电子科技大学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

硕士学位论文

MASTER THESIS

论文题目 推特数据信息的查询扩展方法研究

学科专业 信息与通信工程

学　　号 **201421010814**

作者姓名 马路遥

指导教师 钱峰 副教授

分类号 _____ 密级 _____

UDC[注1] _____

学 位 论 文

推特数据信息的查询扩展方法研究

（题名和副题名）

马路遥

（作者姓名）

指导教师 钱峰 副教授

电子科技大学 成都

（姓名、职称、单位名称）

申请学位级别 硕士 学科专业 信息与通信工程

提交论文日期 2017.3 论文答辩日期 2017.5

学位授予单位和日期 电子科技大学 2017 年 7 月

答辩委员会主席 _____

评阅人 _____

注 1：注明《国际十进分类法 UDC》的类号。

Research On Query Expansion Of Twitter Data Information

A Thesis Submitted to
University of Electronic Science and Technology of China

Major: **Information and Communication Engineering**

Author: **Luyao Ma**

Advisor: **Feng Qian**

School : **School of Communication and Information Engineering**

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据本文所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的材料。与本文一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

签名：_____ 日期：　　年　　月　　日

关于论文使用授权的说明

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后应遵守此规定）

签名：_____ 导师签名：_____

日期：　　年　　月

摘 要

随着互联网的迅猛发展，在全球各地每时每刻都在产生大量的推特数据信息。如何在这大量数据中去筛选满足用户需求的信息变得尤为重要，查询扩展方法在推文检索中广泛应用，可以有效的解决这一问题。

查询扩展主要包含两个部分：一是筛选与原始查询词相关的推文作为语料库；二是筛选语料库中与原始查询最相关的词语作为待扩展词。传统查询扩展方法主要使用 BM25 算法，VSM 算法以及 TF-IDF 算法等对原始查询和推文进行相关性比较，筛选出满足用户需求的推文作为语料库。这种方法存在两个不足：一是含有较少查询词的推文被漏选，二是含有较多查询词的不相关推文被错误的筛选。针对此问题本文在以下几个方面进行研究和创新：

(1) 提出基于推文聚类的查询扩展方法，并对其进行设计和完成。该方法对筛选推文作为语料库这一过程进行改进，并将传统的逐条推文与原始查询词进行相关性比较的推文筛选方法进行优化。该方法先对推文进行聚类，根据与原始查询词的相关性对聚好类的推文进行筛选，得到的推文集合包含了相同语义的所有推文。再比较推文类与原始查询的相关性，筛选出最满足用户需求的推文类。这一方法很好的解决了含有较少查询词的相关推文被漏选的问题。

该方法对比 BM25 算法对两种不同的查询扩展方法在平均准确率(mAP)上分别提升了 11.4%和 12.0%，比 VSM 算法分别提升了 14.9%和 15.3%，比 TF-IDF 算法分别提升了 15.8%和 13.7%。

(2) 提出基于主题划分的查询扩展方法。通过对不相关推文中含有较多查询词而被筛选这一主题偏移问题进行改进，使得含有查询词的不相关推文被有效的过滤。该方法将推文进行主题划分，筛选出满足用户查询的主题下的推文集合作为语料库，有效的去除了含有查询词但并不属于该主题的推文。

该方法对比 BM25 算法对两种不同的查询扩展方法在平均准确率(mAP)上分别提升了 13.2%和 13.9%，比 VSM 算法分别提升了 16.7%和 17.3%，比 TF-IDF 算法分别提升了 17.7%和 15.6%。

(3) 经过分别对主题划分方法和推文聚类方法在查询扩展中的应用进行测试。本文对两种查询扩展方法的优缺点进行分析，发现结合使用两种方法对检索指标有着更高的提升。

关键词：查询扩展，推文反馈，推文聚类，推文检索，主题划分

ABSTRACT

With the rapid development of the Internet, social networks such as Twitter plays an increasingly important role in people's lives. It generates a lot of Twitter data at all times in the world. It is very important to filter the information that satisfies the user's needs in this large amount of data. The query expansion method is widely used in the tweet retrieval, which can solve this problem effectively.

The query expansion mainly consists of two parts: one is to select the tweets associated with the original query word as the corpus, the other is to filter the corpus and the original query is the most relevant words to be extended words, the traditional method mainly uses BM25 algorithm, VSM algorithm and TF- IDF algorithm and so on the original query and tweets to compare the correlation, screening out the user needs to meet the tweet as a corpus. There are two shortcomings in this approach: one is that the tweet containing fewer query words is missed, and the second is that the irrelevant tweets with more query words are mistakenly screened. In view of this problem, this paper studies and innovates in the following aspects:

(1) The thesis proposes the query expansion method based on tweet clustering, designs and completes it. This method improves the process of screening the text as a corpus. This method optimizes the method of comparing the traditional one by one with the original query word. This method first classifies the tweets and filters the tweets of the good class according to the correlation with the original query words. The resulting set of tweets contains all the tweets of the same semantics. And then compare the type of push text and the relevance of the original query, the best to meet the needs of users to promote the class. This method is a good solution to the problem that contains the less frequently asked words.

Compared with the BM25 algorithm, the two methods are improved by 11.4% and 13.9%, which are 14.9% and 15.3% higher than the VSM algorithm, which are higher than the TF-IDF algorithm. 15.8% and 13.7%.

(2) The thesis proposes based on the theme of the query expansion method. This method improves the problem of the subject offset in the irrelevant tweets with more query words, so that the irrelevant tweets containing the query words are effectively filtered. This method divides the tweet into the subject, and selects the set of tweets

under the subject of the user query as the corpus, effectively removing the tweet containing the query term but not the subject.

Compared with the BM25 algorithm, the two methods are improved by 16.2% and 13.9%, which are 16.7% and 17.3% higher than the VSM algorithm, which are higher than the TF-IDF algorithm. 17.7% and 15.6%.

(3) The thesis tests the application of the topic division method and the push text clustering method in the query expansion respectively. In this thesis, the advantages and disadvantages of the two methods are analyzed, and two methods are used to improve the retrieval index.

Keywords: query expansion, tweets feedback, tweets clustering, tweets retrieval

目 录

图目录

表目录

第一章 绪论

1.1 研究背景与意义

推特是一家美国微博客服务网站以及社交网络平台，是全球互联网日访问量最大的网站之一，用户可以在推特上发布 160 个字符之内的信息[1]。在 2006 年由多尔西推出后，推特迅速发展，风靡全球。到了 2016 年的第 4 季度，推特上的月活跃用户已经到达了 3.19 亿[2],而这种活跃人数使得推特上的信息时刻都在更新，其数据量还在不断增加。

相比较传统社交平台，推特拥有更多的用户人群，每个用户都可以发布信息并且通过关注，转发，点赞，评论等进行快速的信息交流。而 160 字的限制使得用户可以大量而且随意的发布原创性内容，因此一些重要的事件都是率先从推特上传播开来，对比传统新闻媒体，推特更具有实时性。推特上的热门话题已经成为了全球关注的焦点，这也使得不同人群的信息获取与交流方式变得更加便捷有效。

但是随着互联网的高速发展，每天更新的推文数据量也在急剧增加，这也使得推文信息的筛选愈发困难。随之而来，在大量的数据中，用户是很难去获取到自己想要的话题信息或者推文内容，本文需要去解决这个问题。传统的推文检索主要是基于在推文中出现的词频频率作为特征，即查询词在推文中的词频越高，推文越满足用户的需求。基于这种思想衍生出了许多常用又经典的用户信息检索的模型。此外检索速度在技术发展中也一直存在瓶颈，使得检索系统需要满足速度性能的前提下再去优化其他性能。

信息检索在日常生活中显得尤为重要，虽然现在存在着许多经典而优秀的检索模型。但是由于查询词的固定，本文一般对查询词与推文的相关性进行分析，对相关性高的查询词给予更高的权重，但是归根结底还是基于查询词中出现的词语对推文进行检索。用户在查询时由于对信息掌握不周全，输入的查询词往往有遗漏甚至是错误的，这就在检索设计时去考虑优化查询词语。

对于上述问题的解决，查询扩展可以简单并且有效的在检索中进行应用。查询扩展是结合了多种信息技术，得到与原始查询相关的待扩展词，并以一定方法加入到原始查询中构成得到新的扩展词用于检索的技术，这样有效的改善了推文检索中的查询结果不全甚至是不正确的问题，进而使得查询结果可以满足用户需求。而根据是否需要用户的参与，查询扩展分为相关反馈和伪相关反馈两种方法。

相关反馈（Relevance Feedback）需要用户的参与，对反馈结果进行人工筛选，

因此反馈结果显然会满足用户需求，但是这带来了复杂的人工参与过程，而伪相关反馈（Pseudo Relevance Feedback）恰恰克服了这个缺点，只要通过自动分析得到反馈结果进而通过比较原始查询和待扩展词的相关性，筛选语言是查询最相关的扩展词，全程不需要人工参与，节约了许多人力成本。显然在大量数据面前，伪相关反馈可以跳过用户进行检索的方法更能得到研究人员的青睐。它对结果的提升是显而易见的，但是其反馈跳过用户这也导致了在初次检索得到的反馈结果不能像相关反馈那样完全满足用户需求，它只是将排序结果的前 k 条推文进行反馈。

虽然伪相关反馈已经很大程度上解决了检索过程中存在的部分问题。但是由于在选取与原始查询最相关的推文这个步骤中获取到的反馈推文还是根据原始查询词与推文的相关性进行匹配的。选取只比较了原始查询与推文的相关性便将相关性最高的 k 条推文进行反馈。而这个相关性不能很好的提供语义信息，这使得反馈结果漏掉了许多相关推文同时也引入无关推文。

对此，本文在反馈时从逐条反馈优化为逐类反馈，选取最满足用户需求的类作为反馈推文，可以有效解决所含查询词不多但是与用户需求相关的推文被漏选的问题。此外基于这种反馈不能去除含有查询词却与主题不相关的推文这种问题，本文使用主题划分筛选出最满足用户需求的主题下的推文进行反馈作为待扩展词的语料库。

1.2 国内外研究现状

上世纪 70 年代，世界上就已经在研究查询扩展的相关方法了。研究较多的是查询扩展优化，即使用一定方法去更好的筛选与原始查询相关的词语与原始查询进行合并得到更满足用户输入要求的新的查询，有效解决了用户输入与用户需求不一致的问题。

最初在 1960 年，Maron 和 Kuhns[3]指明了与原始查询词有着相同描述的词语可以作为查询中的新词语，进而用来查询相关文档。该论文主要研究了增加查询词可以很好的提升检索结果，并未对具体方法做出实现。

在 1965 年，Rocchio[4]在向量空间模型中提出了相关反馈（测试平台为 Smart 系统[5]）,并被广泛应用。该方法使用向量空间模型获取与原始查询最相关的推文作为待扩展词的语料库。Ide[6]在上述研究的基础上对扩展得到的新词语的加权提出了实现方法。

1971 年，Spark Jones[7]提出了基于统计词典的查询扩展方法，它根据不同词语在文档中共同出现的频率将词语归类。从与原始查询词最相关的词语类中筛选

出与原始查询最相关的词语。该方法只通过词语在文档中的共现频率进行词语聚类，效果有限。

1983 年，Croft 和 Harper[10]认为对于不同的原始查询词，其重要程度也不同。获取得到原始查询词在文档集合中的权重，认为词语在文档集合中出现的频率越高，其对文档特征的描述越差，权重也就越低。该方法主要对查询词的重要性进行描述。

1988 年，S.T. Dumais[11]等研究人员提出的隐含的语义分析在文档的检索聚类和分类中被广泛应用。此外在 1999 年 Hofmann 提出了 LSA 模型进一步优化了潜在语义分析，而到了 2003 年，由 Blei, David M.、Ng, Andrew Y.、Jordan[12]等人提出的 LDA 主题模型将文档中的每篇文档的主题以概率形式给出，这对隐含的语义分析提供了更大的进步空间。

1993 年，Qiu 和 Frei[13]使用了全局文档中的相似词语进行扩展。计算了所有文档中词语的相关性。该方法对所有词语相关性进行计算，工作量很大，在实际应用中性能很差。

1996 年，Xu 和 Croft[14]提出在局部文档语料库中的词语与原始查询的相关性，得到新查询词，并认为该相关词将会与原始查询词共现，进而将共现的词语作为待扩展词加入到原始查询，改善检索结果。

相关反馈在大规模的推文检索中不适用，因此 Y. Rui, T.S. Huang, M. Ortega[25]等人在 1998 年提出了不需要人工参与的伪相关反馈方法，伪相关反馈的主要步骤如图 1-1。

伪相关反馈流程包括了通过输入原始查询词（Query）对推文（Tweet）进行排序，将排序靠前的推文作为待扩展词的语料库，从语料库中选取与原始查询最相关的词语作为新查询词（Term），将新查询词加入到原始查询重构新查询对推文进行检索，该方法对于作为待扩展词语料库的筛选十分粗糙，只比较了原始查询词在推文中出现的频率。

V. Lavrenko 和 W.B. Croft[8]等人在 2001 年提出了使用联合概率去描述待扩展词与经过查询词筛选出的排名最高的初始检索结果的相关性，进而去描述待扩展词与查询词的相关性。该方法从联合概率角度说明了筛选待扩展词的方法，但是在进行初始检索的时候使用了 TF-IDF 算法，它主要从查询词在推文中出现的频率去分析查询词和推文的相关性，其检索结果不能很好的满足用户需求。

A. R. Rivas, E. L. Iglesias, and L. Borrajo[37]等人在 2014 年针对生物医学方面做查询扩展方法研究，但其在初次筛选推文时使用 BM25 算法，该方法虽然对 TF-IDF 方法进行了诸多改进，但是在实际应用中还是不能跳出查询词在推文中出

现的频率这一框架，导致初现含有较少查询词的相关推文被遗漏，含有较多查询词的不相关推文被错误的筛选这一问题。

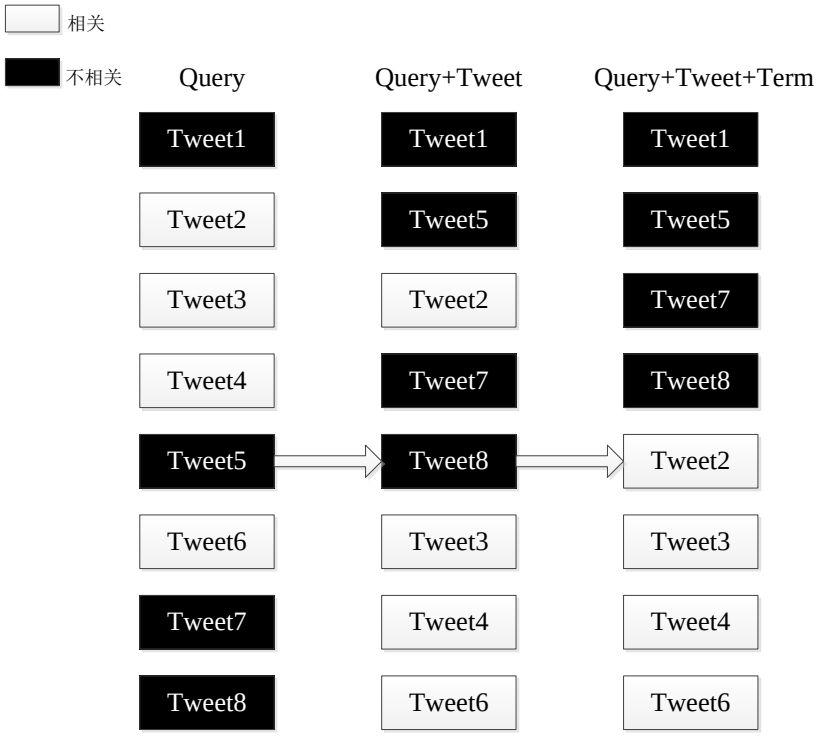

图 1-1 伪相关反馈流程

吴秦[38]等人在 2014 年提出了基于语义词典的查询扩展方法，先根据词频共现以及语义分析获取扩展词，然后对原始查询词和扩展词进行加权，最后计算新查询与文档的相关性获取满足要求的文档。该方法对其他局部扩展将会加入大量非相关词语这一问题做出优化。

吕碧波，赵军[39]等人提出了基于相关文档词建模的查询扩展方法，该方法是基于语言模型的查询扩展方法，同时考虑了扩展词需要具备的相关性和覆盖性。

Xu 和 Croft[15]进一步提出了基于局部上下文分析的方法，它将全局分析与局部分析进行结合。而在 2003 年，Kelly 与 Tcevan 提出了隐氏相关反馈模型，它们分析了用户的查询习惯，挖掘出隐含的查询特征，然后获取新查询。

X. Li 和 W. Croft[9]在 2003 年通过研究发推时间对推文检索结果的影响进行了分析，认为发推时间与查询时间相近则其有着更好的意义，在初次检索中更容易被检索出。在实际应用中，它只是直观的对发推时间和查询时间的差值进行分析，其模型过于简单。

T. Miyanishi, K. Seki, K. Uehara[36]等人在 2014 年提出 two-stage relevance feedback 方法进行查询扩展，对于传统方法筛选待扩展词语料库存在不足的问题

提出使用人工参与方法进行筛选，可以很好的解决问题，虽然简化了纯粹的人工筛选，但是人工参与过程对本文研究还是不满足需求的。

此后主要的研究工作都在针对查询扩展提出了各种优化方法，特别在社会媒体高度发展的今天，结合用户的社交网络对查询扩展的结果有更好的效果。Wang[16]等人对多媒体图像的研究将社交网络中的交互信息按照语义内容进行聚类，Zhou[17]等人针对社交网络的使用者的标签提供了独有的扩展方法，L.Anagnostopoulos[18]等人针对标签中语义关联信息对社交媒体的语义进行扩展。本文不对社交网络对查询扩展方法的影响作具体分析。

而查询扩展方法还从语义相似层面研究了查询扩展方法，它通过人工构建的词典比如 WordNet 等，根据构建规则计算词语间的相似性。也在查询扩展方法中有着广泛的应用，其相似度计算主要基于包含在层次结构“is-a”中的信息[23]，通过结构中不同词语间的连接特性计算词语间的相关性，进而选取满足相关性阈值的词语作为待扩展词。

在相似度的计算方法中，论文[35]认为使用 WordNet 在相关性上的判断主要根据词典构造的网络特征，比如某个节点的密度（拥有兄弟节点的个数），节点的深度（在词典网络中的层次）以及节点间的连接方式（上下位、整体部分）等。

基于语义相似性的 WordNet 词典对于大部分已知词语都作了语义构建，但是对于推文文本，它不能产生良好的效果。首先推文是由用户针对某些事件或者看法而形成的，组成的词语有很大的随意性，甚至针对某些新闻事件产生的新词并不能很快被收入到 WordNet 中。同时即便对于某些词语的近义词在用户的特定需求下将会变成噪音（不符合用户输入的意愿），对查询结果产生负面影响，影响的结果已经经过实验验证，因此本文的查询扩展研究不考虑使用 WordNet。

1.3 本文主要工作

查询扩展已经在信息检索中形成了一套行之有效的方法，它可以完美解决用户输入的信息不完备乃至有错误的问题。检索系统对于信息不匹配也有自己的扩展方法，将筛选出来的待扩展词以一定权重加入到原始查询，它将用户的查询表述更加准确。尤其在短文本检索中，查询扩展与传统的文献检索也有不同，本文主要在检索流程和方法上去进行优化。

综上所述本文的主要工作有：

首先，对于传统的推文检索方法存在的含有较少查询词的相关推文被错误检索的问题提出了基于推文聚类的查询扩展方法，并对结果进行测试分析

其次，对于含有较多查询词的不相关推文被错误筛选的问题提出了基于主题

划分的查询扩展方法，并对结果进行测试分析。

最后，类比上述两种反馈结果，分析其在查询扩展中的实际结果，得出最后结论。

1.4 本文的结构

全文共分为四章，主要章节安排：

第一章主要对论文的研究背景和意义做了介绍，同时针对国内外的研究前景作了展开分析，此外还对本文需要做的内容进行了简单介绍，最后对于文章的章节安排作了简单描述。

第二章对本文所涉及的研究方法和技术做了必要的说明。主要包括推文检索的概念以及一些主要模型，以及现有的一些查询扩展方法，还对检索结果的评价方法也做了介绍。

第三章提出了基于推文聚类的查询扩展方法，分析在查询扩展中的应用结果。对比了通过推文聚类得到待扩展词语料库与传统的 BM25 算法，VSM 算法以及 TF-IDF 算法得到结果存在的优势，并对最终检索结果的影响。

第四章提出了基于主题划分的查询扩展方法，得到与用户主题最相关的推文作为待扩展词的语料库，将结果与使用传统的 BM25 算法，VSM 算法，TF-IDF 算法进行比较，检索结果有着显著提升。同时本章分别对比了使用推文聚类方法和主题划分方法的优缺点，提出结合使用两种方法对结果提升更加明显。

最后是总结与展望。

第二章 相关理论技术基础

2.1 推文检索技术

推文检索技术即针对发布在推特上的海量的无序信息，用户给出查询词并结合检索工具，根据一定的检索方法和规则，将无序信息进行排序建立索引，最终向用户返回所需要的推文的方法。在推特检索中有三个重要概念：用户查询词，检索方法以及检索结果。

用户查询词即用户想要得到查询结果而输入的词语，用户所需要的搜索的内容一般被表示为一个或者几个关键词以提高查询的召回率。

检索方法即为根据用户查询词，搜索工具对无序的推文按照推文与用户相关性排序返回给用户查询结果的方法。相关性的概念为是否满足用户的检索要求，包含实时性，可行度，相关度以及其他条件等。

检索结果即返回给用户的信息，相较于网页检索，一般返回的某个网页即包含了用户所需要的信息，而推文检索由于字数有限，发推随意，一条推文所包含的信息极为琐碎，需要多条推文一起反应用户的需求。

2.1.1 推文检索相关概念

推文检索引申于信息检索，而这个概念最早来自 1948 年 Calvin Mooers 的论文。信息检索是指用户在包含多种信息的信息源中准确找到用户所需要信息的过程。而推文检索是将泛指的信息源锁定在推文上。

信息检索是多种学科（计算机科学，图书情报学等）的交叉，是在大量存储的信息中心筛选自己想要信息的过程。它以计算机为手段使用语言学，认知科学等处理信息。事实上，信息检索主要包含了两个步骤：“整理”和“选取”。“整理”即将大量无序的信息资源进行组织整合建立索引，“选取”是根据用户需求查找资源。

2.1.1.1 推文检索的相关术语

查询：用户为获取推文信息输入的请求条件，对用户而言，查询是满足需求并向计算机提供检索依据的词句，查询根据用户需求不断变化，其检索和结果也会产生各种变化。

推文集合：所有待检索推文构成的集合。

相关性：用户对于检索结果好坏的判断，同一用户在不同条件下也会做出不

同的论断，是一个主观概念。

相关性可以从以下两个角度去定义：

系统角度：检索系统产生的结果，这就将最后相关性优化的重担交给了系统。检索系统得出的核心内容需与用户需求挂钩，同步不断改进检索系统的方法等去满足用户需求。

用户角度：用户主观的去评判检索结果，从自己的需求出发。这是根据用户需要对结果的评价，不能脱离用户去界定。主要通过多用户对检索结果进行人工标记，进而产生结论。

2.1.1.2 推文检索的过程

形式化而言，推文检索的相关性结果R由用户查询Q，推文D，推文集合C影响，f是构建用户查询，推文，推文集合关系的模型，返回一个实数$R = f(Q,D,C)$。

其意义就是通过对用户的原始查询以及推文集合进行相关性匹配，匹配算法和模型在下一章节中细讲，最后返回满足用户需求的推文集合，这便是最基本的推文检索。

2.1.2 检索模型

检索模型是推文检索的重点，是将查询，推文，推文集合以及相关性等要素进行抽象描述并得到相关性计算方法的流程框架。

而布尔模型，向量空间模型和概率模型是信息检索中最为经典的三个模型。

2.1.2.1 布尔模型

布尔模型是对特征项的匹配只有两种结果。这是一个二值变量，如果文本中特征项匹配成功则为“true”，否则为“false”。

在包含 m 条推文的集合C中，每条提问可以表示成$\{d_1, d_2, ...d_m\}$，同时再输入一个包含有 n 个查询词的查询词集合，每个查询词可以表示为$\{t_1, t_2, ...t_n\}$。如果推文d_i中的所有词语的都是对查询词的描述，则C中任意一条推文d_i就可以表示成$\{d_{i1}, d_{i2}, ...\mathrm{d}_{im}\}$，其中$d_{ik}$的值为：

$$d_{ik} = \begin{cases} 1 & \text{如果文献中包含} t_{ik} \\ 0 & \text{否则} \end{cases} \tag{2-1}$$

查询词被表示为关键词的布尔组合，这可以通过多个查询词的交集或者并集

进行组合：

$$q=(t_1\cap t_2)\cup(t_3\cap t_4) \tag{2-2}$$

比如公式 2-2 中是由多个关键词构成的查询表达式，当表达式满足用户设定的需求时，结果才会被检索出来。

布尔模型的优点：对于组合查询，布尔模型可以通过组合方式进行筛选，同时逻辑简单，比较直观。

布尔模型的缺点：由于其特征相匹配只能是个二值，适用情况很窄，同时对于各个特征项的重要性也没有区分。

2.1.2.2 向量空间模型

向量空间模型(VSM)表示推文集合C和查询Q都可以表示成向量的形式，而对比两个向量的相关性也就得到了与原始查询最相关的推文集合，同时对检索出来结果进行处理，从而更好的进行重检索，向量空间模型的主要过程：

(1) 推文向量的构造。

(2) 查询向量的构造。

(3) 查询与推文的匹配函数。

(4) 相似度阈值的确定。

向量空间模型的优点：该模型根据余弦相似度来比较文档与查询的相似程度，完美的克服了布尔模型中的二值限制，它可以根据相似性对相关文档进行排序，同时它也可以根据多个查询词的重要程度对查询词权重做出设计。

向量空间模型的缺点：需要计算每个查询与文档的相关性，计算量过大，此外它不能处理结构化查询。

VSM 广泛应用于文档相似性的计算，其在自然语言处理中有着很好的使用。它的实际意义就是将每个词语作为向量的维度，其值就是向量长度，而比较两个向量的相近程度的方法如下：

Dot：

$$Sim(d,q)=d\bullet q=\sum_i(a_i\times b_i) \tag{2-3}$$

Cosine：

$$Sim(d,q)=\frac{d\bullet q}{\|d\|\times\|q\|}=\frac{\sum_i a_i\times b_i}{\sqrt{\sum_i a_i^2\times\sum_i b_i^2}} \tag{2-4}$$

Dice：

$$Sim(d,q)=\frac{2\times d\bullet q}{\|d\|^2\times\|q\|^2}=\frac{2\sum_i a_i\times b_i}{\sum_i a_i^2\times\sum_i b_i^2} \tag{2-5}$$

Jaccard：

$$Sim(d,q)=\frac{d\bullet q}{\|d\|^2\times\|q\|^2-d\bullet q}=\frac{\sum_i(a_i*b_i)}{\sum_i a_i^2+\sum_i b_i^2-\sum_i(a_i*b_i)} \tag{2-6}$$

比较常用的是公式 2-4 中的余弦相似度方法，而在本文有关向量空间模型的测试中都是使用余弦相似度作为相关性比较的。

当推文中的词语含量变大时，效率会很低，使用降维方法去处理这些问题可以有效提升性能和准确率。

2.1.2.3 概率模型

检索系统中，用户所想要的查询结果不可能能被精确的表示出来，在信息检索中存在了很大的不确定性，不存在一个很完美的过程可以去表征得到的数据结果是否就是用户想要的，而概率模型可以对这种不确定的结果进行表示。

在推文检索中，可以使用概率模型去估计原始查询和每条推文的相关程度然后选取最相关的推文集合。

在概率模型中，$T=\{T_1,T_2,...T_n\}$ 表示构建的一个含有所有表音词的集合，每一条推文 d_j 根据是否存在有集合 T 中的词语将推文集合表示为向量 $x=\{x_1,x_2,...x_n\}$，n 是标引词的数量，$x_i=0$ 或 1 表示推文集合中有没有第 i 个标引词。

根据用户的检索 q，将推文集合中的所有推文分为两类，一类与检所需求 q 相关(R)，另一类不相关($\bar{R}$)。

将查询 q 与每条推文进行相关性比较可以将推文分为相关和不相关：$P(R|\vec{x})$ 和 $P(\bar{R}|\vec{x})$。根据贝叶斯公式：

$$P(R|\vec{x})=p(R|\vec{x})=\frac{p(R)p(\vec{x}|R)}{p(\vec{x})} \tag{2-7}$$

推文 d_j 与查询 q 的相似度函数为：

$$sim(d_j,q)=\frac{p(R\mid\vec{x})}{p(\overline{R}\mid\vec{x})} \tag{2-8}$$

公式 2-8 表示的结果越大，文档 d 与检索 q 的相关性也就越明显。设置阈值 k，要求相似度必须大于 k。

2.1.2.4 语言模型

语言模型（Language Model，LM）对自然语言进行了深入研究，并对其客观规律按照统计知识进行建模。对于各种文档中的语言结构，比如词语间和段落间的信息进行结构上的比较。Ponte 和 Croft[19] 最早把 LM 应用到了检索系统中[26]，他们使用了概率模型去比较文档集合与原始查询的匹配程度语言模型是把与用户认为相关的文档中查询标准是根据选择得到词汇。

语言模型在自然语言处理中有着非常重要的作用，在常用的几种语言模型已经构成了自然语言处理的最基本的研究方法。它使得计算机和稍显主观的语言上的内容统一起来，更好的去描述人类的语言表述。

2.1.2.5 几种模型的比较

布尔模型模型简单，也比较好理解，使用起来很方便。同样的，它太简单，很多场景不适用，它只适合在完全匹配中使用。布尔模型只存在是与不是两种匹配结果，这在推文检索中很不现实，很难精确匹配。

相比于布尔模型，向量空间模型引入了权值计算，除了完全匹配除信息外还能根据用户输入得到满足用户需求但不与查询完全匹配的信息，相比于二值比较，它采用了余弦相似度比较原始查询与推文的相似性，很好的提高了检索结果。同时这样加大了模型的复杂程度。同时他要求索引项两两正交，相互独立，这样也不切实际。

而在空间向量模型和概率模型的比较中，Croft 得出的结果指出概率模型比向量空间模型更好，但是 Salton 和 Buckley 通过实验对上述结果提出了质疑，他们认为向量空间模型远比概率模型要好，而这个结论在一些开发者中得到了普遍的认可。但是也有很多研究者认为两个模型所得到的结果在实验中不能分出好坏，与实验内容息息相关。

2.1.3 BM25 算法

在检索系统中，BM25 算法是一个简单而且常用的算法，它可以对原始查询和推文进行相关性比较得到最满足用户需求的推文，同时在查询扩展中这也用于

对待扩展词语料库的反馈[20]。

BM25 算法将原始查询 Q 解析为一个个查询词 q_i，对于每条推文 d，比较每一个 q_i 与 d 的相关性得分，然后对所有得分求加权和，进而得到原始查询 Q 和推文的相关性。

BM25 算法的公式：

$$Score(Q,d)=\sum_{i}^{n}W_i\cdot R(q_i,d) \tag{2-9}$$

在公式 2-9 中 Q 为原始查询，q_i 为原始查询中的每一个查询词，d 是每一条推文，W_i 是每一个查询词的权重。$R(q_i,d)$ 每个查询词和每一条推文的相关性得分。

对于每个查询词权重 W_i，本文一般采用 IDF 去判断词与推文的相关性权重：

$$IDF(q_i)=\log(\frac{N-n(q_i)+0.5}{n(q_i)+0.5}) \tag{2-10}$$

在公式 2-10 中 N 表示推文集合，$n(q_i)$ 表示了含有词语 q_i 的推文数。

IDF 表明了对于给定的推文集合，包含了 q_i 的推文越多，那么 q_i 的权重越低。即对于经常出现的词语，它对于推文的区分度会变得不明显，进而它来判断推文权重的权值就越不重要。

表达了查询词与推文相关性的 $R(q_i,d)$ 公式：

$$R(q_i,d)=\frac{f_i\cdot(k_1+1)}{fi+K}\cdot\frac{qf_i\cdot(k_2+1)}{qf_i+k_2} \tag{2-11}$$

$$K=k_1\cdot(1-b+b\cdot\frac{dl}{avgdl}) \tag{2-12}$$

在公式 2-11 和 2-12 中 k_1，k_2，b 是调节因子，b 可以调节推文长度对结果的影响，可以理解为越长的推文越有可能包含查询词，区分度越不明显。f_i 为查询词 q_i 在推文中出现的频率，qf_i 为查询词 q_i 在整个查询 Q 中的出现频率。dl 为推文的长度，$avgdl$ 为推文的平均长度。基本上查询 q_i 只会在查询 Q 中出现一次，因此上述公式可以简化为：

$$R(q_i,d)=\frac{f_i\cdot(k_1+1)}{f_i+K} \tag{2-13}$$

BM25 算法广泛用于原始查询与推文的相关性比较，在查询扩展的待扩展词反馈有着很好的效果。

2.1.4 TF-IDF 算法

TF-IDF 算法由 TF（词频）和 IDF（逆文档频率）两部分相乘得到的结果。如果某个词语对推文集合的 TF-IDF 的值越大，表明这个词在集合中越重要，其相关性就越好。

TF 表示的是查询词在推文集合中出现的频率，频率越高它对推文集合越重要，其计算方法如公式 2-14：

$$TF(\text{词频})=\frac{\text{某个词出现在文章的次数}}{\text{文章的总词数}} \tag{2-14}$$

IDF 是逆文档频率，它表达了如果一个词语在各种语料库中是经常出现的，那么它的特征就不是很明显就不能特定的去描绘某个推文集合，其重要性就会下降，常用方法如公式 2-15：

$$IDF(\text{逆文档频率})=\frac{\text{语料库的文档总数}}{\text{包含该词的文档数+1}} \tag{2-15}$$

显然使用 TF-IDF 可以很好的判断查询词与推文集合的相关性，如公式 2-16：

$$TF-IDF=TF\times IDF \tag{2-16}$$

2.2 查询扩展方法

一般在信息检索中，检索系统根据用户输入的查询词进行匹配，只有满足用户的查询规则时才能被检索到。但是，对于用户输入的查询词往往会有许多种相关表示（用户不能很好地去想到所有的查询词句），它们也是对用户需求的表达，因为不在查询词中，它们将不会作为查询依据，这将对查询结果产生很大影响。

查询扩展将很好的去解决这种潜在查询词与用户输入不匹配的问题，它包含了两个主要步骤：扩展查询词，重构查询词。

同时在查询扩展中存在着集中非常经典的方法：基于相关反馈的查询扩展，基于全局分析的查询扩展，基于局部反馈的查询扩展。

2.2.1 基于相关反馈的查询扩展

在相关反馈查询中，检索系统向用户反馈一组推文集合，用户手工标记满足

用户的推文，一般情况下选取最满足用户的 top m 条推文，主要思想就是基于用户反馈的 m 条推文作为带扩展语料库，选取满足用户需求的词句按照一定权重加入到原始查询中进行最终检索。

相关反馈的实际过程就是根据原始查询得到反馈语料库，再对语料库与原始查询进行对比得到新查询再进行检索知道满足用户需求，流程图为：

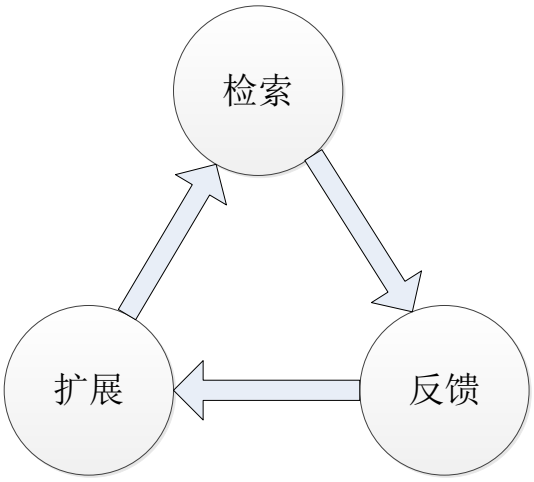

图 2-1 相关反馈的流程图

Rocchio 算法是通过查询的初始匹配文档，对原始查询进行删改优化的方法，是一种经典的相关反馈算法。它将相关反馈信息融合到向量空间模型。

标准 Rocchio 公式：

$$\overrightarrow{q_m} = \alpha\vec{q} + \beta\frac{\sum_{i\in R}\vec{d_i}}{n_R} - \gamma\frac{\sum_{i\notin R}\vec{d_i}}{n_{\bar{R}}} \tag{2-17}$$

在公式 2-17 中 $\overrightarrow{q_m}$ 是修改后的查询向量，它是 $\vec{q}$（原始查询向量），$\frac{\sum_{i\in R}\vec{d_i}}{n_R}$（相关文档的平均文档向量）以及 $\frac{\sum_{i\notin R}\vec{d_i}}{n_{\bar{R}}}$（不相关文档的平均文档向量）的加权求和。$\alpha$，$\beta$，$\gamma$ 为加权权重。几种常用取值方法：$\alpha=\beta=\gamma=1$，或者 $\alpha=\beta=1$，$\gamma=0$。

同时 Rocchio 公式还存在以下两种变形：

$$\overrightarrow{q_m} = \alpha\vec{q} + \beta\sum_{d_i\in R}\vec{d_i} - \gamma\sum_{d_i\notin R}\vec{d_i} \tag{2-18}$$

$$\overrightarrow{q_m} = \alpha\vec{q} + \beta\sum_{d_i\in R}\vec{d_i} - \gamma\max\vec{d_i} \tag{2-19}$$

公式 2-19 中的 $\max \vec{d_i}$ 是不相关文档中和查询 q 最相似的文档向量。

以上三种方法效果相当，都是通过增加相关文档中词语的权重以及减少不相关文档中的词语权重，对词语权值进行优化。相关反馈算法简单易于理解，并在实际应用中有较好的结果，但是，这都是需要用户参与，用户筛选的结果受主观影响[21]。

采用概率模型以及词语加权在相关推文反馈中也被广泛应用。首先对推文集合和原始查询 Q 根据用户需求去判定是否相关。然后将检索词 t_i 出现在相关或者不相关推文中的概率分别为 p_i 和 $\overline{p_i}$，据此，本文对检索词进行权重的变更，如公式 2-20：

$$W_i = \log\{p_i \times (1-\overline{p_i}) / [\overline{p_i} \times (1-p_i)]\} \tag{2-20}$$

根据公式 2-20 中得到的变更后的权重，此时本文可以去计算原始查询 Q 和推文集合的相关性，然后进行排序，如公式 2-21：

$$sim(d_j, q) = \sum_{i=1}^{n} w_{i,q} w_{i,j} (W_i) \tag{2-21}$$

在公式 2-21 中 $w_{i,j}$ 和 $w_{i,q}$ 分别表示了推文 d_i 和 d_j 的关联性以及原始查询 Q 和推文 d_i 的关联性， n 表示的是推文集合中的推文数量是 n 。

基于概率模型的相关反馈并没有将查询词进行增加，它主要根据查询词在相关推文和不相关推文的出现情况对原始查询词的权重进行变更，这样去表征查询词的重要程度。

相关反馈的优点主要是对原始查询进行了加权，这样更好的表示了各个原始查询词在推文检索中的重要程度。但是其缺点也很明显，在反馈中并没有考虑词语在推文中的权值，也没有进行查询扩展只对用户输入的查询词进行循环反馈，反复加权，但是没有在原查询中重新加入新的查询信息。

2.2.2 基于全局分析的查询扩展

全局分析即利用文档集合中的全部文档进行分析，利用全部文档信息计算词语之间的相似度（与原始查询词 q 无关），利用生成好的词向量或矩阵计算与查询 q 的相似度。

传统全局分析需要计算查询 q 的每个词与文档中的每个词的相似度，1990 年以后发展了现在的全局分析，通过计算整个查询 q 与文档中词语的相似度，拥有比传统的全局分析更好的效果。

2.2.2.1 基于语义词典的查询扩展

采用 completelink 算法，对所有对进行计算，采用向量余弦规则，两个簇进行距离计算，使用最近的两个文档距离作为计算值，结果为一个个簇。

簇的选择：选择相似度大于阈值，文档数量大于阈值的簇，同时，选择逆文档频率小于最小逆文档频率的具有区分性的短语。

2.2.2.2 基于相似词典的查询扩展

基于相似词典的查询扩展的主要流程为：

(1) 定义词语之间的相似度，设 N 为文档数目，t 为整个文档集合中的词语数目，$f_{i,j}$ 为词语 k_i 在文档 d_j 中的频度，t_i 为文档 d_j 中的不同词语数目，定义文档 d_j 中的逆词语频率 itf_j。

$$itf_j = \log \frac{t}{t_j} \tag{2-22}$$

(2) 对于所有 N 篇文档，其矩阵表示：

$$A_{m*N} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & w_{2N} \\ \vdots & \vdots & \cdots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mN} \end{bmatrix} \tag{2-23}$$

公式 2-23 中的每个 $w_{i,j}$ 表示的是$[k_i, d_j]$对应的权重，可如公式 2-24 计算：

$$w_{i,j} = \frac{[0.5 + 0.5\frac{f_{i,j}}{\max_j(f_{i,j})}]itf_j}{\sqrt{\sum_{i=1}^{N}[0.5 + 0.5\frac{f_{i,j}}{\max_j(f_{i,j})}]^2 itf_j^2}} \tag{2-24}$$

(3) 将公式 2-24 中的第 i 行看成词语 k_i 的一个向量表示 $\vec{k_i}$，采用内积算法计算 k_u, k_v 之间的相似度，得到词语相似矩阵，其中 u 行 v 列表示为：

$$c_{u,v} = \vec{k_u} \cdot \vec{k_v} = \sum_{\forall d_j} w_{u,j} \times w_{v,j} \tag{2-25}$$

(4) 将原始查询 q 作为文档向量，对于 q 中的每个词语 k_i，计算原始查询 q 中的每个词语 k_i 与文档 d_j 的权重 $w_{i,q}$，原始查询 q 的向量表示为：

$$\vec{q} = \sum_{k_i \in q} w_{i,q} \vec{k_i} \tag{2-26}$$

(5) q 与任意 k_v 之间的相似度：

$$sim(q,k_v)=\vec{q}\cdot\vec{k_v}=\sum_{k_u\in q}w_{u,q}\times(k_u\cdot k_v)=\sum_{k_u\in q}w_{u,q}\times c_{u,v} \tag{2-27}$$

(6) 通过公式 2-27 中获取与原始查询相关性最高的 m 个词语加入到原始查询中，得到新查询 q'，新加入的词语 k_v 的权值设置为：

$$w_{v,q}=\frac{sim(q,k_v)}{\sum_{k_u\in q}w_{u,q}} \tag{2-28}$$

2.2.3 基于局部分析的查询扩展

2.2.3.1 基于局部聚类的方法

局部聚类的方法即低聚部分当中的词语根据一定方法进行聚类，将描述同一类型的词语聚在一起，结果就是聚成表述不同类型的词语簇，通过查询 q 在各个簇中查找与 q 相关的词语进行扩展。

使用不同表示方法，得到的簇的结果也是各不相同。

三种簇的定义：

关联簇，度量簇和标量簇

关联簇：如公式 2-29 将局部文档中的词语构建成相关性矩阵 $a_{ij}=f(t_i,d_j)$ 表示为：

$$A_{m*n}=\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & \cdots & \vdots\\ a_{m1} & a_{m2} & \cdots & a_{mn}\end{bmatrix} \tag{2-29}$$

在公式 2-29 中 AA^T 称为关联矩阵，其中的第 u 行，第 v 列元素 $c_{u,v}$ 表示的是 t_u 和 t_v 的相似度：

$$c_{u,v}=\sum_j f(t_u,d_j)*f(t_v,d_j) \tag{2-30}$$

公式 2-30 中的 $c_{u,v}$ 实际上表示的是 t_u ,t_v 在局部文档中共现的频度。

将 $c_{u,v}$ 进行归一化：

$$c_{u,v}=\frac{c_{u,v}}{c_{u,u}+c_{v,v}-c_{u,v}} \tag{2-31}$$

最后通过对归一化结果进行分析，根据与原始查询的相关性进行降序，选取最相关的几个词语作为扩展词。

度量簇：关联簇中并没有把词语之间的位置距离考虑在内，实际中距离越近的词语拥有更近的相似度。

如果在一篇文档的窗口中出现了两个词语t_u,t_v，则$r(t_u,t_v)$表示了词语t_u,t_v距离信息，若不处于通一篇文档则其距离为无穷大。

构造关联矩阵，其中t_u,t_v的相关度定义如公式 2-32：

$$c_{u,v}=\sum_{t_u}\sum_{t_v}\frac{1}{r(t_u,t_v)} \tag{2-32}$$

对$c_{u,v}$进行归一化：

$$c_{u,v}=\frac{c_{u,v}}{Nu*Nv} \tag{2-33}$$

标量簇：在关联簇和度量簇情况下，本文均能得到归一化相关矩阵，该矩阵的第u行，第v列可以看成t_u和t_v的向量表示$\vec{s_u}$和$\vec{s_v}$。

本文定义关联矩阵的元素为：

$$c_{u,v}=\frac{\vec{s_u}\cdot\vec{s_v}}{\left|\vec{s_u}\right|\cdot\left|\vec{s_v}\right|} \tag{2-34}$$

簇中的不同词语互称为邻居，也称为搜索同义项，区别于语法上的同义词。

2.2.3.2 基于局部上下文分析的方法

由于局部聚类存在以下缺陷：计算了q中的每个词和所有词的相似度而不是计算整个查询q和所有词的相似度。

局部上下文分析将查询q看成一个整体，在局部文档中计算和查询q最相关的词语进行扩展[22]。

局部上下文分析的步骤为：

(1) 第一步，先将所要测试的文档分段，每一段作为查询的目标，对与原始查询q，返回与原始查询最匹配的段落。

(2) 返回的n个段落中，计算原始查询q和段落中的词语c的相关性$sim(q,c)$。

(3) 将与原始查询最为匹配的m个词语加入到原始查询。加入词语的权重为$1-0.9i/m$，i为m个词语在排序时的位置。原始查询词拥有一个较大的初始值。

下面介绍 q 和 c 相似度的计算方法：

(1) 定义 c 和某个词语 k_i 的相似度，其中 $pf_{i,j}$，$pf_{c,j}$ 分别表示在第 j 个段落中 k_i 和 c 的出现次数：

$$f(c,k)=\sum_{j=1}^{n} pf_{i,j} * pf_{c,j} \tag{2-35}$$

(2) 定义 c 和 q 的相似度：

$$sim(q,c)=\prod_{k_i \in q}(\delta+\frac{\log(f(c,k_i)\times idf_c)}{\log n})^{idf_i} \tag{2-36}$$

在公式 2-36 中 idf_i，idf_c 分别表示基于段落计算的 k_i 和 c 的 idf，其计算方法如公式 2-37 和公式 2-38：

$$idfi=\max(1,\frac{\log_{10} N/np_i}{5}) \tag{2-37}$$

$$idfc=\max(1,\frac{\log_{10} N/np_c}{5}) \tag{2-38}$$

δ 是用于平滑的常数，常去近于 0.1 的值，$sim(q,c)$ 可以看成是利用 TF-IDF 进行相似度计算的一个变种，该方法在不同文档中使用的效果大不相同，需要对参数进行一定调节。

2.2.4 基于 WordNet 的查询扩展方法研究

WordNet 是由普林斯顿大学的心理学家，语言学家等按照一定规则构建的词典，是一个免费的软件包，可以根据构建规则计算词语间的相似性。它是人为构建的词典，但是其词语之间有着各种联系，它的心理语言学假设：

(1) 可分离性假设：将所有语言的每个词语从语言中分离开来在不同语境下去研究。

(2) 可模式化假设：一个人不可能掌握某种语言的所有词汇，除非它能利用词意之间存在的系统模式和关系。

(3) 广泛性假设：计算机对于语言的处理方式想跟人类相同，那么它就需要有着像人类一样丰厚的语言储备。

其相似度计算主要基于包含在层次结构“is-a”中的信息[23]。如果两个词语有着共同的祖先那么显然这两个词语的相似性会比其他词语更高。比如“train”和“bus”有着共同祖先“public transpot”，显然这两者的相关程度会比“ceramic”

这个词语高。

此外，概念之间除了相似度的关联还有其他多种方法，比如：整体与部分关系，正反词义，等，比如椅子腿“leg”是椅子“chair”的一部分，正是不同的连接关系共同构成了 WordNet 的语义网络。

在相似度的计算方法中，论文[35]认为使用 WordNet 在相关性上的判断主要根据词典构造的网络特征，比如某个节点的密度（拥有兄弟节点的个数），节点的深度（在词典网络中的层次）以及节点间的连接方式（上下位、整体部分）等，因此某词语与其父节点词语的距离判断如公式 2-39：

$$wt(c,p)=\left(\beta+(1-\beta)\frac{\bar{E}}{E(p)}\right)\left(\frac{d(p)+1}{d(p)}\right)^{\alpha}T(c,p) \tag{2-39}$$

在公式 2-39 中 $d(p)$ 表示了节点的深度，即节点在词典构成的层次越深表达的意义越细致，与父节点的距离越近。$E(p)$ 是节点的密度，词语含有的兄弟节点越多，密度越密，则其表征的意义也越明显，与父节点的距离也越近。此外，$T(c,p)$ 表示了不同连接方式代表的权值。α 和 β 是调节参数。

根据公式 2-40 本文可以比较任意两个节点的距离：

$$Dist(w_1,w_2)=\sum_{c\in\{path(c_1,c_2)-LSuper(c_1,c_2)\}}wt(c,parent(c)) \tag{2-40}$$

公式 2-40 中 path(c_1,c_2)包含了 c_1 到 c_2 之间最短距离的所有点，而 $LSuper(c_1,c_2)$ 表示了 c_1 和 c_2 最低公共父节点。两个节点距离越近，它们的相关性越高，一般在查询扩展中本文会设置一个阈值，只有当相关性高于某个阈值的词语才会被采用。

WordNet 对于大部分已知词语都作了语义构建，但是对于推文文本，它不能产生良好的效果。首先推文是由用户针对某些事件或者看法而形成的，组成的词语有很大的随意性，甚至针对某些新闻事件产生的新词并不能很快被收入到 WordNet 中。同时即便对于某些词语的近义词在用户的特定需求下将会变成噪音（不符合用户输入的意愿），对查询结果产生负面影响，影响的结果已经经过实验验证，因此本文的查询扩展研究不考虑使用 WordNet。

2.3 推文检索评价

在推文检索中，研究人员认为它在两个方面存在着一些缺陷，首先对于给定测试集的相关性判断存在着很大的主观性，不同人对于是否相关的判断有着自己的标准，此外对于多样的测试集也没能给出一个稳定的比较方法。针对不同测试集和不同的研究角度，检索性能不可能有一个普遍认可的比较方法。

检索结果推文检索结果主要根据是否符合根据用户需求进行判断，这就需要用户对检索结果进行判断。针对以上问题，本文使用了已经经过人工标记好的推文集合作为测试集，如果推文满足查询结果标记为 1，否则标记为 0。

推文检索主要评价方法有[24]：

召回率：召回率(recall) = 系统检索到的相关文件/系统所有的相关文总数,它反应的是检索系统得到结果的完全性，如果尽量多的相关结果被检索到则检索系统的性能越好。

准确率(precision) = 系统检索到的相关文件/系统所有检索到的文件总数：它表示的是检索系统得到结果的准确程度。

$F1$：召回率和准确率往往负相关的，随着召回率的增加，准确率会逐渐降低，反之亦然，因此本文综合考虑这两个指标用 $F1$ 值去表示：

$$F1 = 2PR/(R+P) \tag{2-41}$$

平均准确率：平均准确率(mAP)为了解决单一值表述的局限性，它反映检索系统的一个全局指标。

检索系统的召回率和准确率曲线如图 2-2：

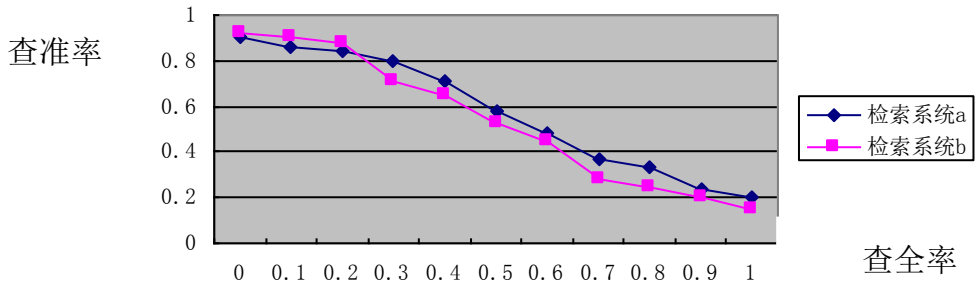

图 2-2 不同检索系统的性能比较

在图 2-2 中的横坐标是查全率，纵坐标是查准率，可以看出检索系统 a 的性能大部分情况下都比检索系统 b 性能好，图中曲线表明，如果检索系统的性能越好，曲线越要上突，曲线与坐标轴围成面积相应的也越大，mAP 的表述内容就是曲线与坐标轴所围成的面积大小。

同时，检索结果的前 n 条推文的准确率通常也被用来评价检索结果，比较普遍的是前 10 条或者前 30 条推文的查准率分别用 P@10(precision at 10)和 P@30(precision at 30)表示。

nDCG：nDCG 是用来衡量排序质量的指标，如公式 2-42：

$$N(n) = Zn\sum_{j=1}^{n}(2^{r(j)} - 1) / \log(1 + j) \tag{2-42}$$

nDCG 的理解：

(1) 与检索需求越相关的推文本文越要将其放在前面使得用户可以一眼就看到它们所需要的结果。

(2) 提供给用户的推文与用户需求越相关越好。

其计算分为两个部分：它是将所有提供给用户的推文权重进行累加，同时，对于推文的排列顺序进行一个权值递减。

2.4 本章小结

本章对推文检索的一些相关技术和方法作了总结，并对在推文检索中广泛应用的 TF-IDF 算法，VSM 算法和 BM25 算法做了具体的介绍。在推文检索中使用的查询扩展方法进行了具体分析，对于相关反馈，全局分析，局部分析等方法的具体流程进行详细说明。此外，对于查询扩展的结果引入了量化结果分析，对几种分析方法做了介绍，比较了它们的应用背景。本文的主要思路就是借鉴了经典的查询扩展方法，并对这些方法做出优化。

第三章 基于推文聚类的查询扩展方法

3.1 研究背景

第二章对查询扩展方法进行了重点的讨论，其意义即在用户输入查询词和推文集合的时候可以有效的检索出满足用户需求的推文。而伪相关反馈查询扩展方法在文本检索中已经有着成熟的运用，如何准确的筛选得到用户所需要的推文是评价检索结果好坏的重要指标。

伪相关反馈不需要人工参与，而是比较查询词和推文的相关性选择相关性高的推文作为待扩展词的语料库，进而比较查询词和待扩展词的相关性，选择新的查询词。这也是伪相关反馈的重要的步骤，现阶段的主要研究内容也是对这两个步骤做优化。

对此，论文[36]认为待扩展词语料库的筛选对推文检索有着很大的影响，因此该论文详细测试了多种不同的伪相关反馈方法在使用人工筛选语料库和传统方法筛选语料库对最终的检索结果产生的影响进行了分析，得出了在相同的查询扩展方法中，加入人工筛选步骤可以有效提升检索结果的结论。

而在现有筛选待扩展词语料库的方法中，论文[8]使用 2.1.4 节的 TF-IDF 算法对推文进行筛选，该方法主要根据查询词在推文中出现的频率以及查询词的权重进行分析的。而论文[9]使用 2.1.3 节中的 BM25 算法对待扩展词语料库进行初次筛选，该方法在 TF-IDF 的算法上引入了不同的调节因子分析了推文长度，查询词在推文中的词频以及查询词的权重对检索结果的影响。论文[4]中使用了向量空间模型算法进行语料库的筛选，该方法根据不同词语在推文和查询中出现的频率将推文和查询词向量化，通过向量间的余弦相似度比较查询词和推文的相关性。

上述方法是筛选查询词与推文相关性的主流方法，但是这类方法存在着严重的问题：

根据查询词在推文中出现的频率来筛选推文并不能有效的得到满足用户需求的结果，如果与用户相关的推文中所含有的查询词较少甚至不存在查询词，显然使用上述方法会将这类推文遗漏。这也导致了许多与用户查询相关的推文集合被过滤，进而导致语料库中含有的待扩展词存在缺失。

例如在某个话题中，含有查询词“Pakistan，diplomat，murder，arrest”，它表示了美国外交官在巴基斯坦枪杀两人被巴基斯坦政府逮捕的事件，如表 3-1：

表 3-1 含有少量查询词的相关推文

Prosecutor says US consular employee suspected in Pakistan shooting deaths will face charge
Pakistani judge orders American held more days LAHORE　Pakistan - A court on Thursday ordered the de gvnc
Pakistan court refuses to release US official denying immunity
BBC News - Lahore deaths accused is diplomat must be freed - US
Lahore Shooting Pakistan Refuses to Release American Who Killed
US diplomat killed intelligence operatives in Lahore shootout
US official Raymond Davis on Lahore murder charges

含有下划线的词语表示原始查询词，表 3-1 中的推文含有查询词较少。而在使用传统方法对推文进行筛选时，主要根据查询词在推文中的词频来衡量推文和用户查询的相关性。这样导致了许多像表 3-1 中的不存在或者存在少量原始查询词但是与用户相关的推文被漏选，而实际上表中的推文都是满足用户需求的。

针对传统的语料库筛选方法存在的含有较少查询词的相关推文不被选取的缺点，本文提出了基于推文聚类的查询扩展方法。

该方法将具有相同语义的推文进行聚类，筛选与用户查询最相关的推文类作为语料库，将传统的逐条推文筛选改为逐类推文筛选。如果含有较少查询词的相关推文存在于被选取的类中，这些推文也可以被筛选。该方法有效解决了与用户相关的推文由于含有查询词较少而被遗漏的问题。

3.2 推文聚类查询扩展方法

正如 3.1 节所述，传统方法对待扩展词语料库的筛选存问题。而基于传统方法筛选待扩展词语料库中存在含有较少查询词的推文会被漏选这一问题，本文提出了基于推文聚类的查询扩展方法，并对该方法的流程进行设计和实现。

推文聚类的查询扩展方法可以有效的对传统方法存在的含有较少查询词的相关推文被漏选的问题进行解决。它主要在传统方法基础上加入推文聚类以及与用户相关的推文类被筛选的过程使得含有较少查询词的推文随着存在该推文的推文类被筛选而被选取，使得得到的反馈语料库中含有查询词较少的相关推文。

推文聚类的查询扩展方法流程图如图 3-1：

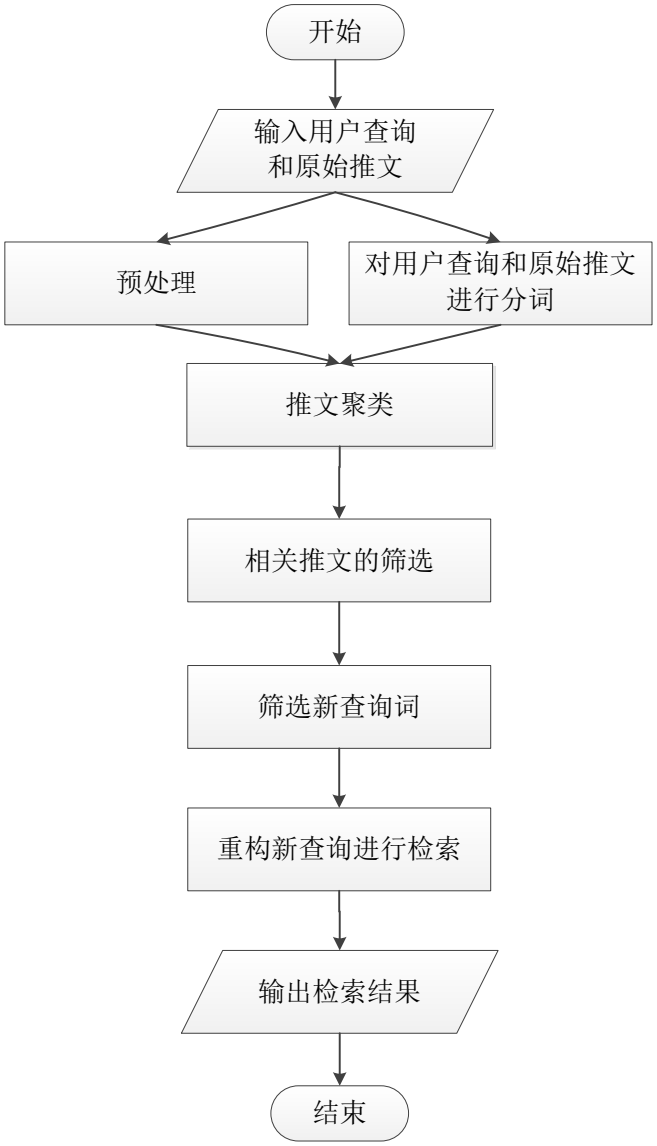

图 3-1 基于推文聚类的查询扩展方法流程图

图 3-1 表示的是基于推文聚类的查询扩展方法流程，它包含了推文文本处理，推文聚类和查询扩展三个步骤。

推文文本处理是一个重要的步骤，由于推文特性，大量推文都是不规范的，因此本文对输入的推文文本和原始查询进行处理，它包含了推文正规化，去除重复推文和多联词的提取等步骤。

正规化的主要任务是根据需要对推文文本进行进一步去噪工作，如去掉与推特文本语义无关的字符，同时提取 URL，用户名，hashtags 等关键信息。去除重复推文也是一个重要的步骤，数据库中含有大量重复的推文，有部分是一模一样的，而有些是转发的推文(含 RT)。本文把经过正规化处理的推文映射为一个 64

位的 CRC 循环校验码，当新读入一条推文时，通过计算该新推文 CRC 校验码，检查是否出现过，从而判断该新推文是否重复。多联词提取的主要任务是从推文中提取出名词性短语。在实际应用中，本文对首字母为大写的相邻 2~3 个单词进行提取。多联词提取的主要思想：首先从推文文本中提取出首字母为大写的连续两个或三个单词，将其作为一个最小分词单元，放入到已建立的多联词词库中，定期检查筛选该词库中的多联词，以保留较高词频的词（词频阈值设为 0.02%），接着以该多联词词库为依据，从推文中划分出这些多联词作为一个分词单元。

此外分词也是一个重要工作，它除了使用 NLTK 中所提供的方法，针对不同步骤使用正则匹配方法进行筛选，对于缩略词还需要进行还原。分词完后存在一些没有意义的词语如：“a”，“an”，“the”，这类词语不表示实际意义，在对推文的后续处理中，这类被称为停止词的词语将对结果产生错误的影响，故将其去除。

推文聚类是对普通伪相关反馈查询扩展方法的改进。它将具有相同语义信息的推文进行聚类，当类中的某些推文由于用户描述的原因含有查询词较少，它也会因为该类中的其他推文含有较多查询词而被选取。通过该过程可以有效的筛选出满足用户需求的推文类作为待扩展词的语料库。它包含了对所有推文进行聚类和用户最相关推文类的选取两个步骤。

通过得到的语料库可以使用伪相关反馈进行推文检索，即通过比较查询词和待扩展词的相关性筛选新查询词，将新查询词加入到原始查询构成新查询对推文进行检索。查询扩展包含了新查询词的筛选和对推文进行重新检索两个步骤。

本文下面对推文聚类在查询扩展方法上的应用和查询扩展方法做具体介绍。

3.3 推文聚类在查询扩展中的应用

推文聚类在查询扩展中进行应用可以更好的得到扩展词的语料库。聚完类后的结果包含了相同语义的推文类，如果含有较少查询词的推文存在该类中，而该类与用户查询的相关性大而被选取，含有较少查询词的推文也会被筛选。该过程首先对推文进行聚类，使推文类中存在的是含有相同语义的推文，然后通过比较查询词和推文类的相关性选取作为待扩展词语料库的推文类。

3.3.1 推文聚类

推文聚类算法常见的有 k-means 聚类，层次聚类，基于密度的聚类和基于网格的聚类。本文的聚类方法不是研究重点，而是通过聚类优化查询扩展的流程，因此采用被广泛使用的 k-means 聚类，它需要所有推文将文本形式转换为数学模

型。

k-means 算法是一种以距离作为相似度评判依据的聚类算法，聚类对象间的距离越近则认为其相似度越高，最终的聚类结果是相似度最高的聚类对象聚成 k 个类。

该算法首先随机选取 k 个聚类中心（质心），计算每个聚类对象与各个质心的距离，选取距离最近的质心作为该聚类对象的簇，这是一次迭代过程。然后通过各质心与各个聚类对象的距离调整质心，计算各个聚类对象与新质心的距离，直到新质心与原质心的距离差值在阈值范围内，便是最终的聚类结果。而初始选取的聚类个数会对聚类结果产生很大的影响。

推文聚类算法流程如下：

(1) 从 N 条推文中随机选取 k 条推文作为质心。

(2) 计算剩余所有推文到质心的距离，将推文划归到与其距离最近的质心的类。

(3) 计算并调整各个质心的位置。

(4) 迭代(2)(3)两步，如果新质心与原质心的距离在阈值内则结束流程。

最理想的聚类结果需要有以下两个特点[31]：

(1) 每个类中的推文需要尽量紧凑。

(2) 不同的类之间需要尽量分开。

如果每条推文中的各个单词的重要程度相似，我就可以说这些推文是相似的。本文使用欧几里德距离(Euclidean Metric)作为推文间相似程度的评判依据，欧几里得距离的定义如公式 3-1：

$$d(X,Y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+...+(x_n-y_n)^2} \tag{3-1}$$

公式 3-1 表示 X,Y 两个元素在欧几里德空间中的集合距离，而本文将 X,Y 分别理解为两条推文，x_i 和 y_i 是两条推文中单词的 TD-IDF 值（2.1.4 节）。这样就可以通过欧式距离计算出两条推文的相似度了。在实际运算中，如何整个推文集合中的词语词频小于 5 则认为这些词语是没有意义的，并将这些低频词过滤。

在重新计算调整质心时将属于该质心的推文向量的各维取平均。

实际应用中考虑 3 个迭代停止的条件：

(1) 比较连续迭代两次的质心距离，如果所有两次迭代质心点的距离差之和小于阈值则算法终止。

(2) 如果 k-means 聚类达到收敛时间过长则需要设置最大迭代次数，如果迭代最大迭代次数后还不收敛则算法终止。

(3) 比较连续迭代两次的质心，所有质心都没有变化则算法终止，这种约束条件非常严苛，很少有满足这种条件的情况。

k-means 文本聚类需要将推文尽量分散，k 值不能取值过小，否则聚类中可能存在将多个类进行合并成一个类，这在类的筛选过程中就会影响筛选的结果。

在实际应用时，本文对 k 值的选取进行测试，k 从 10 依次以差为 5 逐渐增加。

3.3.2 相关推文的筛选

相关推文筛选使用 2.2 节中的 TF-IDF 算法通过比较经过聚类的推文和查询词的相关性得到最满足用户需求的推文类。

TF 表示的是查询词在推文集合中出现的频率，频率越高它对推文集合越重要，其计算方法如公式 3-2：

$$TF(q_i,d)=\frac{f_i}{F} \tag{3-2}$$

f_i 表示查询词在推文类中出现的频率， F 表示推文类中词语的个数。

IDF 是逆文档频率，它表达了如果一个词语在各种语料库中是经常出现的，那么它的特征就不是很明显就不能特定的去描绘某个推文集合，其重要性就会下降。

在实际应用中，对于每个查询词权重 W_i，本文采用 IDF 去判断词与推文的相关性权重：

$$IDF(q_i)=\log(\frac{N-n(q_i)+0.5}{n(q_i)+0.5}) \tag{3-3}$$

在公式 3-3 中 N 表示推文集合，$n(q_i)$ 表示了含有词语 q_i 的推文数。它表示了 q_i 的推文越多，那么 q_i 的权重越低。即对于经常出现的词语，它对于推文的区分度会变得不明显，进而它来判断推文的权值就越不重要。

显然使用 TF-IDF 可以很好的判断查询词与推文集合的相关性，如公式 3-4：

$$TF-IDF=TF\times IDF \tag{3-4}$$

而在每个查询词和推文类的比较中，本文将一个推文类看成一个文档，计算类中查询词的 TF，然后在整个推文集合中计算查询词的 IDF，最后求取查询词和推文集合的 TF-IDF 值。

通过计算每个推文类对所有查询词的 TF-IDF 值的和，作为该推文类与原始查询的相关性，对相关性降序排列，筛选推文集合直到推文集合中的推文数大于 30，得到的结果作为待扩展词的语料库。

为了避免少数推文不与其他推文语义相似，在结果中一条推文作为一个类，而其又含有较多的查询词，故对推文条数少于 4 的推文类进行忽略。

3.4 查询扩展方法

在得到满足用户需求的推文语料库的前提下，本文通过得到的语料库使用伪相关反馈方法进行推文检索，该方法首先比较原始查询和推文的相关性，选取相关性最高的词语作为新查询词，将新查询词加入到原始查询得到新查询，进而对推文进行检索。

查询扩展包含了新查询词的筛选和对推文进行重新检索两个步骤。

3.4.1 筛选新查询词

本文使用的伪相关反馈模型建立于信息检索中的语言处理框架上，此模型认为所有文档按一个后验概率 $P(D|Q)$ 排序，根据贝叶斯规则它可以表示为公式 3-5：

$$P(D|Q) \propto P(Q|D)P(D) \tag{3-5}$$

$P(Q|D)$ 是文档与查询词的相似度，$P(D)$ 是文档 D 与任意查询相关的先验概率。用词频信息来表征文档，本文使用了 uni-gram 模型，将查询词和文档的相关性表示为：

$$P(Q|D) = \prod_{i=1}^{|Q|} P(w_i|D) \tag{3-6}$$

公式 3-6 中 $|Q|$ 是查询中的词语个数，$P(w_i|D)$ 是查询中每一个查询词语 w 对于文档 D 的概率分布，对于短文本处理时，词语在文本中的出现非常有限，本文使用狄利克雷平滑[27]：

$$P(w|D) = \frac{|D|}{|D|+\mu} P_{ml}(w|D) + \frac{\mu}{|D|+\mu} P(w|C) \tag{3-7}$$

公式 3-7 中 $Pml(w|D) = \frac{c(w;D)}{\sum_{w' \in V} c(w';D)}$，$c(w;D)$ 表示文档 D 中 w 出现的次数，C 是一个文档集合，μ 是狄利克雷平滑因子。

词语 w 与原始查询的相关性 $P(w|D)$ 分为两部分[28]：

(1) 文本逻辑相关：即待扩展词与原始查询与推文信息文本相关。

(2) 时序信息相关：即原始查询与推文与推文的发推时间相关。

$P(w|D)$表示如公式 3-8：

$$P(w|Q)=\sum_{D\in R}P(w,D|Q)=\sum_{D\in R}P(w,D_w,D_t|Q)=\sum_{D\in R}P(w,D_w|D_t,Q)P(D_t|Q) \tag{3-8}$$

本文假定时序相关D_t与文本信息D_w是独立的，根据公式 3-8，本文先只考虑文本逻辑相关部分，此外，本文还考虑给定查询包含了所有的查询词。本文得到以下公式：

$$P(w|Q)=\sum_{D\in R}P(w,D_w|Q)\propto\sum_{D\in R}P(w|D_w)P(Q|D_w)P(D_w)=\sum_{D\in R}P(w|D_w)\prod_{i}^{|Q|}P(q_i|D_w) \tag{3-9}$$

$P(w|D_w)P(Q|D_w)P(D_w)$是标准相关反馈模型的一个因子。本文认为这是一个逻辑模型。公式 3-9 中，本文假定先验概率$P(D_w)$是一个归一化结果。实际意义便是衡量待扩展词w与查询Q中的每一个查询词以及反馈回的每一条推文信息的相关程度。结合公式和本文的认知情况，待扩展词与原始查询的相关性与在每条推文中，待扩展词在推文中出现的频率以及原始查询中每个查询词出现的频率有关。

此外，本文还认为在一条推文中，两个词越临近，则它们具有更大的相关[29]，主要原因是构成句子的词语一般由一定的短语组成，相邻的词语很大程度上是一个修饰的过程。因此，对于每一条推文，本文认为其他待扩展词与原始查询词的相关性随距离的增加递减。

在文本逻辑中，原始查询和待扩展词的表示形式为：

$$P(w|Q)=e^{-\lambda d}\sum_{D\in R}P(w|D_w)\prod_{i}^{|Q|}P(q_i|D_w) \tag{3-10}$$

公式 3-10 中的$e^{-\lambda d}$表明随着两个词语间距离的增加，其相关性逐渐减少，λ是一个衰减因子，它调控了距离增加对相关性减少的影响程度，d是待扩展词与全部查询词的平均距离。

d的计算方法如公式 3-11：

$$d=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{d_{ij}}{nm} \tag{3-11}$$

公式 3-11 中的 d 是待扩展词与原查询词语的平均距离，d_{ij} 是某个待扩展词与一个查询词的距离，同时由于待扩展词可能同时在推文中出现，则 m 表示待扩展词出现的次数，同时由于原始查询也存在着多个查询词， n 则表示原始查询中查询词的个数。平均距离可以消除待扩展词与单个原始查询词相近的问题，它衡量了待扩展词与整个查询的距离。

同时本文还引用了时序模型作为获取新查询词的依据，其主要思想认为查询时间对检索结果存在影响。用户的查询时间的某个时间区间内，与用户相关的推文越多，表示该事件越重要，其对检索结果影响越大。

公式 3-10 中的 $P(w|Q)$ 中的 $P(D_t|Q)$ 来表征时序模型。在时序模型中，对于给定的查询 $P(t|Q)$，计算方法如公式 3-12：

$$P(t|Q)=\frac{1}{Z}\sum_{D\in R}P(t|D)P(Q|D) \tag{3-12}$$

在公式 3-12 中，如果推文 D 的发推时间和时间 t 的差值在一定的时间范围内，那么 $P(t|D)$=1，否则 $P(t|D)$=0，而 $P(Q|D)$ 表示原始查询 Q 和推文 D 的相关性。它表示了时间和查询词的关联性，集中在用户查询时间的与查询相关的推文越多，其值越大。

根据公式 3-12 本文将查询 Q 和推文 D 的时序模型分别表示为 $P(t|Q)$ 和 $P(t|Q_D)$，Q_D 即是对推文 D 的一个伪查询，它表示对推文 D 的查询词刚好是由推文 D 中的词语构成[34]。

$P(D_t|Q)$ 的计算方法如公式 3-13：

$$P(Dt|Q)=e^{-\lambda d} \tag{3-13}$$

公式 3-13 中的时序模型由查询 Q 和推文 D 的相关性由变量 λ 和 d 表征，它是一个指数分布。

λ 表征了如果在时间区间内的相关推文数量越多则其值就越大，那么这些推文就是就是越在近期发布的， $P(D_t|Q)$ 将会变化的很快，反之如果推文都不是查询者近期所需要的，那么 $P(D_t|Q)$ 的变化将会很迟缓。

公式 3-13 中的另一个参数 d 形容了 $P(t|Q)$ 和 $P(t|Q_D)$ 两个时序模型的距离，距离越近说明两个模型越近似 $P(D_t|Q)$ 也就越大。本文是用巴塔恰里雅距离（Bhattacharyya Distance）去衡量两个时序模型[30]，其计算方法如公式 3-14：

$$d=-\ln \mathrm{B}(Q,D) \tag{3-14}$$

公式 3-14 中的 $\mathrm{B}(Q,D)$ 是巴氏系数，这个值越大则距离 d 越小，其计算方法如公式 3-15：

$$\mathrm{B}(Q,D)=\sum_{t\in\tau}\sqrt{P(t\mid Q)P(t\mid Q_D)} \tag{3-15}$$

归根结底本文使用了巴氏系数去衡量两个时序模型的相似程度，当且仅当两个时序模型完全相同时，巴氏系数最大，两个时序模型最相近。

时序模型的最终结论就是当 $P(t\mid Q)$ 和 $P(t\mid Q_D)$ 越接近，$P(D_t\mid Q)$ 就越大，当推文越是满足用户的时间需求，$P(D_t\mid Q)$ 的变化速度就越快，时序模型对最终结果的影响就越明显。

结合使用文本逻辑模型和时序模型，本文对语料库中的待扩展词语和输入的原始查询计算其相关性并降序排列，选取相关性最大的词语作为原始查询的扩展词重构新的查询。

3.4.2 重新检索

对于筛选出来的新查询词，本文选取其个数与原始查询词的个数比为 1:1，同时新加入的查询词权值如公式 3-16：

$$w_{v,q}=\frac{sim(q,v)}{\sum_{u\in q}w_{u,q}} \tag{3-16}$$

公式 3-16 中的 $w_{v,q}$ 表示新加入的查询词 v 的权重，$sim(q,v)$ 表示原始查询 q 与 v 的相关性，$w_{u,q}$ 表示原始查询 q 中的每个查询词的权重。

将带有权值的新查询词加入到原始查询中构建新查询，使用 BM25 算法对推文集合进行重新检索。

3.5 测试分析

对于本章的测试结果，本文从测试集，推文聚类结果分析和检索结果分析三方面进行展示。

3.5.1 测试集

本文采用了 http://trec.nist.gov.data/tweets/网站上的 tweets2011 的数据作为测试集进行实验，其数据使用 mongodb 数据库进行存储，其存储格式如图：

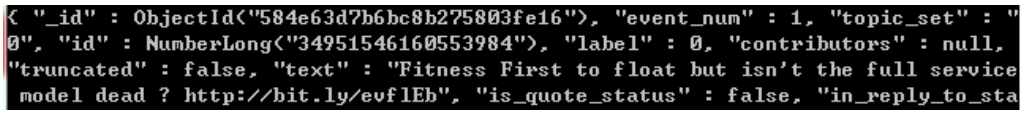

图 3-2 tweets2011 的测试集合

图 3-2 中“topic_set”表明推文属于那个话题，“id”是推文 id，“label”是推文的标记，“create_at”是为推文创建的时间。

测试集分为 TREC2011 和 TREC2012，tweets2011 一共含有 108 个可用话题，1 千多万条推文，每个话题与推文的相关性都已经进行了人工标记，话题与推文强相关推文被标记为“2”，弱相关标记为“1”，不相关则标记为“0”。每个话题的表示如图 3-3：

```
<top>
<num> Number: MB003 </num>
<title> Haiti Aristide return </title>
<querytime> Tue Feb 08 21:32:13 +0000 2011 </querytime>
<querytweettime> 35088534306033665 </querytweettime>
</top>
```

图 3-3 测试集中的话题

图3-3表示了测试集中对于话题MB003给出的话题查询词以及给出查询时的时间。本文使用强弱两种相关的推文都认为是相关的进行测试。此外，本文分别使用 TREC2011 和 TREC2012 中的所有话题进行测试。

所有的量化测试环节都对每个话题测试了 50 次取平均值，并对所有话题进行测试再取平均值进行比较。

3.5.2 推文聚类测试分析

本文得到对于推文聚类的结果，本节将聚完类的推文集合进行比较分析，去判断聚类结果的好坏。

仍然使用图 3-2 中的测试集，将测试集中的 MB003 话题进行测试，话题中包含了“Haiti，Aristide，return”三个词语作为查询词，用户使用上述三个词语进行查询，这三个词语表征了海地总统 Aristide 返回海地这个事件，MB003 话题下的部分推文为：“Donate for the Haiti Relief Effort through the Red Cross Text HAITI to，Haiti to give Aristide passport Officials in Haiti say they are ready to issue ex-president Jean-Bertrand Arist”。

由于推文发布存在随意性，集合中存在由于拼写错误或者表述不明确等与聚类结果不相关的低频词，因此本文将词频小于 5 的低词频词语进行过滤。本文认

为组成不同推文的词语对于这些推文有着相近的重要程度则这些推文语义相近。

例如推文“Haiti OKs giving ex-President Aristide passport to return from exile – AFP”和“Haiti s former president Jean-Bertrand Aristide vows to return”都存在词语“Haiti”，“Aristide”和“return”，而这些词语在推文中有着相近的 TF-IDF 值，推文间的欧几里德距离较小，推文语义相近。

在 k-means 聚类中 k 值的选取会对结果产生很大影响，本文需要使聚类结果中的推文集合所包含的推文都描述着同一个话题，因此 k 值选取不能过小，本文分别使 k 值从 10 以 5 的差值逐渐增加进行测试。

在测试中 k 值为 20，通过 k-means 聚类得到的结果如表 3-2：

表 3-2 k 为 20 时的聚类结果

Cluster14:
Seattle WA Times Haiti Aristide can have passport　hasn t applied
Haitian Politics - Millions Reasons Why Duvalier Came Back To Haiti　Haitian want their money back we afraid
Tell Middle Class Haitians the worst living Haitian political assassin Duvalier is free in Haiti they say What about Aristide
If Duvalier Can Return to Haiti Why Can t Aristide New America Media haiti
Haiti – Aristide His return an international affair
Haitian Politics AFRAID TO FACE HAITI S JUDICIAL SYSTEM ARISTIDE FABRICATES A PASSPORT ISSUE　PRESS RELEASE OF
Haiti - Aristide Passport small administrative problem Jean Bertrand Aristide had written last Monday
Haitian Politics – With Duvalier In Haiti Is Rene Preval AFRAID Map mand tete moin pouqui a Americain yo ac

表 3-2 中的“Haitian Politics - Millions Reasons Why Duvalier Came Back To Haiti　Haitian want their money back we afraid”，“Tell Middle Class Haitians the worst living Haitian political assassin Duvalier is free in Haiti they say What about Aristide”，“Haitian Politics – With Duvalier In Haiti Is Rene Preval AFRAID Map mand tete moin pouqui a Americain yo ac”这 3 条推文表达的是海地前总统 Duvalier 流亡国外的事件。但是其他推文表示的是流亡总统 Aristide 返回海地的事件，这才是满足话题的推文。

因此在推文聚类 k 取值比较少时，类的区分度不明显导致聚类结果不准确。而在使用查询扩展方法时，待扩展词的语料库的准确性是首要的，k 值的选取需

要在满足准确性的前提。

为使聚类结果中的准确率更高，本文需要适当将 k 值以 5 的差值逐渐增加，使得不同类之间的区分更加明显，将一些会对结果产生影响的推文从该类中分离出去，如此有利于后面推文的筛选。随着 k 值的增加经过测试发现 k 值为 40 时，会得到一个比较好的效果。K 值为 40 时得到的聚类结果如表 3-3。

表 3-3 k 为 40 时 cluster38 的聚类结果

Cluster38：
Haiti OKs giving ex-President Aristide passport to return from exile - AFP
Haiti s former president Jean-Bertrand Aristide vows to return
MIAMI Haiti to issue ex-president Aristide with passport clearing way for him to return
Haiti to give Aristide passport Officials in Haiti say they are ready to issue ex-president Jean-Bertrand Arist
Haiti opens door for return of ex-president Aristide
Haiti Aristide His return an international affair Haitilibre com haiti
Haiti allows ex-president Aristide s return from Can Haitian politics get any more interesting
Haiti issues passport to ex-president Jean-Bertrand Aristide enabling him to end his exile - SunSentinel

表 3-3 是 k 为 40 时类 cluster38 中的推文，而这些推文全部与话题 MB003 相关。通过测试分析本文发现对于特征明显的类，当 k 值越高其准确率可以达到 100%，即整个类的推文都是在表述同一个话题。这也表明了在 k-means 聚类中 k 值的选取十分重要，其与该话题下所含有的相同语义集合的个数有关。而对各话题进行测试，k 为 40 时可以得到一个理想的结果。

本文使用图 3-2 表示的数据作为测试集，使用测试集中的 MB007 话题进行测试，该话题包含了“Pakistan，diplomat，arrest，murder”这些词语作为查询词，表示的是美国外交官涉嫌在巴基斯坦用枪谋杀了两个人这个事件。本文通过 BM25 算法比较查询词与每条推文的相关性，得到的排名前 30 的部分推文结果如表 3-4。

表 3-4 是通过 BM25 算法筛选得到的部分推文，下划线表示的是原始查询词。而通过该方法筛选得到的推文都含有较多的查询词，如果与话题相关的推文含有查询词过少则被错误的遗漏。本文使用推文聚类对推文集合进行筛选，分析对比未经聚类和经过聚类得到结果的差异。

本文已经对推文进行了聚类，然后需要对聚类好的推文集合进行筛选，将满足用户需求的推文集合进行反馈，主要采用 TF-IDF 算法。

表 3-4 BM25 算法筛选得到的推文

DTN India US mounts pressure on Pakistan to release illegally detained murder accused diplomat Islamabad
Albanian police arrest in alleged murder plan
Uh oh American diplomat charged with double murder in Pakistan He should have just called in a drone
ISLAMABAD Pakistan AP -- US demands release of diplomat in Pakistan
India Threatens to Cancel Pakistan Meeting over held diplomat we ll see WSJ
Pakistan da diplomat krizi Lahor kentinde rev yapan Amerikal bir diplomat n kendisine sald ran iki ki iyi vu
Pakistan warns US over diplomat s release Growing US demands to free an American official who shot dead two men
Pakistan Officials Move To Arrest Four Officials Of The U S Consulate In Lahore islam islamophobia
Guard Mexican police arrest gang leader blamed for Acapulco killings Jose Lozano behind murder earlier in month

表 3-5 是使用 TF-IDF 对所有聚好类的推文筛选得到的两个类，经过分析本文发现这两个类中的 8 条推文都是与用户查询相关的。但是表中的“Prosecutor says US consular employee suspected in Pakistan shooting deaths will face charge -AP”和“Pakistani judge orders American held more days LAHORE Pakistan - A court on Thursday ordered the de”这两条推文由于含有原始查询词较少在使用 BM25 算法时不能被直接检索。在使用了推文聚类方法后，可以直观的发现含有少量查询词的相关推文随着整个类被检索出来而被筛选。这类推文中含有重要的与用户需求相关的词语，上述两条推文中的下划线词语很好的描述了话题中的事件，有着很高的作为待扩展词的价值。

同时本文还分析了由于被标记为“1”而被认为相关的推文集合使用推文聚类筛选得到的检索结果。不使用推文聚类时，这些被人工标记的推文被随机选取 30 条作为待扩展词的语料库，实际上这种随机选取方法存在缺点。

例如图 3-2 测试集中的 BM019 话题含有的“Cuomo， budget， cuts”查询词表达了纽约州州长科莫对财政支出的削减这个事件，其中有部分被标记为“1”的推文表示如表 3-6。

从表 3-6 中本文发现许多被标记为“1”的推文是推文发表者对某些网站进行了引用，然后自己做的一些评价，由于这些推文只是一些评价性的语言，含有的查询词很少，其推文内容看起来并不与事件内容那么相关。先经过聚类，再使用

TF-IDF 算法筛选可以将这些推文进行过滤。

表 3-5 筛选后的聚类结果

Cluster21:
US official in Pakistan to face murder charge Pakistan will pursue murder charges against a US consular emplo
Prosecutor says US consular employee suspected in Pakistan shooting deaths will face charge -AP
US official in Pakistan to face murder charge AP - AP - Pakistan will pursue murder charges against a US co
DTN World News US official in Pakistan to face murder charges Pakistan will pursue murder charges against a

Cluster27:
Pakistani judge orders American held more days LAHORE Pakistan - A court on Thursday ordered the de
Pakistan court extends detention of US diplomat LAHORE Pakistan Reuters - An American who killed two Pakis
Pakistan court refuses to release US official denying immunity
Overload News Pakistan court extends detention of US diplomat - Reuters DAWN Pakistan court extends detent

表 3-6 部分被标记为“1”的推文

Read This: Why spending's always up http://bit.ly/g2ga2f www.hvagents.info
NYGovCuomo recognizes that reduced state spending is key to solving NY's fiscal crisis http://nyp.st/f7TVD1
Hey I found this! For mid-Hudson's public employees, no more job security: McLaughlin could ... http://bit.ly/huseBc I hope this helped!
NY Times: Taxing Our Patience: "Wealthy families?" Are these folks kidding? Those earning in the $200000+ range ... http://bit.ly/gjeEdu

3.5.3 推文检索测试分析

本文主要对查询扩展方法中的语料库筛选步骤使用推文聚类方法进行创新。在 2.4 节中，本文描述了传统推文检索方法，它们都是将原始查询与推文集合的每条推文进行了相关性的比较，然后将与原始查询最相关的推文筛选出来作为语料库，本节通过比较了通过 BM25 算法以及推文聚类算法得到的反馈结果对最终

的检索结果产生的影响。

使用 k-means 聚类与使用 BM25 算法，TF-IDF 算法和 VSM 算法相比，其反馈的推文准确率（被标记为 1 的推文的概率）更高，结果如表 3-7：

表 3-7 反馈结果比较

k-means 聚类方法准确率	BM25 算法准确率	VSM 算法准确率	TF-IDF 算法准确率
0.825	0.738	0.719	0.711

对于测试集中的各个话题，本文使用推文聚类方法，BM25 算法和经典的 VSM 算法，TF-IDF 算法对推文进行筛选做测试，将筛选得到的前 30 条推文作为待扩展词的语料库。测试集中 TREC2011 的检索结果如表 3-8。

表 3-8 中的 RM(Retrieval Model)，QDRM(Query-Document Dependent Temporal Relevance Model)分别表示的是只使用逻辑模型的检索结果和结合使用逻辑模型和时序模型得到的检索结果，其对语料库的筛选方法是 BM25 算法，而 VSM，TF-IDF 分别对应两种经典的语料库筛选方法，Cluster 表示使用推文聚类对语料库进行筛选。

表 3-8 未使用人工筛选的检索结果

评价指标 / 方法	mAP	P@10	P@30
RM	0.394	0.382	0.335
QDRM	0.417	0.423	0.394
VSM+RM	0.382	0.369	0.323
VSM+QDRM	0.405	0.407	0.378
TF-IDF+RM	0.379	0.372	0.312
TF-IDF+QDRM	0.411	0.413	0.383
Cluster+RM	0.439	0.446	0.398
Cluster+QDRM	0.467	0.512	0.424

为能清晰表现推文聚类方法的效果，对 mAP 检索指标进行分析如图 3-4。图 3-4 表示的是在未对推文集合进行人工筛选时，在 RM 和 QDRM 两种筛选新查询词的方法下，使用推文聚类方法与传统的 BM25 算法，VSM 算法以及 TF-IDF 算

法在 mAP 检索指标下的结果对比。通过分析发现使用推文聚类方法在 RM 和 QDRM 方法下 mAP 检索指标比 BM25 算法提升了 11.4%，12%，比 VSM 算法提升了 14.9%，15.3%，比 TF-IDF 算法提升了 15.8%，13.7%。推文聚类方法比较传统语料库筛选方法在检索结果上都有着明显的提升。

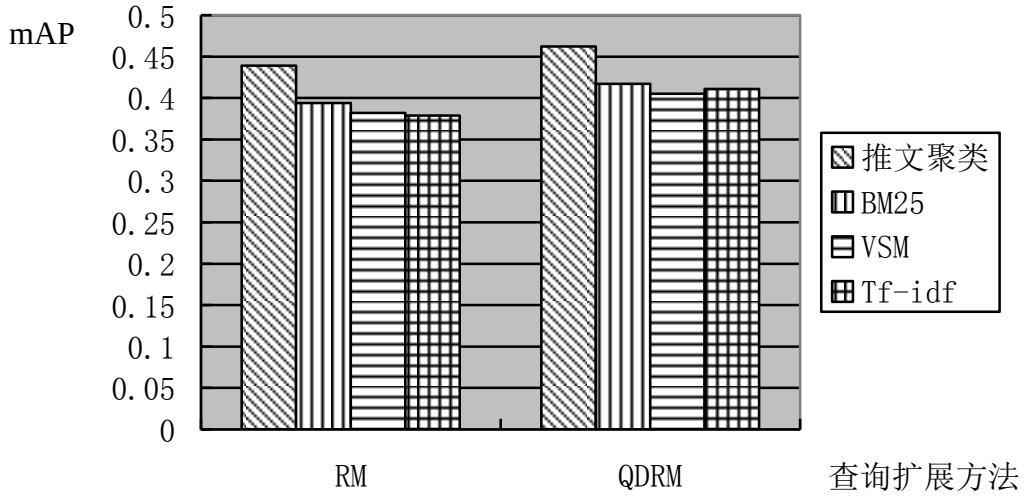

图 3-4 推文聚类对检索结果的影响

表 3-6 表明了对经过人工筛选的推文也存在包含的词语与原始查询不相关的问题，可能会对检索结果产生影响。现对经过人工筛选的推文进行聚类以及 TF-IDF 算法的筛选其检索结果如表 3-9。表中清晰表示了使用人工筛选推文做语料库时，推文聚类也能有效的提升检索结果。

使用人工筛选作为语料库时，使用推文聚类方法在 RM 和 QDRM 两种查询扩展方法上分别提升了 5.3%和 4.5%。

表 3-9 使用人工筛选推文作为语料库时的检索结果

评价指标 / 方法	mAP	P@10	P@30
RM	0.493	0.644	0.548
QDRM	0.521	0.669	0.567
Cluster+RM	0.513	0.658	0.554
Cluser+QDRM	0.536	0.677	0.573

3.6 本章小结

由于传统查询扩展方法存在筛选语料库时含有较少查询词的相关推文被错误的漏选的问题，本章提出了基于推文聚类的查询扩展方法，并对该方法的模型和流程做了详细的介绍。本章具体分析了聚类结果以及通过聚类得到的待扩展词语料库的优越性。最后本章还对比了推文聚类方法与 BM25 算法，VSM 算法和 TF-IDF 算法在两种不同的查询扩展方法上的检索指标，发现通过推文聚类可以有效的提升检索结果。

第四章 基于主题划分的查询扩展方法

4.1 研究背景

通过第三章对推文进行聚类得到了与原始查询最为相关的推文集合作为查询扩展的语料库。通过类的筛选，它找出了某些含有查询词较少但是与用户查询相关的推文，有效的解决了含有较少查询词的相关推文被遗漏的问题。

如果聚类结果中包含有较多的查询词，通过比较原始查询和推文的相关性，该类会被检索出来，但是如果该类只是含有查询词但是并不与用户相关，检索出的结果就产生了主题偏移（主题不满足用户需求）。这类推文作为待扩展词的语料库时对检索结果会产生负面影响。

而传统的 TF-IDF 算法，VSM 算法以及 BM25 算法对推文进行检索的时候也共同存在这样的问题，如果推文中含有较多的查询词，该推文不与用户相关也会被错误的筛选出来。

例如某话题下的部分推文如表 4-1，它包含了“Pakistan，diplomat，murder，arrest”等词语作为查询词。

表 4-1 含有多个查询词的不相关推文

US embassy cables Pakistan committed to defending conviction of Omar Saeed Sheikh for Daniel Pearl murder via
Murder arrest after body is found A man is arrested on suspicion of murder after police discover the body of a
Guard Mexican police arrest gang leader blamed for Acapulco killings Jose Lozano behind murder earlier in month
Mass murder suspect arrest in Acapulco

表 4-1 含有下划线的词语是原始查询词，表中推文含有较多的查询词，但是描述的并不是用户所需的推文，使用推文聚类也会错误的将其筛选。把此类推文作为语料库将严重影响检索结果和质量。因此本文对每条推文进行主题归类[33]，只筛选出符合用户需求主题的推文作为待扩展词的语料库。

两个主题的查询词存在交叉时得到满足用户需求的主题，当主题明确的时候，构成该主题的关键词也是基本确定的，通过用户输入的查询词找到满足用户需求的主题，进而筛选出与该主题最为相关的推文。

4.2 主题划分查询扩展方法

基于含有较多查询词的不相关推文被错误的筛选作为语料库这一问题，本文提出了基于主题划分的查询扩展方法，并对其流程进行设计和实现。在初次筛选推文作为语料库时，使用满足用户主题的推文作为待扩展词的语料库进行反馈，有效避免了含有较多查询词的推文被错误筛选这一问题。

该方法的流程如图 4-1：

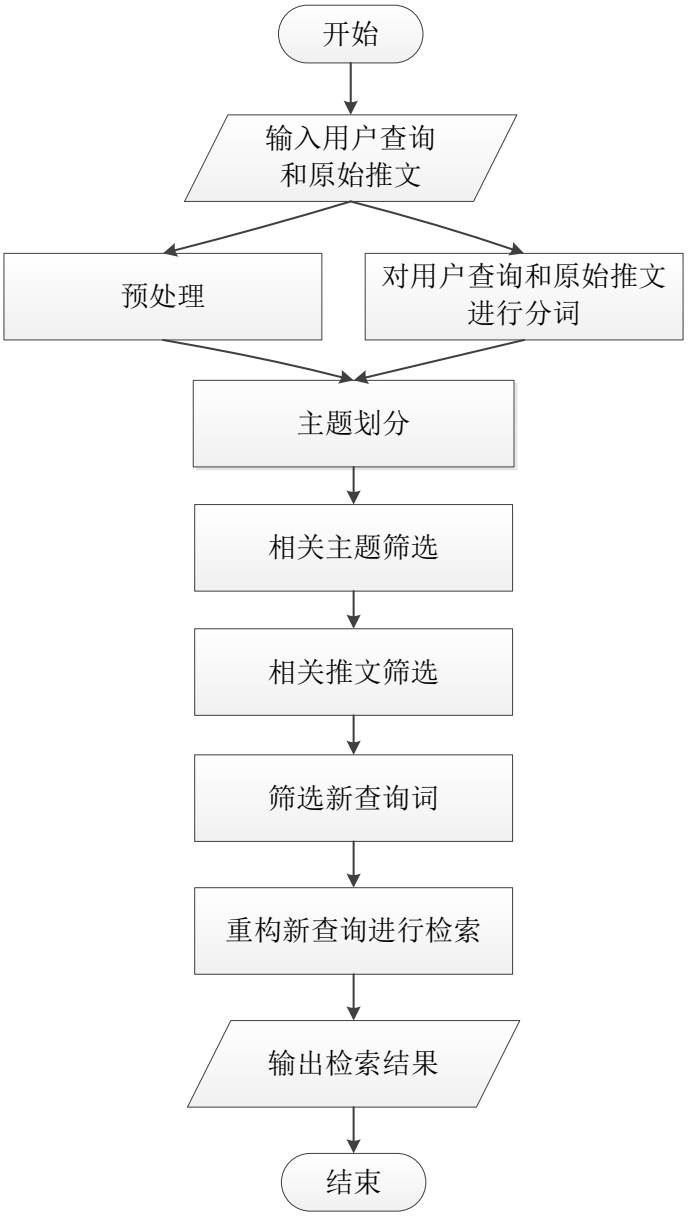

图 4-1 基于主题划分的查询扩展方法

图 4-1 表示了通过主题划分进行查询扩展的主要流程，它是对第三章描述的查询扩展流程使用主题划分方法进行优化。

(1) 通过潜在语义分析获取关键词在主题中的分布情况以及推文在主题中的分布情况。

(2) 比较原始查询与个主题下关键词的相关性去筛选满足用户的主题。

(3) 再通过推文在各个主题下的分布筛选出用户主题下的推文集合。

(4) 将该推文集合作为待扩展词语料库，比较语料库中词语与原始查询的相关性，筛选出最满足原始查询的词语。

(5) 将第(4)步筛选的词语与原始查询词组合成新查询词。

(6) 对新查询使用 BM25 算法对推文集合进行筛选，检索出满足用户需求的推文。

该方法包含了推文文本处理，主题划分和查询扩展。而推文文本处理和查询扩展与第三章内容相同不再赘述。而如何对主题进行划分，并选取满足用户需求的主题进而得到与用户主题最为相关的推文集合作为语料库是本章研究的重点。

4.3 主题划分在查询扩展中的应用

主题划分表示的是对于主题推文的反馈，本文首先将推文集合划分为若干个主题，假定每个主题由多个关键词构成，通过关键词与用户查询词筛选出满足用户的主题。它包含了主题划分，与用户最相关主题的筛选和与该主题最相关的推文筛选三个部分。

4.3.1 主题划分

传统方法对于文档相似程度判断已经在第三章中做了聚类方法的介绍，它将推文以集合的形式进行反馈，这一定程度上体现了上下文的语义信息。局部分析方法一定程度上可以解决缺少语义信息的问题。其主要思想是获取推文集合构成的语料库中与查询词最相关的词语进行扩展，同时认为待扩展词与原始查询词在推文中共现频率越高，则待扩展词与原始查询词的相关性越好。

但是这也默认了原始查询与待扩展词在反馈语料库中是共现的，事实上并非如此。比如：“斯特恩终于退休了。”和“NBA 的格局是否会发生变化”，很显然仅仅通过文档中的共现词语的判断并不能得出两句话是相似的，而事实上两条推文之间有着很高的关联性。同时经过了对聚类推文的筛选，本文发现推文集合的反馈只解决了相关推文存在查询词不足的问题，并不能将存在查询词的相关推文进行过滤。

因此本文认为所有推文都满足某个主题分布，而一个主题是通过许多关键词进行描述的，本文通过筛选出满足用户的主题进而将满足用户的推文进行聚类。

这里本文通过隐式狄利克雷分布（Latent Dirichlet Allocation，LDA）主题模型进行语义分析。

4.3.1.1 LDA 模型作用

如上所述，在推文反馈中存在通过原始查询与推文集合的相关性筛选出不满足用户需求的结果，就是明明与用户描述的内容不相关，但是由于用户查询词输入不准确等原因导致该集合存在许多原始查询词而被反馈，这将严重影响最后的检索结果。故本文找到与用户相关的主题去解决因查询词相近导致的主题偏差。2003 年 D.Blei 提出的隐含狄利克雷分配（LDA）模型可以很好的去提取潜在语义中的主题信息。LDA 是一个词袋模型，它认为某一条推文通过一定概率选择了某一个词袋（主题），然后在词袋中以另一种概率选择了某一个词语，推文的各个位置选择了不同词语便组合成了一条推文。

使用 LDA 模型生成一篇文章的主要流程为：

(1) 分别从狄利克雷多项式分布中抽取文档对主题的分布 θ 以及主题对词语的分布 β。

(2) 假设对文档 i 的主题分布为 θ_i，其表示文档 i 在主题中的出现频率。那么通过 θ_i 可以确定每个主题 z 的概率分布 $p(z|\theta_i)$。

(3) 通过主题 z 中的单词概率分布抽取得到文档中的每一个单词。

生成推文时每个词语的概率分布如公式 4-1：

$$p(\text{词语}|\text{文档})=\sum_{\text{主题}} p(\text{词语}|\text{主题})\times p(\text{主题}|\text{文档}) \tag{4-1}$$

其概率公式可以用矩阵作如图 4-2：

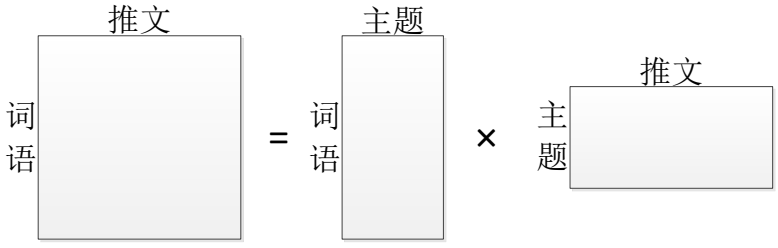

图 4-2 LDA 生成推文的方法

在图 4-2 中推文词语矩阵表示每条推文的词语词频，主题词语矩阵表示每个主题中的词语分布，推文主题矩阵表征推文在每个主题中的分布。

由公式 4-1 和图 4-2 可知，LDA 对主题的提取实际上就是推文构建的逆过程，通过 LDA 建模，可以获取主题下语料库中每个词语的概率分布，通过对主题词

语层的研究进而得到每条推文在各个主题中的概率分布。

其模型表示图 4-3：

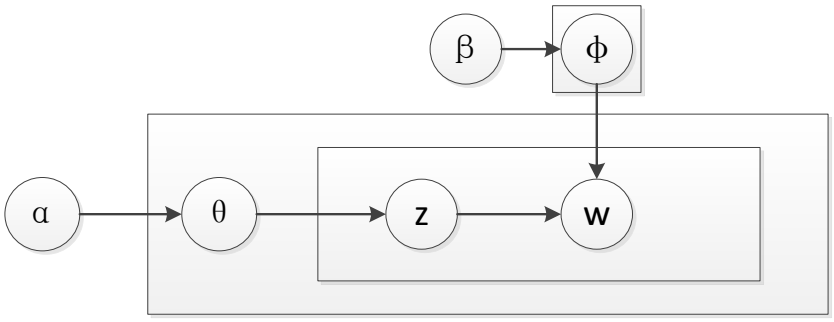

图 4-3 LDA 联合概率模型图

其模型图可以分为两个主要过程，α 到 θ 到 z 的过程表示在生成某一条推文的时候，通过狄利克雷先验分布获取 θ，然后从 θ 到 z 的过程就是通过多项分布得到推文中某个单词的主题编号 $z_{a,b}$。

第二个过程就是从 β 到 φ 再到 w 的过程，其主要作用与生成某条推文中的某个词语，同第一个过程类似，通过狄利克雷先验分布得到的所有主题词语模型中筛选出主题编号为 $z_{a,b}$ 的主题，然后在这个主题中通过多项分布再去得到该主题下的词语 $w_{a,b}$。

4.3.1.2 吉布斯采样

LDA 中的每一篇文档的主题分布都是不确定的，以及每个主题下的词语分布也是不确定的。LDA 中两者的似然函数都满足多项分布，故两者均加上了狄利克雷先验分布，文档对应的主题分布为 θ，以及主题对应的词语分布为 φ，因此生成文档的主要过程可以概括为：

(1) 对文档中每一个需要加入的词语先得到其主题。

(2) 从主题中选取需要加入的词语。

本文需要估算 θ 和 φ 两个分布。

综上实际上本文就是通过吉布斯采样算法对联合分布 $p(w,z)$ 进行采样，同时 w 是推文语料库的词语，是个已知的值，而 z 是一个隐含变量，实际上本文真正需要的概率分布是一个条件概率，即 $p(z|w)$，通过已知 w，通过采样获取 z。

吉布斯采样实际上就是，将抽取到的词语 w 从当前语料库中分离，该词语的分离并不会改变狄利克雷-多项分布的共轭结构，因此 θ 和 φ 的后验分布也是狄利克雷分布。通过去除该词语后的主题分布 θ 和 φ 来计算该词语在其他主体中的概率分布。

LDA 的吉布森采样公式为：

$$p(z_i = k \mid z_{\bar{i}}, w) \propto \frac{n_{m,\bar{i}}^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_{m,\bar{i}}^{(t)} + \alpha_k} \cdot \frac{n_{k,\bar{i}}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\bar{i}}^{(t)} + \beta_t} \tag{4-2}$$

公式 4-2 主要表示的就是 $p(topic \mid doc) \cdot p(word \mid topic)$，表达的就是从文档到主题再到词语这个路径的采样概率，如果含有 K 个主题，吉布森采样就是在这 K 条路径上进行了采样，其过程如图 4-4：

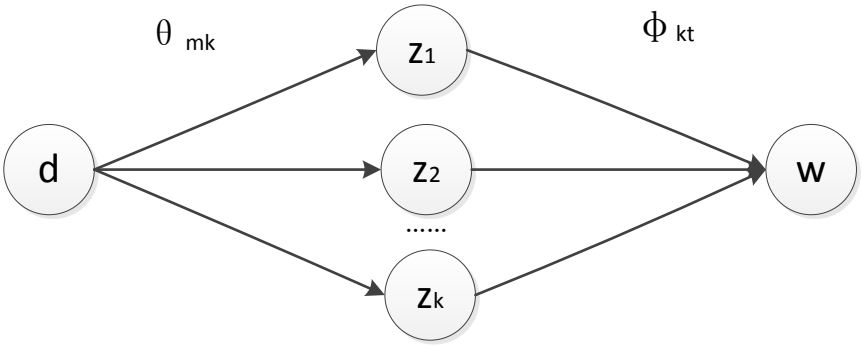

图 4-4 吉布斯采样的主要过程

4.3.1.3 主题划分主要流程

本文首先对 LDA 模型进行训练，并对模型的训练结果进行分析，进而用于主题划分。流程为：

(1) 对于与每篇推文的每个词语随机分配到某个主题中，由于非周期马尔科夫链中无论初始状态如何最后的阶级分布都能够收敛，基于此虽然 LDA 中的推文主题服从狄利克雷先验分布，故每个词语的主题初始状态可以随机分布。

(2) 对于每个词语根据吉布斯采样重新采样它的主题。

(3) 重复 2 过程直到吉布斯采样收敛。

(4) 最后对主题词语的共现频率进行统计，统计的结果就是 LDA 模型。

通过上述获得的 LDA 模型，本文可以计算主题中每个词语在主题下的概率分布 $p(w \mid t)$。在吉布斯采样收敛之后，本文对每条推文在相应主题下的每个词语的概率分布进行统计，可以得到每条推文在相应主题下的概率分布 $p(d \mid t)$。

4.3.2 与用户相关的主题获取

输入主题个数和狄利克雷先验分布的参数和每个主题下的关键词个数，通过 LDA 对主题进行划分可以得到所有主题，主题下的关键词及其概率分布，将测试

集中 MB007 话题下的推文划分为 8 个主题，每个主题有 25 个关键词，部分主题下的词语及其概率分布如表 4-2：

表 4-2 8 个主题下的词语及其概率分布

Topic 6th:		Topic 2th:		Topic 1th:	
词语：	概率分布：	词语：	概率分布：	词语：	概率分布：
pakistan	0.105188	pakistan	0.068010	murder	0.111412
india	0.012954	diplomat	0.044976	charged	0.025285
charged	0.010241	killed	0.019061	man	0.020978
cricket	0.010241	release	0.018485	charge	0.011750
cricketers	0.008884	two	0.017910	trial	0.009905
uk	0.007528	court	0.017910	arrest	0.008059
afp	0.007528	official	0.016182	wife	0.008059
source	0.007528	american	0.012727	death	0.007444
three	0.006850	lahore	0.012151	york	0.006829
world	0.006172	face	0.011575	murdering	0.005598
usa	0.006172	shooting	0.010423	accused	0.005598
february	0.005493	pakistani	0.009847	guilty	0.004983
kill	0.005493	charges	0.009847	murders	0.004983
mohammad	0.005493	arrest	0.008696	charges	0.004983
geo	0.004815	suicide	0.005204	years	0.004368

表 4-2 中包含了话题编号，该话题下的关键词以及关键词在该话题下的分布概率，本文选取了话题下含有原始查询词的个数，以及这些查询词的概率分布作为衡量原始查询与话题相关性的依据。主题中查询词的概率分布表征了该关键词与主题的相关性。

引入查询词的个数可以避免由于单个词语对主题影响过大的缺点。比如 topic1 中的“murder”这个词语与原始查询词相同，而且由于该主题描述的就是 murder，和 arrest，故它在该主题下的概率分布达到 0.119471，这比 topic2 中三个与原始查询相关的词语的概率分布和 0.126466 很接近，但是 topic2 更接近用户查询的描述，这也从侧面验证了本文需要引入主题中与原始查询相同的词语个数作为影响相关性的变量。

经过测试和分析，原始查询与主题的相关性可以表示为：

$$R_{qt} = n \cdot \sum_{i=1}^{n} p_i \tag{4-3}$$

在公式 4-3 中 R_{qt} 表示的是原始查询与主题的相关程度，它与 n 和 p_i 相关，n 表示的是原始查询与主题中词语相同的个数， p_i 则表示主题中与原始查询相同词语的概率分布。

4.3.3 相关推文的筛选

经过 4.2 节本文知道可以通过对所有推文进行主题划分，同时通过计算原始查询与各个主题的相关性得到满足用户需求的主题。而在 4.1 节中本文也介绍了对 LDA 建模得到主题以及主题下的概率分布的同时本文还能得到每条推文在各个主题中的概率分布。同样对于话题 MB007，每条推文在不同主题下概率分布结果如表 4-3：

表 4-3 推文在主题中的概率分布

推文 / 概率 主题 / 分布	1	2	3	4	5
1	0.464286	0.033333	0.388889	0.045455	0.031250
2	0.035714	0.033333	0.166667	0.045455	0.093750
3	0.250000	0.033333	0.166667	0.136364	0.031250
4	0.035714	0.300000	0.055556	0.318182	0.343750
5	0.035714	0.033333	0.055556	0.318182	0.156250
6	0.035714	0.033333	0.055556	0.045455	0.156250
7	0.107143	0.433333	0.055556	0.045455	0.031250
8	0.035714	0.100000	0.055556	0.045455	0.156250

表 4-3 中每一行代表了每条推文，而每一列对应了一个主题，其中的数值表示的是该推文在这个话题中的概率分布。该表表示的是 5 条推文在 8 个主题上的概率分布。

通过 4.2.2 节可以提取与原始查询最相关的主题，比较所有推文在该主题的概率分布，通过降序排列得到该主题下概率分布最高的 30 条推文作为满足用户需求的推文集合作为待扩展词的语料库。

4.4 测试分析

本节分别对主题划分的结果及推文检索结果进行测试和分析。

4.4.1 主题划分测试分析

本节分别测试了最佳主题的选取，主题个数选取对主题划分结果的影响以及被作为语料库的推文筛选结果。

对于所有 108 个话题下的推文进行主题划分，通过 4.2.2 节方法选取与用户最为相关的主题，对于得到所有与原始查询最相关的主题通过人工发现其准确率达到 100%，也是由于测试集中的话题中的查询词均能很好的描述话题内容。因此在满足用户输入的查询词可靠的前提下最佳主题筛选方法十分可靠。

由于推文主题个数需要用户输入，因此本文测试了主题个数对主题划分的影响。在实际测试中，本文选取了主题个数为 8，12 时主题划分的结果，其结果如表 4-2 和表 4-4，对图 4-2 测试集中的话题 MB007 的结果进行测试分析。

表 4-4 12 个主题下的词语及其概率分布

Topic 1th:		Topic 4th:		Topic 9th:	
词语：	概率分布：	词语：	概率分布：	词语：	概率分布：
pakistan	0.079499	pakistan	0.106937	murder	0.085353
diplomat	0.055886	cricket	0.012150	charged	0.034820
release	0.028694	world	0.011288	man	0.022187
killed	0.024401	cricketers	0.011288	charges	0.019029
court	0.022254	corruption	0.009565	charge	0.019029
pakistani	0.015098	charged	0.009565	two	0.018239
lahore	0.015098	source	0.009565	face	0.015081
american	0.014383	team	0.008703	official	0.013502
two	0.012236	revolution	0.007841	arrested	0.012712
held	0.010805	three	0.006980	shooting	0.011133
accused	0.010805	make	0.006980	men	0.009554
reuters	0.008658	afp	0.006980	death	0.009554
demands	0.008658	mohammad	0.006980	ap	0.008764
official	0.007943	today	0.006118	four	0.006396
shot	0.007943	join	0.006118	suspect	0.004816

表 4-2 表示的是该话题下的 3 个主题以及包含了描述这个主题的关键词和每个关键词在主题上的概率分布。很显然，topic2 表达的是美国外交官在巴基斯坦谋杀的事件，topic1 描述的是与谋杀这个主题的相关内容，topic6 描述的是各个国家之家的事情。在各个主题下，分别用了不同的关键词去描述了这些主题，而这些关键词在主题下的概率分布也各不相同。

表 4-4 表示的是 12 个主题下的主题及其词语分布，同样的本文选取了与表 4-2 中描述相似内容的主题，而在 topic1 中的词语与表 4-2 中的 topic2 的词语很相近，两者排名前 15 的词语都分别在对方中出现。表 4-4 中的 topic4 与表 4-2 中的 topic6 排在前 15 的词语也基本相同表述的是巴基斯坦板球队参加当时举办板球世界杯，而表 4-4 中的 topic9 和表 4-2 中的 topic1 描述内容相近表示的是“murder”有关的内容。

经过测试本文发现事件明确时，主题个数的变化不会对主题中的词语带来很大的影响，能描述主题的词语几乎不变。而对所有话题进行测试发现主题个数为 8 时主题划分结果较好，文中对主题划分方法的量化测试均使用 8 个主题。

同时使用 4.2.3 节方法得到的推文集合进行测试和分析，其结果如表 4-5：

表 4-5 与用户主题最相关的部分推文

Pakistani judge orders American held more days LAHORE Pakistan - A court on Thursday ordered the de
Pakistan court extends detention of US diplomat LAHORE Pakistan Reuters - An American who killed two Pakis
US presses Pakistan s president to free American ISLAMABAD - The US ambassador
US demands release of diplomat who killed The United States demanded the immediate release of an Amer
FoxNews US Demands Release of Diplomat in Pakistan US embassy says official who killed two robbers
US accused barred from leaving Pakistan A Pakistani court orders the authorities not to release a US citizen who killed two men in L
killed in Pakistan suicide attack A suicide attack in northwest

表 4-5 中表示的是所有推文中在 topic2 中的概率分布最大的部分推文，通过分析，本文发现其中的大部分都是在描述美国外交官在巴基斯坦涉嫌杀人这件事以及其他的后续事件，比如第 2 条推文描述的是巴基斯坦法院对这件事的审判，第 3 条推文表现的是美国政府敦促巴方释放外交官。

但是其中的最后一条推文描述的在巴基斯坦发生的自杀性袭击事件，而在topic2 中的主题词语分布中，“suicide”这个词语是被包含的，因此在推文主题不明确的时候，该推文在这个主题下的概率分布达到了 0.76，该推文被错误的划归到这个主题中。

同时如果某类主题也含有大量的原始查询，而通过筛选主题可以将其排除，进而将此类影响到结果的推文排除。经过对推文的集合进行分析本文还发现主题划分方法可以有效过滤部分含有查询词的不相关推文，其结果如表 4-6。

表 4-6 表示的是部分含有原始查询词但是由于不属于该话题而未被反馈的推文集合。第 1，2，3 条推文表示的是巴基斯坦和板球运动相关的内容，其都含有“Pakistan”这个词语。第 4，5，6，7 条推文表示的是与谋杀相关的内容，包含有“murder，arrest”等词语。最后两条推文表示的是外交官有关的内容，含有“diplomat”这个词语，这些推文在使用主题划分方法时可以有效的被过滤，从而得到更好的反馈推文作为语料库。

表 4-6 部分被过滤的推文

DTN Cricket Haider ready to sue Pakistan cricket board Pakistan wicketkeeper Zulqarnain Haider says he is re
Pakistan cricket players charged with corruption Source economictimes indiatimes com --- Friday February
Pakistan cricketers charged with corruption AFP - British prosecutors charged Pakistan cricketers Mo
US embassy cables Pakistan committed to defending conviction of Omar Saeed Sheikh for Daniel Pearl murder
Murder arrest after body is found A man is arrested on suspicion of murder after police discover the body of a
Murder arrest after body is found A man is arrested on suspicion of murder after police discover the body
Katy traffic stop leads to arrest of teens in game-room murder According to Harris County auth Harris County
Just Listed Monaco Diplomat - Grass Valley CA Monaco Diplomat PDQ Quad Slide-Out Diesel Pusher
Waww asik bgt rif RT Perdana menuju Gedung Pancasila sbg cln diplomat RI to meet the Minister msh ky mimpi

4.4.2 推文检索测试分析

4.3.1 节对主题划分筛选语料库的各个步骤和结果进行了测试分析。而该方法最后是用于查询扩展对推文进行检索，本节对使用主题划分方法对检索结果的影响进行测试分析。

在 TREC2011 中，对 4.3.1 节得到的 30 条推文作为语料库使用伪相关反馈查询扩展方法得到新查询词，然后用新查询词重构查询进行推文检索。

本文将使用主题划分方法得到的待扩展词语料库分别与传统的 BM25 算法，VSM 算法和 TF-IDF 算法比较，分别对相关的推文检索指标进行评价，得到最后的检索结果如表 4-7：

表 4-7 使用主题推文反馈的检索结果

评价指标 / 方法	mAP	P@10	P@30
RM	0.394	0.382	0.335
QDRM	0.417	0.423	0.394
VSM+RM	0.382	0.369	0.323
VSM+QDRM	0.405	0.407	0.378
TF-IDF+RM	0.379	0.372	0.312
TF-IDF+QDRM	0.411	0.413	0.383
TP+RM	0.446	0.482	0.412
TP+QDRM	0.475	0.524	0.435

RM(Retrieval Model)，QDRM(Query-Document Dependent Temporal Relevance Model)分别表示的是只使用逻辑模型的检索结果和结合使用逻辑模型和时序模型得到的检索结果。其对语料库的筛选方法是 BM25 算法，而 VSM，TF-IDF 分别对应两种经典的语料库筛选方法，TP(Topic Partition)表示使用主题划分对语料库进行筛选。

为能清晰表现主题划分方法的效果，对 mAP 检索指标进行分析如图 4-5，图 4-5 表示评价标准为 mAP 时，使用主题划分方法筛选带扩展词作为语料库与其他经典方法筛选语料库对应 RM 和 QDRM 两种查询扩展方法得到的检索结果。

对图 4-5 进行分析发现使用主题划分方法在 RM 和 QDRM 方法下 mAP 检索指标比 BM25 算法提升了 13.2%，13.9%，比 VSM 算法提升了 16.7%，17.3%，

比 TF-IDF 算法提升了 17.7%，15.6%。主题划分方法比较传统语料库筛选方法在检索结果上都有着明显的提升。

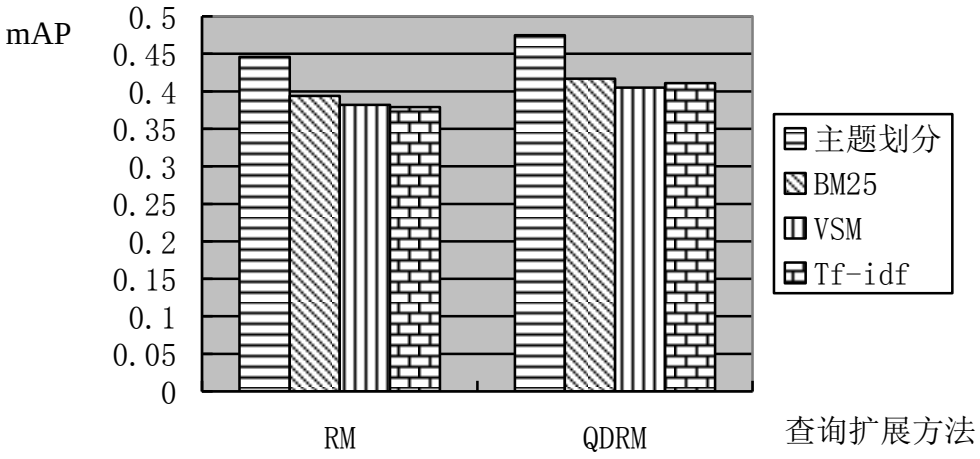

图 4-5 主题划分对检索结果的影响

通过主题划分方法选取语料库可以将含有原始查询词的无关推文进行过滤，优化了待扩展词语料库，进而有效的提升推文的检索结果。

4.5 推文聚类和主题划分方法比较

通过主题划分得到的反馈推文与通过聚类得到的反馈推文进行比较，在 TREC2011 上进行测试，每次取得反馈的前 30 条推文比较正确率，每个话题测试 50 次对正确率取平均值，然后对所有话题的正确率（筛选得到被标记为 1 的推文/总推文）取平均值，其结果的如表 4-8：

表 4-8 通过推文聚类和主题划分得到的准确率

使用推文聚类得到的正确率	使用主题划分得到的正确率
0.825	0.834

通过表 4-8 研究本文发现对比推文聚类方法，通过主题划分得到的反馈结果的正确性比推文聚类的准确率略高，差异不大。

此外本文分别比较了使用推文聚类方法和主题划分方法对语料库筛选存在的优点和缺陷。

使用推文聚类方法的优点：通过推文聚类，得到相似语义的推文类，当推文类与查询词相关，整个类会被筛选，含有较少查询词但存在类中的相关推文会被

筛选，如表 4-10，由于类中的其他词语含有较多查询词，因此整个类会被筛选，含有较少查询词的推文随着类的筛选而被选取。

使用推文聚类方法存在的缺点：语义相似的类中的推文与原始查询无关，但其含有较多查询词会被错误的选取，如表 4-9，表中含有原始查询词中的“arrest”和“murder”被错误筛选，将其作为语料库对检索结果产生错误的影响。

表 4-9 不相关推文聚类结果

Cluster20:
Guard Mexican police arrest gang leader blamed for Acapulco killings Jose Lozano behind murder earlier in month
Mexican police arrest gang leader for murder
Acapulco Mexican officers arrest gang leader for murder

表 4-10 筛选后的聚类结果

Cluster21:
US official in Pakistan to face murder charge Pakistan will pursue murder charges against a US consular
RT Prosecutor says US consular employee suspected in Pakistan shooting deaths will face charge - AP
DTN World News US official in Pakistan to face murder charges Pakistan will pursue murder charges against a

Cluster27:
Pakistan court blocks handing over of US diplomat - Reuters
Pakistani judge orders American held more days LAHORE Pakistan - A court on Thursday ordered the de gvnc
Pakistan court refuses to release US official denying immunity
Overload News Pakistan court extends detention of U S diplomat - Reuters DAWN Pakistan court extends detent

使用主题划分方法的优点：在推文聚类中由于含有较多查询词的不相关推文被错误筛选，由于其属于不与用户相关的主题而被过滤，保证了被反馈推文的准确性。表 4-4 表示的是主题划分的结果，表 4-9 中的推文会被划归到 Topic1 中被过滤，从而不会影响检索结果。

使用主题划分方法存在的缺点：如果推文主题不明确或者描述该类内容的推

文较少，如“killed in Pakistan suicide attack A suicide attack in northwest。”存在的原始查询词只有一个“Pakistan”，但它在用户所需主题的概率分布达到了 0.76 而被错误的筛选，这种推文在使用推文聚类方法时并不会被筛选。

鉴于使用推文聚类和主题划分方法进行查询扩展存在着自己的缺点，都有可能引入无关推文作为语料库影响检索结果。本文分别选取通过推文聚类方法得到的 50 条推文和通过主题划分得到的 50 条推文进行比较，当两种方法结合使用，被筛选的推文需要同时满足两种方法的筛选结果，被筛选的语料库准确率达到了 90.4%，对两种方法单一使用时的准确率有着明显的提升。

此外推文聚类和主题划分两种查询扩展方法都是对待扩展词语料库的筛选进行优化。对已做处理的推文文本可以同时进行主题划分和推文聚类处理，其程序运行性能满足需求。在此前提下，结合使用两种方法可以更加有效的提升检索结果。

结合两种方法筛选待扩展词语料库进行查询扩展，最后得到的检索结果如表 4-11：

表 4-11 结合推文聚类和主题划分得到的检索结果

评价指标 / 方法	mAP	P@10	P@30
RM	0.394	0.382	0.335
QDRM	0.417	0.423	0.394
VSM+RM	0.382	0.369	0.323
VSM+QDRM	0.405	0.407	0.378
TF-IDF+RM	0.379	0.372	0.312
TF-IDF+QDRM	0.411	0.413	0.383
TP+RM	0.446	0.482	0.412
TP+QDRM	0.475	0.524	0.435
Cluster+RM	0.439	0.446	0.398
Cluster+QDRM	0.467	0.512	0.424
Combine+RM	0.482	0.623	0.518
Combine+QDRM	0.509	0.645	0.542

RM(Retrieval Model)，QDRM(Query-Document Dependent Temporal Relevance

Model)分别表示的是只使用逻辑模型的检索结果和结合使用逻辑模型和时序模型得到的检索结果，其对语料库的筛选方法是 BM25 算法，而 VSM，TF-IDF 分别对应两种经典的语料库筛选方法，Combine 表示结合使用推文聚类和主题划分对语料库进行筛选。

为使结果更加清晰绘制图 4-6，前面分别已经描述了使用推文聚类和主题划分对语料库的筛选比传统方法存在的优势，本节对结合使用两种方法对比单一使用的结果进行分析。

对图 4-6 进行分析发现结合使用两种方法在 RM 和 QDRM 方法下 mAP 检索指标比使用推文聚类方法提升了 9.8%，9.0%，比使用主题划分方法提升了 8.0%，7.2%，结合使用两种方法比单一使用在结果上有着明显的提升。

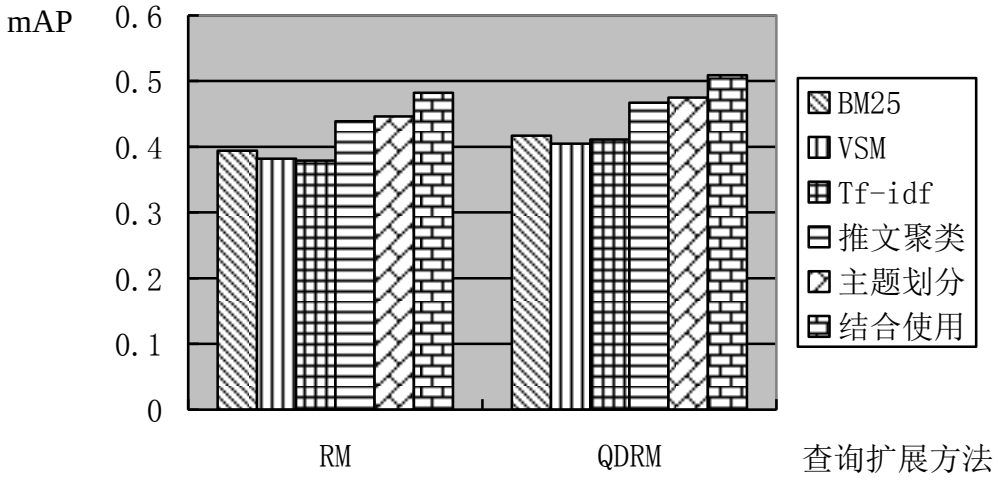

图 4-6 多种方法检索结果比较

4.6 本章小结

由于传统查询扩展方法存在筛选语料库时含有较多查询词的不相关推文被错误筛选的问题，本章提出了基于主题划分的查询扩展方法，并具体介绍了该方法的模型和流程。本章具体分析了如何通过主题划分得到待扩展词语料库，并对得到的语料库做直观评价。本章还对比了主题划分方法与 BM25 算法，VSM 算法和 TF-IDF 算法的检索指标，发现主题划分方法可以有效提升检索结果。最后本章比较了使用推文聚类方法和主题划分方法进行查询扩展的优势和缺点，并发现结合使用两种查询扩展方法对检索结果提升更加明显。

第五章 总结与展望

5.1 工作总结

推文检索方法在如今信息爆炸时代可以得到满足用户需求的信息，而正因为推文数据量的爆炸，有效的推文检索方法十分重要。而查询扩展方法在推文检索中有着广泛的应用，针对此背景，本文主要对查询扩展方法在推文检索中的应用进行分析，并做了以下四个工作：

(1) 伪相关反馈方法是被广泛应用的查询扩展方法，而它包含了两个主要步骤：筛选满足查询词的推文作为待扩展词的语料库和筛选与查询词相关的推文作为新查询词。将新查询词加入到原始查询构成新查询再对推文进行检索。经过测试发现待扩展词语料库的筛选将会对检索结果产生很大的影响。而传统的使用 BM25 算法，VSM 算法，TF-IDF 算法等都不能有效的筛选出满足用户需求的待扩展词语料库。针对如何更好的筛选待扩展词语料库本文提出了基于推文聚类的查询扩展方法和基于主题划分的查询扩展方法。

(2) 基于推文聚类的查询扩展方法是针对传统方法筛选语料库存在将含有较少查询词的相关推文被遗漏的问题进行改进的。该方法包含了三个步骤：推文文本处理，推文聚类和查询扩展。推文中存在大量不标准的写法和格式，本文对推文进行正规化，多联词的提取以及正则匹配处理。推文聚类是本章的关键，通过将具有相同语义的推文使用 k-means 方法进行聚类，然后采用 TF-IDF 方法筛选出满足用户需求的相关类使得含有较少查询词的推文随着含有该推文的类被筛选而被选取，有效的优化了待扩展词的语料库。通过语料库中的词语与原始查询词进行相关性对比，得到与原始查询最相关的部分词语作为扩展词，进而可以对推文进行重新检索，本文使用了 RM 和 QDRM 两种筛选新查询词的方法。通过测试分析本文发现使用推文聚类的查询扩展方法在 RM 和 QDRM 两种方法下平均准确率对比 BM25 算法分别提升了 11.4%和 12.0%，比 VSM 算法分别提升了 14.9%和 15.3%，比 TF-IDF 算法分别提升了 15.8%和 13.7%。

(3) 基于主题划分的查询扩展方法是针对传统方法筛选语料库存在将含有较多查询词的不相关推文被错误筛选的问题进行改进的。该方法在推文文本处理和查询扩展方法上与第三章相同。主题划分首先通过 LDA 得到主题以及主题中词语的概率分布和推文在每个主题中的概率分布，然后比较每个主题中的词语与原始查询的相关性比较获取最满足用户需求的主题，最后根据每条推文在各个主题中的概率分布得到与该主题最相关的推文作为结果进行反馈，将含有部分查询

词但不与主题相关的推文进行排除。通过测试分析本文发现使用推文聚类的查询扩展方法在 RM 和 QDRM 两种方法下平均准确率对比 BM25 算法分别提升了 13.2%和 13.9%，比 VSM 算法分别提升了 16.7%和 17.3%，比 TF-IDF 算法分别提升了 17.7%和 15.6%。

(4) 本文还分别比较了推文聚类查询扩展方法和主题划分查询扩展方法的优缺点。本文发现通过推文聚类可以将含有较少查询词但存在类中的相关推文筛选。但是使用推文聚类方法，如果语义相似的类中的推文与原始查询无关，但其含有较多查询词会被错误的选取。使用主题划分方法将不属于与用户相关主题的推文进行过滤，保证了被反馈推文的准确性。但是如果推文主题不明确或者描述该类内容的推文较少，这类推文可能被错误的划归到该主题下。当两种方法结合使用，被筛选的语料库准确率达到了 93.4%，对两种方法单一使用时的平均准确率有着很好的提升。

5.2 工作展望

查询扩展方法在推文检索中有着广泛的应用，随着信息科技的发展，计算机处理数据的能力得到了飞速的提升。因此先前需要先满足计算速度的瓶颈有了很大的突破，这对推文检索的实现提供了先决条件。

在查询扩展方法研究中，针对反馈的语料库和待扩展词与原始查询的相关性比较是两个研究重点，因此在这两方面也可以有更多的改进。

(1) 本文使用推文聚类方法优化待扩展词的语料库，推文聚类方法的研究是现在的热门，更好的推文聚类方法可以有效的提升最后的检索结果。

(2) 在对比待扩展词和原始查询的时候，本文只是将其分为文本模型和时序模型进行研究，这终将缺少人为因素。现在的基于人工构建的词典如 WordNet 等也在查询扩展中起到很好的作用。虽然经过测试本文发现其在推文的查询扩展中并不佳，相信可以有一种很好的结合方法。

致谢

蓉城七载，终散筵席，笔锋所触，别绪所及。旧影往事，愈发明晰，周遭事物，尽是情意。

成电吾校，大学实谓，潜研有处，论道课堂。同砚以学，辩理于师，得教庠序，长铭于心。钱峰吾师，考研初识，巧为同乡，久处愈知，科研勤勉，育人有方，学业疑难，不吝相教，琐事怅惘，持诚详谈。胡师光岷，治学精谨，律己甚严，私所慕往，高山景行！费师高雷，主于项目，善教循循，良言所获，似拨迷云！余虽不敏，诸师倾力！余之所得，厥功不忘！

良友诸多，倾盖如故，每有闲暇，同闹同游，及至术业，相学相促，得友如此，幸甚之至！致安琪硕士，陈坦硕士，许乔若硕士，初入研门，代码尚亏，常予指点，乃入正轨。致杨阳硕士，罗杰男硕士，文友栃硕士，胡馨月硕士，孙小田硕士，喜事同庆，难处互帮，旧事隐隐，常起思量！

入校之初，尚未加冠，岁及丁酉，二十有六，白驹过隙，止于一瞬，韶华之逝，昼夜不分。所求种种，圆梦寥寥，可恨肉膘，何时可掉！离象牙塔，入社会流，半分憧憬，半分是愁。挥别往昔，捡拾记忆，茫茫前路，砥砺且行。去校不返，初心不忘，孤蓬万里，此处扬帆！

参考文献

[1] K. Semertzidis, E. Pitoura, P. Tsaparas. How people describe themselves on Twitter[C] Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks. ACM, 2013: 25-30.

[2] Number of monthly active Twitter users worldwide from 1st quarter 2010 to 4th quarter 2016. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/,2017-1-1/ 2017-3-15.

[3] M.E. Maron, J.L. Kuhns. On relevance, probabilistic indexing and information retrieval[J]. Journal of the ACM (JACM), 1960, 7(3): 216-244.

[4] G. Salton. The SMART retrieval system—experiments in automatic document processing[J]. 1971.

[5] Rocchio J. Relevance feedback in information retrieval[J]. Computer Science, 2000:313-323.

[6] E. Ide. New experiments in relevance feedback[J]. The SMART retrieval system, 1971: 337-354.

[7] K.S. Jones. Automatic keyword classification for information retrieval[J]. 1971.

[8] Lavrenko V, Croft W B. Relevance based language models[C] Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001: 120-127.

[9] Li X, Croft W B. Time-based language models[C] Twelfth International Conference on Information & Knowledge Management. 2003:469-475.

[10] W.B. Croft. What do people want from information retrieval[J]. D-Lib magazine, 1995, 1(5).

[11] G.W. Furnas, S. Deerwester, S.T. Dumais, et al. Information retrieval using a singular value decomposition model of latent semantic structure[C] Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1988: 465-480.

[12] D.M. Blei, A.Y. Ng, M.I. Jordan. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[13] Y. Qiu, H.P. Frei. Concept based query expansion[C] Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1993: 160-169.

[14] J. Xu, W.B. Croft. Query expansion using local and global document analysis[C] Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1996: 4-11.

[15] J. Xu, W.B. Croft. Cluster-based language models for distributed retrieval[C] Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1999: 254-261.

[16] S.Y. Wang, W.S. Liao, L.C. Hsieh, et al. Learning by expansion: Exploiting social media for image classification with few training examples[J]. Neurocomputing, 2012, 95: 117-125.

[17] D. Zhou, S. Lawless, V. Wade. Improving search via personalized query expansion using social media[J]. Information retrieval, 2012, 15(3-4): 218-242.

[18] I. Anagnostopoulos, V. Kolias, P. Mylonas. Socio-semantic query expansion using Twitter hashtags[C] Semantic and Social Media Adaptation and Personalization (SMAP), 2012 Seventh International Workshop on. IEEE, 2012: 29-34.

[19] S. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer[J]. IEEE transactions on acoustics, speech, and signal processing, 1987, 35(3): 400-401.

[20] S.E. Robertson, Walker S, Jones S, et al. Okapi at TREC-3[J]. Nist Special Publication Sp, 1995, 109: 109.

[21] 孙坦, 周静怡.近几年来国外信息检索模型研究进展[J].图书馆建设.2008.3

[22] C. Buckeley, G. Salton, J. Allan, et al. Automatic query expansion using SMART[C] Proceedings of the 3rd Text Retrieval Conference. 1994: 69-80.

[23] G. Varelas, E. Voutsakis, P. Raftopoulou, et al. Semantic similarity methods in WordNet and their application to information retrieval on the web[C] Proceedings of the 7th annual ACM international workshop on Web information and data management. ACM, 2005: 10-16.

[24] D. Ellis. New horizons in information retrieval[M]. Amer Library Assn, 1990.

[25] Y. Rui, T.S. Huang, M. Ortega, et al. Relevance feedback: a power tool for interactive content-based image retrieval[J]. IEEE Transactions on circuits and systems for video technology, 1998, 8(5): 644-655.

[26] J.M. Ponte, W.B. Croft. A language modeling approach to information retrieval[C] Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998: 275-281.

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

[28] M.H. Peetz, D.M. Rijke. Cognitive temporal document priors[C] European Conference on Information Retrieval. Springer Berlin Heidelberg, 2013: 318-330.

[29] 李海芳, 史俊冰, 段利国, 陈俊杰. 一种基于含糊同义词的查询扩展方法[J]. 计算机应用与软件, 2011, 28(12):41-43.

[30] F. Diaz. Integration of news content into web results[C] Proceedings of the Second ACM International Conference on Web Search and Data Mining. ACM, 2009: 182-191.

[31] T. Kanungo, D.M. Mount, N.S. Netanyahu, et al. An efficient k-means clustering algorithm: Analysis and implementation[J]. IEEE transactions on pattern analysis and machine intelligence, 2002, 24(7): 881-892.

[32] T. Joachims. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization[R]. Carnegie-mellon univ pittsburgh pa dept of computer science, 1996.

[33] S. Huang, Q. Zhao, P. Mitra, et al. Query Expansion Using Topic and Location[C] icdm. IEEE Computer Society, 2007:619-624.

[34] M. Efron, P. Organisciak, K. Fenlon. Improving retrieval of short texts through document expansion[C] Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012: 911-920.

[35] J.J. Jiang, D.W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy[J]. arXiv preprint cmp-lg/9709008, 1997.

[36] T. Miyanishi, K. Seki, K. Uehara. Improving pseudo-relevance feedback via tweet selection[C] Proceedings of the 22nd ACM international conference on Information & Knowledge Management. ACM, 2013: 439-448.

[37] A.R. Rivas, E.L. Iglesias, L. Borrajo. Study of query expansion techniques and their application in the biomedical information retrieval[J]. Scientific World Journal, 2014, 2014(1):132158.

[38] 吴秦, 白玉昭, 梁久祯. 一种基于语义词典的局部查询扩展方法[J]. 南京大学学报(自然科学), 2014, 50(4):526-533.

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

攻读硕士学位期间取得的成果

[1] 科研项目：道驱动执行控制及主控软件 3.0 版本功能研发.