基于词汇时间分布的微博查询扩展

韩中元[1),2)] 杨沐昀[1)] 孔蕾蕾[2)] 齐浩亮[2)] 李 生[1)]

[1)](哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

[2)](黑龙江工程学院计算机科学与技术学院 哈尔滨 150050)

摘 要 该文提出了一种面向微博检索的基于词汇时间分布的查询扩展方法.该方法利用扩展词与查询词的时间分布的相似性来度量扩展词与查询词之间的相关度,建立了基于词汇时间分布的查询模型.具体而言,该文在提出词汇时间分布的定义和估计方法的基础上,给出了查询词与扩展词的时间分布相似性的度量,以此作为它们的相关度,完成扩展词的选择和查询模型的重估.该文方法利用时间信息而不是内容来扩展查询,避免了基于内容的查询扩展方法因微博内容短而无法准确估计扩展词的不足.由 TREC 2011 和 TREC 2012 微博检索评测数据上的实验结果表明,基于词汇时间分布的查询扩展模型有效地提高了微博检索的性能,不仅显著优于经典的基于内容的查询扩展模型,而且优于其他利用时间进行查询扩展的方法.

关键词 微博检索;查询扩展;查询模型;词汇时间分布;时间;社交网络;社会媒体

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2016.02031

Query Expansion Based on Term Time Distribution for Microblog Retrieval

HAN Zhong-Yuan[1),2)] YANG Mu-Yun[1)] KONG Lei-Lei[2)] QI Hao-Liang[2)] LI Sheng[1)]

[1)](*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

[2)](*School of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin* 150050)

Abstract In microblog retrieval, content-based query expansion methods are not adequate for expanding queries since the relevant microblog messages are too short to provide reliable term distribution information. Most of the existing time-based query expansion methods exploit time profile to shift the prior probability of relevant microblogs. In essence, these methods still could not avoid the restrictions of short texts since the relevance between expansion terms and query is still based on the content of microblogs. To address the problem, this paper proposes a query expansion method based on the time distribution of terms, in which the relevance between query terms and expansion terms is measured by their time distribution similarity. First, the changes of term frequency in different time segments are analyzed, the term time distribution is defined and the estimation methods are illustrated. Then a similarity estimation approach of term time distribution is presented to estimate the relevance of query terms and expansion terms, so as to decide the expansion terms in the re-estimated query model. Two query expansion strategies are given to estimate the query expansion model according to the relevance of expansion terms and query. Finally, by integrating the query expansion model and original query model, the term time distribution query model is presented. The effort to use only time profile to establish the relevance between query terms and expansion terms avoids the drawbacks of the classical content-based

收稿日期:2015-06-26;在线出版日期:2016-01-27.本课题得到国家自然科学基金(61370170,61402134,61173074)、国家社科基金(14CTQ032)资助.**韩中元**,男,1977 年生,博士研究生,副教授,中国计算机学会(CCF)会员,主要研究方向为信息检索、信息过滤.E-mail:hanzhongyuan@gmail.com.**杨沐昀**(通信作者),男,1971 年生,博士,副教授,主要研究方向为信息检索、机器翻译.E-mail:ymy@mtlab.hit.edu.cn.**孔蕾蕾**,女,1979 年生,博士研究生,讲师,主要研究方向为信息检索、抄袭检测.**齐浩亮**,男,1972 年生,博士,教授,主要研究领域为信息检索、信息过滤.**李 生**,男,1943 年生,教授,博士生导师,主要研究领域为自然语言处理、信息检索.

query expansion approaches due to the length limit in microblog. Experiments were carried on TREC 2011 and TREC 2012 microblog retrieval collection. Several state-of-the-art baselines are chosen for comparing with our method, including the classical language model, the content-based query expansion method and the time-based query expansion method. The experimental results show that the term time distribution query model outperforms the content-based as well as the time-based approaches.

Keywords microblog retrieval; query expansion; query model; term time distribution; time; social networking; social media

1 引 言

查询相关文档时往往使用不同的词语来描述相同的事物,而这些相关文档是无法通过与查询词的匹配被检索到的,这就是长期影响信息检索性能的词不匹配问题(Vocabulary Mismatch Problem)[1].在微博检索中,词不匹配问题更为严重,原因主要来源于两方面.一方面,用户的查询非常短,例如Twitter查询的平均长度为1.64个词,只有Web查询的50%[2],不足以准确的描述用户的信息需求;另一方面,由于字数限制,微博也非常短,例如,Twitter的长度被限制在140个字符内,新浪微博限制字数为140个汉字,这加剧了词不匹配问题.

查询扩展是解决词不匹配问题的主要方法之一.查询扩展将与查询相关的词加入到用户的原始查询中,丰富查询模型,缓解词不匹配的问题,从而提升检索性能[3-5].

在各种查询扩展方法中,基于内容的查询扩展是最主要的解决策略[5].其基本思路是采用相关反馈方式,首先将查询的初始检索结果展现给用户,然后通过与用户交互反馈获得与查询相关的文档(被称为反馈文档),从反馈文档中提取查询扩展词形成新的、更有效的查询[4],最后使用新查询再次进行检索,以期获得更好的检索效果[5].然而,在微博检索中,基于内容的查询扩展受到了一定的影响:作为反馈文档的微博过短,并且大多数词只出现一次,严重的数据稀疏无法提供可靠的信息来建立查询和扩展词之间的关系[6].

近期研究发现,相关微博在时间上具有新近性[7-9]和爆发性[10-16]的特点,这些特点有助于改善微博检索的性能,利用时间的查询扩展受到越来越多研究者的关注[7,10-12,17].目前,微博检索中利用时间的查询扩展主要包括两类方法.第一类方法是基于新近性的查询扩展,这类方法认为越新发布的文档越符合用户的查询需求,在作为反馈文档时应该给予更高的权重.例如,Li和Croft[17]的研究中给予距离查询时间更近的文档以更高的先验概率,使来自这些文档的词有更高的概率成为扩展词.第二类方法可称为基于爆发性的查询扩展,这类方法利用相关微博的发布在时间上具有爆发性(聚集性)的特点来调整反馈文档的权重,给予在微博爆发期内的文档以更高的先验概率.例如,Keikha等人[10]根据反馈文档在时间上的聚集来估计微博的爆发期,给予爆发期内的文档以更高的先验概率,使扩展词有更高的概率来自于爆发期内发布的文档.

这些基于时间的查询扩展方法的共同特点是利用时间信息来提高相关微博的先验概率,进而提高来自这些微博的词作为扩展词的概率.然而,在这些方法中,扩展词与查询的相关性依然依赖两者在相关微博内容上的分布信息建立,仍未摆脱微博短的限制.

为了进一步解决这一问题,本文尝试从微博内容之外,直接利用时间信息来建立扩展词与查询词的关系.本文的出发点是观察到扩展词和查询词在不同时间段内被使用的频率具有同增同减的现象.这是由于在微博环境下,一个话题会在较短的时间内被大量转发和评论,与该话题的查询相关的词(简称相关词①)在这些爆发期内被大量的使用,当话题的热度消散以后,这些相关词的使用频率也随之下降,这意味着相关词的使用频率在时间上具有同增同减的趋势(详见3.1节).本文利用词的时间分布的相似性来刻画相关词在时间上的这种同增同减的现象,与查询词的时间分布越相似的词是查询的相关词的概率越大.据此,本文提出了基于词汇时间分布的查询扩展.

具体而言,本文定义了词的时间分布及其估计方法,给出了度量查询词与扩展词的时间分布相似

① 本文中特指在相关微博中出现的与查询相关的词,不包括查询词本身.

性的方法，利用这种相似性估计查询词与扩展词的相关性，据此选择扩展词，建立基于词汇时间分布的查询模型. 本文提出的基于词汇时间分布的查询扩展方法在 TREC 2011 和 TREC 2012 微博检索评测数据上进行了实验. 结果表明，本文提出的方法优于基于内容的查询扩展方法和多个基于时间的查询扩展方法.

本文第 2 节介绍相关工作；第 3 节定义词的时间分布，提出用时间分段语言模型估计词的时间分布的方法，给出了词的时间分布相似性的度量；第 4 节提出基于词的时间分布相似性重估查询模型的方法；第 5、6 节给出实验设计、实验结果及分析；第 7 节是对本文工作的总结和展望.

2　相关研究

在微博检索中，时间信息逐渐受到重视，越来越多的研究利用时间信息来扩展查询[7,10-12,17]. 这些基于时间的查询扩展模型大多建立在语言模型的框架下，围绕相关反馈开展. 本节首先介绍语言模型检索框架，然后介绍了语言模型框架下的基于相关反馈的查询扩展方法，最后介绍了以相关反馈为基础的利用时间信息进行查询扩展的相关研究.

2.1　语言模型检索框架

在信息检索领域，文档语言模型(Document Language Model)简称为语言模型(Language Model)，该模型以文档语言模型 θ_D 生成查询(query)的概率作为文档和查询相关度的依据[18]. Lafferty 和 Zhai[19] 在风险最小化模型下发展了语言模型，为查询和文档集合中的每一篇文档都建立一个统计语言模型，并通过一个损失函数建模用户的信息需求，将信息检索问题转化为一个风险最小化问题. 在确定相关文档时，查询语言模型 θ_Q 和文档语言模型 θ_D 的相关度应用 KL(Kullback-Leibler)距离度量[19]：

$$KL(\theta_Q|\theta_D)=\sum_{w\subset V}P(w|\theta_Q)\log\frac{P(w|\theta_Q)}{P(w|\theta_D)} \quad (1)$$

其中 V 是整个词汇表；w 是 V 中的词项；θ_Q 与 θ_D 分别是 w 在查询和文档的语言模型；$P(w|\theta_Q)$ 和 $P(w|\theta_D)$ 分别是在查询语言模型 θ_Q 和文档语言模型 θ_D 中的概率.

由于一篇文档包含词的数量有限，大部分都不在词汇表 V 中，存在着严重的数据稀疏问题，因此在对文档 D 建立语言模型时，通常使用平滑方法调整词在文档上的概率分布[20]. 本文在实验中采用 JM 平滑，文档语言模型可以估计为[21]

$$P(w|\theta_D)=(1-\lambda)P_{ml}(w|D)+\lambda P_{ml}(w|C) \quad (2)$$

其中 D 表示待估计的一篇文档；C 表示整个文档集合；$P_{ml}(w|D)$ 和 $P_{ml}(w|C)$ 分别表示应用最大似然估计，w 在 D 和 C 中的概率；λ 用来控制平滑项的影响. 对于本文的实验，D 是一篇单独的微博，C 是全部微博的集合，即 Tweets2011 数据集全部微博.

在查询扩展中，$P(w|\theta_Q)$ 的估计通常采用原始查询的最大似然估计与查询扩展模型结合[5]，形式为

$$P(w|\theta_Q)=(1-\lambda)P_{ml}(w|Q)+\lambda P(w|\theta_{QE}) \quad (3)$$

其中 $P_{ml}(w|Q)$ 为原始查询 Q 的最大似然估计；θ_{QE} 表示查询扩展模型. 不同的查询扩展方法演化出众多的查询扩展模型. 其中，基于时间的查询扩展的研究大多在基于相关反馈的查询扩展方法下展开. 基于风险最小化的语言模型使得检索系统的优化不仅可以从文档建模的角度进行，也可以通过对查询建立语言模型来改进检索性能. 本文的工作围绕查询建模开展，下面综述这方面的工作.

2.2　基于相关反馈的查询扩展

基于(伪)相关反馈的查询扩展是查询扩展的主流方法之一. 在相关反馈中，每个查询的相关文档需要人工反馈才能获得，代价过大，因此通常将初始检索结果的前 n 篇文档视为相关文档，这种方法称为伪相关反馈，检索结果的前 n 篇文档通常被称为伪相关反馈文档. 在不引起歧义的情况下，本文随后的相关文档均指通过伪相关反馈确定的反馈文档.

基于伪相关反馈的查询扩展的基本步骤可以描述如下：

(1) 用初始查询进行检索，得到检索结果列表；

(2) 假定列表中前 n 篇文档是相关文档；

(3) 将相关文档集中的词视为候选扩展词，根据某种策略对候选扩展词排序，选出排序最高的 k 个词作为扩展词，建立查询扩展模型；

(4) 根据原始查询模型和扩展查询模型重新估计一个查询模型，执行检索，得到最终的检索结果.

在语言模型框架下，基于相关反馈的查询扩展方法中，Lavrenko 和 Croft[22] 提出的相关模型(Relevance Model)是使用最广泛的查询扩展模型之一. 相关模型根据查询 Q 与备选扩展词 w 在与查询相关的文档集合 R 上共现的概率来估计查询模型，扩展后的查询模型为

$$P(w|\theta_{QE}) \propto \sum_{\theta_d \subset R} P(w,Q|\theta_d)P(d) \quad (4)$$

其中：w 是候选扩展词；Q 是查询；R 是相关文档的集合，d 是 R 中的一篇文档，θ_d 是文档 d 的语言模型；$P(w,Q|\theta_d)$ 是 w 和 Q 在 θ_d 中出现的联合概率；$P(d)$ 是 d 的先验概率.

2.3 基于时间的查询扩展

在相关反馈中，由于缺少估计相关文档的先验概率 $P(d)$ 的信息，$P(d)$ 常被赋予相同的值. 在微博环境下，相关微博在时间上的分布并不均匀[14]，因此，研究者利用这一特点来估计 $P(d)$，从而获得更好的查询扩展词.

基于时间的查询扩展主要分为两类：一是假设“时间越新文档越重要”，即利用新近性估计 $P(d)$，本文称其为基于新近性的查询扩展；二是假设“爆发期内的文档更重要”，即利用时间的爆发性调整相关文档的先验概率，本文将这类方法称为基于爆发性的查询扩展.

2.3.1 基于新近性的查询扩展

在基于新近性的查询扩展中，代表性的工作是 Li 和 Croft[17] 提出的基于新近性的相关模型. 该方法认为，越新的文档越可能是相关文档，他们利用文档发布的时间信息，以指数分布估计文档的先验概率，给予发布时间离查询时间越近的文档以越大的先验概率：

$$P(d) = P(d|t_d) = \lambda e^{-\lambda|t_Q - t_d|} \quad (5)$$

其中 t_Q 是用户提交查询的时刻；t_d 是文档发布的时刻；λ 是指数分布的参数.

通过对相关模型的文档先验概率的重新估计，该方法定义查询扩展模型为

$$P(w|\theta_{QE}) \propto \sum_{\theta_d \subset R} P(w,Q|\theta_d)\lambda e^{-\lambda|t_Q - t_d|} \quad (6)$$

其中 θ_d 是文档 d 的语言模型. 在基于新近性的查询扩展中，那些在“新”文档中的词将有更高的概率成为扩展词.

2.3.2 基于爆发性的查询扩展

基于新近性的查询扩展方法不能很好的处理那些爆发期远离查询时间或者具有多个爆发期的查询[16,23]. 因此，研究者提出了利用微博发布时间的爆发性来进行查询扩展. Keikha 等人[10] 将时间因素引入到相关模型的框架中，提出了基于时间的相关模型. 该方法将查询的生成过程看成是一个抽样的过程，首先以概率 $P(t|Q)$ 选择一个时间，然后以概率 $P(w|t,Q)$ 从这个时间的相关文档中抽样选择词 w，$P(w|t,Q)$ 经过一些假设变换后等价于词在文档上的分布 $P(w|d)$ 之和，最终的模型如式(7)所示.

$$P(w|\theta_{QE}) \propto \sum_{t} P(t|Q) \sum_{d \subset R_t} P(w|d) \quad (7)$$

其中 R_t 表示的是在时间段 t 内的相关微博集合；d 表示相关微博；$P(t|Q)$ 表示每个时间段对于一个给定查询的重要程度，应用给定查询的检索结果的得分来估计.

$P(t|Q)$ 本质上是对相关文档的权重进行了调整. 类似的工作包括：Choi 和 Croft[11] 利用时间段内相关文档的转发数来确定重要的时间段，提出了一种利用用户的行为信息估计 $P(t|Q)$ 的查询扩展方法，Peetz 等人[12] 通过不同时间段内的反馈文档的数量来确定爆发期，距离爆发期越近的相关文档越可靠，利用这些相关文档来选择查询扩展词.

无论是利用新近性的查询扩展还是利用爆发性的查询扩展，时间特性仅被用来调整反馈文档的先验概率 $P(d)$，本质上查询扩展词和查询的关系还是在利用词在反馈文档内容上的分布 $P(w|\theta_D)$ 建立. 然而，在微博检索环境下，作为反馈文档的微博过短，并且大多数词只出现一次，使得文档模型 $P(w|\theta_D)$ 估计并不充分，无法为查询扩展提供可靠的分布信息[6]. 仅依赖于时间对 $P(d)$ 调整，不能从根本上解决反馈文档模型 $P(w|\theta_D)$ 自身估计的不足. 这意味着时间信息的作用受到了限制，没有充分发挥作用. 为此，本文尝试不依赖于反馈文档的模型 $P(w|\theta_D)$，只通过词的时间特性建立查询词与扩展词之间的关系，实现查询扩展.

3 词汇时间分布

微博作为人们交流的载体，其快速转发的特点使得受到关注的事件会在短时间之内引起人们大量的转发和讨论，与其相关的微博往往集中在一个或多个时间段内集中发布[13-16]. 这意味着，在一个话题的爆发期内，人们在讨论这个话题时会频繁地使用与这个话题相关的词. 随着时间的流逝，这个话题逐渐淡出人们的视线，与这个话题相关的词的使用频率也随之下降. 因此，词在不同时间段被使用的频率随着话题热门程度的改变，在时间上呈现出一定的波动性，且相关词在不同时间段上的使用频率存在同增同减的一致性趋势. 本节将对这种现象进行分析，给出词汇时间分布的定义以及词汇时间分布相似性的度量方法，进而利用词的时间分布相似性来度量扩展词与查询词的相关程度.

3.1　词汇时间分布相似性

在现有公开数据中，TREC 微博评测提供了规模最大的微博标注数据. 其组织者从 Twitter 网站抽样了 2011 年 1 月 23 日至 2011 年 2 月 7 日之间的 1000 多万条微博，给出了 110 个查询(topic 1～110)，并标注了这些查询的相关微博[24]. 本节以 TREC 微博评测的数据为例，对词的时间分布的相似性进行分析.

首先，我们以“天”为单位，对不同时间段内的相关微博数量进行分析. 图 1 展示了与查询(topic 1、4、14、20)相关的微博的分布情况.

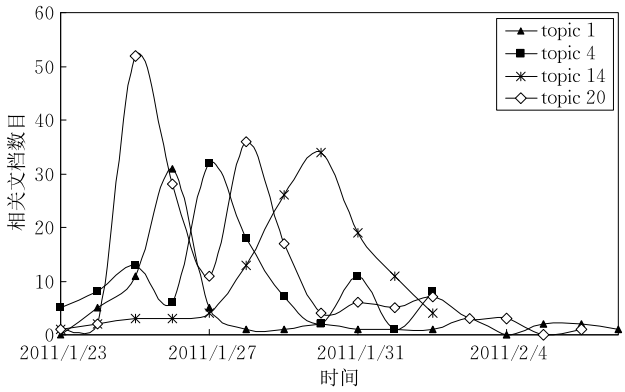


图 1　不同时间段内相关微博的数量

从图 1 中可以看出，不同查询的相关微博数量在时间上的分布是不均匀的. 但相同的是大多数的相关微博都集中在一个或几个“波峰”时间内. 这意味着在不同的时间段内，相关词的使用频率也是不均匀的.

为此，我们进一步以词为统计粒度进行了分析. 结果表明，词汇在不同时间段上的使用频率同样存在着差异，而其中某个查询的相关词在不同时间段上的使用频率又存在一定的相似性.

图 2 展示了查询 1 中的词、与查询相关的词及与查询无关的词在不同时间上使用频率.

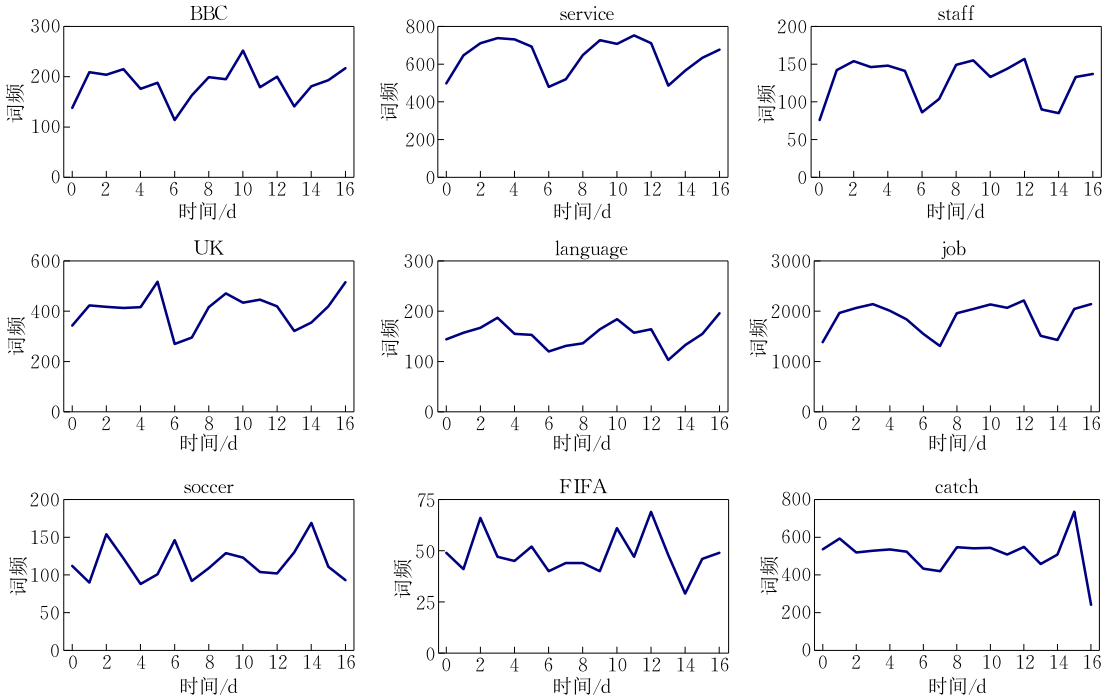


图 2　不同时间段内查询词、与查询相关的词及与查询无关的词的数量

查询 1 关注的是 2011 年 1 月**英国**广播公司(BBC)新闻部门裁员这一新闻相关的微博，主要内容是“**BBC** close five of its 32 World Service **language services** and **staff** have been informed that up to 650 **jobs** will be lost from a workforce of 2400 over the next three years.”图 2 给出了 3 个查询词“BBC”、“service”和“staff”，3 个与查询相关的词“UK”、“language”和“job”，3 个与查询无关的词“soccer”、“FIFA”和“catch”(来自查询 2，关注的是 2022 年世界杯)在以“天”为单位的时间段上的使用频率. 其中，y 轴是词频，x 轴是时间(以“天”为单位)，0 对应的是 2011 年 1 月 23 日.

观察图 2 可以发现，尽管词频的数值不同，但相关词的词频变化趋势是相似的. 从图形上看，词“BBC”和“UK”、“service”和“language”、“staff”和“job”的图形具有一定的相似性，当某个时期内查询词的使用频率增加，与之相关的词的使用频率往往也跟着增加，反之，当某个时期查询词的使用频率减

少,与之相关的词的使用频率也相应的减少,这意味着相关词在不同时间段内的使用频率在一定程度上具有同增同减的一致性趋势.作为对比,3 个与查询无关的词“soccer”、“FIFA”和“catch”与查询词的图形差异度则相对较大.

3.2 词和查询的时间分布

从上面的分析可以看出词在不同时间段内的使用频率是变化的.从离散型概率分布的角度,这种变化可以使用词的时间分布来刻画,为描述方便使用符号 $P(T|w)$来表示,T 是表示时间段的随机变量.对于某个具体的时间段 t_i,使用 $P(t_i|w)$表示 $P(T=t_i|w)$. $P(t_i|w)$描述了人们在时间段 t_i使用该词的频繁程度,反映了词在时间段 t_i的热门程度.其中 t_i表示的是按照某固定时间间隔划分的时间片段,即微博集合的时间可以离散的表示为$\{t_1,t_2,\cdots,t_n\}$,t_i代表一个具体的时间片段(本文中以自然天(24 h)为单位).在本文随后的内容中,无特殊说明的情况下,所提时间均在以一天为单位划分的时间段上加以分析和讨论.

直接估计词的时间分布 $P(t_i|w)$比较困难,根据贝叶斯定理:

$$P(t_i|w)=\frac{P(w|t_i)P(t_i)}{P(w)} \tag{8}$$

假定时间段相互独立且每个时间段的先验概率 $P(t)$相同,由全概率公式:

$$P(w)=\sum_{t_i\in T}P(w|t_i)P(t_i) \tag{9}$$

则显然

$$P(t_i|w)=\frac{P(w|t_i)}{\sum_{t_j\in T}P(w|t_j)} \tag{10}$$

通过式(10),估计 $P(t_i|w)$被转化为估计 $P(w|t_i)$. $P(w|t_i)$表示从时间段 t_i内发布的微博所含的全部词的集合中抽样出词 w 的概率. $P(w|t_i)$越大,表示在时间段 t_i上词 w 被使用的越频繁.换言之,$P(w|t_i)$度量了词 w 在某一时间段 t_i内的热门程度,既表示在时间段 t_i上词 w 被使用的概率,可以应用这段时间内发布的全体微博的语言模型来度量.本文将每个时间段内全体微博建立的语言模型称之为时间分段语言模型,用 θ_i表示,采用最大似然估计的一元语言模型估计.则 $P(w|t_i)$可以估计为

$$P(w|t_i)=P(w|\theta_{t_i})=\frac{c(w,t_i)}{\sum_{w'\subset V}c(w',t_i)} \tag{11}$$

其中 V 为整个词表;w 为 V 中的词,$c(w,t_i)$为 w 在时间段 t_i上发布的全部微博中出现的总次数.

根据式(10)和(11),则词的时间分布可以用式(12)估计为

$$P(t_i|w)=\frac{P(w|\theta_{t_i})}{\sum_{t_j\in T}P(w|\theta_{t_j})} \tag{12}$$

词的时间分布反映词在不同时间段的热门程度,类似的,我们用查询的时间分布来反映这组查询词在不同时间段的热门程度,用 $P(t_i|Q)$表示.与 $P(t_i|w)$的推导同理,其估计方法如式(13)所示:

$$\begin{aligned}P(t_i|Q)&=\frac{P(Q|t_i)P(t_i)}{P(Q)}=\frac{P(Q|\theta_{t_i})}{\sum_{t_j\in T}P(Q|\theta_{t_j})}\\&=\frac{\prod_k P(q_k|\theta_{t_i})}{\sum_{t_j\in T}\prod_k P(q_k|\theta_{t_j})}\end{aligned} \tag{13}$$

3.3 时间分布相似性的度量

从 3.1 节观察到的现象出发,假定相关词的使用频率在时间上具有同增同减的一致性趋势,本文将这种一致性称为词的时间分布相似性.本文认为,这种相似性从一定程度上反映了词的相关程度,可以利用这种时间分布上的相似性来度量词的相关度.考虑到度量两个分布之间距离的常用 KL 距离具有不对称性,本文应用常见闵可夫斯基距离[25]($p=1$)来进行度量.则词 w 和 q_i的时间分布 $P(T|w)$和 $P(T|q_i)$的相似性可计算为

$$\begin{aligned}&S(P(T|w),P(T|q_i))=\\&\quad\sum_{t_i\in T}|P(t_i|w)-P(t_i|q_i)|\end{aligned} \tag{14}$$

在单位时间上,式(14)等价于

$$\begin{aligned}&S(P(T|w),P(T|q_i))=\\&\quad\sum_{t_i}|P(t_i|w)-P(t_i|q_i)|\Delta t\end{aligned} \tag{15}$$

其中 t_i 代表一个时间段,Δt 是单位时间(在微博环境下,通常以自然天(24 h)作为单位[10-12]).

式(15)的物理意义是扩展词和查询词的时间分布所围成的面积. $S(P(T|w),P(T|q_i))$的值越小则两者围成的面积越小,意味着 w 和 q_i越相似.

以 3.1 节描述的查询 1 为例,图 3 分别展示了查询词“job”的时间分布与相关的词“staff”和不相关的词“soccer”的时间分布所围成的面积.

从图 3 中可以看出,两个相关词所围成的图形的面积明显小于两个不相关词所围成的图形的面

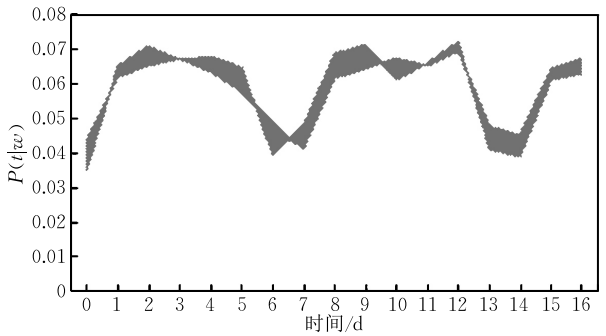
积.式(16)给出了扩展词 w 和查询词 q_i 的基于词汇时间分布相似性的相关度的估计：

$$rel(P(T|w),P(T|q_i))=N(S(P(T|w),P(T|q_i))) \quad (16)$$

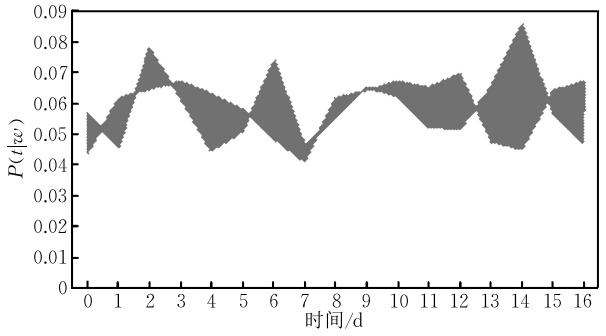
其中 q_i 是查询 Q 中的任意一个词；w 是反馈文档中的词；$N(x)$ 是归一化函数，且 $N(x)=(max-x)/(max-min)$，max 和 min 分别为样本数据的最大值和最小值.由于两个分布围成的面积小于等于 2 大于等于 0，因此本文将 max 设置为 2，min 设置为 0. $N(x)$ 对 x 进行了线性变换，将 x 转换为[0,1]之间的值，并使两个词的时间分布所围成的面积与两个词的相关度成反比关系.即 2 个词的时间分布围成的面积 S 越小，两者的相关度越大.

类似的，定义扩展词 w 和查询 Q 的相关度为

$$rel(P(T|w),P(T|Q))=N(S(P(T|w),P(T|Q))) \quad (17)$$



(a) 查询词“job”和相关词“staff”围成的图形



(b) 查询词“job”和无关词“soccer”围成的图形

图 3　查询词和相关词的时间分布围成图形与查询词和不相关词的时间分布围成图形的差异

4　基于词汇时间分布的查询扩展

根据上述分析，相关词和查询词在不同时间段内的使用频率具有同增同减的一致性趋势.利用时间分布相似度可以度量扩展词和查询(或查询词)的相关程度.本节给出基于词汇时间分布的查询扩展，简称为 TTDM(Term Time Distribution Model).

与其他查询扩展方法类似，TTDM 将查询扩展模型与原始查询结合得到最终查询模型，如式(18)所示：

$$P(w|\theta_Q)=(1-\lambda)P_{ml}(w|Q)+\lambda P(w|\theta_T) \quad (18)$$

其中 θ_T 是基于词汇时间分布的查询扩展模型；w 是原始查询词以及应用词的时间分布相似性获得的查询扩展词.与其他查询扩展方法类似，w 是从反馈文档中选出的相关程度最高的 n 个词.

θ_T 估计如下：

$$P(w|\theta_T)=\frac{score(w,Q)}{\sum_{w'\subset V} score(w',Q)} \quad (19)$$

其中 $score(w,Q)$ 是扩展词与查询之间的相关度得分；w 是扩展词；V 是查询扩展词的集合.

根据扩展词与查询之间的相关度进行查询扩展主要包括两种策略：一对一的策略和一对多的策略[5].所谓一对一的策略指查询扩展词的生成和排序依赖于备选扩展词和查询中的某个词的关联关系，而一对多的查询扩展方法中，查询扩展词的生成和排序则依赖于备选扩展词与查询整体的关系.根据这两种策略，本文定义了两种计算扩展词 w 和查询 Q 的相关度得分方法，记为 $score(w,Q)$，利用该得分选择扩展词和估计查询扩展模型.

第 1 种方法利用一对一的策略，估计扩展词与单个查询词的相关度.该方法倾向于将与查询中某个查询词最相关的词扩展到查询中，即 w 和 Q 的相关程度取决于 w 与 Q 中相关度最高的词：

$$score(w,Q)=\max_{q_i\in Q}(rel(P(T|w),P(T|q_i))) \quad (20)$$

其中 q_i 是查询 Q 中的词；$rel(P(T|w),P(T|q_i))$ 用式(16)估计.

第 2 种方法利用一对多策略，计算扩展词和查询整体的时间相关度.应用式(17)，定义 $score(w,Q)$ 为

$$score(w,Q)=rel(P(T|w),P(T|Q)) \quad (21)$$

其中 $P(T|w)$ 是词 w 的时间分布，由式(12)估计；$P(T|Q)$ 是查询整体的时间分布，由式(13)估计.

对于 TTDM，采用一对多策略会存在一些问题：一是单个查询词的时间分布与查询整体的时间分布存在较大差异.仍然以 TREC 微博评测的查询 1 为例，对比图 1 和图 2 可以看出查询 1 的时间分布和查询词的时间分布存在较大差异.因此，通过查询整体的时间分布来选择扩展词存在一定的风险；二是与查询相关的词在时间上可能与一个查询词相关的，不一定与整个查询相关.例如：BBC 和 UK 存在着相

关性，BBC 的域名是 http://www.bbc.co.uk/，从图 2 中也可以看出 BBC 与 UK 的时间分布是相似的，这种相似是客观的，并不依赖于“BBC cut”这一查询建立起来的. 使用查询整体的时间分布会弱化每个查询词的时间分布，进而影响检索性能. 因此，在 TTDM 方法上，利用扩展词与单个查询词的时间分布相似性进行查询扩展将优于基于扩展词与查询整体的时间分布相似性的查询扩展，第 6 节中的实验结果也验证了这一判断.

5 实验设置

5.1 实验数据

实验使用的是 TREC 微博评测的数据. TREC 评测是信息检索领域影响力最大的评测，其数据集受到研究界的普遍认可. 本文使用 TREC 2011 和 TREC 2012 微博数据集[24]进行实验. 该数据集①来自于 2011 年 1 月 23 日至 2011 年 2 月 7 日 Tweets 发布的微博 1%的抽样. 使用 TREC 微博检索组织方提供的爬虫②，下载了 10 397 336 条微博. 按照 TREC 的评测要求，对数据集做了如下预处理[26]：(1)移除了空的微博；(2)对于含有“RT”标志的转发微博，如果“RT”前面没有任何消息，则删除该微博，如果“RT”标志前面含有信息，将此微博中这部分信息保留，认为是用户在转发此微博时添加的有用信息，删除 RT 后面的内容；(3)使用 Nutch③过滤了非英语微博；(4)使用 Indri toolkit④，应用 Porter 算法进行了词干提取，去除了停用词. 预处理后，剩余 5 058 404 条微博. 实验使用了 TREC 2011 微博检索的 49 个查询(查询 1～49，查询 50 没有相关微博，所以将其剔除)和 TREC 2012 微博检索的 60 个查询(查询 51～110). 每个查询都包含了一个表示该查询被发布时间的标签. TREC 微博检索总共标注了 114 K 个微博与查询的相关性，给出了 4 类标签：垃圾微博、不相关微博、相关微博和高度相关微博. 对于检索任务，将相关微博和高度相关微博均视为相关微博，垃圾微博和不相关微博均视为不相关微博.

根据 TREC 微博检索任务的要求，某个查询提出时间之后的数据，即未来数据，对于该查询的检索是不可见的. 因此，实验中我们为每个查询单独建立了索引，每个索引只含有查询时间之前的微博以满足 TREC 评测的要求.

5.2 基线方法和模型训练

本文选择了 4 个基线方法作为对比. 它们分别如下：

(1)语言模型(Language Model，LM)[18]. 使用原始查询的语言模型，文档模型和查询模型均使用了一元语言模型，文档模型的平滑方法使用了 JM 平滑(如式(2)所示)，设置参数 $\lambda=0.5$(Indri 默认值). LM 不使用查询扩展，本文提出的方法及其他基线方法在不引入查询扩展时均为 LM.

(2)相关反馈模型(Relevance Model，RM)[22]. 相关模型是典型的基于内容的查询扩展模型. 它通过计算扩展词和查询在相关文档中出现的联合概率来估计一个相关模型，如式(4)所示. 实验中，反馈文档数目设置为 50，扩展词数目设置为 20(根据文献[14]在 TREC 微博评测数据集上的参数设置).

(3)基于新近性的相关模型(Recency-Based Relevance Model，RBRM)[17]. Li 和 Croft 将文档的新近性引入到 RM 中，根据反馈文档发布时间距离查询时间的远近给予反馈文档不同的先验概率，增大“新”反馈文档对查询扩展模型的影响(如式(5)所示，根据文献[16]在 TREC 微博评测数据集上的参数设置，λ 设为 0.3)，扩展词数目和反馈文档数取训练数据的最优值.

(4)基于爆发期的相关模型(Burst-based Relevance Model，BBRM)[10]. Keikha 等人根据不同时间段内的反馈文档数量的不同，给予反馈文档不同的权重，处于爆发期内的反馈文档将对查询扩展模型产生较大的影响(如式(7)所示)，扩展词数目和反馈文档数取训练数据的最优值.

上述各个基线方法和本文所提方法的检索模型均是在语言模型框架下实现，查询和文档的相似度使用 2.1 节描述的 KL 距离度量(见式(1))，详情可以参考文献[19]. 实验中，以“天”作为时间单位(参照文献[10-12]在 TERC 微博数据集上的设置).

根据 TREC 评测的数据集设置，分别使用 TREC 2011 和 TREC 2012 两组数据，采用交叉检验的方式来训练参数. 实验中首先使用 TREC 2011 微博检索的 49 个查询作为训练数据训练参数，应用 TREC 2012 微博检索的 60 个查询作为测试数据，实验结果用 TREC 2012 标注. 然后使用 TREC

① http://trec.nist.gov/data/tweets
② https://github.com/lintool/twitter-tools
③ http://nutch.apache.org
④ http://www.lemurproject.org/indri/

2012 的 60 个查询作为训练数据训练参数，应用 TREC 2011 的 49 个查询测试，实验结果标注为 TREC 2011. 测试结果的反馈文档数、扩展词数以及与原始查询结合的插值系数均在训练数据上获得. 参数训练中，扩展词数的取值设定为 10、15、20、30、40、50、75 及 100，反馈文档数的取值设定为 10、20、30、40、50、100 及 150，插值系数取值范围为 [0,1]，步长为 0.1.

5.3　评价指标

实验中，本文使用 TREC 微博检索任务[24]的主要评价指标 $P@30$ 来度量性能，并提供 $P@20$ 的评价结果供参考. $P@30$ 反映了前 30 个检索结果中相关微博的数目. 由于微博很短，检索结果容易阅读，因此，微博检索并不像传统检索强调相关微博在检索结果前几位的重要性，而只是强调可以快速浏览的前 30 条微博中相关微博的数量.

评价指标 $P@k$ 定义为

$$P@k=\frac{1}{k}\sum_{j=1}^{k}r_j \tag{22}$$

其中 k 是前 k 个检索结果，如果检索结果的第 j 个位置是相关的，则 $r_j=1$，否则，$r_j=0$.

6　结果及分析

6.1　实验结果

表 1 展示了我们的方法与 5.2 节描述的基线方法在 Tweets2011 数据集上的结果，将我们的方法中，采用与单个查询词相关的查询扩展策略标记为 TTDM-q，采用与查询整体相关的查询扩展策略标记为 TTDM-Q. 表 1 列出了我们的方法与基线方法在评价指标 $P@30$ 上的值以及各种方法相对于语言模型的提升幅度. 显著性检验采用 Wilcoxon 校验 ($p<0.05$). "&"表示本文的查询模型显著优于 LM，"*"表示显著优于 RM，"+"表示显著优于 RBRM，"#"表示显著优于 BBRM.

从实验结果可以看出，全部查询扩展方法比不使用查询扩展的语言模型 LM 都有提升，基于时间的查询扩展优于仅使用内容相关反馈模型的查询扩展 RM，本文所提出的基于词汇时间分布的查询扩展（TTDM-Q、TTDM-q）优于其他基于时间的查询扩展方法（RBRM 和 BBRM）. 这一实验结果从一定程度上验证了本文提出的相关词在不同时间段内的使用频率在一定程度上具有同增同减的一致性趋势假设的合理性.

表 1　模型检索性能对比

	TREC 2011		TREC 2012	
	$P@30$	$P@20$	$P@30$	$P@20$
LM	0.4136 (−)	0.4490 (−)	0.3350 (−)	0.3610 (−)
RM	0.4503 (+8.9%)	0.4939 (+10%)	0.3661 (+9.3%)	0.4017 (+11.3%)
RBRM	0.4714 (+14.0%)	0.5204 (+15.9%)	0.3836 (+14.5%)	0.4297 (+19.0%)
BBRM	0.4510 (+9.0%)	0.5041 (+12.3%)	0.3757 (+12.1%)	0.4068 (+12.7%)
TTDM-Q	$0.4741^{\&,*}$ (+14.6%)	$0.5133^{\&}$ (+14.3%)	$0.4096^{\&,*,+,\#}$ (+22.3%)	$0.4475^{\&,*,\#}$ (+24.0%)
TTDM-q	$0.4769^{\&,\#}$ (+15.3%)	$0.5265^{\&}$ (+18%)	$0.4186^{\&,*,+,\#}$ (+25.0%)	$0.4500^{\&,*,\#}$ (+24.7%)

与前文的判断一致，采用与单个查询词相关的查询扩展策略 TTDM-q 优于采用与查询整体相关的查询扩展策略 TTDM-Q. 鉴于 TTDM-q 优于 TTDM-Q 方法，因此下文的详细分析只针对 TTDM-q 进行.

6.2　扩展词分析

表 2 展示了 TTDM-q 对 4 个查询（2,29,30,61）扩展的部分查询词. 表 2 中倾斜加重标注的扩展词是 TTDM-q 额外扩展的其他基线方法都没有扩展到查询中的词，其他词为 TTDM-q 与基线方法均选择作为扩展词的词. 因进行了词干提取，在括号中将单词补全.

表 2　TTDM-q 对 4 个主题选择的查询扩展词

2022 FIFA soccer	global warming and weather	Keith Olbermann new job	Hu Jintao visit to the United States
World	climate	Current	presid(ent)
Cup	cold	Tv	China
Qatar	destroi(y)	career	relat(ion)
Held	caus(e)	Msnbc	diplomat(ic)
Summer	feel	Rt	Obama
Plan	***circumst(ance)***	Talk	***Korea***
footbal(l)	***worldwid(e)***	Host	***North***
Sport	***extrem(e)***	***Join***	***Chicago***
…	…	…	…

以查询 2"2022 FIFA soccer"为例，该主题表述的是卡塔尔（Qatar）经国际足联（FIFA）投票确定举办 2022 年世界杯. 从表 2 中不难发现，不论是 TTDM-q 还是基线方法，词"world"、"cup"、"qatar"等与查询相关的词均被扩展到查询中. 此外，TTDM-q 还额外将"football"和"sport"等其他有关的词扩展了进来. 类似的，查询 29"global warming and weather"关注全球变暖和天气问题，额外将"circumstance"（环境）、"worldwide"（全球）、"extreme"（极端）等相关词扩充了进来. 查询 30"Keith Olber-

mann new job"指的是 Keith Olbermann 离开 msnbc 加入(join)了 Current TV 这件事,TTDM-q 额外将"join"扩展进了查询. 对于查询 61"Hu Jintao visit to the United States",关注的是 2011 年胡锦涛访美的事件,朝鲜是此行的一个关注的议题. 此外,访美期间胡锦涛到访了芝加哥. TTDM-q 相比其他方法,额外将"north"、"korea"、"chicago"等词扩展了进来.

从上面的例子可以看出,TTDM-q 能够在基线方法之外进一步扩展一些与查询相关的词,进而获得性能的提升.

6.3 查询扩展对检索性能的影响

在查询扩展中通常将查询扩展模型和原始查询模型结合来估计最终的查询模型,其权重通过 λ 来调节(见式(3)). 本节对 λ 的取值对检索性能的影响进行了分析,结果如图 4 所示.

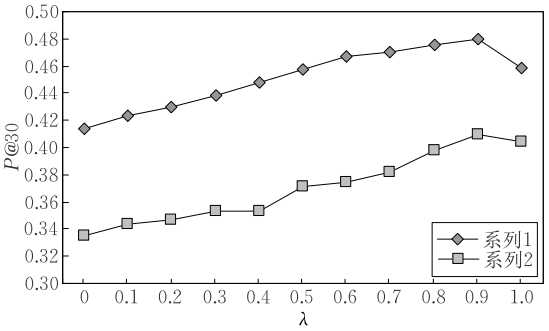


图 4 查询扩展对检索性能的影响

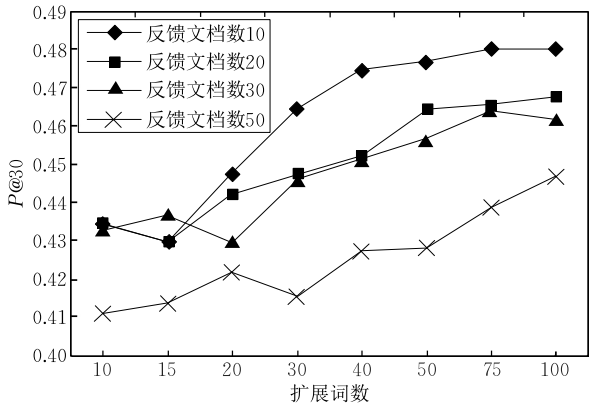
从图 4 中可以看出,随着查询扩展模型的权重 λ 逐渐增大,检索性能逐渐提升,当查询扩展模型的权重为 0.9 时检索性能最佳,此时原始查询的权重只有 0.1.

该实验结果还表明,仅使用原始查询(即 $\lambda=0$)难以取得令人满意的结果,而仅使用查询扩展模型($\lambda=1$)能够在一定程度上代替原始查询来完成检索任务. 但是,一个有效的查询模型应该结合原始查询与扩展查询,并合理分配二者的权重.

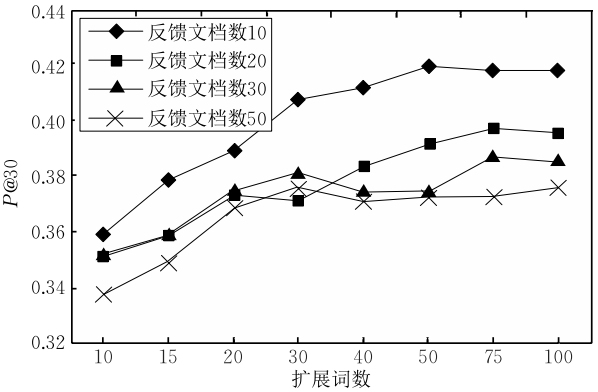
6.4 扩展词数量对 TTDM 的影响及分析

和其他查询扩展方法类似,本文方法也同样存在两个重要的参数:查询扩展词数和反馈文档数. 本文余下部分将对这两个参数对 TTDM-q 的影响进行分析.

扩展词的数量决定了使用多少扩展词来扩展查询. 图 5 显示了本文模型在取不同相关文档数的情况下查询扩展词数对检索性能的影响.



(a) TREC2011 查询1～49



(b) TREC2012 查询51～110

图 5 扩展词数量对 TTDM-q 的影响

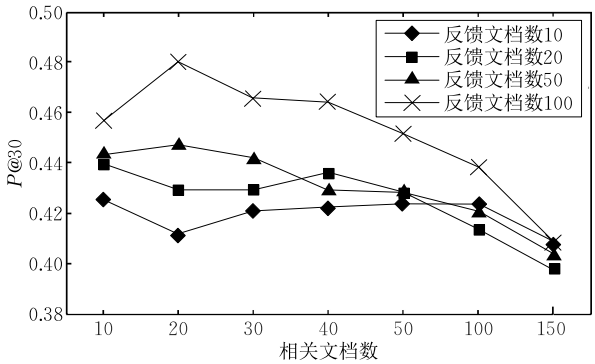
从图 5 中可以看出,不同反馈文档下,随着扩展词数的增加,TTDM-q 的性能逐渐增加并趋于稳定,这与基于内容的查询扩展方法存在着较大的差异:基于内容的查询扩展通常使用数量较少的查询扩展词(例如文献[14]在 TREC 微博数据上设置为 20). 而我们在实验中也发现较大数量的查询扩展词会引起基于内容的查询扩展方法的性能下降. 其原因在于,在基于内容的查询扩展方法中,估计查询模型时利用的是扩展词和查询在反馈文档中的分布. 由于微博较短,仅利用词在反馈文档中的分布信息难以准确的建立查询词与扩展词的关系,随着扩展词数目的增多,扩展词与查询词的相关性估计变得更加不可靠,从而更难于准确的估计查询扩展模型. 因此,基于内容的查询扩展在设置较大扩展词数的时候,检索性能有所下降.

相对而言,TTDM-q 随着扩展词数的增加,性能依然稳定. 这是因为 TTDM-q 中扩展词与查询词的关系利用了每个时间段上全体微博的语言模型(本文实验数据上每天平均微博数约为 30 万条),大量微博被用来计算扩展词和查询词的时间分布的相

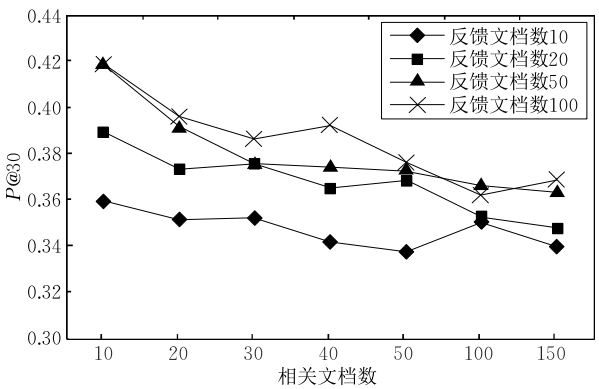
似性，避免了基于内容的方法利用少量的微博估计扩展词和查询关系的不足. 因此，TTDM-q 可以较为稳定的估计扩展词和查询的关系，在查询扩展模型中容纳更多的扩展词而不会对检索结果产生过于明显的影响. 在 TTDM-q 中，那些与查询弱相关和不相关的词被赋予了很小的权重，增大扩展词数虽然会将这部分词扩展进来，但由于其较低的权重而不会对检索结果产生太大的影响. 因此，应用 TTDM-q，扩展词数可以设置为一个相对较大的值.

6.5　反馈文档数量对 TTDM 的影响及分析

另一个影响查询扩展的重要参数是反馈文档数量. 图 6 展示了在固定反馈词数的情况下，TTDM-q 方法随反馈文档数增加的检索性能变化.



(a) TREC2011 查询1～49



(b) TREC2012 查询51～110

图 6　反馈文档数对 TTDM 的影响

从图 6 可以看出，当设置较大的反馈文档数时，TTDM-q 的检索性能会迅速降低，这一点和基于内容的查询扩展方法存在一定的差异：在基于内容的查询扩展中，通常需要一定的反馈文档数量(例如文献[14]在 TREC 微博数据上设置为 50)来保证可以提供足够的扩展词和查询词的分布信息，太少的反馈文档会产生严重的数据稀疏问题，影响模型的估计. 而 TTDM-q 随反馈文档数量增加而性能下降的原因，是由于同一时间段内会有多个不同的话题一起被讨论，存在两个不同热门话题具有相似时间分布的可能性. 由于候选扩展词集合来源于与查询相关的伪反馈文档，当伪反馈文档数量较少时，另一个与查询无关的热门话题中的词在候选扩展词集合中出现的可能性较小，所以基本上不会对当前查询扩展产生影响，但反馈文档数目过多，与查询无关的热门话题中的词出现在反馈文档中的概率会增大，从而使查询扩展受到其他话题的干扰. 因此，TTDM-q 倾向于使用较少的反馈文档，来避免噪声.

6.6　时间片段大小对 TTDM 的影响及分析

由于每个词的时间分布并不符合一些常见的分布形式(如正态分布等)，很难构造出连续的概率密度函数来表示时间分布. 现有基于时间的查询扩展方法大多是基于离散的时间分段展开研究，本文也采用了离散的时间分段，词汇时间分布利用时间片段上的语言模型估计.

本节分析离散的时间片段大小对 TTDM-q 的影响. 图 7 展示了时间片段为 1h、2h、3h、4h、5h、6h、12h、24h 和 48h 的时候，时间片段大小对 TTDM-q 检索性能的影响.

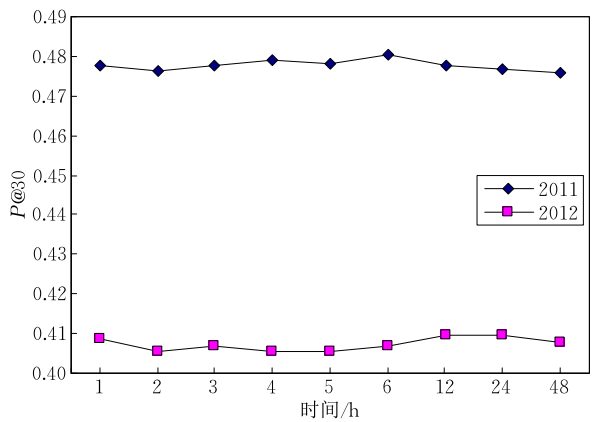


图 7　时间分段大小对检索性能的影响

从图 7 可以看出，时间片段的大小对 TTDM-q 的检索性能的影响并不明显. 在 TTDM-q 中，扩展词和查询词的相关度是由两者词的时间分布所围成的面积决定的，尽管不同的时间片段大小影响的是面积的计算准确程度，但总体上并不影响相关词和查询词的分布面积和不相关词和查询词的分布面积的比较. 理论上，太大的时间片段会弱化时间的影响，较小的时间片段有助于更准确的计算两个分布所围成的面积. 但在实际应用中，过小的时间分段会使得每个分段上微博数量较少，有可能影响时间分

段语言模型的准确度.在实验中,以"天"作为单位性能较好.但本文数据仅仅是 Twitter 的海量微博的1%抽样,在实际应用中,每小时甚至每分钟都会有大量的微博发布,这使得较小的时间片段上也能得到相对精确的语言模型,因此,我们认为实际应用中应选择更小的时间片段.

7 结论和未来工作

本文提出了面向微博检索的基于词汇时间分布相似性的查询扩展方法,该方法在时间维度上建立了查询词与扩展词之间的关系,避免了传统基于内容的查询扩展方法因反馈文档短引起的数据稀疏问题,有助于建立更为可靠的扩展词和查询之间的关系.

具体的,我们在分析词在不同时间段上使用频率的特点基础之上,提出了用词的时间分布来反映人们在不同时间段使用该词的频繁程度,并给出了应用时间分段上的语言模型来估计词的时间分布的方法.进一步,提出了利用词汇时间分布相似性度量扩展词和查询(词)的相关度的方法.最后,建立基于词汇时间分布的查询模型.

本文提出的方法在 TREC 2011 和 TREC 2012 数据上进行了验证.实验结果显示,基于词汇时间分布相似性的查询扩展优于 4 个基线方法:语言模型、基于相关模型的查询扩展、基于新近性的查询扩展和基于爆发期的查询扩展.

与以往基于时间的查询扩展仅利用时间来调整反馈文档权重不同,本文直接利用时间关系来挖掘扩展词与查询词之间的关系.这是本文创新性的尝试,尚存在一些改进空间.研究中我们发现,由于话题数量巨大,并不排除个别情况下两个无关话题的时间分布也可能是相似的,这会对基于时间的文档扩展带来不利影响.虽然扩展词来源于反馈文档中出现的词,保证了无关话题的词成为候选扩展词的可能性较小,但仍存在受到其他话题干扰的可能.未来的工作将探索如何将时间相关和内容相关统一到一个扩展模型中,来避免单一使用 TTDM 方法带来的影响.此外,对词汇时间分布的估计、时间分段语言模型、时间相似度的度量等也有待进一步的完善和发展.

尽管本文提出的方法是面向微博检索的查询扩展,但这个方法同样适用于其他对时间敏感的检索任务,例如微信以及其他即时通讯信息的检索任务.此外,利用时间挖掘词之间的关系并不仅限于应用在查询扩展中,对其他挖掘词与词之间关系的研究也具有一定的参考价值.

参 考 文 献

[1] Gao J, Xu G, Xu J. Query expansion using path-constrained random walks//Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 2013: 563-572

[2] Teevan J, Ramage D, Morris M R. #TwitterSearch: A comparison of microblog search and web search//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. Hong Kong, China, 2011: 35-44

[3] Bai J, Song D, Bruza P, et al. Query expansion using term relationships in language models for information retrieval//Proceedings of the 14th ACM International Conference on Information and Knowledge Management. Bremen, Germany, 2005: 688-695

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

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

[6] Metzler D, Cai C, Hovy E. Structured event retrieval over microblog archives//Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Montreal, Canada, 2012: 646-655

[7] Efron M, Golovchinsky G. Estimation methods for ranking recent information//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 495-504

[8] Dong A, Zhang R, Kolari P, et al. Time is of the essence: Improving recency ranking using twitter data//Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 331-340

[9] Cheng S, Arvanitis A, Hristidis V. How fresh do you want your search results?//Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management. San Francisco, USA, 2013: 1271-1280

[10] Keikha M, Gerani S, Crestani F. Time-based relevance models//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 1087-1088

[11] Choi J, Croft W B. Temporal models for microblogs//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Maui, USA, 2012: 2491-2494

[12] Peetz M H, Meij E, de Rijke M, Weerkamp W. Adaptive temporal query modeling//Proceedings of the 34th European Conference on Information Retrieval Research. Barcelona, Spain, 2012: 455-458

[13] Dakka W, Gravano L, Ipeirotis P G. Answering general time-sensitive queries. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(2): 220-235

[14] Efron M, Lin J, He J, et al. Temporal feedback for tweet search with non-parametric density estimation//Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. Gold Coast, Australia, 2014: 33-42

[15] Lin J, Efron M. Temporal relevance profiles for tweet search//Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval Workshop on Time-Aware Information Access. Dublin, Ireland, 2013

[16] Wei Bing-Jie, Wang Bin. Time-aware mixed language model for microblog search. Chinese Journal of Computers, 2014, 37(1): 229-237(in Chinese)
(卫冰洁，王斌. 面向微博搜索的时间感知的混合语言模型. 计算机学报，2014，37(1)：229-237)

[17] Li X, Croft W B. Time-based language models//Proceedings of the 12th International Conference on Information and Knowledge Management. New Orleans, USA, 2003: 469-475

[18] Ponte J M, Croft W B. A language modeling approach to information retrieval//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, 1998: 275-281

[19] Lafferty J, Zhai C. Document language models, query models, and risk minimization for information retrieval//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, Louisiana, USA, 2001: 111-119

[20] Zhai C, Lafferty J. A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems, 2004, 22(2): 179-214

[21] Zhai C, Lafferty J. Two-stage language models for information retrieval//Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland, 2002: 49-56

[22] Lavrenko V, Croft W B. Relevance based language models//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, Louisiana, USA, 2001: 120-127

[23] Miyanishi T, Seki K, Uehara K. Combining recency and topic-dependent temporal variation for microblog search//Proceedings of the 35th European Conference on Information Retrieval. Moscow, Russia, 2013: 331-343

[24] Soboroff I, Ounis I, Lin J, et al. Overview of the TREC-2012 microblog track//Proceedings of the 21st Text REtrieval Conference. Gaithersburg, USA, 2012

[25] Merigó J M, Casanovas M. A new Minkowski distance based on induced aggregation operators. International Journal of Computational Intelligence Systems, 2011, 4(2): 123-133

[26] Han Z, Li X, Yang M, et al. Hit at TREC 2012 microblog track//Proceedings of the 21st Text REtrieval Conference. Gaithersburg, USA, 2012

HAN Zhong-Yuan, born in 1977, Ph. D. candidate, associate professor. His research interests include information retrieval and information filtering.

YANG Mu-Yun, born in 1971, Ph. D., associate professor. His research interests include information retrieval and machine translation.

KONG Lei-Lei, born in 1979, Ph. D. candidate. Her research interests include information retrieval and plagiarism detection.

QI Hao-Liang, born in 1972, Ph. D., professor. His research interests include information retrieval and information filtering.

LI Sheng, born in 1943, professor, Ph. D. supervisor. His research interests include natural language processing and information retrieval.

Background

Short messages, such as microblog, which are prevailing with the development of mobile Internet, bring new challenges to information retrieval by its 140-letter length. In fact, microblog retrieval has become an important sub-field of information retrieval.

Focused on improving microblog retrieval, this paper

proposes a term temporal distribution based query expansion approach, so as to alleviate the deficiency of classical content based pseudo feedback approach.

In recent years, exploiting the temporal profile to boost the performance of microblog retrieval has become one of the hot issues in the field of information retrieval. A typical practice in existing studies is to re-estimate the term weight by introducing a temporal prior to the traditional pseudo relevance feedback model, which will inevitable inherit the defects of it in microblog retrieval.

In contrast, this paper presents a novel query expansion approach named Term Time Distribution Model(TTDM). It evaluates the relevance between the query terms and the expansion terms according to their correlation in time distribution. Essentially, this method attempts to establish a time relevance model for query expansion, which provides a new alternative to query modeling independent from classical content relevance. Experiments on TREC 2011 and TREC 2012 microblog retrieval track collections show that TTDM outperforms both a standard baseline and other state-of-the-art temporal query expansion models.

The research team of the authors is devoted to information retrieval and other natural language processing researches. From 2010, they began to pursue microblog retrieval, achieving the first place in the microblog real-time adhoc task and microblog real-time filtering task of TREC 2012. This paper reports their findings in short text modeling: exploring the time distribution similarity as a substitute of the content relevance.

This work is supported by National Natural Science Foundation of China (Grant Nos. 61370170, 61402134, and 61173074) and Youth National Social Science Foundation of China (Grant No. 14CTQ032).