ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.10, October 2008, pp.2611−2619
DOI: 10.3724/SP.J.1001.2008.02611

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于权重查询词的 XML 结构查询扩展*

万常选[1,2+], 鲁 远[1,2]

[1](江西财经大学 信息管理学院,江西 南昌 330013)

[2](江西财经大学 数据与知识工程江西省高校重点实验室,江西 南昌 330013)

Structural Query Expansion Based on Weighted Query Term for XML Documents

WAN Chang-Xuan[1,2+], LU Yuan[1,2]

[1](School of Information and Technology, Jiangxi University of Finance and Economics, Nanchang 330013, China)

[2](Jiangxi Key Laboratory of Data and Knowledge Engineering, Jiangxi University of Finance and Economics, Nanchang 330013, China)

+ Corresponding author: E-mail: wanchangxuan@263.net

Wan CX, Lu Y. Structural query expansion based on weighted query term for XML documents. *Journal of Software*, 2008,19(10):2611−2619. http://www.jos.org.cn/1000-9825/19/2611.htm

Abstract: The main reason of low precision in information retrieval (IR) is that it is difficult for the users to submit a precise query expression for their query intensions. Furthermore, XML documents have characteristics not only in the content, but also in its structure. Therefore it is more difficult for users to submit precise query expressions. In order to solve this problem, this paper puts forward a new query expansion method based on relevance feedback. It can help users to construct a content and structure query expression which can satisfy users' intentions. This method includes two steps. The first step is to expand keywords for finding the weighted keyword which can represent the user's intentions. The second step is structural expansion based on the weighted keywords. Finally a full-edged content-structure query is formalized. Experimental results show that the method can obtain better retrieval results. The average precision of prec@10 and prec@20 is 30% higher than the original query.

Key words: XML; information retrieval; structural semantics; structural query expansion; relevance feedback

摘 要: 文本文档信息检索中检索质量不高的一个主要原因是用户难以提出准确的描述查询意图的查询表达式.而 XML 文档除了具有文本文档的内容特征外,还具有结构特征,导致用户更难以提出准确的查询表达式.为了解决这一问题,提出一种基于相关反馈的查询扩展方法,可以帮助用户构建满足查询意图的"内容+结构"的查询表达式.该方法首先进行查询词扩展,找到最能代表用户查询意图的权重扩展查询词;然后在扩展查询词的基础上进行结构查询扩展;最终形成完整的"内容+结构"的查询扩展表达式.实验结果表明,与未进行查询扩展相比,扩展后 prec@10 和 prec@20 的平均准确率提高 30%以上.

关键词: XML;信息检索;结构语义;结构查询扩展;相关反馈

中图法分类号: TP311 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant No.60763001 (国家自然科学基金); the National Social Science Foundation of China under Grant No.07BTQ025 (国家社会科学基金); the Key Science-Technology Project of the Education Department of Jiangxi Provincial of China under Grant No.[2006]320 (江西省教育厅重点科技项目)

Received 2007-09-10; Accepted 2008-05-29

随着 XML 文档在数字图书馆、科学数据库以及互联网等领域的广泛应用,对于 XML 文档检索的研究已逐渐成为信息检索领域的一个重要研究方向.信息检索中检索质量不高的一个很重要的原因就是用户往往不能准确地描述自己的查询意图,而 XML 文档除了具有文本文档的内容特征外,还具有结构特征,因此准确描述用户查询意图的查询表达式不仅要包含查询词,还要包含结构信息,这对用户来说具有很大的难度,所以帮助用户形成准确的查询表达式,对于 XML 信息检索来说是非常有必要的.

在信息检索的研究过程中,研究者们发现反馈技术能够帮助用户构建较为准确的查询表达式,从而提升检索质量.信息检索中的反馈技术根据反馈信息获取途径的不同分为 3 类:相关反馈、伪反馈和隐式反馈.学术界通常认为相关反馈的研究源于 Rocchio,1971 年他针对文本信息检索基于向量空间模型在 SMART 系统中完成了最初的相关反馈实验.无论是早期实验还是近几届的 TREC 实验都说明相关反馈可以提高检索效率[1].国内也开展了相关反馈技术的研究[2−5],研究结果都显示了相关反馈技术能够有效提高查准率、查全率.伪反馈又称为局部反馈、盲反馈,它假设初始检索结果的前面若干篇文档是相关的,然后利用标准的相关反馈过程进行查询扩展.多次 TREC 评测会议表明,伪反馈是一种简单但十分有效的查询扩展技术.伪反馈方法的一个重大缺点就是可能在查询扩展期间产生“查询主题漂移”.目前主要采用两种思路解决查询漂移问题,一是对初始检索结果进行调整,使用于反馈的文档都是相关文档[6,7];二是先检索结果聚类再进行查询扩展[8−11].国内对此问题也进行了积极的研究[12−14].隐式反馈方法中用户不主动参与反馈,但是系统仍需要从用户的浏览行为中分析得到一些有用的信息用来确定用户兴趣模式,从而推理出描述用户查询需求的表达式,并据此进行检索.在查询扩展中利用隐式反馈是让用户行为中体现的兴趣信息来影响查询扩展词的选取和查询词权重的设置.目前的研究主要思路是将现有的检索算法,如向量空间模型、概率模型中的查询词扩展和加权公式加以改造,引入反映用户兴趣的因子[15−20].

1 研究现状

XML 文档与传统的文本文档的最大区别在于 XML 文档是半结构化文档,它具有结构特点.这导致它的查询扩展比传统的文本文档更为复杂,根据从反馈信息中抽取信息的类别不同,XML 查询扩展可分为:基于文本内容的查询扩展、基于结构信息的查询扩展和同时基于文本内容和结构信息的查询扩展.由于 XML 文档的信息检索才刚刚起步,而在 XML 文档的信息检索中引入反馈机制是近两三年才开始的,所以研究成果还不多[21−31],并且其中大部分是基于相关反馈的文本内容(即查询词)的查询扩展[21−24],少数几篇提出了基于相关反馈的“内容+结构”的查询扩展方法[25−30],其中 Schenkel 和 Theobald 的 3 篇文献比较成熟[26−28].基于伪反馈 XML 查询扩展的研究极少[31],而基于隐式反馈的 XML 查询扩展的研究目前还没有发现相关成果.

1.1 基于相关反馈的XML查询扩展

目前,基于相关反馈的 XML 内容查询扩展方法主要是将文本文档中查询词扩展加权方法,如向量空间模型和概率模型的语词加权方法引入到 XML 文档中[21−24],这种思路只是传统信息检索的简单延伸,而没有考虑 XML 的结构化特征.

文献[25−30]考虑从用户相关反馈结果中抽取结构信息来进行查询扩展最终形成“内容+结构”的查询扩展表达式.其中文献[25]分两种情况讨论结构查询扩展:如果用户提出的是 CO(content-only)(查询表达式中只有查询词)查询,则用户判断结果元素的相关性,系统从用户标记为相关的元素中抽取其到根节点的路径,将路径中的最大公共部分作为扩展查询的结构信息;如果用户提出的是 CAS(content and structure)(查询表达式中包含查询词和路径信息)查询,则用户判断结果元素的相关性,并给出不同的权重,系统从用户标记为相关的元素中抽取其到根节点的路径,找出路径中的所有公共部分,并根据用户给出的权重计算每条公共路径的得分,取分数最大的路径作为扩展查询的结构特征.但是作者并没有通过实验来验证其正确性.

文献[26−28]提出:结构特征包括文档特征(*D*)、祖先特征(*A*)和祖先后裔特征(*AD*).其中,文档特征为包含用户所标识元素的文档中所有 *tag-term*(元素标记-词项)对及其得分;祖先特征为用户所标识元素的祖先中的所有 *tag-term* 对及其得分;祖先后裔特征为用户所标识元素的祖先的所有后裔的 *tag-term* 对及其得分.内容特征(*C*)

从用户所标识的元素中抽取.扩展后的查询形式为//*ancestor-tag*[*A*+*AD* constrains]//*[*keyword*+*C*+*D* constrains].

文献[29,30]为国内文献,其思路是要求用户从检索结果中标识相关的文档及相关的结构信息,检索器根据用户提供的信息进行严格的结构匹配找到最终结果.但此方法除了要用户选择相关文档还需要用户选择相关的结构信息,这给不了解文档集结构信息的普通用户增加了难度.

1.2 基于伪反馈的XML查询扩展

目前只有文献[31]利用了伪反馈进行查询扩展,主要针对科技文献检索进行了研究,其思路是去除XML文档的结构信息,使之成为无结构的文档,检索系统根据查询进行排序,假定前 *n* 篇为相关文档,系统自动抽取前 *n* 篇文档的关键词和主题词将其扩充到初始查询中,形成新的查询.由于去除了文档的结构特征,该方法得到的仍然只是基于内容的查询扩展表达式,而没有考虑XML文档的结构特征.

本文提出了一种基于用户相关反馈文档中的权重查询词的结构查询扩展方法,实验表明,利用此方法能够较好地实现“内容+结构”的查询扩展.

2 基于权重查询词的结构查询扩展

当用户提出CAS查询时,查询中既有内容信息又有结构信息.我们认为CAS查询表达式中主要为内容信息,其次才是结构信息.这是因为,用户在信息检索中主要是搜索在内容上符合自己查询意图的信息,而内容主要是通过查询词表达出来的;至于查询中的结构信息,应该将它理解为对内容的限定,即在内容都满足的情况下,用户需要符合这种结构信息的结果.

基于对内容信息与结构信息相互关系的这种理解,最终得到的查询扩展表达式中的查询词必须能最大限度地体现用户对于查询内容的需求,结构必须能够最准确地体现用户对于查询内容的限定.因此,本文提出了一种基于权重查询词的XML结构查询扩展方法.经此方法得到的扩展查询表达式中扩展查询词为用户选择的相关文档中权重最大的词项,因为这些查询词在内容上最能代表用户的查询意图;而结构扩展分支为用户选择的所有相关文档中公共的结构分支,因此结构部分也能对内容起到最大程度的限定.具体实现时,本文采取了两步走的方法:首先进行查询词扩展,找到最能代表用户查询意图的扩展查询词;其次,在扩展查询词的基础上进行结构查询扩展,最终形成完整的“内容+结构”的查询扩展表达式.

2.1 查询词扩展

用户反馈的相关元素或相关文档中包含的所有词项(去除停用词)的集合即为扩展查询词的候选集合,该集合中权重高的词项将被用来进行查询扩展.因此相关元素或相关文档中词项的权重计算是选择扩展查询词的核心问题.传统的信息检索中词项的权重计算主要考虑词项在文档中出现的频率因素及文档的大小因素.例如,当前流行的TF×IDF方法.目前基于反馈的XML信息检索研究成果不多,并且基本上只是沿用了此词项加权方法.XML文档与传统的文本文档相比,最大的不同之处就是XML文档具有结构信息,因此,结构信息对于词项权重的影响是本文的研究重点.

本文主要考虑以下几点对于词项权重的影响因素:词项所属元素的语义权重;词项所属元素的层次;词项与初始查询词间的距离关系.

2.1.1 元素语义权重对词项权重的影响

XML文档可以建模为一棵标记树,如图1所示,每个元素或属性表示为一个节点,元素-子元素或元素-属性之间的关系用相应节点间的实线边来表示,“…”代表文本内容.虚线框节点表示直接包含文本内容的元素节点(称为文本元素节点),文档中的文本内容都出现在文本节点中.因为一般意义上的信息检索都是对文档的文本内容进行检索,所以我们在考虑词项所属元素的语义权重时,主要考虑的是文本元素节点的语义权重设置.

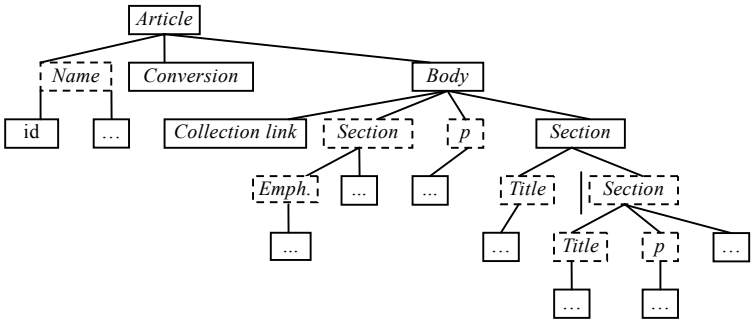


Fig.1 Tree of XML document
图 1 XML 文档的树形建模图

我们认为,对文档主题贡献大的元素,所包含的词项对文档的主题贡献也比较大.例如,文档标题 *name* 中出现的词项,要比章节 *section* 中出现的词项更能反映文档的主题;章节标题 *title* 中出现的词项,要比段落 *p* 中出现的词项更能反映文档的主题;强调 *emph* 中的词项,要比 *section* 中的词项内容更能反映文档的主题.所以我们对文档中的文本元素设置元素语义权重,以反映信息检索中元素的重要程度,并根据词项所属元素的语义权重来判断词项的重要程度.根据元素对文档主题贡献的大小,图 1 中的元素语义权重由大到小的顺序为 *name*,*title*,*emph*,*section* 或者 *p*.

2.1.2 元素层次对词项权重的影响

在 XML 文档中存在元素嵌套情况,最普遍的情况是 *section* 节点中嵌套 *section* 节点,而上层 *section* 节点中 *title* 节点与下层 *section* 节点中 *title* 节点中的词项,对于文档主题的重要程度是不同的.所以仅仅设置元素语义权重是不够的,还要考虑元素在文档中的层次,层次高的元素中词项的权重要大于层次低的元素中词项的权重.

2.1.3 词间位置距离对词项权重的影响

在进行查询词扩展时还应该充分重视用户的初始查询词,它是用户查询意图的最真实的体现.但是,普通用户提出的查询词经常只有 1~2 个,而通常认为短语或词组查询的准确率会高于单关键词查询.如何将单个的初始查询词扩展为短语,最直接的想法就是考虑与初始查询词相邻的词项是否能与之组成短语.例如用户的初始查询词为“熊猫”,在只考虑语义权重和层次因素时可以得到相关文档中“彩电”和“电视机”两个词项的权重相同,但在相关文档中,“彩电”经常与“熊猫”组合在一起形成“熊猫彩电”短语,而“电视机”则不常和“熊猫”组合在一起出现,这种情况下,我们认为“彩电”与“电视机”相比,“彩电”更应该成为扩展查询词,所以“彩电”的权重应该比“电视机”的权重大.因此,我们引入了距离因子概念,用它来反映词项与用户初始查询词位置距离的远近,与初始查询词越近的词项的距离因子越大,表示其越有可能与初始查询词结合在一起组合成一个具有完整意义的短语查询.

2.1.4 查询词扩展公式

基于以上考虑,本文提出改进的词项加权公式为

$$W_{kd}=\sum_{e\in d}\frac{w_e}{h_e}\times\gamma_{ke}\times tf_{ke}\times idf_k \tag{1}$$

其中,W_{kd} 为词项 k 在文档 d 中的权重;w_e 为元素 e 的语义权重因子;h_e 为元素 e 在文档树中所处的层次;γ_{ke} 为距离因子,体现元素 e 中词项 k 与最近的初始查询词 q_i 间的距离对权重的影响,具体计算如下:

$$\gamma_{ke}=\begin{cases}1 & & d(q_i,k)=1\\ \alpha_1 & (0\le\alpha_1\le 1) & d(q_i,k)=2\\ \alpha_2 & (0\le\alpha_2\le\alpha_1) & d(q_i,k)=3\\ \dots & & \\ \alpha_{m-1} & (0\le\alpha_{m-1}\le\alpha_{m-2}) & d(q_i,k)=m-1\end{cases},$$

其中,$d(q_i,k)$为文档 d 中词项 k 与初始查询词 q_i 的距离,$d(q_i,k)=1$ 意味着 k 和 q_i 相邻.如果词项间距离大于3,一般认为这两个词项不能组合在一起构成一个完整的意义.

tf_{ke} 为词项 k 在元素节点 e 中出现的频率;idf_k 为词项 k 的逆元素频率,$idf_k=\log(N/n_k)$,其中,N 为文档 d 中所有元素的数目,n_k 为文档 d 中包含有词项 k 的元素数目.

根据公式(1),可以计算出用户选择的相关文档中所有词项的权重,从而可以选择权重最大的几个词项作为候选查询扩展词.

2.2 结构查询扩展

在“内容+结构”的查询扩展表达式中内容与结构信息应以内容为主、结构为辅.所以查询扩展表达式中的结构信息必须是与内容信息紧密相关的,并能进一步对内容信息进行结构上的限定.一方面,XML 文档中的内容信息是由词项所体现的,与词项最紧密相关的结构信息就是词项所属的元素信息.另一方面,由于结构信息是对内容的进一步限定,所以用于查询扩展的结构信息应为用户选择的相关文档中的公共结构特征.基于这两点考虑,结构扩展方法是:对于在相关文档中共同出现的每一个权重查询扩展词 $term_i$,分别在相关文档中找到公共的语义权重最大的元素 $tagw_i$,则 $tagw_i$-$term_i$ 对作为最终结构查询扩展的一个分支结构;文档根元素节点作为结构查询扩展的根节点,由这些分支和根节点共同形成用户的扩展查询表达式.用INEX的查询语言NEXI表示的扩展查询表达式的形式如下:

$$//Dr[\text{about}(.//tagw_1,term_1)\ \text{and}\ \text{about}(.//tagw_2,term_2)\ \text{and}\ldots]$$

其中,Dr 为查询文档集中的文档根元素节点,$term_i$ 为扩展查询词,$tagw_i$ 为相关文档中 $term_i$ 所属的公共的语义权重最大的元素.

假设 doc 为查询文档集中的 XML 文档,$rdoc$ 为用户在初始检索结果中选择为相关的文档,相关文档集合 $D=\{rdoc_i|i\in[1,m]\}$,其中 m 为用户选择为相关文档的数量;对于 $rdoc_i$ 文档,经解析形成若干个描述文档信息的表(集合),词项信息表记为集合 $termrdoc_i$,元素信息表记为集合 $elemrdoc_i$,元素-词项对信息表记为集合 $elemtermdoc_i$;元素语义权重记为 $weight_{tag}$.本文提出的结构查询扩展方法可描述如下:

1) 利用公式(1)得到词项权重 $term_{weight}$.
2) 设定词项权重阈值为α,得到候选查询扩展词集合 $EKC=\{term|term_{weight}>\alpha\}$.
3) 得到查询扩展词集合 $EK=\{term|\forall i\in[1,m],term\in termrdoc_i\}$.
4) 得到 $tag\text{-}term$ 对集合 $TT=\{tag\text{-}term|term\in EK\wedge(\exists i\in[1,m](tag\text{-}term\in elemtermdoc_i))\}$.
5) 得到候选结构扩展分支集合 $ESC=tag\text{-}term|tag\text{-}term\in TT\wedge(\forall i\in[1,m](tag\text{-}term\in elemtermdoc_i))\}$.
6) 得到结构扩展分支集合 $ES=\{tagw\text{-}term|tagw\text{-}term\in ESC\wedge(\forall tag\text{-}term\in ESC\ (weight_{tagw}>weight_{tag}))\}$.
7) 最后,完整的扩展查询表达式由文档根节点 Dr 与所有的结构扩展分支形成:$//Dr[\text{about}(.//tagw_1,term_1)\ \text{and}\ \text{about}(.//tagw_2,term_2)\ \text{and}\ldots]$.

3 体系结构

体系结构如图2所示,图中虚线部分为本文重点研究的“内容+结构”查询扩展模块.

首先,用户提出初始查询词,在XML搜索引擎上进行检索,本实验利用了INEX 2006提供的TopX搜索引擎实现对用户初始查询的检索;然后,将初始检索结果返回用户浏览,用户选择符合自己查询需求的几篇文档作为相关文档,作为查询扩展模块的输入;第三,系统对相关文档进行解析,生成查询扩展所需的索引结构文件;第四,在查询词扩展模块中人工设置节点语义权重,系统根据节点语义权重、词频、层次、词项与初始查询词间的距离关系计算相关文档中词项的权重,根据查询扩展词权重计算公式系统得到扩展查询词候选集合,选取权重最大的几个候选查询词作为结构扩展模块的输入;第五,经结构查询扩展后得到的最终扩展查询表达式,提交给XML 搜索引擎进行进一步的检索,将检索结果返回用户.该过程为可循环过程,即用户若对扩展查询检索结果不满意,还可在检索结果中再次选取相关文档作为扩展模块的输入,再次进行查询扩展和再次对扩展查询进行检索.

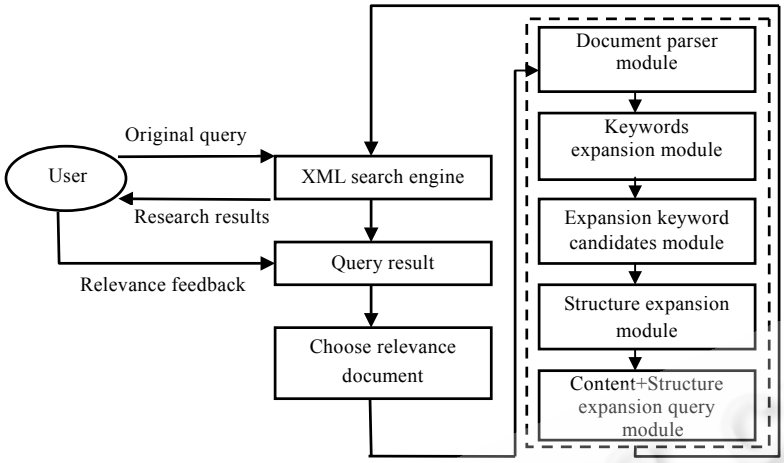


Fig.2 System structure of query expansion based on relevance feedback

图 2 基于用户相关反馈的“内容+结构”查询扩展体系结构

4 实 验

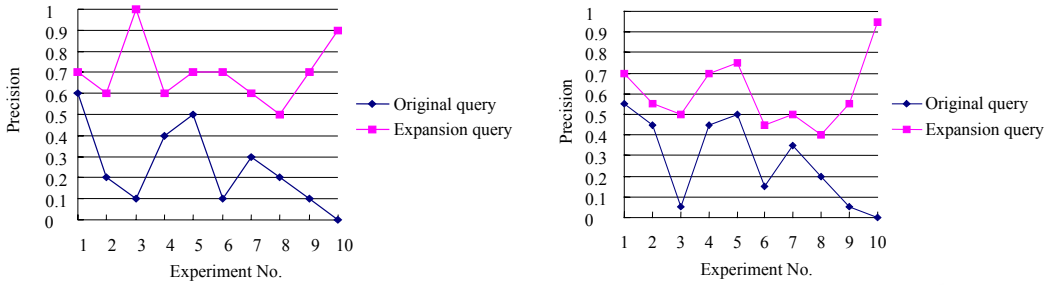
实验平台是 Pentium 3.2G CPU,1G 内存,Windows XP 操作系统,程序在 VC6.0 下编译通过.我们使用了 XML 信息检索中著名的 INEX benchmark,测试数据集为 INEX 2006 的 Wikipedia XML 文集,共包含 659 388 篇文章.实验中对查询词的检索排序采用了 INEX 2006 提供的 TopX 检索排序引擎[32].实验中参数的设定如下:元素权重 W_e 的值分别为 W_{name}=1.0,W_{title}=0.5,W_p=0.05,W_{emph}=0.2,$W_{section}$=0.05.实验中只考虑了 $d(q_i,k)$=1 的情况.

在检索过程中力图模拟普通用户的行为:初始查询词一般选择一个或两个,因为根据调查表明 80%的用户在进行信息检索时查询词不超过两个;在返回的初始检索结果中,选择的相关文档为 1~3 篇,这也是充分考虑了用户的浏览习惯,用户一般不会打开太多的文档进行浏览.本实验共进行了 10 组比较,主要评测指标为 Prec@10 和 Prec@20,Prec@X 指的是针对某个查询 Q,在检索出的 X 篇文档中的平均准确率.Prec@10 和 Prec@20 在搜索引擎中通常反映了第 1 页和前两页的检索结果的准确率.对于 10 个测试实验,它们的初始查询词、选择的相关文档以及得到的“内容+结构”扩展查询表达式见表 1,扩展前后检索结果准确率的对比如图 3 所示.

Table 1 Original query, relevance documents and expansion queries

表 1 实验初始查询词、相关文档及扩展查询

No.	Original query	Relevance documents	“content+structure” expansion queries
1	“dvd”	8275 “DVD” 548444 “HD DVD”	//article[about(.//title,“dvd”) and about (.//p, format hd history disc video)]
2	“cpu”	19553 “microprocessor” 5218 “central processing unit”	//article[about (.//p, central microprocessor cpu unit process design)]
3	“algorithm”	40254 “Genetic algorithm” 901149 “Human based genetic algorithm”	//article[about(.//title,“genetic algorithm”) and about (.//p, ga human solute)]
4	“web services”	93483 “web service” 705779 “web services resource frame”	//article[about(.//title,“web service”)]
5	“software”	10635 “free software” 2025551 “open source vs. free software”	//article[about(.//title,“free software”) and about (.//p, open)]
6	“retrieval”	15271 “Information retrieval” 1315248 “Full text search”	//article[about (.//p, retrieval document text search precise)]
7	“Huffman”	13883 “Huffman coding” 66139 “Prefix coding”	//article[about(.//title,“coding”) and about (.//p, Huffman Prefix)]
8	“dragon”	66203 “Chinese dragon” 254956 “dragon king” 156154 “culture of Chinese”	//article[about (.//p, dragon Chinese culture)]
9	“Music”	15613 “Jazz” 99040 “Free jazz”	//article[about(.//title,“jazz”) and about (.//p, music style improvise)]
10	“language”	64750 “Computer language” 189845 “Low-level programming language” 701685 “Fifth-generation programming language”	//article[about(.//title,“language”) and about (.//p,program)]



(a) Prec@10 (b) Prec@20

Fig.3 Precision for original query and expansion query

图 3 扩展前后的检索结果准确率

对于 Prec@10,扩展查询检索结果的平均准确率为 70%,而初始查询的平均准确率为 25%;对于 Prec@20,扩展查询检索结果的平均准确率为 61%,而初始查询的平均准确率为 28%.实验结果表明,本文提出的查询扩展方法是有效的.

Schenkel 提出的结构查询扩展方法[26−28]针对 IEEE 数据集进行了实验,当用户反馈的相关元素为 1~3 个时,对于 Prec@10 的检索结果准确率与 TopX 基准查询准确率基本相同;只有当用户反馈的相关元素大于 3 个时,才显示出结构查询扩展的优势,对于 Prec@10 的平均准确率提高幅度为 28%.而我们实验中的检索结果准确率是在用户反馈结果为 1~3 个的情况下产生的,对于 Prec@10 的平均准确率的提高幅度为 45%.

5 结束语

本文主要的贡献有以下几点:

(1) 提出了基于相关反馈的 XML 内容(即查询词)查询扩展的新方法.现有的 XML 文档信息检索的内容查询扩展方法,在计算扩展词权重时,大多考虑相关文档中词项的频率因素,而没有考虑 XML 文档的结构信息对扩展词权重的影响.本文提出的计算候选查询扩展词权重的方法,除考虑了频率因素外,还考虑了文档的元素语义权重、元素在文档中所在层次及词项与初始查询词的距离关系等因素.

(2) 提出了基于相关反馈的 XML 结构查询扩展的新思路.目前对于 XML 结构查询扩展的研究极少[25−30],比较成熟的是 Schenkel 提出的结构查询扩展方法[26−28],该方法是将传统文本文档信息检索的关键词查询扩展方法引入到 XML 文档中,将原来基于关键词频率计算扩展查询词权重的公式,改造为基于 *tag-term* 对的频率计算扩展结构查询权重的公式.此方法考虑的是 *tag-term* 对的频率因素,始终将元素 *tag* 与词项 *term* 作为一个整体考虑,同等对待,并且只考虑 *tag-term* 对的频率因素而没有考虑 *tag* 语义对于 *term* 的影响.实验结果[28]表明,只有当用户反馈的相关元素大于 3 个时,此方法才显示出结构查询扩展的优势,其平均准确率提高幅度为 28%(针对 IEEE 数据集).

要求在浏览结果中标注至少 4 个相关结果,无疑给用户带来很大的困难,所以,我们设计的反馈过程中力图降低用户行为的复杂度,要求用户只标识 1~3 个相关结果,实验中以标识 2 个相关结果为主,且检索结果准确率 Prec@10 的提高可以达到 45%(针对 Wikipedia XML 数据集).本文与 Schenkel 的查询扩展思路的根本不同点在于 Schenkel 方法中将扩展表达式中结构与内容同等对待,始终将元素 *tag* 与词项 *term* 作为一个整体来考虑,通过统计 *tag-term* 对的频率决定查询扩展表达式中的结构分支. 而本文对于扩展查询表达式的定位为内容为主、结构为辅,因此我们首先进行查询词扩展,在查询词扩展中并不单纯考虑频率因素,而是将结构因素对于词项权重的影响也充分考虑了,在找准扩展查询词的基础上再进行结构抽取,实验结果表明,本文提出的方法能够有效地提高检索准确率,并且准确率的提高幅度要大于 Schenkel 方法.

(3) 进行了大量的实验测试,实验结果表明,本文提出的查询扩展方法得到的扩展表达式,与初始查询相比,检索准确率有大幅度的提高,达到了本文的研究目的.

References:

[1] Voorhees E, Harman D. Overview of the 6th text retrieval Conf. In: Proc. of the 6th text retrieval Conf. 1997. 1−24. http://trec.nist.gov/pubs/trec6/t6_proceedings.html

[2] Lu BB, Zhao J. Query expansion based on modeling of relevant documents pool. Journal of Chinese Information, 2006,20(3): 78−83 (in Chinese with English abstract).

[3] Liu SH, Wu GS, Zhang FY. Applying relevance feedback to information retrieval using keyword and weight algorithms. Journal of the China Society for Scientific and Technical Information, 2002,21(6):668−673 (in Chinese with English abstract).

[4] Huo H, Feng BQ, Zhao SS. Retrieval algorithm combining multi-query data fusion with positive relevance feedback. Journal of Xi'an Jiaotong University, 2005,39(8):820−823 (in Chinese with English abstract).

[5] Song LL, Cheng Y. About term ranking methods in relevance feedback. New Technology of Library and Information Service, 2004, (8):44−47 (in Chinese with English abstract).

[6] Mitra M, Singhal A, Buckley C. Improving automatic query expansion. In: Proc. of the 21st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. 1998. 206−214. http://www.informatik.uni-trier.de/~ley/db/conf/sigir/sigir98.html

[7] Amo P, Ferreras FL, Cruz F, Rosa M. Smoothing functions for automatic relevance feedback in information retrieval. In: Mohamed TI, Josef K, Norman R, eds. Proc. of the 11th Int'l Workshop on Database and Expert Systems Applications. Heidelberg: Springer-Verlag, 2000. 115−119.

[8] Lu A, Ayoub M, Dong J. Ad hoc experiments using EUREKA. In: Proc. of the 5th Text Retrieval Conf. 1996. 229−240. http://trec.nist.gov/pubs/trec5/t5_proceedings.html

[9] Buckley C, Mitra M, Walz J, Cardie C. Using clustering and SuperConcepts within SMART. In: Proc. of the 6th Text Retrieval Conf. 1997. 107−124. http://trec.nist.gov/pubs/trec6/t6_proceedings.html

[10] Claudio C, Giovanni R, Vittorio G. Improving retrieval feedback with multiple Term2Ranking function combination. ACM Trans. on Information Systems, 2002,20(3):259−290.

[11] Makoto I. Relevance feedback with a small number of relevance judgements: Incremental relevance feedback vs. document clustering. In: Nicholas JB, Peter I, Mun-Kew L, eds. Proc. of the 23rd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2000. 10−16.

[12] Ding GD, Bai S, Wang B. Local co-occurrence based query expansion for information retrieval. Journal of Chinese Information, 2006,20(3):84−91 (in Chinese with English abstract).

[13] Huang XJ, Xia YJ, Wu LD. Text filtering system based on vector space model. Journal of Software, 2003,14(3):435−442 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/435.pdf

[14] Yue W. Information-Retrieval algorithm based on query expansion and classification. Journal of System Simulation, 2006,18(7): 1926−1934 (in Chinese with English abstract).

[15] Ryen WW, Ian R, Joemon MJ. The use of implicit evidence for relevance feedback in Web retrieval. In: Fabio C, Mark G, Rijsbergen CJ, eds. Proc. of Conf. of 24th BCS-IRSG European Colloquium on IR Research. Heidelberg: Springer-Verlag, 2002. 93−109.

[16] Sugiyama K, Hatano K, Yoshikawa M. Adaptive Web search based on user profile constructed without any effort from users. In: Proc. of the 13th Int'l Conf. on World Wide Web. 2004. 675−684. http://www.iw3c2.org/WWW2004/docs/1p675.pdf

[17] Shen XH. Context-Sensitive information retrieval using implicit feedback. In: Ricardo AB, Nivio Z, Gary M, *et al*., eds. Proc. of the 28th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2005. 43−50.

[18] Wang ZJ, Yu C. Technology and implementation of personal information retrieval based on implicit feedback. Computer Engineering, 2003,29(6):158−192 (in Chinese with English abstract).

[19] Chui H, Wen JR, Li MQ. A statistical query expansion model based on query logs. Journal of Software, 2003,14(9):1593−1599 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/1593.pdf

[20] Chao HL, Zhu X, Yu Y. SDQ E: A semantic query optimization in P2P Sys tem. Journal of Shanghai Jiaotong University, 2005, 39(10):1706−1760 (in Chinese with English abstract).

[21] Crouch CJ, Mahajan A, Bellamkonda A. Flexible XML retrieval based on the extended vector model. In: Norbert F, Mounia L, Saadia M, Zoltán S, eds. INEX Workshop Proc. Heidelberg: Springer-Verlag, 2005. 232−302.

[22] Mass Y, Mandelbrod M. Relevance feedback for XML retrieval. In: Norbert F, Mounia L, Saadia M, Zoltán S, eds. INEX Workshop Proc. Heidelberg: Springer-Verlag, 2005. 303−310.

[23] Sigurbjornsson B, Kamps J, Rijke MD. The university of amsterdam at INEX 2004. In: Norbert F, Mounia L, Saadia M, Zoltán S, eds. INEX Workshop Proc. Heidelberg: Springer-Verlag, 2005. 104−109.

[24] Pan H. Relevance feedback in XML retrieval. In: Elisa B, Stavros C, Dimitris P, *et al*., eds. Proc. of the EDBT 2004. Heidelberg: Springer-Verlag, 2005. 187−196.

[25] Hlaoua L, Boughanem M. Towards context and structural relevance feedback in XML retrieval. In: Proc. of the Workshop on Open Source Web Information Retrieval. 2005. http://www.emse.fr/OSWIR05/

[26] Schenkel R, Theobald M. Relevance feedback for structural query expansion. In: Norbert F, Mounia L, Saadia M, Zoltán S, Gabriella K, eds. INEX 2005 Workshop. LNCS 3997, Heidelberg: Springer-Verlag, 2006. 344−357.

[27] Schenkel R, Theobald M. Structural feedback for Keyword-Based XML retrieval. In: Mounia L, Andy M, Stefan M, *et al*., eds. Proc. of the 28th European Conf. on IR Research. LNCS 3936, Heidelberg: Springer-Verlag, 2006. 326−337.

[28] Schenkel R, Theobald M. Feedback-Driben structural query expansion for ranked retrieval of XML data. In: Yannis E, Marc H, Joachim W, *et al*., eds. Proc. of the 10th Int'l Conf. on Extending Database Technology. LNCS 3896, Heidelberg: Springer-Verlag, 2006. 331−348.

[29] Li XH. Design of a XML information retrieval system with feedback. Journal of Xiamen University of Technology, 2006, 14(1):33−36 (in Chinese with English abstract).

[30] Li JB, Li XH. Research about XML information retrieval system based on feedback. Journal of Information, 2005,24(10):72−74 (in Chinese with English abstract).

[31] Doucet A, Aunimo L, Lehtonen M, Petit R. Accurate retrieval of XML document fragments using EXTIRP. In: INEX Workshop Proc. 2005. 73−80. http://www.cs.helsinki.fi/u/doucet/papers/inex03.pdf

[32] http://inex.is.informatik.uni-duisburg.de/2006/

附中文参考文献:

[2] 吕碧波,赵军.基于相关文档池建模的查询扩展.中文信息学报,2006,20(3):78−83.

[3] 刘绍翰,武港山,张福炎.基于词条权值的相关反馈算法在 Web 信息检索中的应用.情报学报,2002,21(6):668−673.

[4] 霍华,冯博琴,赵深深.基于多查询数据融合和正相关反馈的检索算法.西安交通大学学报,2005,39(8):820−823.

[5] 宋玲丽,成颖.相关反馈技术中的检索词排序算法.现代图书情报技术,2004,(8):44−47.

[12] 丁国栋,白硕,王斌.一种基于局部共现的查询扩展方法.中文信息学报,2006,20(3):84−91.

[13] 黄萱菁,夏迎炬,吴立德.基于向量空间模型的文本过滤系统.软件学报,2003,14(3):435−442. http://www.jos.org.cn/1000-9825/14/435.pdf

[14] 岳文,陈治平,林亚平.基于查询扩展和分类的信息检索算法.系统仿真学报,2006,18(7):1926−1934.

[18] 王志军,于超.基于隐式反馈的个人信息检索技术及实现.计算机工程,2003,29(6):158−192.

[19] 崔航,文继荣,李敏强.基于用户日志的查询扩展统计模型.软件学报,2003,14(9):1593−1599. http://www.jos.org.cn/1000-9825/14/1593.pdf

[20] 曹华梁,朱星,俞勇.适用于 P2P 的系统查询扩展优化方法.上海交通大学学报,2005,39(10):1706−1710.

[29] 李小华.一种带反馈的 XML 信息检索系统设计与研究.厦门理工学院学报,2006,14(1):33−36.

[30] 李剑波,李小华.基于 XML 的反馈式信息检索系统研究.情报检索,2005,24(10):72−74.

万常选(1962−),男,江西新建人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为 XML 数据管理,信息检索,数据挖掘.

鲁远(1974−),女,硕士生,主要研究领域为数据库,信息检索.