(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103207905 B
(45)授权公告日 2015.12.23

(21)申请号 201310105450.2
(22)申请日 2013.03.28
(73)专利权人 大连理工大学
地址 116024 辽宁省大连市凌工路 2 号
(72)发明人 孔祥杰 宋秀苗 夏锋
(74)专利代理机构 大连理工大学专利中心 21200
代理人 关慧贞 梅洪玉
(51)Int.Cl.
G06F 17/30(2006.01)
(56)对比文件
CN 101833579 A,2010.09.15, 全文 .
US 7734627 B1,2010.06.08, 全文 .
TW I317488 B,2009.11.21,
闫亮 李先国 . 基于网页特征关键词的近似检测算法 .《科学技术与工程》.2009, 第 9 卷 (第 4 期), 第 919-923 页 .
杨建武 . 基于倒排索引的文本相似搜索 .《计算机工程》.2005, 第 31 卷 (第 5 期), 第 1-3 页 .
审查员 刘申

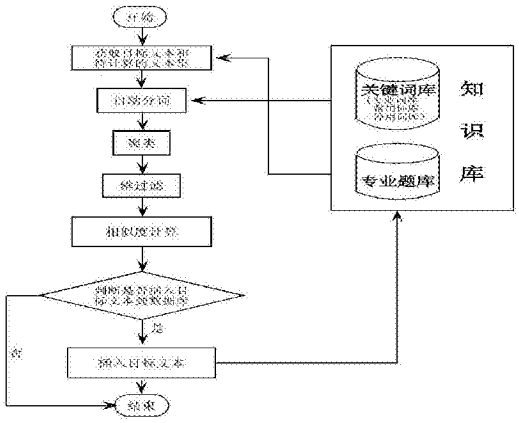
权利要求书1页 说明书4页 附图1页

(54)发明名称

一种基于目标文本的计算文本相似度的方法

(57)摘要

本发明公开了一种基于目标文本的计算文本相似度的方法。为了克服现有文本相似度算法单纯地考虑算法的准确率和效率，本发明结合项目实际情况综合考虑算法的准确率和效率两方面，保证在一定准确度的前提下提高算法的效率。在提高算法准确度方面，本发明充分考虑专业词汇对相似度计算的影响，采用了对不同类别的关键词加权的方式计算相似度；而在提高算法效率方面，本发明采用倒排索引聚类方法和维过滤方法。本发明既保证了文本之间的专业相关度计算的准确度，同时又兼顾了计算效率。



CN 103207905 B

1.一种基于目标文本的计算文本相似度的方法,其特征在于包括以下步骤:

(1)获取目标文本和待计算相似度的文本集合:获得目标文本和文本集合D后组成一个新的文本组合textSet,首先将所有的文本进行唯一id编号处理,目标文本用id号区分,其他文本按照输入的先后顺序依次编号;

(2)自动分词获取各文本的特征向量,包括以下步骤:

(2.1)去停用词;

(2.2)匹配专业关键词和常用关键词;

(2.3)同义词转换;

(2.4)统计各关键词在各文本中出现的频率,关键词的词性;

(2.5)计算各关键词的权值,创建特征向量;

关键词的权值计算公式为:$\omega(T_i) = \alpha \cdot \beta \cdot TF(T_i) \cdot IDF(T_i)$

其中 $\omega(T_i)$ 为关键词 T_i 的权值,$TF(T_i)=N/M$,其中N为 T_i 关键词在含M个关键词的文本中出现的次数,$IDF(T_i)=\log(D/D_w)$ 其中D为文章总数,D_w 为 T_i 关键词出现过的文章数;专业词汇IDF的计算以该词汇所属的专业领域的文章总数和该关键词出现在该专业领域内的文章总数;α 为关键词类别决定关键词权值的一个因子,其中专业词汇>常用词汇;β为关键词词性决定关键词权值的另一因子,其中名词>形容词>副词;

(3)聚类:首先为待计算文本集合创建倒排索引文件,然后以目标文本的专业词汇向量中的专业词汇为基础,搜索倒排索引文件,由于倒排索引文件中的关键词是按照关键词拼音的字母顺序排列的,采用二分法查找将倒排索引文件中含有目标文本中专业词汇的文件找出来,并将区分这些文本的唯一标识id放到集合C中;

(4)维过滤:首先为目标文本和集合C中的文本建立一个共同的倒排索引文件,然后根据建立的倒排索引文件创建一个存储各文本相应的关键词权值的矩阵M;矩阵M的列数为倒排索引文件中关键词的个数加1,矩阵M的行数为集合C中文本的个数加1,矩阵的第一列存储文本的id号,矩阵的第一行存储目标文本的特征向量;把目标文本中权值为0的列全部去掉,得到一个新矩阵M′,统计其他文本去掉的列中非0的列数并保存;

(5)计算相似度:利用步骤(4)维过滤后得到的矩阵M′,其中矩阵M′中的一行就代表维过滤后的某一文本的特征向量,然后计算目标文本向量即矩阵中的首行向量和其他各行向量之间的相似度。

一种基于目标文本的计算文本相似度的方法

技术领域

[0001] 本发明涉及信息检索和数据挖掘领域，尤其涉及一种基于目标文本的相似度计算方法。

背景技术

[0002] 随着互联网时代的到来，信息的爆炸式增长已经将人们淹没在信息的海洋中，人们再也不用担心互联网上没有自己想要的资源，但是如何才能找到这些资源成为了摆在信息检索专家面前的难题。文本相似度计算理论在信息检索和数据挖掘领域一直占据着非常重要的位置，而且在现实中也有很好的应用。

[0003] 学生作业抄袭检测，使用文本相似度计算方法可以很好的发现学生作业的抄袭现象，整治不良学风。

[0004] 保护知识产权，使用文本相似度计算方法检测是否含有剽窃他人研究成果的，以此来判断知识产权是否遭到侵犯。如若发现知识产权遭到剽窃等非法行为，可以对剽窃者实施必要的惩罚措施，通过这种方式更好的保护知识产权。

[0005] 网页的去重，通过文本相似度计算找到近似的网页并去除。去除重复网页不仅能够提高用户搜索效率，还能为用户提供很好的搜索体验。

[0006] 然而目前现有的文本相似度算法除了过于追求准确度，就是单纯的追求提高算法效率，根本没有考虑到具体的应用场景以及文本所涉及的专业领域。如果两个文本根本不属于同一专业领域，那么这两个文本就没有什么相似度可言。发明内容

[0007] 本发明正是鉴于上述技术问题而提出了一种基于目标文本的文本相似度计算方法，该方法包括以下几个步骤：

[0008] (1)获取目标文本 targetText 和待计算相似度的文本集合 D

[0009] (2)自动分词获取各文本的特征向量

[0010] (3)聚类

[0011] (4)维过滤

[0012] (5)计算相似度

[0013] 步骤(1)获得目标文本和文本集合 D 后组成一个新的文本组合 textSet，首先将所有的文本进行唯一 id 编号处理，目标文本可以用特定的 id 号区分，如目标文本 id 为 0，其他文本按照输入的先后顺序依次编号。

[0014] 步骤(2)的自动分词获取各文本的特征向量又通过以下几步完成：

[0015] (2.1)去停用词

[0016] (2.2)匹配专业关键词和常用关键词

[0017] (2.3)同义词转换

[0018] (2.4)统计各关键词在各文本中出现的频率、关键词的词性

[0019] (2.5)计算各关键词的权值，创建特征向量

[0020] 该步骤主要通过调用停用词库、常见词库还有专业词库提取出文本集合 textSet

中各文本的特征向量。对 textSet 中的任一文本首先进行去停用词处理，即将和停用词库中匹配的停用词从文本中去掉，然后再进一步匹配专业词库中的专业词汇，匹配成功的专业词汇，经词频统计、同义词转换，并进一步计算出相应专业关键词的权值后存储到专业关键词向量中，常用词库和文本的匹配处理和专业词库的类似，最后我们得到文本的两个特征向量－专业关键词向量和普通关键词向量。在提取目标文本特征向量的过程中的同义词转换，可以是将英文关键词转化为相应的中文关键词。另外步骤（2.5）中关键词的权值计算公式为：$\omega(T_i)=\alpha \cdot \beta \cdot TF(T_i) \cdot IDF(T_i)$

[0021] 其中 $\omega(T_i)$ 为关键词 T_i 的权值，$TF(T_i)=N/M$，（其中 N 为 T_i 关键词在含 M 个关键词的文本中出现的次数），$IDF(T_i)=\log(D/D_w)$ 其中 D 为文章总数，D_w 为 T_i 关键词出现过的文章数。专业词汇 IDF 的计算以该词汇所属的专业领域的文章总数和该关键词出现在该专业领域内的文章总数。α 为关键词类别决定关键词权值的一个因子，其中专业词汇＞常用词汇；β 为关键词词性决定关键词权值的另一因子，其中名词＞形容词＞副词。

[0022] 步骤（3）首先为待计算文本集合创建倒排索引文件，然后以目标文本的专业词汇向量中的专业词汇为基础，搜索倒排索引文件，由于倒排索引文件中的关键词是按照关键词拼音的字母顺序排列的，本发明采用二分法查找将倒排索引文件中含有目标文本中专业词汇的文件找出来，并将区分这些文本的唯一标识 id 放到集合 C 中。

[0023] 步骤（4）维过滤，首先为目标文本和集合 C 中的文本建立一个共同的倒排索引文件，然后根据建立的倒排索引文件创建一个存储各文本相应的关键词权值的矩阵 M（矩阵 M 的列数为倒排索引文件中关键词的个数加 1，矩阵 M 的行数为集合 C 中文本的个数加 1，矩阵的第一列存储文本的 id 号，矩阵的第一行存储目标文本的特征向量），把目标文本中权值为 0 的列全部去掉并统计其他文本中去掉的列中非 0 的列数并保存，得到一个新矩阵 M′。

[0024] 其中步骤（3）和步骤（4）中都用到的倒排索引文件的建立，输入的是文本的集合，输出的是文本集合中的所有关键词的倒排索引文件。创建成功的倒排索引文件中含有关键词列、（文件 id，在文本 id 中出现的频率）两列。其中关键词列按照关键词的拼音字母顺序排列，关键词后面对应的是关键词出现在各文本中的统计信息。

[0025] 步骤（5）计算相似度利用步骤（4）维过滤后得到的矩阵 M′，其中矩阵 M′ 中的一行就代表维过滤后的某一文本的特征向量，然后计算目标文本向量即矩阵中的首行向量和其他各行向量之间的相似度。

[0026] 其计算公式为：$\cos(D_1,D_2)=\frac{d_1 \cdot d_2}{\|d_1\| \cdot \|d_2\|}$，其中 d_1, d_2 分别表示文本 D_1 和 D_2 的特征向量。

[0027] 该方法得到的结果我们可以找到和目标文本专业最相关的文本。如果相似度的计算结果中有相同的，我们还可以根据该方法维过滤步骤中的删除的非 0 列数进一步判断哪个文本和目标文本更相似。

[0028] 克服现有的文本相似度计算方法单纯的注重提高计算的准确度和计算效率的缺点，本发明提出了一种基于目标文本的文本相似度计算方法，该方法在充分考虑项目实际的情况下，既能保证一定的准确度，又能提高计算效率。

附图说明

[0029] 图 1 为该发明某具体应用的流程图。

[0030] 图 2 为自动分词模块的流程图。

具体实施方式

[0031] 本发明的实际应用环境是在向某一专业题库中插入试题前，首先通过该发明中的计算方法找到专业题库中和预插入的试题即目标文本最接近的试题集，然后根据计算得到的相似度结果最终决定是否将该试题插入专业题库。

[0032] 本发明在实际项目中具体实施时，如图 1 所示包括以下几个步骤：

[0033] (1)获取目标文本 targetText 和待计算相似度的文本集合 D

[0034] (2)自动分词获取各文本的特征向量

[0035] (3)聚类

[0036] (4)维过滤

[0037] (5)计算相似度

[0038] (6)根据计算结果判断目标文本是否插入数据库并执行相应的操作

[0039] 步骤(1)通过将输入的问题的题干和答案整合得到目标文本，待计算的文本集合通过从专业题库中获取，获得目标文本和文本集合 D 后组成一个新的文本组合 textSet，对所有的文本进行唯一 id 编号处理，目标文本可以用特定的 id 号区分，如目标文本 id 为 0，其他文本按照输入的先后顺序依次编号。

[0040] 如图 2 所示，以目标文本的自动分词获取特征向量为例，文本自动分词获取特征向量通过以下几步完成：

[0041] (2.1)去停用词

[0042] (2.2)匹配专业关键词和常用关键词

[0043] (2.3)同义词转换

[0044] (2.4)统计各关键词在各文本中出现的频率、关键词的词性

[0045] (2.5)计算各关键词的权值，创建特征向量

[0046] 该步骤主要通过调用停用词库、常见词库还有专业词库提取出文本集合 textSet 中各文本的特征向量。对 textSet 中的任一文本首先进行去停用词处理，即将和停用词库中匹配的停用词从文本中去掉，然后再进一步匹配专业词库中的专业词汇，匹配成功的专业词汇，经词频统计、同义词转换，并进一步计算出相应专业关键词的权值后存储到专业关键词向量中，常用词库和文本的匹配处理和专业词库的类似，最后我们得到文本的两个特征向量－专业关键词向量和普通关键词向量。在提取目标文本特征向量的过程中的同义词转换，可以是将英文关键词转化为相应的中文关键词。另外步骤(2.5)中关键词的权值计算公式为：$\omega(T_i)=\alpha \cdot \beta \cdot TF(T_i) \cdot IDF(T_i)$ 其中 $\omega(T_i)$ 为关键词 T_i 的权值，$TF(T_i)=N/M$，(其中 N 为 T_i 关键词在含 M 个关键词的文本中出现的次数)，$IDF(T_i)=\log(D/D_w)$ 其中 D 为文章总数，D_w 为 T_i 关键词出现过的文章数。专业词汇 IDF 的计算以该词汇所属的专业领域的文章总数和该关键词出现在该专业领域内的文章总数。α 为关键词类别决定关键词权值的一个因子，其中 α (专业词汇)＞α (常用词汇)；β 为关键词词性决定关键词权值的另一因子，其中 β (名词)＞β (形容词)＞β (副词)。在本实例中我们取 α (专业词汇)=8，α

(常用词汇)=2；β (名词)=3，β (形容词)=2，β (副词)=1。

[0047] 步骤(3)首先为待计算文本集合创建倒排索引文件，然后以目标文本的专业词汇向量中的专业词汇为基础，搜索倒排索引文件，由于倒排索引文件中的关键词是按照关键词拼音的字母顺序排列的，本发明采用二分法查找将倒排索引文件中含有目标文本中专业词汇的文件找出来，并将区分这些文本的唯一标识 id 放到集合 C 中。

[0048] 步骤(4)维过滤，首先为目标文本和集合 C 中的文本建立一个共同的倒排索引文件，然后根据建立的倒排索引文件创建一个存储各文本相应的关键词权值的矩阵 M (矩阵 M 的列数为倒排索引文件中关键词的个数加 1，矩阵 M 的行数为集合 C 中文本的个数加 1，矩阵的第一列存储文本的 id 号，矩阵的第一行存储目标文本的特征向量)，把目标文本中权值为 0 的列全部去掉并统计其他文本中去掉的列中非 0 的列数并保存，得到一个新矩阵 M′ 。

[0049] 其中步骤(3)和步骤(4)中都用到的倒排索引文件的建立，输入的是文本的集合，输出的是文本集合中的所有关键词的倒排索引文件。创建成功的倒排索引文件中含有关键词列、(文件 id，在文本 id 中出现的频率)两列。其中关键词列按照关键词的拼音字母顺序排列，关键词后面对应的是关键词出现在各文本中的统计信息。

[0050] 步骤(5)计算相似度利用步骤(4)维过滤后得到的矩阵 M′ ，其中矩阵 M′ 中的一行就代表维过滤后的某一文本的特征向量，然后计算目标文本向量即矩阵中的首行向量和其他各行向量之间的相似度。

[0051] 其计算公式为：$$\cos(D_1,D_2)=\frac{d_1\cdot d_2}{\|d_1\|\cdot\|d_2\|}$$

[0052] 其中 d_1，d_2 分别表示文本 D_1 和 D_2 的特征向量。

[0053] 在计算完目标文本和超集中各文本之间的相似度后，找到相似度中的最大值，如果相似度集中的最大值超过了我们为实际项目的阈值 F，我们就放弃将目标文本所代表的试题插入专业题库，如果相似度集中的最大值小于阈值 F，我们将目标文本所代表的试题插入目标文本，在本实例中阈值 F 的取值范围为 0.95。

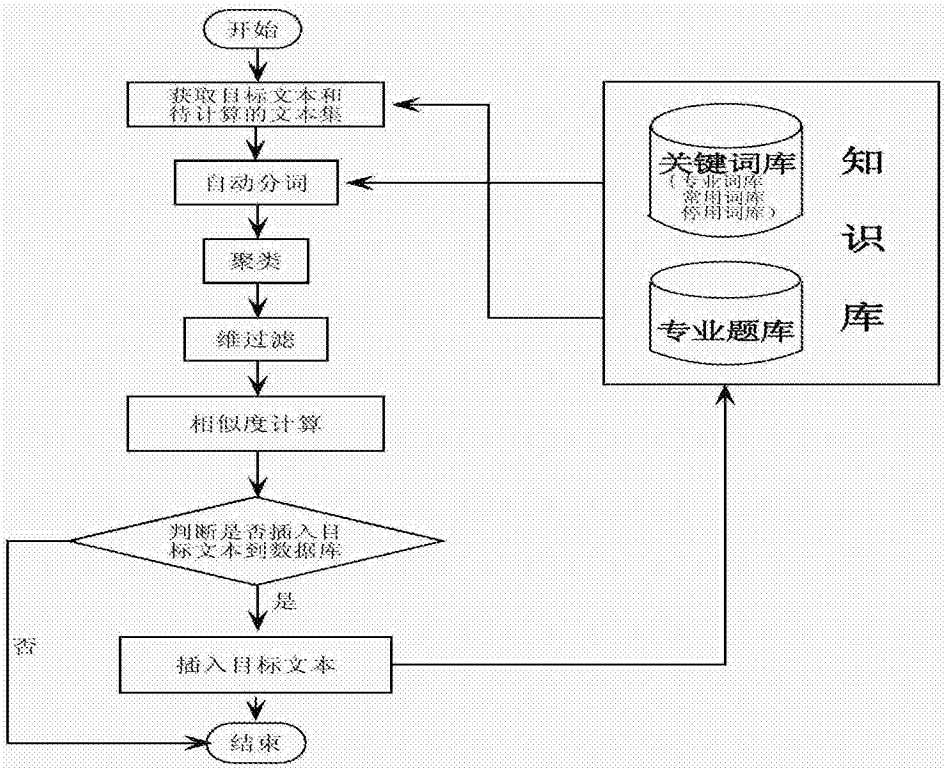

开始
获取目标文本和待计算的文本集
自动分词
聚类
维过滤
相似度计算
判断是否插入目标文本到数据库
是
否
插入目标文本
结束
关键词库
（专业词库
常用词库
停用词库）
专业题库
知识库

图 1

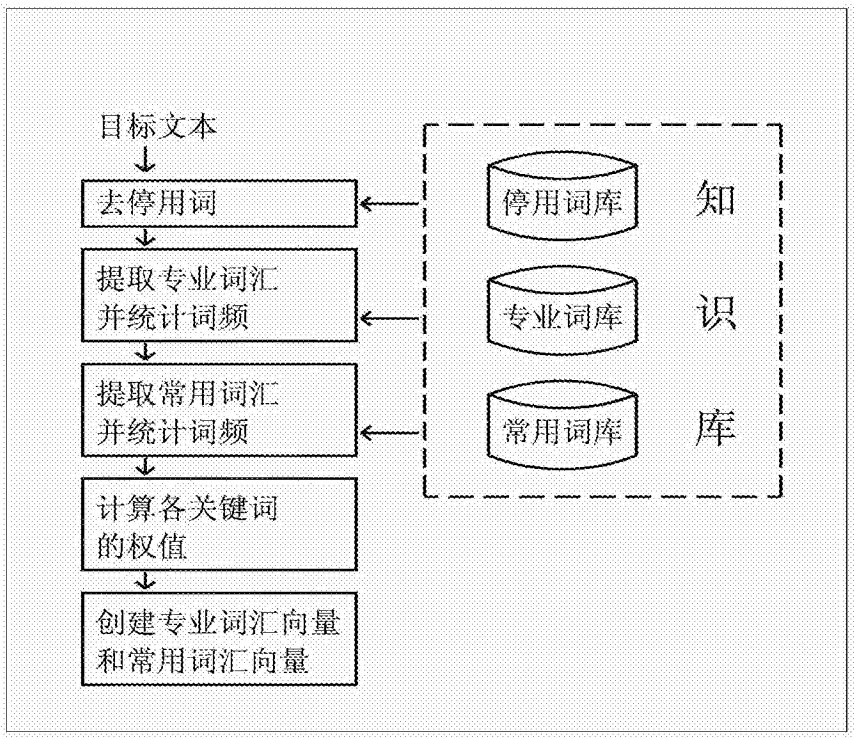

目标文本
去停用词
提取专业词汇并统计词频
提取常用词汇并统计词频
计算各关键词的权值
创建专业词汇向量和常用词汇向量
停用词库
专业词库
常用词库
知识库

图 2