doi:10.3969/j.issn.1005-3026.2015.07.005

基于搜索日志与局部上下文的查询扩展方法

张书波,马安香,张 斌,孙达明
(东北大学 信息科学与工程学院,辽宁 沈阳 110819)

摘 要:当搜索日志存在查询词稀疏性和时效性问题时,基于搜索日志的查询扩展方法无法保证查询扩展的质量,无法满足具有时效性查询请求的需求.提出了基于搜索日志与局部上下文的查询扩展方法,通过发掘搜索日志中用户查询词和相关文档的关联关系,抽取查询扩展词,并用局部上下文从相关文档集中提取出频率较大,同时具有时效性的查询扩展词.通过对查询扩展词的查询性能和时效性的计算,该方法为原始查询补充更高质量的查询扩展词.实验结果表明,该方法能够有效地提升准确率和召回率,使原始查询获得更好的查询性能.

关 键 词:信息检索;查询扩展;搜索日志;局部上下文;查询性能
中图分类号:TP 391 **文献标志码**:A **文章编号**:1005-3026(2015)07-0933-04

A Query Expansion Method Based on Search Log and Local Context

ZHANG Shu-bo, *MA An-xiang*, *ZHANG Bin*, *SUN Da-ming*
(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: ZHANG Bin, professor, E-mail: zhangbin@mail.neu.edu.cn)

Abstract: When search logs encounter over sparse and timeliness problems, the quality of extensions cannot be guaranteed by the query expansion methods based on the search log, resulting in the impossibility to meet the needs of timeliness information. A query expansion method based on the search log and local context was introduced. By exploring relationships between user queries and relevant documents in search logs, candidate expansion terms were extracted, and local contexts from related documents were concentrated to extract words with higher frequencies and timeliness. By considering query performances and timeliness of expansions, expansions with higher qualities were provided with this method. Experimental results showed that precisions and recalls could be effectively improved, and better query performance to original queries could be obtained.

Key words: information retrieval; query expansion; search log; local context; query performance

搜索引擎使人们能从丰富的信息资源海洋中快速找到所需信息,但存在词典匹配问题[1].查询扩展技术是解决这一问题的有效方法.目前,查询扩展技术根据扩展词的来源和提取方法不同可分为基于全局分析[2]、基于局部分析[3]、基于用户查询日志[4]和基于语义资料[5],实验结果表明,在一定的条件下,这些方法可以提高查询质量,改善搜索性能.

前人研究结果表明,当搜索日志中存在大量关联查询时,基于用户查询日志的查询扩展方法能够提高查询质量,但是基于用户查询日志的查询扩展存在两个影响查询扩展效果的问题.首先,数据稀疏问题,导致基于搜索日志进行查询扩展的方法无法提取有效的相关查询词,致使查询扩展性能不佳.其次,数据时效性问题,导致扩展出的查询词没有反映当前文档集的真实状态;搜索

收稿日期:2014-05-22
基金项目:国家自然科学基金资助项目(61100090);中央高校基本科研业务费专项资金资助项目(N110204006,N120804001,N110604002,N120604003).
作者简介:张书波(1973-),男,山东烟台人,东北大学博士研究生;张 斌(1964-),男,辽宁本溪人,东北大学教授,博士生导师.

日志和当前文档集的时效性关联度不够,出现找不到新文档的现象,影响查询性能.

本文针对数据稀疏问题和时效性问题,提出了基于搜索日志与局部上下文的查询扩展方法.该方法合并基于搜索日志提取的候选扩展词和结合时效性改进的局部上下文查询扩展方法抽取的候选扩展词,通过判断候选扩展词的查询性能,提取查询性能高的 M 个扩展词,加入到原查询中构成最终的查询.实验结果表明,本文的查询扩展方法有效地解决了存在的数据稀疏问题和时效性问题,提高了查询性能.

1 基于搜索日志与局部上下文的查询扩展方法

近几年,越来越多的学者对查询扩展方法进行了研究.文献[6]分别基于搜索日志和组合方法进行了研究,提出了相应的查询扩展方法,提高了查询性能.但当搜索日志存在数据稀疏和时效性问题时,提取的扩展词较少或很旧,将会很大地影响查询扩展的质量.在用户查询词较少时会出现查询扩展效果不佳,在搜索日志过旧时,会遗漏新的文档信息,尤其对新闻类的信息检索影响更明显.针对这一问题,本文提出了基于搜索日志与局部上下文的查询扩展方法(query expansion method based on search log and local context QEMSLLC),从局部文档集提取查询候选扩展词来弥补基于用户日志候选扩展词较少和时效性不强的缺陷,提高查询效能.

1.1 QEMSLLC 的基本思想

本文方法的基本思想:依据用户输入的查询词,利用用户日志挖掘的查询扩展方法获得候选扩展词集.如果存在数据稀疏问题,导致扩展词数量很少时,继续判断是否存在时效性问题,若存在,则利用处理时效性问题的局部上下文查询扩展方法来选取扩展词,若不存在,则通过不处理时效性问题的局部上下文查询扩展方法选取扩展词,合并两个候选扩展词集,计算扩展词权重,并依据扩展词选取规则提取扩展词作为最后的扩展词集.如果不存在数据稀疏问题,但搜索日志数据很旧,时效性差,即存在数据时效性问题时,利用处理时效性问题局部上下文查询扩展方法来选取扩展词,合并两个候选扩展词集,计算词项时效价值度和权重,并依据扩展词选取规则提取扩展词作为最后的扩展词集.QEMSLLC 的总体流程如图 1 所示.

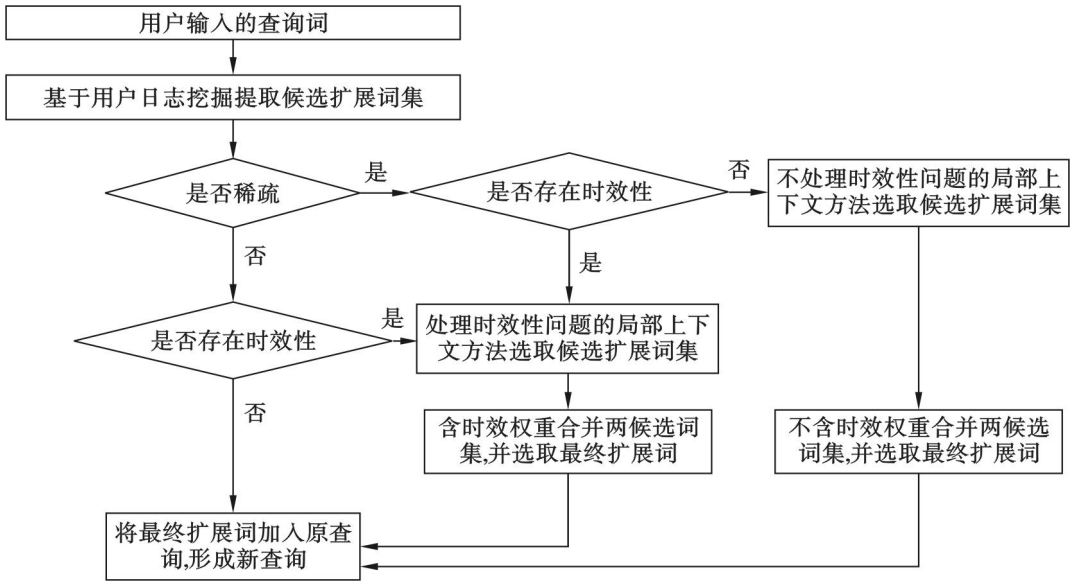

图 1 QEMSLLC 的总体流程

Fig. 1 Overall process of QEMSLLC

在本文方法的总体流程中有 3 个关键问题需要解决:一是如何判断是否存在数据稀疏问题;二是如何判断是否具有数据时效性问题;三是两个扩展词集如何有效合并问题.在本文查询扩展方法中采用了基于用户日志挖掘的方法、局部上下文分析的查询扩展方法[7]和查询性能预测方法[8],这些方法为本文的研究工作提供了基础.同时依据以下思想,分别解决上述 3 个问题.

1)对于数据稀疏的判断问题.如果词汇在搜索日志中出现较少,则无法直接从搜索日志中获取有价值的信息实现查询扩展.此时,单纯基于搜索日志的查询扩展方法便无法取得高质量的结

果. 针对此类搜索请求，考虑从局部上下文中提取相关概念，补充从稀疏的搜索日志中选取的扩展词，实现查询扩展. 由于词汇的出现情况为连续值，以固定阈值的方式判断词汇出现是否稀疏的方法并不科学合理，因此考虑定义比例函数$f(x)$，其中 x 为词汇在搜索日志中出现的次数. 以此为基础，融合分别从搜索日志和局部上下文提取的扩展词结果集，实现对具有稀疏性词汇的有效查询扩展.

2）对于时效性的判断问题. 搜索日志具有时效性. 在用户查询后出现的新的相关文档包含时效性更强的相关扩展词，而基于用户日志的查询扩展方法无法获得这些扩展词，进而降低查询性能. 为获取时效性强的扩展词，提出词项时效价值度来衡量候选扩展词的时效性强弱，词项时效价值度值越大，词项的时效性越强，反之越低. 融合分别从搜索日志和局部上下文提取的扩展词结果集，实现对具有时效性词汇的有效查询扩展.

3）对于扩展词集的合并问题. 为从基于搜索日志提取的候选扩展词集和基于局部上下文提取的候选扩展词集中提取查询性能高的扩展词，根据扩展词判定规则，通过计算扩展词的查询性能，从大到小选取具有高查询性能的前 m 个扩展词，作为最终扩展词.

1.2　数据稀疏和时效性问题判定规则

1.2.1　数据稀疏性问题判定规则

如果查询词在搜索日志中出现次数比较少，则一定无法扩展出好结果，因此可以用局部上下文选取扩展词进行补充.

因此，假设使用跳跃函数

$$f(x)=\begin{cases}1, & x>5;\\ 0, & x\leqslant 5.\end{cases}$$

并以查询词出现次数 $x=5$ 为标准判断是否稀疏，那么，稀疏性可以定义为

$$f(x)\times \text{日志}+(1-f(x))\times \text{文档}.$$

其意思是如果 $x>5$，就全用基于搜索日志的扩展词集，否则就用基于搜索日志和局部上下文的两种扩展词集.

接下来，考虑一个光滑的函数

$$f(x)=1/(1+\mathrm{e}^{-(x-5)}),$$

其意思是，当查询词出现 5 次的时候，这个概念既是稀疏的，又是不稀疏的，因此当 $x<5$ 时就倾向于稀疏，$x>5$ 时就倾向于不稀疏. 此时进一步定义为

$$f(x)\times \text{日志}+(1-f(x))\times \text{文档}.$$

这是第二次定义，唯一的改变是修改了 $f(x)$ 的形式. 因此，可以令 Q 为搜索日志中的查询词集，q 为查询，x 为 q 在 Q 中出现的次数，比例函数 $f(x)=1/(1+\mathrm{e}^{-(x-\alpha)})$，当 $f(x)<0.5$ 时，称为搜索日志对 q 是稀疏的，α 为阈值.

1.2.2　数据时效性问题判定规则

文档的创建时间离当前系统时间越近，文档的时效性越强，文档中的词项时效价值越高，反之越低. 在搜索日志之后出现与查询词相关的仅包含新概念的文档，这时搜索日志会存在时效性问题. 因此，可以定义搜索日志对查询词存在时效性问题.

令 Q 为搜索日志中的查询词集，q 为查询，t_{d} 为 q 在 Q 中出现的最后时间，t'_{d} 为与 q 相关的新文档出现的最新时间，当 $t'_{\mathrm{d}}>t_{\mathrm{d}}$ 时，称为搜索日志对 q 存在时效性问题.

同理可以定义词项时效价值度，利用以 e 为底，文档的时间属性与当前系统时间的比值为幂的指数函数来度量词项的时效价值，即词项时效价值度(timeliness)：

$$\text{timeliness}(t)=\mathrm{e}^{\frac{\text{number}(d)}{\text{number}(s)}}. \tag{1}$$

其中：d 表示词项 t 所在的文档最后修改时间；s 表示当前系统时间.

词项 t 与查询 Q 的相关度可以用增加词项时效价值度的改进局部上下文分析的相关度公式来计算，即

$$W(t,Q)=\text{timeliness}(t)\times f(t,Q). \tag{2}$$

1.3　基于扩展词权重计算的候选集合并方法

定义扩展词权重：对来源于搜索日志里的候选扩展词集 T 和文档集的候选扩展词集 T'，设 $\lambda\in(0,1)$ 为权重调节因子. 当扩展词 t 在词集 T 和 T' 中同时出现时，加大其权重值，当扩展词 t 仅在词集 T 中出现时，降低其权重值，当扩展词 t 仅在词集 T' 中出现时，保持其权重值. 即

$$\left.\begin{array}{l}\text{当 } t\in T\cap T' \text{ 时},\text{weight}(t)=(1+\lambda)\times\text{IDF}(t);\\ \text{当 } t\in T \text{ 且 } t\notin T' \text{ 时},\text{weight}(t)=\lambda\times\text{IDF}(t);\\ \text{当 } t\in T' \text{ 且 } t\notin T \text{ 时},\text{weight}(t)=\text{IDF}(t).\end{array}\right\} \tag{3}$$

则最终扩展词权重为

$$\text{wtl}(t)=\text{timeliness}(t)\times\text{weight}(t). \tag{4}$$

两个候选扩展词词集的合并规则：

1）当搜索日志存在稀疏问题时，若不存在时效性问题，则利用式(3)计算候选扩展词集 T 和 T' 中的扩展词权重，若存在，则利用式(4)计算候选扩展词集 T 和 T' 中的扩展词权重，选择权重值大的前 m 个词作为最终扩展词加入初始查询.

2）在搜索日志存在时效性问题时，可以利用式(4)计算候选扩展词集 T 和 T' 中的扩展词权重，选择结果值最大的前 m 个词作为最终扩展词

加入初始查询.

本文查询扩展方法与其他基于用户查询日志的查询扩展方法最大的不同是,解决了搜索日志存在数据稀疏或时效性问题时,查询扩展质量不高问题. 针对搜索日志查询词数据稀疏问题,通过搜索日志扩展后的查询搜索到相关文档,依据局部上下文的相关度计算提取候选扩展词补充扩展词. 针对搜索日志的时效性问题,通过选择查询后的新文档,依据局部上下文的相关度和词项时效价值度计算提取时效性强的扩展词. 然后通过计算扩展词的查询性能选取查询性能高的词作为最终扩展词.

2 实验与分析

本文实验采用标准的 MAP,Prec@10,Prec@20 指标,对比分析查询扩展方法的查询性能. 实验的基线(baseline)是没用任何查询扩展方法的向量空间模型系统搜索结果. 在实验后,可在全部的查询集合上,综合对比分析基准系统结果(baseline)、基于用户日志挖掘取得的结果(QE on log)以及基于搜索日志与局部上下文查询扩展后的结果(QEMSLLC).

本文实验采用了 Sogou 搜索引擎提供的 2012 年的用户搜索日志. 为避免影响实验效果,保证实验结果客观实用,在实验时将系统时间调整为 2012 年,并且不选择用户搜索日志截止时间以后的文档作为局部上下文查询扩展方法中的文档.

在实验中,使用与用户初始查询相关的前 50 篇文档作为提取查询扩展词的来源. 选择前 20 个与用户初始查询关联度最高的词为最终查询扩展词. $\lambda = 0.6, \alpha = 20$.

使用“方太”、“闹钟”、“卤菜”、“山吉树”、“紫檀木”、“山语城”、“美丽人生”7 个用户查询词作为测试用查询,当搜索日志存在数据稀疏问题时,查询扩展性能对比结果见表 1.

表 1 存在数据稀疏问题时,查询扩展性能对比结果
Table 1 Query expansion performance comparison with data sparse problem

方法	MAP	Prec@10	Prec@20
baseline	0.261 5	0.543	0.514
QE on log	0.335 0	0.658	0.626
QEMSLLC	0.358 0	0.758	0.736

使用“最新电影”、“美国总统”、“流行歌曲”、“年终总结”、“马年运势”、“足球世界杯”、“元旦祝福语”7 个用户查询词作为测试用查询,当搜索日志存在数据时效性问题时,查询扩展性能对比结果见表 2.

表 2 存在数据时效性问题时,查询扩展性能对比结果
Table 2 Query expansion performance comparison with data timeliness problem

方法	MAP	Prec@10	Prec@20
baseline	0.261 5	0.543	0.514
QE on log	0.353 0	0.732	0.712
QEMSLLC	0.365 0	0.766	0.748

从表 1,表 2 中可以看出,查询扩展比没有查询扩展提高了搜索性能. 当搜索日志存在数据稀疏问题时,本文方法比基于搜索日志的查询扩展方法提高了搜索性能,其中 MAP 提高了 8.8%. 当搜索日志存在数据时效性问题时,本文方法比基于搜索日志的查询扩展方法提高了搜索性能,其中 MAP 高出 4.6%.

3 结 论

本文提出的基于搜索日志与局部上下文的查询扩展方法有效地解决了基于用户查询日志的查询扩展方法存在的数据稀疏和时效性问题. 实验结果表明本文方法虽然查询执行平均时间稍长,但在可接受范围内,并取得了很好的查询效果,提高了查询性能.

参考文献:

[1] Furnas G W, Landauer T K, Gomez L M, et al. The vocabulary problem in human-system communication[J]. *Communication of ACM*, 1987, 30(11): 964 – 971.

[2] Runkler T A, Bezdek J C. Automatic keyword extraction with relational clustering and Levenshtein distances[J]. *Institute of Electrical and Electronics Engineers*, 2002, 9(2): 636 – 640.

[3] Buckley C, Salton G, Allan J, et al. Automatic query expansion using SMART[C]//Proceedings of the 3rd Text Retrieval Conference. Gaithersburg: Comell University, 1994: 69 – 80.

[4] Cui H, Wen J R, Nie J Y, et al. Probabilistic query expansion using query logs[C]// Proceedings of the 11th International Conference on World Wide Web. New York, 2002: 325 – 332.

[5] Fang H. A re-examination of query expansion using lexical resources[C]//Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics. Clumbusohio, 2008: 139 – 147.

[6] Pal D, Mitra M, Datta K. Query expansion using term distribution and term association[J]. *The Computing Research Repository*, 2013, 21(3): 1 – 19.

[7] Xu J X, Croft W B. Improving the effectiveness of information retrieval with local context analysis[J]. *ACM Transactions on Information Systems*, 2000, 18(1): 79 – 112.

[8] He B, Ounis I. Inferring query performance using pre-retrieval predictors[C] //The 11th International Conference on String Processing and Information Retrieval. Padova, 2004: 43 – 54.