|  |  |  |
| --- | --- | --- |
| 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. **背景技术**

全球互联网中存在大量数据，特别是社交媒体上，每天更是产生了数以千亿计的海量数据，有效的分析、处理并获取有用的数据是现在研究的热点。其中，推特这个全球性质的社交媒体中的信息值得研究。推特是一家美国微博客服务网站以及社交络平台，全球互联日访问量最大的网站之一，用户可以在推特上发布 160 个字符之内的短文本[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]，而这么多的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。

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），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几种非常经典的方法：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。

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文本集合，用户手工标记满足用户的文本，一般情况下选取最满足用户的前m条文本，主要思想就是基于用户反馈的m条文本作为带扩展语料库，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1-1：



图1-1 相关反馈流程图

1. 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。其中基于语义词典的查询扩展，采用了completelink 算法，对所有查询词进行计算，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，结果为一个个的簇。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。

1. 基于局部分析的查询扩展

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[3]。

局部上下文分析的步骤为：（1）先将所要测试的文档分段，每一段作为查询的目标，对比原始查

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；（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。

* 1. **与本发明相关的现有技术**
     1. **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文本进行相关性比较得到最满足用户需求的文本，同时在查询扩展中这也用于对待扩展词语料库的反馈[4]。

BM25算法将原始查询Q解析为一个个查询词，对于每条文本d，比较每一个与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：

(1-1)

在公式1-1中Q为原始查询，为原始查询中的每一个查询词，d是每一条文本，是每一个查询词的权重。每个查询词和每一条推文的相关性得分。

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

* + 1. **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，起相关性就越好。

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1-2：

(1-2)

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1-3：

(1-3)

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1-4：

(1-4)

# 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）**

对于传统的查询扩展方法，具体流程图如2-1所示。在相关文本筛选的时候，一般使用TF -IDF算法对文本进行筛选，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。接着使用伪相关反馈建立模型，依据文本逻辑相关筛选出新的词语，将原有词进行扩展。最后使用BM25算法对带扩展词语料库进行筛选，该方法在TF -IDF的算法上引入了不同调节因子分析文本长度，查询词在文本中的词频以及查询权重对检索结果影响。最终将用户可能关心的文本输出。



图2-1 传统扩展查询流程图

上述方法是筛选查询词与文本相关性的主流方法，但是这类方法存在严重的问题，即根据查询词在文本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，如与用户相关文本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文本遗漏。也导致了许多与用户查询相关的文本集合被过滤，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。下面这条文本：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文本进行筛选时，主要是根据查询词在文本中的词频来衡量文本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文本。

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文本不被选取的缺点，本专利提出了基于文本聚类的查询扩展方法。该方法将具有相同语义的文本进行聚类，筛选与用户查询最关作为语料库，将传统的逐条文本筛选改为逐类。如果含有较少查询词的相关文本存在于被选取的类中，这些文本也可以被筛选。该方法有效解决了与用户相关的文本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）**

现在我们已经知道传统扩展查询方法对待扩展词语料库的筛选存在问题。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文本会被漏选这一问题，本专利提出了基于文本聚类的查询扩展方法，并对该方法的流程进行设计和实现。

文本聚类的查询扩展方法流程图如2-2图所示：它包含了文本处理、文本聚类、相关文本筛选、查询词扩展和收敛性条件判定这五个步骤。

* + 1. **文本数据预处理**

文本处理是一个重要的步骤，由于社交网络上文本的不确定与不规范特性，大量文本都是不规范的，因此本专利对输入的文本和原始查询进行处理，它包含了文本正规化，去除重复文本和多联词的提取等步骤。

正规化的主要任务是根据需要对文本进行进一步去噪工作，通过对比大量的社交网络文本，发现了许多对语义造成干扰的字符，如文本中的URL链接、标签、@用户名等信息。同时去除重复推文也是一个很重要的步骤，我们都知道，在社交网络中，当某个特定事件发生时，会有大量人转发这一事件，造成大量重复性的推文，这时需要把这一类的推文精简，去重。本专利通过将正规化后的文本映射为一个64位的CRC循环校验码，当新读取一条文本时，通过计算它的CRC检查是否出现，从而达到去重的目的。而多联词提取的主要任务是从文本中提取出名词性短语。在实际应用中，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：首先从文本中提取出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库中定期检查筛选该词库中的多联词，以保留较高频的词（词频阈值设为0.02 %）， 接着以该多联词库为依据，从文本中划分出这些多联词作一个分词单元。

同时我们也需要对文本进行分词，在社交网络文本中存在一些没有太多意义但是词频又十分高的词语如：“a”，“an”，“the”等，这一类词在后续处理的过程中将会对结果产生错误影响，故需要将它们去除。



图2-2 文本聚类的查询扩展流程图

* + 1. **文本聚类**

对文本预处理与分词结束后，我们已经得到了许多正规化无重复的文本，这些文本中可能包含有查询词，也有可能没有包含查询词，所以为了使查询词扩展更加准确，我们需要将那些语义相近但仅仅包含少量查询词的文本放入语料库中。要达到这个目的，本专利使用了文本聚类方法。

文本聚类可以更好的得到扩展词的语料库，在聚类完后的结果包含了相同语义的文本类，如果含有少量查询词的文本存在在该类中，且该类与用户查询的相关性很大，这时这种含有少量查询词的文本也将会被筛选，弥补了传统扩展查询的漏洞。因为聚类方法不是本专利研究的重点，因此选择了最常用的聚类方法——k-means聚类方法。

* + 1. **相关文本筛选**

相关文本的筛选就是将上一步的聚类后的文本进行相关性筛选的过程，本专利使用TF-IDF算法通过比较经过聚类的文本和查询词的相关性得到最满足用户需求的文本类。

TF表示的是查询词在文本集合中出现的频率，频率越高说明它对文本集合越重要，其计算方法如公式2-1：

表示查询词在文本类中出现的频率，F表示文本类中词语的个数。

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，这样就不能特定的区描绘某个文本集合，同时重要性就会下降。

相比与传统的扩展查询方法将一条文本作为一个文档，本专利是将一个文本类作为一个文档，计算类中查询词的TF，然后在整个文本集合中计算查询词的IDF，最后求取查询词和文本集合的TF-IDF值。经过实际测试，本专利通过计算每个文本类对所有查询词的TF-IDF值的和，作为该文本类与原始查询的相关性，对相关性降序排序，筛选出文本集合直到文本集合中的推文数大于30，将这个结果作为带扩展词的语料库。同时本专利为了避免少数文本不与其它文本语义相似而又含有较多查询词，这类文本会对结果产生不好的影响，故对文本条数少于3的文本类进行忽略。

* + 1. **查询词扩展**

现在我们已经获得了相关性较大的文本语料库，需要对查询词扩展，获得新的查询词。本专利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，认为词语与原始查询的相关性分为两个部分：

1. 文本逻辑相关：

即带扩展词与原始查询与文本信息文本相关，在实际中，待扩展词与原始查询的相关性是同待扩展词在文本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的。同时在一条文本中，两个词越临近，则它们具有更大的相关性[5]，因为构成句子的词语一般有一定的短语组成，相邻的词语很大程度上是一个修饰的过程。因此，对于每一条文本，本专利认为待扩展词与原始查询词的相关性随距离的增加而递减。

1. 时序信息相关

即原始查询与文本与文本创建时间相关。本专利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。比如，在用户查找2017年4月7日的叙利亚空袭事件。“The people still living in #Mosul are surrounded by ruins: Drone footage shows the extent of damage caused by recen…”这条文本是4月5日发送的，而“Hillary is now calling for US airstrikes in Syria. I mean... what”这个文本是4月7日创建的，很显然用户是希望获取后面一条信息。

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算法对文本集合进行重新检索。

* + 1. **收敛性条件判定**

在进行了查询词扩展后，会重新检索出新的文本，但是有可能会获得较多的文本，这说明用户需要的真正信息还需要进一步挖掘。在结合实际的情况下，本专利认为，当重新检索后文本只有5条以内或者进行6次扩展词查询后，所检索出来的文本才是用户最关心的文本。

* 1. **本发明技术方案带来的有益效果**
     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。

在测试集中包含了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。本专利通过BM25方法比较查询词与每条文本的相关性，得到的排名前30的部分文本结果如表2-1所示。

表2-1 BM25算法筛选得到的文本

|  |
| --- |
| 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2-1中下划线表示的是原始查询词，我们可以清楚得看见，传统的方法筛选出来的文本都含有较多的查询词。表2-2是使用TF-IDF对所有聚类后的文本筛选得到的结果，这个类有4条推文，全部与用户查询相关，而且表中的“Prosecutor says US consular employee suspected in Pakistan shooting deaths will face charge -AP”这句话中由于含有原始查询词较少，在使用传统的方法的时候不能被直接检索出来，但是使用了文本聚类方法后，可以直观的发现这一类含有少量原始查询词的文本可以随着类的检索而被检查所出来。

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文本聚类方法，在这点上进行创新。在传统检索方法中，它们都是将原始查询与文本集合的每条文本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算法以及文本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文本，有时不一定能精确的获取用户想要得到的文本信息。

使用k-means聚类方法与使用BM25算法，TF-IDF算法和VSM算法相比，聚类方法反馈的文本的准确率更高，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。

（3）收敛判定的条件是通过大量数据测试得出的较好的判定标准，它影响着最终的结果，需要保护。

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. C. Buckeley，G. Salton，J. Allan，et al. Automatic query expansion using SMART[C] Proceedings of the 3rd Text Retrieval Conference. 1994:69-80.
4. S.E. Robertson，Walker S，Jones S，et al. Okapi at TREC-3[J]. Nist Special Publication Sp，1995，109:109.
5. 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.