|  |  |  |
| --- | --- | --- |
| 技术交底书V1.0  （以下由知识产权部填写）  案号：  日期： | 申请人：  邮编： ，电话： | 请勿泄漏 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 交底书名称 | 基于英文的社交媒体短文本查询扩展方法 | | | |
| 发明人 | 费高雷，杨杰，胡光岷 | | | |
|  |  |  | （以下由知识产权部填写） | |
| 交底书撰写人 |  |  | 专利类型 |  |
| 联系电话 |  | 知识产权部负责人 |  |
| 传真号码 |  | 电话/传真 |  |
| E－mail |  | E－mail |  |

版权所有 侵权必究

# 相关技术背景（背景技术），与本发明最相近的现有实现方案（现有技术）

**1.1 背景技术（2.2 3点）(**基于相关反馈的查询扩展、基于全局分析的查询扩展、基于局部分析的查询扩展**)**

**1.2 与本发明相关的现有技术（BM25算法、k-means聚类算法）**

# 2、本发明技术方案的详细阐述（发明内容）

**2.1 本发明所要解决的技术问题（发明目的）（3.1 即传统的查询扩展的不足，以及需要使用什么技术改进）**

对于传统的查询扩展方法，在筛选待扩展词语料库的时候，一般使用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权重进行分析的。接着使用BM25算法对带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调节因子分析推文长度，查询词在推文中的词频以及查询权重对检索结果影响。最后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，通过向量间的预选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，但是这类方法存在严重的问题，即根据查询词在推文中出现的频率来筛选并不能有效得到满足用户需求的结果，如与用户相关推文中所含有查询词较少甚至不存在显然使用上述方法会将这类推文遗漏。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,murder,arrest”，它表示了美国外交官在巴基斯坦枪杀两人被逮捕的事件。下面这条推文：Prosecutor says US consular employee suspected in Pakistan shooting deaths will face charge。仅仅只包含了Pakistan这一个关键词，如果使用传统的方法对推文进行筛选时，主要是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，这样会遗漏一些满足用户需求但是包含查询词较少的推文。

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，本专利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，筛选与用户查询最关作为语料库，将传统的逐条推文筛选改类。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问题。

**2.2 本发明提供的完整技术方案（发明方案）（3.2-3.4** 推文聚类查询扩展方法、推文聚类在查询扩展中的应用、查询扩展方法**）**

正如2.1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，本专利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如下图所示：它包含了推文文本处理、推文聚类、相关推文筛选、筛选新查询词和重构检索这五个步骤。

* + 1. 推文文本处理

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。

* + 1. 推文聚类

推文聚类算法常见的有k-means means聚类，层次基于密度的和网聚类，层次基于密度的和网格的聚类。在本专利中聚类方法不是研究重点，而是通过优化查询扩展的流程，因此采用被广泛使的k-means means聚类，它需要所有推文将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。

* + 1. 相关推文筛选

相关推文筛选使用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。

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。?

* + 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。。。

词语w与原始查询的相关性P(w| D)分为两部分：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。



推文聚类的查询扩展流程图

* + 1. 重构检索

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：

公式中的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，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。

* 1. **本发明技术方案带来的有益效果（3.5.3——推文检索测试分析，即k-means与BM25方法的比较）**
     1. 推文聚类测试分析

本专利采用了http://trec.nist.gov.data/tweets/网站上的tweets2011的数据作为测试集进行实验。

表2-1 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 |
| Uh oh American diplomat charged with double murder in Pakistan He should have just called in a drone |
| 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 |

表2-2 筛选后的聚类结果

|  |
| --- |
| 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 |

* + 1. 推文检索测试分析

本专利对查询扩展方法中的语料库筛选步骤使用推文聚类方法，在这点上进行创新。在传统推检索方法中，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筛选出来作为语料库，通过比较了BM25算法以及推文聚类得到的反馈结果对最终的检索结果产生的影响，这样将会出现遗漏的推文，有时不一定能精确的获取用户想要得到的推文信息。

使用k-means聚类与使用BM25算法，TF-IDF算法和VSM算法相比，聚类算法反馈的推文的准确率更高，结果如下表所示。

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |

从上表可以看出k-means聚类方法的准确率明显高于其他3个方法，这验证了本专利的有效性。

# 3、针对2中的技术方案，是否还有别的替代方案同样能完成发明目的

没有别的替代方案。

# 4、本发明的技术关键点和欲保护点

（1）在预处理分词阶段需要去除a、an、the等没有意义的高频词  
（2）在正规化推文时，需要进行去噪工作，如去掉hashtags标签，用户名以及重复的推文，如何快速去重以及如何去噪的方法影响着扩展查询的精确度，因此需要保护。

参考文献