基于分类模型的查询扩展方法

李维银 石玉龙 陈 杰 施重阳

(北京理工大学计算机学院 北京 100081)

摘 要 查询扩展作为查询优化的重要组成部分,对改善信息检索系统的性能起到了至关重要的作用。传统的伪相关反馈查询扩展方法虽然在一定程度上提高了检索性能,但选择的扩展词中会包含一部分与原查询不相关的词语,这对检索性能的提升产生了不利影响。提出了一种基于分类模型的查询扩展方法,该算法综合候选扩展词的统计信息和多种特征,采用朴素贝叶斯分类模型对初次得到的候选扩展词进行再次分类选择,进一步去除与查询词相关性小的扩展词。在TREC 2013数据集上的实验结果表明,提出的查询扩展方法能够有效提高用户查询的查准率和查全率。

关键词 查询扩展,分类模型,信息检索,伪相关反馈

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.6.004

Query Expansion Based on Classification Model

LI Wei-yin SHI Yu-long CHEN Jie SHI Chong-yang

(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract As a key component of query optimization, query expansion plays an important role in improving the performance of information retrieval systems. Traditional query expansion methods on pseudo-relevance feedback improve the performance of retrieval to some extent. However, the selected expansion terms will also include some irrelevant ones, which leads to adverse effect. In this paper, a novel query expansion method based on classification model was proposed. Combining with statistical information and various features of the candidate expansion terms, this method employs Naive Bayes classification model to reselect the candidate expansion terms so as to further filter the irrelevant ones. Experimental results on TREC 2013 datasets show that the proposed query expansion method can efficiently improve the precision and recall of user queries.

Keywords Query expansion, Classification model, Information retrieval, Pseudo relevance feedback

1 引言

搜索引擎在一定程度上帮助用户快速、准确地从"信息海洋"中获取到自己想要关注的信息,但信息检索的困难之一就是如何根据用户提交的信息精确地判断用户所要表达的查询意图。传统的信息检索系统是根据用户请求中的关键词进行文档信息的检索,即如果文档中包含用户查询的关键词,信息检索系统就会将该文档作为结果返回给用户。但是在很多情况下,用户使用信息检索系统时并不能明确地表达自己的查询请求,并且进行检索时输入的查询关键词过于短小和简单,检索系统无法正确理解用户真正的查询意图,导致检索系统返回的结果与用户的查询需求之间存在很大的差异。据统计,在英文检索中,用户输入的查询关键词个数平均为2到3个[1,2],而在中文检索中情况更加严重,检索词的平均长度仅为1.8,长度在3个词以下的查询占总查询的93%[3]。除此之外,检索过程中还存在"一义多词"和"一词多义"的难题。因此,学者们提出了查询扩展技术来对查询进行优化。

查询扩展技术是信息检索领域查询优化的一种重要方法,主要用来改善信息检索中词不匹配的问题。传统的基于伪相关反馈的查询扩展方法使用用户提交的初始检索词,利用相关模型(如BM25)对资料库进行检索,从返回的结果文档集中采用合适的相关性计算方法选择扩展词对初始查询进行扩展。但选择的扩展词中通常会包含一部分与原查询不相关的词语,对检索性能的提升产生不利影响。

本文提出的基于分类模型的方法在对扩展词进行初步筛选以后,使用通过自动生成的训练数据集学习的分类模型对扩展词进行进一步筛选,确定最终的查询扩展词,从而提升了选择扩展词的精度。我们采用的分类器是贝叶斯分类模型,将扩展词与查询词的相关性信息以及一些扩展词的统计特征作为分类器的输入特征,从而对扩展词进行分类筛选。实验证明,本文提出的基于分类模型的查询扩展方法能使检索系统的性能获得显著提升。

到稿日期:2014-07-25 返修日期:2014-10-29 本文受中国科学院自动化研究所复杂系统管理与控制国家重点实验室开放课题(99S9021F4D),国家自然科学基金(61472034),教育部新世纪优秀人才支持计划(NCET-13-0041),北京理工大学基础研究基金资助。

李维银(1989-),男,硕士,CCF学生会员,主要研究方向为信息检索,E-mail:liweiyin@bit.edu.cn;**石玉龙**(1988-),男,硕士,主要研究方向为信息检索;**陈 杰**(1987-),男,博士,主要研究方向为信息检索;**施重阳**(1980-),男,博士,讲师,主要研究方向为信息检索,E-mail:cy_shi@bit.edu.cn(通信作者)。

2 相关工作

按照查询扩展词的来源可以将查询扩展分为全局分析方法、局部分析方法和基于用户查询日志的方法。基于伪相关反馈的扩展方法属于局部分析方法，候选扩展词的获取相对容易，而且在提高信息检索系统性能方面相对于其他方法也毫不逊色，因此目前的查询扩展研究大多是基于伪相关反馈的方法。

Imran 和 Sharan[4] 提出一种基于局部分析的查询扩展框架，该方法的实质是一种伪相关反馈的局部分析自动查询扩展方法，把文档及其摘要同时作为扩展词的来源，使用 KLD (Kullback-Liebler Divergence) 模型对扩展词进行排序和选择，而后对原查询进行扩展。Carpineto 和 Mori[5] 等人同样使用 KLD 方法进行了查询扩展的研究，他们将 KLD 方法与其它几种方法进行了比较，充分证明了 KLD 方法在查询扩展中的优势；同时还分析验证了不同的数据集扩展词的数量对查询扩展结果产生的影响。Xu 和 Croft[6] 等人通过使用局部上下文分析 (Local Context Analysis, LCA) 方法来计算文档词和查询词之间的共现程度，得到它们之间的相关性信息，并以此作为扩展词的选取依据。Pal 和 Mitra[7] 通过综合基于分布和基于共现信息的方法对查询扩展进行了研究，指出在大多数情况下，相对于基于共现信息的方法而言，基于分布的方法在发现扩展词方面具有更好的效果。Luo 和 Meng[8] 等人使用 Google 相似距离来计算查询词与抽取出的关键词的相似度，取得了不错的效果。

基于伪相关反馈的查询扩展方法在一定程度上提高了信息检索系统的性能。然而，该方法仍然存在一定的局限性，因为其基于这样一个假设：根据用户初始查询词返回的文档都是与查询相关的文档，因此从这些文档中选出的词与查询词都是相关的。但是情况并非如此，Cao 和 Nie[9] 等人在研究中指出通过相关性计算方法从伪相关反馈文档中选出的扩展词并不总是与查询词相关，如果将这些扩展词加入到原始查询中，会对检索系统性能的提升产生不利影响。

3 基于分类模型的查询扩展

本文在传统的伪相关反馈查询扩展方法的基础上，提出了一种可以提高扩展词选取精度的方法——基于分类模型的查询扩展方法。

3.1 基于分类模型的查询扩展策略及流程

基于分类模型的查询扩展方法是在传统的基于伪相关反馈的查询扩展方法的基础上进行改进的。该方法的提出主要基于以下几点考虑：

(1)基于伪相关反馈的查询扩展方法将伪相关反馈文档作为查询扩展词的来源，但是当这些文档中存在不相关信息时，会把一些与原查询不相关的词语加入到扩展词中，影响了信息检索的效果[10]。

(2)在传统的基于伪相关反馈的研究中，往往只考虑候选扩展词的一种特征，比如扩展词与查询词的共现信息或者扩展词在文档中的分布特征等，而没有将各种特征综合考虑[11]。

(3)传统的基于伪相关反馈的查询扩展中，对扩展词的评价往往依赖于启发式的经验公式，这些公式中的参数需要在大量的实验中进行调整，最终确定合适的值。

本文提出的查询扩展方法在这几个方面进行了改进，利用有监督的学习方法训练一个可以判断扩展词好坏的分类模型对候选扩展词进行进一步的筛选，从中选出与原查询真正相关的词语，同时在模型中综合考虑了扩展词的各种特征，从多方面对扩展词进行评价；而且由于以可靠的训练数据作为基础，因此扩展词的选取不再完全依赖于启发式经验公式。

整个查询扩展过程如图 1 所示。

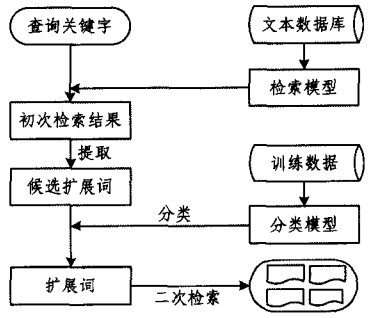

图 1 基于分类模型的查询扩展工作流程

第一步 根据用户提交的初始查询词对语料库进行检索，返回最相关的前 N 篇文档。该过程使用 Okapi 的 BM25 模型对文档进行评分。

第二步 从上一步返回的文档中抽取关键词，统计关键词的各种特征，并使用关键词的概率分布信息对关键词进行初步筛选。

第三步 使用训练好的分类模型对候选扩展词进行进一步筛选，并确定扩展词的权重，将扩展词及其权重信息与初始查询词进行结合，形成新的查询条件。

第四步 使用新生成的查询条件进行第二次检索，返回排序的相关文档。

3.2 候选扩展词的选取

由于直接从前 N 篇文档中抽取出的词语非常多，并且其中会包含大量无关的候选词，因此，为了提高查询扩展的效率，在本文提出的基于分类模型的查询扩展方法中，首先根据候选词在前 N 篇文档中的概率分布情况对候选词进行初步筛选。文献[7]指出使用词的概率分布情况比使用局部上下文信息在提高检索性能方面更有优势。KL 距离 (Kullback-Liebler Distance) 就是用于衡量相同事件空间里的两个概率分布的差异情况，因此，先使用式(1)计算文档词的 KL 距离：

$$KLD(t)=[p_r(t)\ p_c(t)\times\log\frac{p_r(t)}{p_c(t)}] \tag{1}$$

其中，$p_r(t)$ 和 $p_c(t)$ 分别表示词语在相关文档集 R 和整个文档集 C 中的概率分布，$p_r(t)$ 和 $p_c(t)$ 可以由式(2)和式(3)计算得出：

$$p_r(t)=\frac{\sum_{d\in R}tf(t,d)}{\sum_{d\in R}\sum_{t'\in d}tf(t',d)} \tag{2}$$

$$p_c(t)=\frac{\sum_{d\in C}tf(t,d)}{\sum_{d\in C}\sum_{t'\in d}tf(t',d)} \tag{3}$$

从中挑选出 50 个得分最高的作为候选扩展词，等待使用已经训练好的分类模型进行进一步筛选。

3.3 扩展词特征提取

常用的伪相关反馈查询扩展方法有基于词分布的方法和局部上下文分析的方法，但是这两种方法中都仅使用了扩展

词的一个特征。基于词分布的方法使用文本关键词在文档中的概率分布情况计算其与查询词的相关性，而局部上下文分析使用共现的频率作为特征来计算相关性。在本文提出的查询扩展方法中，将两者同时作为关键词分类的依据来筛选扩展词，同时还加入了扩展词其它方面的特征，例如关键词频率特征、关键词的 BM25 值等，以增加分类依据的多样性。

3.3.1 关键词的频率信息

除了停用词外，关键词在文档集中出现的频率可以反映出该关键词在文档中的重要性，从而可以间接地表示该关键词和查询词之间的相关性。因此，本文提出的分类模型使用关键词的频率信息作为候选扩展词的特征之一，在伪相关反馈文档集上的频率特征计算公式如式(4)所示：

$$f_1(t)=\frac{\sum_{d_i\in R} tf(t,d_i)}{\sum_{d_i\in R}|d_i|} \tag{4}$$

其中，R 表示初次检索返回的相关文档集，$|d_i|$ 表示文档 d_i 包含的关键词的总数。

本文还使用在整个文档集合上的频率信息作为关键词的一项特征：

$$f_2(t)=\frac{\sum_{d_i\in C} tf(t,d_i)}{\sum_{d_i\in C}|d_i|} \tag{5}$$

3.3.2 关键词在相关文档集上的权重

关键词在某篇文档中的权重可以反映一个词在这篇文档中的重要程度。权重信息可以用该词的 TF-IDF 信息来计算，而关键词在整个相关文档集上的重要程度可以用该关键词在相关文档集上的权重进行描述。这个特性可以通过将关键词在每篇文档上的权值累加得到：

$$f_3(t)=\sum_{d_i\in R} tf(t,d_i)\times\log\frac{|C|}{df(t)} \tag{6}$$

其中，$df(t)$表示包含关键词 t 的文档个数，而$|C|$表示整个语料库中文档的数量。

3.3.3 关键词与查询词的共现文档频率

通常情况下，当两个词频繁地出现在相同文档中时，可以认为这两个词之间具有较高的相关性。因此，在文档集上，词语之间的共现情况也可以作为判断一个词是否为查询扩展词的标准。在本文中，将关键词与全部查询词的共现文档频率作为分类器的一个特征，用式(7)计算共现文档的频率：

$$f_4(t)=\log(df(Q\cap t)+0.5) \tag{7}$$

3.3.4 关键词在文档集中的概率分布

在基于词分布的查询扩展中，常用的筛选扩展词的方法是根据文档词与查询词之间的 KL 距离计算它们之间的相关性。这种方法的思想是基于关键词在整个文档集和相关文档集中不同的分布情况来判断关键词与查询词之间的重要性，以此来区分扩展词之间的好坏。与查询词相关性较高的关键词在第一次检索返回的前 N 篇文档中出现的概率比在其它文档集中出现的概率大。

由式(1)得到 KL 距离，在选出合适数量的扩展词以后，需要将扩展词和原查询词进行合并，以形成新的查询条件进行第二次检索过程，此时需要对查询条件中的关键词重新计算权值。原始查询词和扩展词的权重分别使用式(8)和式(9)进行正则化：

$$weight_{orig}(t)=\frac{1+\log tf(t,Q)}{1+\max_{t'\in Q} tf(t',Q)} \tag{8}$$

$$weight_{\exp}(t)=\frac{KLD(t)}{\max_{t'\in d,d\in R} KLD(t')} \tag{9}$$

其中，Q 表示原始查询，关键词的最终权重可以由式(8)和式(9)计算出的权重进行相加得到。

使用式(1)计算出的关键词的 KL 距离作为分类模型的一个特征。

3.3.5 扩展词与查询词的统计相关性

在基于伪相关反馈的查询扩展研究中，另一种计算扩展词与查询词之间相关性的方法是通过两者之间的共现频率挖掘它们之间的统计相关性。局部上下文分析(LCA)是该领域比较有代表性的方法，但文献[12]指出，在某些情况下采用传统的方法计算的词与词之间的共现度并不合适。本文使用一种改进的局部上下文分析方法来挖掘候选扩展词和查询词之间的这种相关性。

$$co(t,q_i)=\sum_{d\in R}\left(tf_{t\cap q_i}\times idf\times\frac{sim(d,Q)}{\max_{d'\in R} sim(d',Q)}\right) \tag{10}$$

$$tf_{t\cap q_i}=\min(tf(t,d),tf(q_i,d)) \tag{11}$$

$$idf=\max(idf_{t\cap q_i},0) \tag{12}$$

其中，$tf_{t\cap q_i}$是指在文档 d 中文档词 t 和查询词 q_i 出现频率较小的词的频率值，而 $idf_{t\cap q_i}$是指频率值较小的词的逆文档频率。

最终，文档词 t 与整个查询 Q 的相关性可由式(14)计算得出：

$$codegree(t,q_i)=\frac{\log(co(t,q_i)+1)}{\log(n)} \tag{13}$$

$$S(t)=\sum_{i=1}^{m} idf_{q_i}\times\log(\delta+codegree(t,q_i)) \tag{14}$$

其中，n 为第一次检索返回的相关文档的数量，m 为用户查询中包含的关键词数量，δ 是调节参数，根据文献[13]的建议将其设置为 0.1。

3.3.6 关键词在相关文档集上的 BM25 权重

本文将关键词在伪相关反馈文档集上的 BM25 权重之和作为扩展词筛选的一个特征。BM25 模型加入了文档权值和查询项权值，扩展了二元模型的得分函数，是从将信息检索视为分类问题的模型中演化来的有效排序算法[14]。为了防止某一特征中特征值的差异过大导致结果不精确，在使用各种特征之前，首先对特征使用式(15)进行归一化处理，处理后的特征值都分布在[0,1]范围之内。

$$f_i'(t)=\frac{f_i(t)-\min_i}{\max_i-\min_i} \tag{15}$$

其中，$\min_i$ 和 $\max_i$ 分别表示第 i 个特征上的最小值和最大值。

3.4 分类模型的训练

本文提出的基于分类模型的查询扩展算法与普通的基于伪相关反馈的查询扩展方法的区别是：在得到了查询词的候选扩展词以后，使用一个通过有监督学习方法训练得到的朴素贝叶斯分类模型对扩展词进行进一步评分、筛选，并确定最终使用的查询扩展词。

在对样本进行分类时，最常用的一个规则是把样本划分到后验概率最大的那个类别中，即基于最大后验概率(Maximum a Posteriori，MAP)的朴素贝叶斯分类模型[15]。相应的

分类器的定义如式(16)所示：

$$classify(f_1, f_2, \cdots, f_n) = \arg\max\ p(C=c)\prod_{i=1}^{n} p(F_i = f_i | C=c) \quad (16)$$

其中，$p(F_i|C)$称为类 C 对特征 F_i 的似然函数，它表示类别 C 中的样本取特征 F_i 的概率。

本文中使用的分类模型就是基于最大后验概率的朴素贝叶斯分类模型，分类器使用的特征为 3.3 节中提到的各种文档词的特征值。使用分类模型不仅仅是为了得到候选扩展词的类别，还要从分类模型中得到候选扩展词的评分，按照评分的高低选择合适的查询扩展词。对扩展词进行筛选的贝叶斯分类模型如图 2 所示。

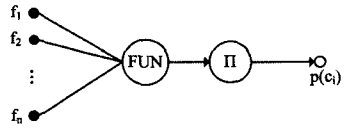

图 2　筛选扩展词的贝叶斯分类模型

用于筛选查询扩展词的朴素贝叶斯算法如表 1 所列。

表 1　筛选查询扩展词的贝叶斯算法描述

LWARN_NAIVE_BAYES_QUERYTERM(Terms,C)

Terms 为一组候选扩展词以及它们的目标类别，C 表示所有类别的集合，该函数可以用来学习概率项 $p(f_k|c_j)$，它描述了从一个类别 c_j 中随机抽取一个扩展词，该扩展词的第 k 个特征为 f_k 的概率。

1. 收集 Terms 中所有的扩展词及其特征
   - Vocabulary ←Terms 中所有的扩展词及其特征项集合
2. 计算所需的概率项 $p(c_j)$和 $p(f_k|c_j)$
   - $Count(c_j)$ ←Terms 中目标值为 c_j 的扩展词数
   - $p(c_j) \leftarrow \frac{Count(c_j)}{|Terms|}$
   - 对于 Examples 中的每个扩展词：
     - n_k←具有特征 f_k 的扩展词出现在类 c_j 中的次数
     - $p(f_k|c_j) \leftarrow \frac{n_k+1}{Count(c_j)}$

CLASSIFY_NAIVE_BAYES_QUERTTERM(Term)

对扩展词 Term 返回其估计的目标值。其中，f_i 表示扩展词 Term 的第 i 个特征。

返回 $p(c|t)$

$$p(c|t) = \arg\max_{c_j \in C} p(c_j) \prod_{f_i \in F} p(f_k|c_j)$$

4　实验及结果分析

4.1　实验数据

本文采用 TREC 2013 任务的子集 Category B 作为实验数据集，将 Adhoc 检索任务[16]的前 40 个话题作为训练数据，后 10 个话题作为测试数据。在训练分类模型时，将前 40 个话题随机平均分成 5 份，轮流将其中 4 份作为训练数据对分类模型进行训练，剩余 1 份为测试数据验证算法的正确率。在进行 5 次交叉验证之后，将得到的正确率取平均值作为训练模型的正确率估计。在实验过程中引入没有进行查询扩展的检索结果作为基准(Baseline)，并采用基于词分布和词共现的查询扩展算法作为对比方法，来验证本文中提出的查询扩展方法对检索性能的提升效果。

4.2　实验结果分析

文献[11]指出，扩展词的个数对于查询扩展的结果会有一定的影响，并且对于不同的数据集，最佳扩展词个数的选取也不同。因此，本文首先通过实验的方法确定在本文中用到的 TREC 数据集上，当扩展词个数选取多少时检索性能能够到达最优。

实验中，首先使用基于 KLD 模型的方法从初检结果的前 30 篇相关文档中抽取扩展词，得到相关性排名靠前的 50 个词作为候选扩展词，然后使用本文中提出的基于分类模型的扩展方法确定最终的扩展词的数量。图 3 描述了扩展词个数不同时平均精度(MAP)的分布情况，其中 NBC 表示基于分类模型的查询扩展方法。

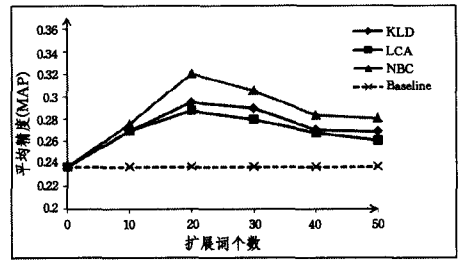

图 3　不同扩展词数目对查询精度的影响

如图 3 所示，当扩展词数目为 20 左右时，检索性能提高最为显著。而且，在扩展词数目一定的情况下，基于分类的查询扩展性能要优于其余两种方法，而基于 KLD 的查询扩展总体性能优于基于局部文档分析的方法，这一结果与其他文献中的研究结果保持了一致。

表 2 给出了在 Category B 数据集上几种查询扩展算法进行检索的结果对比。其中 Baseline 表示未进行任何扩展的情况下的检索结果，KLD 表示使用 KLD 模型进行查询扩展的检索结果，LCA 表示使用局部上下文分析方法的检索结果，NBC 表示使用基于分类模型的查询扩展方法后的检索结果。

表 2　各种查询扩展方法结果比较

扩展方法	P@20	P@30	MAP	Improvement
Baseline	0.470	0.451	0.2371	—
KLD	0.521	0.502	0.2947	24.3%
LCA	0.512	0.491	0.2873	21.2%
NBC	0.556	0.537	0.3201	35.0%

如表 2 所列，与基线方法相比，基于 KLD 和 LCA 的扩展方法在检索性能上都有较大提升，提升程度在 20%以上，而基于 KLD 方法的总体表现优于 LCA 方法。本文提出的基于分类模型的查询扩展方法相对于未扩展的情况有 35%的性能提升，相对于 KLD 方法在平均精度方面也有 8.6%的提升。

在实验中，对每个查询主题的检索结果也进行了统计。通过分析实验结果，可以得到在每个查询主题上不同查询扩展方法对于检索性能的提升情况。每个查询主题的 P@20、P@30 精度统计结果如图 4、图 5 所示。

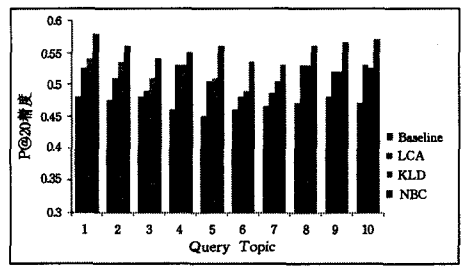

图 4　查询主题 1－10 的 P@20 精度

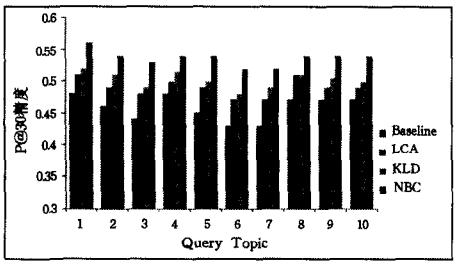

图 5　查询主题 1－10 的 P@30 精度

如图 4、图 5 所示，从每个查询主题的查询结果来看，无论在 P@20 精度还是 P@30 精度方面，基于分类模型的查询扩展方法都要优于其他几种查询扩展方法，表明了基于分类模型的查询扩展方法对检索系统性能的提升显著。

图 6 是几种查询扩展方法的 11 点精度曲线图。

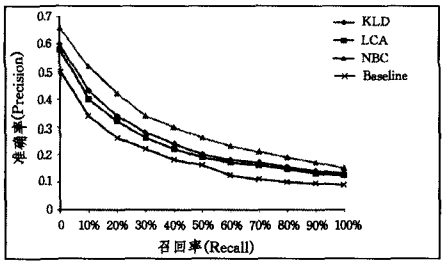

图 6　各种查询扩展方法的 11 点精度曲线

通过分析图 6 的精度曲线可以看出，在每个查全率空间中，基于分类模型的扩展方法的检索精度都高于其余两种方法；而且相对于基于 KLD 模型的扩展方法，本文提出的扩展方法在 11 点召回率上平均精度提高了 21%。

总体来讲，在平均精度(MAP)、前 20 精度(P@20)、前 30 精度(P@30)以及 11 点精度曲线 4 个评价标准中，本文提出的基于分类模型的查询扩展方法相对于传统的伪相关反馈查询扩展方法在检索性能上都有很大提升。这也说明通过分类模型对候选扩展词进一步筛选以后，确实可以过滤掉其中一部分与原查询不相关的扩展词，从而提高检索的精度。相对于未扩展和伪相关反馈的扩展方法，基于分类模型的查询扩展方法能在传统的伪相关反馈扩展方法的基础上进一步提高检索性能。

结束语　针对传统伪相关反馈方法存在的问题，本文提出了一种基于分类模型的查询扩展算法，利用从标准评价结果集中生成的训练数据，通过训练分类模型，对候选扩展词进行进一步筛选，提高了扩展词选择的准确率。在分类筛选扩展词时综合考虑了扩展词的各种统计特征，从多方面对扩展词进行评价。在 TREC 数据集上的实验表明，本文提出的查询扩展方法相对于传统的局部反馈方法，在平均检索精度上提高了 8.6%。在下一步的研究中可以对更多的特征进行分析，更加全面地对查询扩展词进行评价。

参 考 文 献

[1] Jansen B J, Spink A, Saracevic T. Real life, real users, and real needs: a study and analysis of user queries on the web[J]. Information Processing & Management, 2000, 36(2): 207-227

[2] Ogilvie P, Voorhees E, Callan J. On the number of terms used in automatic query expansion[J]. Information Retrieval, 2009, 12(6): 666-679

[3] 余慧佳，刘奕群，张敏，等. 基于大规模日志分析的搜索引擎用户行为分析[J]. 中文信息学报，2007，21(1)：109-114
Yu Jia-hui, Liu Yi-qun, Zhang Min, et al. Research in Search Engine User Behavior Based on Log Analysis[J]. Journal of Information Processing, 2007, 21(1): 109-114

[4] Imran H, Sharan A. A framework for automatic query expansion [M] // Web Information Systems and Mining. Springer Berlin Heidelberg, 2010: 386-393

[5] Carpineto C, De Mori R, Romano G, et al. An information-theoretic approach to automatic query expansion[J]. ACM Transactions on Information Systems (TOIS), 2001, 19(1): 1-27

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

[7] Pal D, Mitra M, Datta K. Query expansion using term distribution and term association[J]. arXiv preprint arXiv: 1303.0667, 2013

[8] Luo J, Meng B, Tu X, et al. Selecting good expansion terms based on Google similarity distance[C] // 2010 2nd International Conference on Future Computer and Communication (ICFCC). IEEE, 2010, 2: V2-710-V2-714

[9] Cao G, Nie J Y, Gao J, et al. Selecting good expansion terms for pseudo-relevance feedback[C] // Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2008: 243-250

[10] Collins-Thompson K. Reducing the risk of query expansion via robust constrained optimization[C] // Proceedings of the 18th ACM Conference on Information and Knowledge Management. ACM, 2009: 837-846

[11] Carpineto C, Romano G. A survey of automatic query expansion in information retrieval[J]. ACM Computing Surveys (CSUR), 2012, 44(1): 1-50

[12] Pal D, Mitra M, Datta K. Query expansion using term distribution and term association[J]. arXiv preprint arXiv: 1303.0667, 2013

[13] Cummins R. A Standard Document Score for Information Retrieval[C] // Proceedings of the 2013 Conference on the Theory of Information Retrieval. ACM, 2013: 24

[14] 范晨熙，黄理灿，李雪利. 基于 Lucene 的 BM25 模型的评分机制的研究[J]. 工业控制计算机，2013，26(3)：78-79
Fan Chen-xi, Huang Li-can, Li Xue-li. Research on Scoring Mechanism of BM25 Model Based on Lucene[J]. Industrial Control Computer, 2013, 26(3): 78-79

[15] Rish I. An empirical study of the naive Bayes classifier[J]. IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001, 3(22): 41-46

[16] Dean-Hall A, Clarke C L A, Kamps J, et al. Overview of the TREC 2012 contextual suggestion track[C] // 21st Text REtrieval Conference. Gaithersburg, Maryland, 2012