**基于相似文本反馈的社交媒体文本查询扩展方法**

**技术领域**

本发明属于数据挖掘和数据检索技术领域，具体涉及一种社交媒体文本查询扩展方法。

**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。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，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。

传统扩展查询方法文本筛选时一般使用TF-IDF算法对文本进行筛选，该方法主要根据查询词在文本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，最终将用户可能关心的文本输出。上述方法是筛选查询词与文本相关性的主流方法，但是这类方法存在的问题是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。

**发明内容**

为解决现有技术存在的上述问题，本发明提出了一种基于相似文本反馈的社交媒体文本查询扩展方法。

本发明的具体技术方案为：一种基于相似文本反馈的社交媒体文本查询扩展方法，具体包括如下步骤：

步骤S1. 输入原始查询词与原始文本；

步骤S2. 对原始文本进行预处理与分词；

步骤S3. 对处理后的文本进行聚类；

步骤S4. 相关文本筛选，

步骤S5. 对筛选出这些文本进行查询词扩展，一旦达到收敛条件就输出结果。

本发明的有益效果：本发明的方法将具有相同语义的文本进行聚类，筛选与用户查询相似的文本类作为语料库，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似文本存在于被选取的类中，这些文本也可以被筛选，可以获取到相似性最高的文本进行扩展，并结合使用时序模型对已知的查询词进行扩展，能够有效的提高文本和查询词扩展的准确性。本发明的方法有效解决了与用户相似的文本由于含有查询词较少而被遗漏问题，使得最终结果的精确度有了大幅的提高。

**附图说明**

图1为本发明的基于相似文本反馈的文本查询扩展方法流程图。

**具体实施方式**

下面结合附图对本发明的实施例做进一步的说明。

本发明的方法是基于相似文本反馈的社交媒体文本的查询扩展方法进行研究，具体流程如图1所示：包含了输入原始查询词与原始文本、文本处理、文本聚类、相关文本筛选、查询词扩展及收敛性条件判定这5个步骤。首先用户输入原始查询词与原始文本，接着对原始文本进行预处理与分词，为文本聚类做准备，然后对处理后的文本进行聚类，可以得到语义相似的文本，筛选出相似文本后对这些文本进行查询词扩展，一旦达到收敛条件就输出结果。具体说明如下：

步骤S1. 输入原始查询词与原始文本：

步骤S2. 对原始文本进行预处理与分词：

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，因此需要对输入的原始文本进行处理，它包含了文本正规化，去除重复文本等步骤。

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，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；同时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。

步骤S3. 对处理后的文本进行聚类：

对文本预处理与分词结束后，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，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，本实施例具体使用文本聚类方法。

文本聚类可以更好的得到语义相似的扩展词语料库，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。本实施例选择k-means聚类方法，该聚类方法为本领域现有技术，在此不再详细描述。

步骤S4. 相关文本筛选：

相关文本的筛选就是将步骤S3聚类后的文本进行相关性筛选的过程，在此具体使用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，其计算公式如下：

(1)

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，具体如下：

(2)

相比与传统的扩展查询方法将一条文本作为一个文档，本实施例的方法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，本实施例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。

步骤S5. 对筛选出这些文本进行查询词扩展，一旦达到收敛条件就输出结果：

现在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。本实施例使用的伪相关反馈模型（伪相关反馈，也称之为盲式相关反馈，提供的是一种自动局部分析方法，它可以自动化相关反馈的手动操作部分，因此用户可不用参与额外的交互也可以获得更好的检索性能）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：

即待扩展词与原始查询词集合以及文本的相关性，在实际中，待扩展词与原始查询的相关性是与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在文本中的频率有关的，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，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，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

即原始查询与文本创建时间相关，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，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下：

表示新加入的查询词v的权重，sim(q,v)表示原始查询词集合q与v的相关性，表示原始查询词集合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。

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，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。

传统的相似文本检索方法是将原始查询与文本集合的每一条文本进行了相关性比较，然后将与原始查询最相关的文本筛选出来作为相似文本。而本专利主要是对查询扩展方法中的相似文本筛选步骤使用了文本聚类方法进行创新，这种方法相比于传统的扩展方法，不在使用逐条文本比较，而是使用逐类比较的方式，这样获取的文本类将会扩大文本的范围，将本来不能够获取的文本也包含进来，提高了准确性。通过实验，使用k-means聚类方法后，反馈结果的准确率高达0.825比使用BM25算法的0.738与TF-IDF算法的0.711还要高出0.1，这是一个巨大的提升，也证明了本发明方法提出的聚类方法的有效性。

同时在查询词扩展时，传统方法只是使用了RM（Retrieval Model），即只使用逻辑模型的检索结果，而发明的方法利使用了QDRM（Query-Document Dependent Temporal Relevance Model），即结合使用逻辑模型和时序模型得到的检索结果。对于mAP指标，即目标检测中衡量识别精度的指标，RM的值是0.394，而QDRM的值是0.417，提升了0.023，说明查询词扩展时使用时序模型将有助于获取更加准确的扩展词。



图1