每个初始查询词构造概念树，生成语义扩展词候选集；第三步，在多个初始查询词中，用其中一个查询词调整其他查询词的语义森林中的各概念节点，计算出各节点的语义权重，同时进行剪枝操作；第四步，结合统计扩展词候选集，计算概念树各节点扩展词的最终权重；从语义扩展词候选集和统计扩展词候选集中选择满足一定权重要求的扩展词构成最终的查询扩展词集合参与查询。

3 OLCA 方法剪枝概念树的查询扩展算法

3.1 OLCA 方法构造统计扩展词候选集

定义 1(统计扩展词候选集) 由 OLCA 方法初次检索出的扩展词集，记为 $Sta\_Candidates\{T_1, T_2, \cdots, T_m\}$，其中 T_i 是与初始查询词最相关的 m 个候选查询词。这里采用 Xu J 等人的实验[1]，使用与原查询最相关的前 100 篇文章作为查询扩展的基础。由于在此选择的只是备选候选词而不是最终扩展词，因此选择的扩展词数是 Xu J 等人的实验数据的 3 倍，即选择前 120 个与原查询关联度最高的词加入统计扩展词候选词集，即 $m=120$。

为了体现不同权重的关键字对产生潜在扩展词的不同影响，OLCA 方法修改了 LCA 方法：从初始查询词获得的前 N 个文档中提取的概念与原始查询词相关度的计算公式。LCA 中的原公式如下：

$$bel(Q,c)=\prod_{t_i\in Q}(\delta+\log(af(c,t_i))idf_c/\log(n))^{idf_i} \quad (1)$$

其中，

$af(c,t_i)=\sum_{j=1}^{j=n} ft_{ij}\cdot fc_i$

$idf_i=\max(1.0,\log10(N/N_i)/5.0)$

$idf_c=\max(1.0,\log10(N/N_c)/5.0)$

公式中的 c 是文章中的一个扩展词；ft_{ij} 是初始查询词 t_i 在第 i 篇文中出现的次数；fc_i 是扩展词 c 在第 i 篇文章中出现的次数；N 是选定的文章数；N_i 是包含初始查询词 t_i 的文章数；N_c 是包含扩展词 c 的文章数。

根据初始查询词在查询中所处的位置，分配不同权重来修正原公式中初始查询词 t_i 对应的项。实验中尝试添加一个 δ 的加法项来体现不同查询词的权重，也可以采用将权值作为一个系数添加到每个对应项中。为了较好地体现各个初始查询词的不同权重，本文采用了后一种办法。

修正后的 LCA 公式，OLCA 公式为：

$$bel(Q,c)=\prod_{t_i\in Q}(\delta+W_i\cdot\log(af(c,t_i))\cdot idf_c/\log(n))^{idf_i} \quad (2)$$

其中，W_i 是初始查询词的权重，其他各参数意义同 LCA 方法中相应参数。将查询词的权重引入计算公式中，体现了不同权重的查询词对潜在扩展词的影响作用。

定义 2(统计候选词权重) 统计扩展词候选集中各个候选扩展词的权重，记为 $Weight\_LCA$。为了体现扩展词和统计扩展词候选集中的其他扩展词的比较，对 LCA 方法计算得到的表示潜在查询扩展词和原始查询紧密程度的结果 $bel(Q,c)$ 进行归一化：

$$Weight\_LCA(C)=\frac{bel(Q,c)-\mathrm{Min}(bel(Q,c_i))}{\mathrm{Max}(bel(Q,c_i))-\mathrm{Min}(bel(Q,c_i))} \quad (3)$$

其中，$\mathrm{Max}(bel(Q,c_i))$和 $\mathrm{Min}(bel(Q,c_i))$分别表示选定的排名前 m 位备选扩展词与原始查询关联度的最大值和最小值。

3.2 利用 WordNet 构造语义扩展词候选集

定义 3(语义扩展词候选集) 利用语义词典 WordNet 为初始查询词构造的概念树，记为 $CForest\{STree_1, STree_2, \cdots, STree_n\}$，其中 $STree_i$ 表示由查询词的一个词义生成的概念树。

为初始查询词的每个词义构造该词在该词义上的语义子树 $STree_i$，由于每个初始查询词有多种语义，因此初始查询词的语义概念树实际上是一个概念森林，记为 $CForest\{SenseTree_1, SenseTree_2, \cdots, SenseTree_n\}$。

如图 1 中 Changjiang 的语义森林由一个语义子树构成，而 Bank 的语义森林由 10 个语义子树构成，不失一般性，为了便于描述，本文图 1 中只列出了 bank 的其中 3 棵语义树。每个子树的根节点是初始查询词，根节点的第一个子节点是关系类型节点 Relation，表示其上的各个子节点是查询词在这种关系上的扩展，在概念树上，用椭圆节点表示 Relation。根据 WordNet[3]的定义，Relation 包括同义关系(synonymy)、反义关系(antonymy)、上位关系(hypernym)、下位关系(hyponym)、部分整体关系(meronymy)、整体部分关系(holonymy)。Relation 节点下的 $Concept_i$ 节点是初始查询词在这种关系上的扩展，$Concept_i$ 在概念树中用方框表示。每个概念

的同义词，在概念树上用一个节点表示。

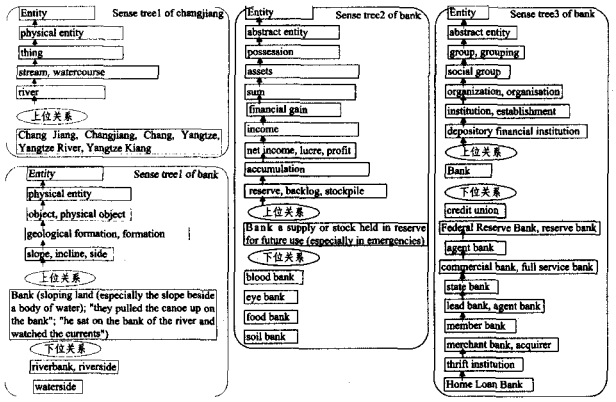

图 1　查询词“Changjiang bank”的概念树

3.3　原始扩展词剪枝概念树

在该步处理中，利用多个初始查询词及其权重对概念树上各个节点的语义权重进行更新，并剪枝，分 5 步操作。

定义 4（概念权重）　根据各词在概念树中位置的不同来定义的权重，记为 $Weight\_Sem(C_i)$。该权重在语义方面体现了扩展词和初始查询词的紧密程度。Leacock[2] 给出了一种语义相似度算法。本文采用概念树节点和原始查询概念间的语义距离来计算概念权重。概念节点的初始语义权重记为 $Weight\_Sem_1(C_i)$。

初始语义权重的计算公式为：$Weight\_Sem_1(C_i)=1/\mathrm{distance}(Root,C_i)$，其中 $Root$ 是初始查询词，$\mathrm{distance}(Root,C_i)$ 是初始查询词到 $Concept_i$ 的最短距离。这里概念间的距离为两个节点间的节点数，不是节点间的边数。需要说明的是，两个概念属于同一个同义词集合，则在概念树上是一个节点，概念间的距离为 1，不为 0。

1）粗略剪枝，去掉概念树最末节点 physical entity，Entity 等，因为这些节点对概念的查询不能提供足够多的信息，尽早剪枝以减少后续基于概念树的复杂计算。

2）根据每个初始查询词的权重，调整树中各节点的语义权重。修正后的概念权重为：$Weight\_Sem_2(C_i)=Weight\_Sem_1(C_i)\cdot W_i$，其中 W_i 为初始查询词 T_i 的权重。

3）如果某初始查询词的权重较高，则由其生成的概念树上的各概念节点的语义权重也会相应较高。

利用式(4)，根据初始查询词的语义树相互影响的程度来重新计算语义权重，即用初始查询词的语义森林中的每一个语义树上的节点对其他初始查询词的语义树上的节点的权重进行修正。

$$Weight\_In(C_i)=\frac{\sum_{t_m\in Q\wedge m\neq i}\ \sum_{STree_k\in CForestm}\ \sum_{(c_j\in STree_k)\wedge(c_j\wedge c_i)} Weight\_Sem_2(C_j)}{\sum_{tm\in Q\wedge m\neq i}\ \sum_{STree_k\in CForestm}\ \sum_{c_j\in STree_k} Weight\_Sem_2(C_j)}\cdot Weight\_Sem_2(C_i) \tag{4}$$

其中，$Weight\_In(C_i)$ 表示概念节点 C_i 通过与其他初始查询词概念树上的节点进行运算而获得的语义权重增加值。式(4)中 $c_j\wedge c_i$ 表示两个语义树的节点间有重合的内容。如上例中，Changjiang 的描述中出现了 river，其上位关系的第一个节点也出现了 river，而 bank 的第二个概念树中也存在包含 river 的节点，则初始查询词 Changjiang 的语义树中的 Changjiang 节点和 river 节点通过式(4)产生语义权重增加值。同样 bank 节点的语义权重也会增加。

4）计算第 i 个初始查询词 T_i 的第 j 个概念树上各个节点获得的权重增加值的总和为：

$$Weight(T_{ij})=\sum_{c_k\in STree_{ij}} Weight\_In(C_k) \tag{5}$$

选择其中增加值最大的语义分支为该查询词在本次查询中的词义。如例中 bank 的 10 个语义子树中，表示“河岸”的语义子树获得了最高的权重加分，即通过原始查询的上下文判断得出 bank 在本次查询中表示河岸的意思，排除其他意思。

5）用获得的语义权重增加值更新选定的概念树上的每一个节点，得到概念节点的语义权重为：

$$Weight\_Sem(C_i)=Weight\_Sem_2(C_i)+Weight\_In(C_i) \tag{6}$$

通过步骤 5）的处理，每个初始查询词只剩下一棵概念树，概念树上各节点的语义权重也体现了初始查询词的权重。步骤 5）是用初始查询词修正概念树的关键部分，用多个原始查询词中的每一个词的含义来推测理解其他查询词的含义，以此消除歧义。在这个例子中，bank 的银行、赌博基金和条形堆等其他 9 种意思都被剪枝去掉。语义判断出这里的 bank 表示的是河堤的意思，显然，在该词义上的扩展是正确的。

3.4　使用统计扩展词候选集修正概念树

定义 5（查询选扩展词集）　从语义扩展词候选集 $CForest$ 和统计扩展词候选集 $Sta\_Candidates$ 中选择满足条件的扩展词组成的最终扩展词集合，记为 $Query\_Exp\{T_1,T_2,\cdots,T_t\}$。

在本步处理中，利用统计扩展词候选集来修正语义扩展词候选集，从语义扩展词候选集中选择满足一定条件的扩展词加入到查询选扩展词集 $Query\_Exp$ 中。步骤如下：

1）选择在阈值范围内的概念节点和其周边一定距离的概念节点加入到查询扩展词集 $Query\_Ext$ 中。对每一个子树 $STree_i$，从根节点开始按照广度优先的原则遍历，找到最后一个满足下列条件的节点：$Weight\_Sta(C_{ij})<r_1, 1<j<n_i$，其中 n_i 为 $STree_i$ 上的节点个数，取 $r_1=0.35$。设该概念节点为 C_{ik}，则选择的扩展词为：在根节点 Root 指向 C_{ik} 方向上，距离根节点不大于 $1.5*distance(Root,C_{ik})$ 的所有节点。

2）利用统计扩展词候选集对语义扩展词候选集进行剪枝，在步骤 1）生成的概念节点中删除与实际查询结果相关度低的概念节点。计算方法为：当 $Weight\_Sta(C_{ij})<r_2, 1<j\leqslant n_i$ 时，从概念森林中删除该概念节点。其中 n_i 是概念树 $STree_i$ 上的节点个数，取 $r_2=0.05$。

对于例子中查询 Changjiang bank，通过语义扩展 Chang Jiang bank，Changjiang bank，Yangtze bank，Yangtze River，Yangtze Kiang bank，Changjing side，Yangtze side 等可以被检索出来。

3.5　从统计候选词集中选择扩展词

从统计扩展词候选集 $sta.candidates$ 中选择满足下列条件的扩展词，加入到查询扩展词集 $Query\_Exp$ 中。

1）$Weight\_LCA(Term_i)\geqslant r_3$，取 $r_3=0.60$，表示与初始查询词相关度较高的词。

若在 WordNet 中，$Term_i$ 不是一个多义词，并且在 $Term_i$ 的定义描述中出现了初始查询词，则将该词添加到查询扩展词集中。

例如在图 1 的例子中，查询“Changjiang”时，“Shanghai”是统计扩展词候选集中与初始查询词相关度高的词，且“Changjiang”的第一个定义“flows eastward from Tibet into the East China Sea near Shanghai”中出现了“Shanghai”，同时 WordNet 中“Shanghai”只有一个词义，则将“Shangha”加入到查询扩展词集。

2)$Weight\_LCA(Term_i)\geqslant r_4$，其中 $r_4>r_3$，取 $r_4=0.80$，表示与初始查询词极度相关的词，则不考虑概念树中是否包含该词，也不考虑该词在概念树中的语义权重，均将该词加入到查询扩展词集。这种新词，可能与初始查询词没有语义相关，但在一定时间或特定领域内，它们高度相关，则这种词会被选中加入到扩展集中。

例如在图 1 的例子中，查询“Changjiang bank”时，如果 LCA 的前若干个结果中都包含了“Three Gorges”，则尽管这些词或短语不在 Changjiang 或 bank 的语义树中出现，也会被加入到查询扩展词集中。

3.6 计算查询扩展词集 *Query_Ext* 中每个扩展词的权重

最后计算上述步骤得出的扩展词集中各扩展词的权重。

定义 6(扩展词权重) 综合概念权重和统计权重，对扩展词的权重重新计算，记为：

$$Weight(C_i)=\frac{(1+\alpha)*(Weight\_Sem+\delta)*(Weight\_Sta+\delta)}{\alpha*(Weight\_Sem+\delta)+(Weight\_Sta+\delta)} \tag{7}$$

其中，$\delta=0.1$，为了防止出现两个权重值中的一个可能为 0 的情况，α 为两个权值之间的调节因子，调整两个权值对综合权值的贡献程度。

3.7 结果集排序

利用初始查询词和 OLCA 方法得出的查询扩展集 *Query_Exp* 中的扩展词进行查询。按照传统的空间向量模型对结果集进行排序，其中初始查询词的权重为 1，其他扩展查询词权重用 OLCA 方法算出，即由 3.6 节中的式(7)计算出 *Query_Exp* 中各扩展词的权重。

4 实验

试验架构如图 2 所示。

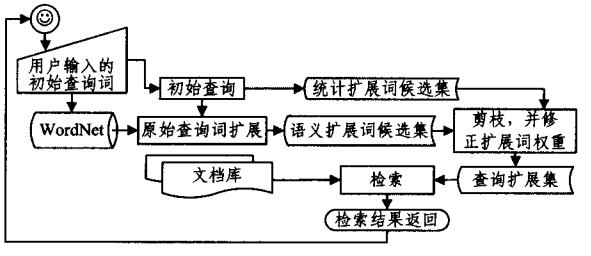

图 2 查询系统流程示意图

4.1 实验数据

采用一个在线的信息发现和学习引擎 Wizag 的语料[12]，文章类别包括体育 Sports、技术 Tech、商业 Business、视频 Video 和热门博客 Blog 5 大类。需要说明的是，文章选自热门 Feeds，内容是 Feeds 文件中抽取的描述信息，不是文章的全文。

4.2 评测标准

测试时采用 WordNet 提供的接口 WordNet. Net[4] 来构造概念树。评测时，采用了传统的考察查询质量的两个指标：查准率和查全率。本文在计算查准率和查全率时，借鉴了标准指标 MAP(mean average precision)，描述前 k 个结果文档中相关文档的准确率平均值，本文 $K=20$。同时采用文本检索会议 TREC 提出的评测指标 T10F 来进行评价。F 值是查准率和查全率的函数，定义为：

$$F=\frac{(\beta^2+1)p*r}{\beta^2 p+r}$$

其中，p 为查准率，r 为查全率，β 为控制查准率和查全率权重关系的参数，TREC 取 β 为 0.5。

4.3 实验结果和分析

为了测试本文所提出的 OLCA 剪枝概念树方法的查询质量，分别对非扩展的查询、LCA 方法和结合使用 WordNet 的 OLCA 剪枝概念查询进行测试对比，结果如表 1 所列。

表 1 多关键字查询中 3 种方法的查询质量对比

方法	查准率 p	查全率 r	F-measure
非扩展	0.422	0.293	0.369
LCA 方法	0.407	0.412	0.393
OLCA 剪枝概念树	0.483	0.514	0.436

结果表明，采用扩展的查询比非扩展的查询方法在查全率上都有了很大的提高。但查准率提高不大，甚至有小幅下降。使用 OLCA 剪枝概念树的查询扩展在提高了查全率的同时，也提高了查准率。这一点原因比较明显，本文提出的 OLCA 方法中，从两个角度来确定扩展词的选择，一方面使用多个查询词提供的上下文信息，结合语义词典确定每一个查询词的准确语义，在该语义上进行扩展；另一方面，使用初始查询词基于统计的信息，对语义扩展词进行修正和补充，使选择的扩展词更准确，更有针对性，并据此进行扩展词的权重再分配，使得查询结果更能反映用户的意图，从而很好地解决了用词不一致、查询信息不全造成的结果偏差和遗漏。

使用 OLCA 剪枝概念树的查询扩展，较之分别单独采用两种扩展方法，在提高了查全率的同时，也提高了查准率。这是因为引入的查询扩展词不但通过概念树考虑了语义方面的扩展，也通过 LCA 方法考虑了实际语料的特征，因而扩展词的选择更有针对性。这一点在新词较多、语义较丰富的技术 Tech 领域文章中表现尤为明显。OLCA 方法从本质上说，与其他的局部上下文扩展法一样，均采用两次查询的方法解决扩展问题，不同在于扩展词的选择上，考虑了初始查询词间的相互语义影响。

结束语 信息查询系统中，由于查询词表达的多样性给基于关键词的传统查询带来了许多语义理解错误(词语问题)，因此基于概念的语义扩展查询已经是一种公认的解决方法[7]。在基于概念的语义查询扩展中，扩展词的选择和权重的分配，是决定查准率和查全率的关键因素。本文提出的 OLCA 方法，从两个角度来确定扩展词的选择，一方面使用多个查询词提供的上下文信息，结合语义词典确定每一个查询词的准确语义，在该语义上进行扩展；另一方面使用初始查询词的基于统计的信息，对语义扩展词进行修正和补充，使选择的扩展词更准确、更有针对性，并据此进行扩展词的权重再分配，使得查询结果更接近用户查询意图，从而很好地解决了用词不一致、查询信息不全造成的结果偏差和遗漏，找到与查询语义主旨匹配度高的扩展词，从而提高查询效果。实验结果

(下转第 162 页)

Web服务描述语言)协议的基础上,采用Java、JSP、Servlet、密码算法、安全协议、数字签名、CA证书等技术对UDDI和WSDL进行安全增强。它通过以下几点来实现应用层次上的信任和授权:①客户端与服务器端的双方认证。当用户登录安全增强UDDI的服务注册模块时,模块将验证用户所提交的机构证书的合法性和有效性,而客户端也将验证可安全增强UDDI模块的证书,在客户端与服务器端的证书相互认证通过以后,用户才可以进入安全增强UDDI模块。②操作权限的控制。政务机构可以为自己定制操作权限,在安全增强UDDI服务注册中心的用户管理功能中,政务机构可以给用户的公钥证书赋予安全增强UDDI模块的操作权限;在使用安全增强UDDI模块中的其它功能时,用户必须提交拥有该操作权限的有效的、合法的证书,当服务器验证该证书通过后,用户才可以使用安全增强UDDI模块的功能。③关键信息的安全传输。为了确保数据在传输过程中的机密性和完整性,安全增强UDDI在传输过程中对所有的数据进行签名,更会对一些重要的数据信息进行加密。当数据传输到安全增强UDDI时,它对数据进行验证签名和解密。验证签名失败的数据不处理,并给出反馈信息;通过验证签名的数据,才允许进行相应的操作。④安全日志。安全日志将每个用户所进行的关键操作进行记录,而且对每一个记录进行签名,之后存入数据库,为审计工作提供依据。

需提及的是,WSDL是实现协同能力的关键之一,它提供了一份契约用于与各种政务应用之间交互,使得各个政务组织可以将标准的制定集中在Service的外部接口,而不用考虑各政务组织的具体实现。

结束语 集群政务协同业务平台已经成功地在某省(市)投入应用。运行实践表明,该平台取得了多项集成创新,能够最大限度地整合利用省(市)级协同政务平台的网络、服务存储平台、软件资源和政务业务信息资源,使农村区县直接基于省(市)级平台构建各自的政务平台,实现了政务平台的城乡统筹建设和维护,大大降低了全省(市)政务建设和维护成本,显著提高了政务管理和协同办公效率。

参 考 文 献

[1] Kaliontzoglou A, Sklavos P, Karantjias T, et al. A Secure e-Government Platform Architecture for Small to Medium Sized Public Organizations [J]. Electronic Commerce Research and Applications, 2005, 4(2): 174-186

[2] Velez I P, Velez B. Lynx: An Open Architecture for Catalyzing the Deployment of Interactive Digital Government Workflow-based Systems[C]// Proceedings of the 2006 International Conference on Digital Government Research. San Diego, California, USA: ACM, 2006: 309-318

[3] A Lightweight Decentralized Authorization Model for Inter-domain Collaborations[C]//Proceedings of the 2007 ACM Workshop on Secure Web Services. Fairfax, Virginia, USA: ACM, 2007: 83-89

[4] Janssen M, Chun S A, Gil-Garcia J R. Building the Next Generation of Digital Government Infrastructures[J]. Government Information Quarterly, 2009, 26(2): 233-432

[5] 王宁. 电子政务中信息资源整合的建模与应用研究[D]. 大连: 大连理工大学, 2005

[6] Bass L, Clements P, Kazman R. Software Architecture in Practice, Second Edition [M]. Addison-Wesley, 2003

[7] 林泊, 周明惠, 刘天成, 等. 一个J2EE应用服务器的Web容器集成框架[J]. 软件学报, 2006, 17(5): 11958-1203

[8] 李长云, 阳爱民, 满君丰, 等. 一种面向按需集成服务的业务模型构造方法[J]. 计算件学报, 2006, 29(7): 1095-1104

(上接第135页)

表明,与未经查询扩展的概念检索方法和仅基于LCA的概念检索方法相比,本文提出的OLCA方法在查准率和查全率上均有较大程度的提高。

为了更好地理解用户的查询意图,使用用户日志为用户建立查询模型来修剪概念树,能提供更好的个性化查询,这将是我们后继研究的一个重点问题。

参 考 文 献

[1] Xu J X, Croft W B. Query expansion using local and global document analysis[C]// Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Zurich, Switzerland, 1996: 4-11

[2] Leacock C, Chodorow M. Combining local context and wordnet similarity for word sense identification in WordNet: An Electronic Lexical Database [M] // Christiane Fellbaum, ed. MIT Press, 1998: 265-283

[3] http://wordnet.princeton.edu/

[4] http://wordnet.princeton.edu/linksJHJ.NET

[5] Xu Y, Papakonstantinou Y. Efficient Keyword Search for Smallest LCAs in XML Databases[C]// Proceedings of SIGMOD' 2005. Baltimore, Maryland, USA

[6] Tran T, Wang H, Rudolph S, et al. Top-k Exploration of Query Candidates for Efficient Keyword Search on Graph-Shaped (RDF) Data[C]// ICDE. 2009: 405-416

[7] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval [M]. NewYork: Addison-Wesley-Longman, 1999

[8] Furnas G W, Landauer T K, Gomez L M, et al. The vocabulary problem in Human-System communication[J]. Communications of the ACM, 1987, 30(11): 964-971

[9] Qiu Y G, Frei H P. Concept based query expansion[C]// Korfhage R, Rasmussen E, Willett P, eds. Proc. of the 16th annual Int'l ACM SIGIR Conf. on research and development in information retrieval. Pittsburgh: ACM Press, 1993: 160-169

[10] Chang Y, Ounis I, Kim M. Query reformulation using automatically generated query concepts from a document space[J]. Information Processing and Management, 2006, 42: 453-468

[11] 赵军, 金千里, 徐波. 面向文本检索的语义计算[J]. 计算机学报, 2005, 28(12): 2068-2078

[12] 张超盟, 李战怀. 局部上下文分析法剪枝概念树的查询扩展技术[J]. 计算机工程, 2009, 35(14): 45-48

[13] 李新叶, 苑津莎. 一种快速的语义检索算法[J]. 电子学报, 2007: 2220-2225

[14] 万常选, 鲁远. 基于权重查询词的XML结构查询扩展[J]. 软件学报, 2008: 2611-2619

[15] 田萱, 杜小勇. 语义查询扩展中词语-概念相关度的计算[J]. 软件学报, 2008: 2611-2619