基于用户日志分析的查询扩展研究

李东园，白宇，蔡东风
（沈阳航空工业学院 知识工程中心 辽宁 沈阳 110034）
Email：li_dongyuan@126.com

摘 要： 查询扩展是信息检索中关键问题之一，查询扩展的有效性决定其检索性能。用户查询日志是大量用户长期查询行为的记录，通过挖掘用户查询与用户日志之间的联系，构建相关词表，从而实现查询扩展。本文提出了一种结合局部上下文分析与用户行为分析的查询扩展方法，该方法在选择相关用户日志时加入用户查询行为信息，获取相关词表时采用局部上下文分析方法。在54000篇语料上的测试表明，该方法相对于未扩展时准确率提高50%以上。

关键词： 查询扩展，用户日志，用户行为分析

A Study of Query Extension Based on Query Log Analysis

Li Dongyuan，Bai Yu，Cai Dongfeng
(Knowledge Engineering Center，Shenyang Institute of Aeronautical Engineering，Shenyang，Liaoning，110034)
Email：li_dongyuan@126.com

Abstract: Query extension is a critical point in information retrieval, the efficiency of query extension detennines information retrieval performance. Query log is a record of query behavior by a great quantity of users, it can find the related word list through mining the relation between query and query log, thus to realize the query extension. In this paper, we propose a method for query extension which combines local context analysis with query behavior. User query behavior is used in selecting related query logs, and local context analysis is used to getting related word list The testing based on 54,000 papers shows that the precision has about 50% improvement after using this method.

Key words: query extension; query log; query log analysis

1 引言

随着Internet的广泛应用和普及，使用检索系统的用户从原来单一的专业人员拓展到不同层次、拥有不同知识背景的普通用户。这些用户的查询请求往往只包含少量的关键词，通常不能准确的描述自身的需求，结果导致检索返回的查询结果往往不能完全满足用户的需要。同时，由于自然语言中还存在着大量的同义现象，相同的概念通常以不同的形式表现出来，产生用户查询中词语与文档集合中的词语不匹配的现象。如：用户输入关键词“元宵节”，但在文档集合中可能以“正月十五”形式存在，正是由于这种不匹配现象导致检索效率低下甚至检索失败。因此，解决用户查询用词与文档集合用词之间的匹配性问题是信息检索领域中的一个关键问题，查询扩展是解决这种不匹配性问题的一个有效的方法。

查询扩展，即在用户所给查询用词的基础上，按照一定的扩展策略，构建相关词语列表，从而在检索时能够返回更多的相关文档，提升检索性能。查询扩展即是挖掘词语之间联系，由于

作者简介：李东园（1980—），女，硕士研究生，主要研究方向为自然语言处理、信息检索；白宇（1982—），男，硕士，助教，主要研究方向为自然语言处理、问答系统；蔡东风（1958—），男，博士，教授，主要研究方向为人工智能理、自然语言处理。

用户查询日志是大量用户长期查询行为的记录，通过分析多个用户对同一查询的不同行为，能够建立用户查询与用户文档中词语之间的关联，从而达到扩展的目的。

本文主要讨论如何更加充分、有效的利用用户日志实现查询扩展，通过分析用户查询和用户行为，初步确定一个相关文档集合，然后对相关文档集合进行过滤，得到与原始查询相关度更高的文档集合，最后在该文档集合中采用局部上下文分析方法抽取相关词语，形成相关词语列表。在54000篇文章的数据集上进行了实验，并分析了实验结果。

本文的组织结构如下：第二节介绍现有的查询扩展方法；第三节介绍用户查询日志；第四节详细阐述如何利用用户查询日志进行查询扩展；第五节是实验及实验结果分析；最后是本文研究工作的总结。

2 查询扩展方法

查询扩展方法基本可分成三类：基于语义资源的方法、全局分析方法、局部分析的方法。

2.1 基于语义资源的查询扩展方法

借助于语义资源，如：WordNet、HowNet等，进行查询扩展。扩展时选择同义词、近义词、同位词、上下位词等高度相关的词。此方法[1]可准确地找到相关词，引入噪声较小，但语义资源中手工分类的粒度不同，在应用中难以把握扩展的力度；并且，目前语义资源中实例有限，存在大量未登录词问题，因此直接使用语义资源进行查询扩展很难取得良好的扩展效果。

2.2 全局分析方法

全局分析是在初始查询条件提交给检索系统前就实行扩展的一种查询扩展方法。其原理是对整个文档集合的词语进行相关分析，得到每对查询用词与文档用词之间的关联程度，并构造相关词表，最后从相词表中选择与原查询关联程度较高的词作为扩展用词。目前常见的全局分析方法包括词聚类(Term Clustering)[2]、潜在语义索引(Latent Semantic Indexing)[3]、相似性词典[4]等方法。词聚类方法的主要缺陷是不能处理查询词的歧义问题，歧义词被分配到多个不同的簇中去，造成扩展词语义过于宽泛，导致查询性能下降；潜在语义索引使用检索词的共现信息进行奇异值分解来发现检索词之间的关联关系，以减少向量空间的维数，该方法能够较好的解决同义词问题，部分解决多义词问题，其缺点是开销大，以牺牲准确率来提高召回率；相似性词典包括基于概念的查询扩展技术和Phrasefinder技术，用来消除查询词的歧义，它部分解决了歧义性问题，但其主要缺点是需要计算每一对词的共现率来产生概念，生成伪文档，造成计算要求较高，导致查询效率下降。

2.3 局部分析方法

局部分析过程中，首先将查询条件提交给检索系统，得到初次检索结果后，对初检结果中的局部信息进行分析，从中找出与初始查询条件相关的特征或直接使用由用户所提供的相关信息来优化查询条件。代表方法有局部反馈[5]、局部上下文分析[6]。局部相关反馈是从用户认为相关的初检文档中选择重要的词语，加重其权重，从而提高检索目标的得分，其缺点是必须依赖于初检结果的相关性判断。局部上下文分析方法是从初检出的文档中选出与原查询词共现的概念，计算每一个概念与整个查询的相似度并排序，排在前面的概念作为扩展词；其缺点是扩展效果高度依赖于初次检索结果，如果初次检索返回的多数文档与原查询无关，将会有大量的无关词加入新

查询，导致检索精度大大降低。

3 用户查询日志

用户查询日志是用户向搜索引擎提交的查询以及相关数据的记录。比如查询时间、查询内容、点击的链接等等。这些数据往往能够反映出用户的兴趣、查询用词的特点、查询内容与用户点击结果链接之间的关系等。本文采用搜狗提供的用户查询日志作为处理对象，结构见表1。

表1 搜狗用户查询日志

查询	session	排名	点击序号	点击链接	时间
元宵节	17899	1	1	www.china.com.cn/ch-jieri/yuanxiao/1.htm	12:00
元宵节	20533	2	1	baike.baidu.com/view/1949.htm	12:00
正月十五	10578	2	1	www.china.com.cn/ch-jieri/yuanxiao/1.htm	12:01
元宵	35461	1	3	www.gio.gov.tw/info/festival_c/glue/glue.htm	19:14

从表 1 中可以看出，查询、点击链接、session 之间是多对多的关系。由于系统为每个用户分配唯一的 session ID，所以相同 session 中是同一用户的查询行为。相同 session 相同查询的查询日志集合实际是用户对该查询的一个多次选择过程，此过程中的每个用户行为都与该查询相关。因此，可将相同 session 相同查询的查询日志记录合并成为一条记录。为了更加清楚的描述问题，本文对以下概念加以说明：

（1）用户文档：用户点击链接对应的页面内容，记作 d_i。

（2）用户日志：相同 session 相同查询包含的所有用户查询日志记录，记作：h_j。

（3）用户查询：初始查询，记作：1。

4 利用用户日志进行查询扩展

假设查询关键词为“元宵节”，用户A向搜索引擎提交了此查询，从返回结果中点击链接为 url_A 的文档，另一用户B也向搜索引擎提交了相同的查询，同样也点击链接为 url_A 的文档，经过长期积累，则可以认为“元宵节”和链接为“url_A”的文档之间存在联系。从用户查询和用户文档间的联系中找到用户查询与用户文档中词语之间的关联，从而实现对用户查询的扩展。下面将详细介绍这一过程。

4.1 将用户查询映射到用户日志中的相关查询

原始用户查询映射到用户日志中相关的查询是将尽可能多的用户文档添加到相关文档集合中。用户日志中相关查询应满足以下两个条件：从内容上来看，若用户查询与用户日志中查询之间的相似度大于某一阈值，则认为用户查询与用户日志中查询是相关的；从用户行为上，多个用户输入某一查询词后，若多个用户选择相同文档的频率大于某一阈值，则认为这两个用户查询之间存在这一定的关联（在这里暂时不考虑用户操作错误这种情况，对于用户操作错误的情况将在下一步中进行处理）。例如：用户 A 输入查询“元宵节”，从返回结果中选择文档 A，另一用户 B 输入查询“正月十五”，从返回结果中也选择文档 A，那么就认为查询“元宵节”和查询“正月十五”之间存在一定的关系。

基于上述思想，采用表 2 所示策略来获取相关文档集合。

表 2 用户查询到用户日志中相关查询的映射策略

设用户查询 I，用户日志 h∈H= {h_1，h_2，…，h_n}， 相关文档 d∈D={Φ}，相关查询 q∈Q={Φ}，相关点击链接 url∈U={Φ}，用户查询与用户日志中查询相似度阈值 λ_1，频率阈值 λ_2。

输入：用户查询 I，用户日志集合 H，用户查询与历史查询相似度阈值 λ_1=2，频率阈值 λ_2=5；

输出：相关文档集合 D，相关查询集合 Q；

步骤：

for h_i in H

Step1：设用户日志中查询为 q_i，用编辑距离计算用户查询 I 与用户日志中查询 q_i 之间的相似度 $sim(I,q_i)$；

Step2：若 $sim(I,q_i)>\lambda_1$ AND q_i 在用户日志中的出现频率> λ_2 则

（1）设 q_i 对应的点击链接集合为 url，q_i 对应的文档集合为 doc，分别将 doc、q_i、url 添加到相关文档集合 D、相关查询集合 Q、相关点击链接集合 U 中；

（2）for sub-url in url

S1：在用户日志集合 H 中查找具有相同 sub-url 的查询集合 H_q；

S2：若 $h_j \in H_q$ 中的点击链接频率大于阈值 λ_2，则将 h_j 中的查询添加到相关查询集合 Q 中去，h_j 中的点击链接添加到相关点击链接集合 U 中；

S3：重复 S2，直到遍历 H_q；

4.2 对相关文档集合过滤

对相关文档集合 D 过滤出于以下两方面考虑：第一，文献[4]研究显示：用户的查询任务包括导航类、信息类和事物类三类，并分别给出了定义。导航类是以直接搜索一个网站入口为目的的查询任务；信息类是在一个或多个网页中查找所需信息为目的的查询任务；事物类是以执行网络媒介活动为目的的查询任务。根据上述定义，可以看出只有信息类查询在查询扩展中是有效的；第二，在3.1节中尽可能多的引入相关文档，同时也会带入一些不相关文档，对于采用局部上下文分析方法抽取相关词语会带来一定的影响。因此对相关文档集合进行过滤是十分必要的。

相关文档过滤基于假设：如果一个文档与该查询高度相关，则该查询中的每个词在这篇文档中均以较高的频率出现。这里采用 OKAPI[7]的 BM25 公式计算原始查询 I 与相关用户文档 d_j 之间的相似度 $sim(Q,d_j)$（3-1）：

$$sim(Q \mid d_j)=\sum_{t \in d_j \cap Q} \frac{(k_1+1)\times tf(t \mid d_j)}{k_1[(1-b)+b\times \frac{dl(d_j)}{avdl(D_c)}]+tf(t \mid d_j)} \times \log \frac{N-df(t \mid D_c)+0.5}{df(t \mid D_c)+0.5} \times tf(t \mid d_j) \tag{3-1}$$

实验表明，用户查询中词项在文档中出现频率对文档相关性的贡献更大，因此本文在原有计算公式基础上，再加入词频特征。公式 3-1 中 $dl(d_j)$ 为相关文档 d_j 的长度，$avdl(D)$ 为相关文档集合 D 的所有文档的平均长度，$tf(t \mid d_j)$ 为相关文档 d_j 中词项 t 的频率，$df(t \mid D)$ 为包含词项 t 的文档数目，N 为相关文档集合 D 中文档数目。在实验中参数 k_1、b 分别设置为 1.2 和 0.75。

4.3 构建相关词语列表

构建相关词语列表是从相关文档集中抽取相关词语，组成相关词表。在抽取相关词语中，采用局部上下文分析方法，取排序前 m 位的词语加入原始查询作为扩展结果。查询与文档中每

个词语的相似度计算公式（3-2）如下：

$$sim(I,c) = \prod_{t \in I} [\frac{\log((\sum_{j=1}^{n} f_{t,j} \times f_{c,j}) \times idf_c)}{\log n}]^{idf_t} \quad (3\text{-}2)$$

其中，$f_{t,j}$ 表示查询词 t 在第 j 个文档中出现的频率，$f_{c,j}$ 表示文档中词语 c 在第 j 个文档中出现的频率，n 表示文档总数，idf_c 表示出现词语 c 的文档数，idf_t 表示出现查询词 t 的文档数。

5 实验及结果分析

5.1 测试语料与实验设置

下载 sogou 提供的 2007 年 3 月的用户查询日志，经过过滤、合并等预处理后形成 247M 的用户查询日志，其中包含 98 万条用户日志记录，将用户日志作为训练数据；收集了联合早报中的 54,000 篇文章，作为该实验的测试数据，并从中提取了 8 个测试问题：

表 3 测试问题集列表

序号	问题	序号	问题
1	元宵节习俗	5	小三通
2	青藏铁路	6	克林顿绯闻
3	生物芯片	7	开放式基金
4	春运	8	意甲

使用向量空间模型（VSM）作为检索算法，将不加扩展的向量空间模型作为 Baseline。将基于用户日志的查询扩展与不加查询扩展以及局部反馈方法进行比较。

基于局部反馈的查询扩展中最经典的为 Rocchio 公式，这里采用标准 Rocchio 公式（4-1）

$$q_m = \alpha q + \frac{\beta}{|D_r|} \sum_{\forall d_j \in D_r} d_j + \frac{\gamma}{|D_n|} \sum_{\forall d_j \notin D_n} d_j \quad (4\text{-}1)$$

其中 D_r 是指用户判定的相关反馈文档集；D_n 是用户判定的不相关文档集；使用与 Salton 相同的实验参数，$\alpha = 1.0, \beta = 0.75, \gamma = 0.25$，取排序后的前 30 个词作为扩展词语加入到原查询中。

在比较查询性能时，采用准确率和召回率作为评价指标。

$$\text{准确率} = \frac{\text{检索结果中的相关文档数}}{\text{检索结果中的全部文档数}} \times 100\%$$

$$\text{召回率} = \frac{\text{检索结果中的相关文档数}}{\text{文档集合中的相关文档数}} \times 100\%$$

5.2 实验结果及分析

对于没有进行查询扩展的检索结果，加入局部反馈后的检索结果以及本文提出的方法的检索结果进行比较。

表 4 查询性能比较

准确率(%) / 召回率(%)	Baseline	局部反馈查询扩展（相对 Baseline 提高）	用户日志查询扩展（相对 Baseline 提高）
20	41.35	59.30(+43.41)	67.93(+64.35)
30	35.82	49.78(+48.97)	55.24(+80.45)
40	30.63	44.97(+46.81)	52.49(+71.37)
50	27.29	38.26(+40.20)	44.80(+64.20)
60	24.92	33.77(+35.51)	39.94(+60.28)
70	21.88	29.50(+34.82)	36.23(+74.74)
80	18.27	24.13(+32.07)	28.77(+51.99)
90	15.92	19.10(+19.97)	25.63(+60.99)
平均值	27.74	37.35(+37.24)	43.88(+60.23)

表 4 可以看出，在查询扩展中使用用户日志信息能够获得比较好的检索效果，相对于无查询扩展的检索结果，最大可提高 80.45%，平均提高 43.88%；相对于局部反馈方法最大提高 22.81%，平均提高 17.48%。

在相关文档集合过滤中，不同的相关文档数目对检索性能造成一定的影响。实验表明，从前 40 篇相关文档中抽取相关词语时，平均检索性能最好，如图 1 所示。这是由于过多的相关文档会使得最终的扩展词语中加入许多无关的词汇，而较少的相关文档则使得扩展词语中包含的信息量较少，使得查询召回率较低。

在查询扩展时，扩展用词的数量对检索性能产生一定的影响，向原始文档中加入 20 个查询词时查询性能达到最好，加入 25 个查询词时性能变化不大，在加入 35 个查询词时性能有较大幅度降低，如图 2 所示。这是由于在进行查询扩展时，一些扩展词与原始查询词语的相关度较低，加入查询中反而会增加噪声，使检索性能下降。

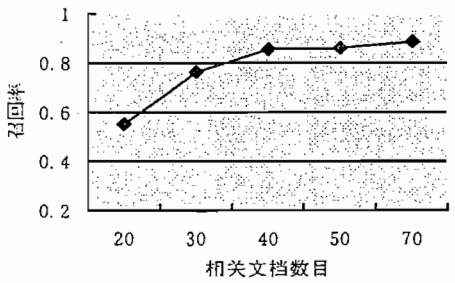

图 1 相关文档数对检索性能影响

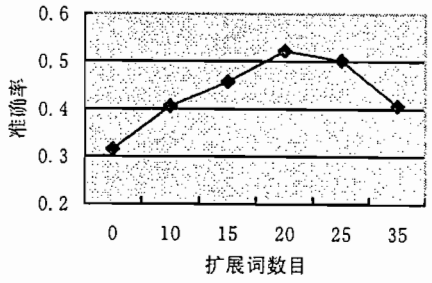

图 2 扩展词数对检索性能影响

对相关文档内容比较分散的查询主题来说，传统的查询扩展方法仅仅在测试语料上进行反馈，因此，对于某一查询主题来说，会存在对于这一主题中某一子主题的反馈文档相对较多、另一子主题的反馈文档相对较少的现象，在选择相关词语时，反馈相对较多的词语的相关度会高于反馈相对较少的词语，根据打分进行筛选时，反馈少的那部分词语被放入相关词表的可能性就比较小，导致相关词表中词语的不平衡（这里的不平衡指该主题中各子主题的没有均匀的扩展），造成返回结果中某一子主题相关的文档集中在前面，而与其他子主题相关的文档在后面，在取其前 n 个文档时可能会将与其它子主题相关的文档舍弃掉，造成召回率的下降。采用本文使用的方法时，如："青藏铁路"扩展后的结果为"青藏、铁路、青海、拉萨、海拔、冻土、高原、格

尔木、西宁、隧道，建设、部队，印度”等，可以近似认为每个子主题的反馈文档具有相同的比例，在构建相关词表时，词表中的词语能够比较均匀的表示该主题各个方面的信息，因此，利用扩展后的词表进行检索时，将出现频率较低的文档包含到检索结果中，从而性能得到提升。

在测试语料中，缩写词、专有名词占有比较大的一部分比例，利用测试文档集合从中查找这些相关词是比较困难的，而在用户日志中对于相同的文档不同用户会使用不同的查询词来检索，可以认为它们之间是存在关联的，因此，对缩写词、专有名词的同义扩展也是提高检索效果的一个因素。

6 结论

用户日志是大量用户长期查询行为的记录，在某种程度上可以看作用户的一种“隐式反馈”，利用用户日志信息能够有效的建立相关词语列表，从而提高检索结果性能。本文不仅考虑文本内容而且在一定程度上也考虑到了用户行为信息，更大限度的利用了用户日志。实验表明，此方法在查询扩展中可以更好的构建相关词语集合，有效的提高检索性能。

下一步还需要对该方法进行更加深入的研究，如：针对原始查询的不同语义选择不同类别的扩展词，从而更进一步缩小扩展用词的范围；在大规模的测试集上验证其有效性；深入挖掘话题的关联词表，确保算法的稳定性。

参 考 文 献

[1]. A.F. Smeaton and C. Berrut. Thresholding postings lists, query expansion by word-word distances and POS tagging of Spanish text. In: Proceedings of the 4th Text Retrieval Conference, 1996.

[2]. Wen JR, Nie JY, Zhang HJ. Clustering user queries of a search engine. In: Proceedings of the 10th International World Wide Web Conference (WWW10), 2001：162~168.

[3]. Deerwester S, Dumai ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of ACM Transactions on Information Systems, 2000, 18(1)：79~112.

[4]. Qiu Y, Frei H. Concept based query expansion. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1993：160~169.

[5]. Buckley C., Singhal A., Mitra M., Saltom G. New retrieval approaches using SMART. In: proceedings of the 4th Text Retrieval Conference (TREC-4), 1995：25-48.

[6]. Xu JX, Croft WB. Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems, 2000, 18(1)：79~112.

[7]. Robertson S EWalker S, Jones G J.F. , Hancock-Beaulieu, GatfordM. Okapi at TREC-3. In: proceeding of the Third Text Retrieval Conference(TREC-3), 1995：109-206.

[8]. 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.

[9]. Xuanhui Wang, Chengxiang Zhai. Learn from Web Search Logs to Organize Search Results. In SIGIR'07, 2007.

[10]. 余慧佳，刘奕群，张敏，茹立云，马少平. 基于大规模日志分析的搜索引擎用户行为分析. 中文信息学报，2007，21(1)：109-114. .

[11]. Andrei Broder. A taxonomy of web search. In SIGIR Forum, 2002, 36(2)：56-62.