(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104866573 A
(43)申请公布日 2015.08.26

(21)申请号 201510267849.X

(22)申请日 2015.05.22

(71)申请人 齐鲁工业大学
地址 250353 山东省济南市西部新城大学科技园

(72)发明人 耿玉水 杨涛 杨振宇

(74)专利代理机构 济南信达专利事务所有限公司 37100
代理人 孟峣

(51)Int.Cl.
G06F 17/30(2006.01)

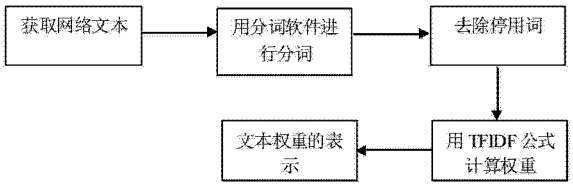
权利要求书1页 说明书5页 附图1页

(54)发明名称

一种文本分类的方法

(57)摘要

本发明公开了一种文本分类的方法，其具体实现过程为：首先获取网络中的文本；对文本进行预处理，提取特征词，对网络中的文本进行分词，然后去除停用词；计算出网络文本中各特征词的权重，并将文本用向量模型进行表示。该文本分类的方法与现有技术相比，具有很强的适应性，能满足大部分不同文本的分类要求，有利于文本分类，实用性强。



CN 104866573 A

1.一种文本分类的方法，其特征在于，其具体实现过程为，

首先获取网络中的文本；

对文本进行预处理，提取特征词，对网络中的文本进行分词，然后去除停用词；

计算出网络文本中各特征词的权重，并将文本用向量模型进行表示。

2.根据权利要求1所述的一种文本分类的方法，其特征在于，所述特征词的选取过程为：

构造评估函数，对特征集合中的每个特征进行评估，并对每个特征打分，使每个词语都获得一个评估值，即权值；

然后将所有特征按权值大小排序；

提取预定数目的最优特征作为提取结果的特征子集。

3.根据权利要求1所述的一种文本分类的方法，其特征在于，所述特征词的权重计算通过改进的TFIDF算法完成，该改进的TFIDF的算法中加入一个可变常量，来对选取的特征词的权重进行调整，剔除干扰特征性在内间的影响，达到为选取的特征词赋予更加合适的权重，提高文本分类的精确度。

4.根据权利要求3所述的一种文本分类的方法，其特征在于，所述改进的TFIDF的算法的具体内容为：

IDF = logn×log(N/(n+k)+0.01)，其中 $n \in N^+$, 求 $n+k \neq 0$；

其中，总文档文本数为N，包含特征词条 t_i 的文档数为n，k为任意参数，该k为上述可变常量，对选取的特征词 t_i 的权重进行调整，在该公式中，当含特征词条 t_i 的文档数为n逐渐增大时，特征词 t_i 的文档区分能力逐渐增强，当n达到某一值时，特征词 t_i 的文档区分能力应随着n的增大而逐渐减少，在IDF公式中，IDF先增后减，且n趋向于1和n趋向于N时，IDF都趋向于0。

5.根据权利要求4所述的一种文本分类的方法，其特征在于，所述改进的TFIDF的算法中还增加一个类内离散度的新的权值来观察所选特征词在类内的分布情况，该类内离散度CD的计算公式如下：

$$CD=\frac{\sqrt{\frac{\sum_{j=1}^{n}(tf_{ij}-\overline{tf}_{ij})^2}{m-1}}}{\overline{tf}}$$

其中，$\overline{tf}_{ij}=\frac{1}{n}\sum_{j=1}^{n}tf_{ij}$；m为类内总的文档数，$tf_{ij}$表示特征词 t_i 在第j篇中出现的次数；$\overline{tf}$ 是特征词 t_i 在类内各个文档中出现的次数的平均值；当类内的离散度CD取1或接近于1的值时，表示特征词只在少数的文档中出现，其分类能力差；当类内离散度取0或接近于0的值时，表示特征词在类内文档中每篇文档的TF值相等或大致相等，其分类能力好。

一种文本分类的方法

技术领域

[0001] 本发明涉及云计算大数据技术，具体地说是一种实用性强的文本分类的方法。

背景技术

[0002] 随着网络技术的快速发展，海量的信息资源以文本的形式存在。人们迫切的希望能从爆炸式的信息浪潮中快速有效的找到自己感兴趣的内容。文本分类作为信息处理的重要研究方向，是解决文本信息发现的常用方法。在文本分类的过程中，关键词的权重起到决定性的作用，它能快速反映一篇文档主题内容或与文档所在领域高度相关的词语，帮助人们在搜寻所需的信息时能够迅速地定位到相应的文档。

[0003] 目前获取关键词或特征词的方式有 4 种 :(1) 用映射或变换的方法把原始特征变换为较少的新特征 ;(2) 从原始特征中挑选出一些最具代表性的特征 ;(3) 根据专家的知识挑选最有影响的特征 ;(4) 用数学的方法进行选取，找出最具分类信息的特征，这种方法是一种比较精确的方法，人为因素的干扰较少，尤其适合于文本自动分类挖掘系统的应用。

[0004] 针对该数学的方法，在国外 1973 年，Salton 结合了 JONES K S 的思想首次提出了 TFIDF(Term Frequency&Inverse Documentation Frequency) 算法。此后他又多次论证了该算法在信息检索中的有效性，并在 1988 年将特征词和权重运用到文献检索中，并详细阐述了实验的情况，进而他得出 TFIDF 算法具有以下思想 : 如果某个词或短语在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类 ;一个词在一篇文档中出现的范围越广 , 说明它区分文档内容的属性越低 (IDF)。1999 年 Roberto Basils 提出了改进的 TF×IWF×IWF 算法，该算法提高了在大量文档出现的频率较低的特征词的权重，有利于多文档的区分，但是它没有考虑到当某一类文档在总文档数所占比例较高时，对该类文档进行区分时，无形降低了可以区分该类文档的特征词的权重，导致该类文档不能有效区分出来 ;另一方面，该算法大大提高了一些不具有区分能力单出现次数较少的特征词权重。因此该算法适用范围具有一些局限性。2004 年，Bong Chih How 和 Narayanan K 根据不同类别的文档数可能存在数量级的差距提出了用 Category Term Descriptor(CTD) 来改进 TFIDF，以解决了不同类别的文档数目对 TFIDF 算法的影响。

[0005] 在国内，也有很多研究学者对 TFIDF 算法进行研究和改进，且取得了很多显著的成果。2006 张玉芳等人为了解决特征性项在类间和类内的分布情况，对 TFIDF 公式进行了修改。该算法考虑到了特征项在内间的分布情况，提高了那些在某一类文档大量出现而在其他类文本含量较少的特征项的权重，能够较精确的区分出该类文档。但当某一类文本 c 所含关键特征项 t 的文档数量较小时，其关键特征项 t 的权重会随着其他类中包含特征项 t 的文档数量的增大而减小 (在一定程度内，如果不含关键特征项的文本数量过大，特征词 t 也就不能成为区分文本的关键词)，故有一定的局限性 ;同时该算法只考虑到特征项在内间的分布情况而没注意到其在类内的分布情况。

[0006] 更为具体的，现有的 TFIDF 算法存在以下不足 :

[0007] 1)IDF 没有考虑到特征词在内间的分布信息。

[0008] 如果某一类 c_i 中包含词条 t 的文档数为 m，而其它类包含 t 的文档总数为 k，显然所有包含 t 的文档数 n = m+k，当 m 大的时候，n 也大，按照 IDF 公式得到的 IDF 的值会小，则表示该词条 t 类别区分能力不强。但是实际上，m 大，说明词条 t 在 c_i 类的文档中频繁出现，就说明 t 词条能够很好地代表 c_i 类的文本特征，应该赋予较高的权重并选作该类文本的特征词。这就是 IDF 没有考虑特征词在类间分布的一个方面；另一方面，虽然包含 t 的文档数 n 较小，但是如果其均匀分布在各个类间，这样的特征词不适合用来分类，应该赋予较小的权重，可按照传统的 TFIDF 算法计算其 IDF 值却很大。

[0009] 2)TFIDF 没有考虑特征词不完全分类的情况。

[0010] 实际使用的已分类的训练文本集通常是不完全的分类。即有些类别的文档集还可以继续划分出更细的类别。如，计算机类一般来说至少可以再细分出计算机硬件、计算机软件两个子类。在这种不完全的分类条件下，各个子类文章所占的比重是不均衡的。可能在某个计算机类的文本集中，软件类的占了 80％，硬件类的只有 20％的比例。在这个训练集中，属于计算机硬件类的特征词也应该作为判别计算机类文章的特征词。如果某些词在一类文章中整体出现频率较低，但是在本类中一定数量的文章中出现较频繁，那么这些词也应该对分类来说具有较多的信息量。这就是不完全分类的情况。

[0011] 3)TFIDF 没有考虑特征词在类内的分布信息。

[0012] 同样是集中分布于某一类别的不同特征项，类内分布相对均匀的特征项的权重应该比分布不均匀的要高。

[0013] 基于此，现提供一种基于改进的 TFIDF 算法的文本分类的方法，该方法结合文本分类的实际情况，结合传统的特征词权重的计算方法，分析了传统 TF-IDF 算法在特征词权重计算上的不足，即传统的 TFIDF 算没有考虑特征词在类内和内间的分布，导致一些区分度不强的特征词赋予了较大的权重。针对传统的 TFIDF 算法的不足，结合特征词权重对文本分类的实际影响，本发明对传统 TFIDF 算法公式进行了修改，剔除干扰特征性在内间的影响，同时加入了类内离散度的概念，实现了文本分类精确度的要求。

发明内容

[0014] 本发明的技术任务是针对以上不足之处，提供一种实用性强、文本分类的方法。

[0015] 一种文本分类的方法，其具体实现过程为：

[0016] 首先获取网络中的文本；

[0017] 对文本进行预处理，提取特征词，对网络中的文本进行分词，然后去除停用词；

[0018] 计算出网络文本中各特征词的权重，并将文本用向量模型进行表示。

[0019] 所述特征词的选取过程为：

[0020] 构造评估函数，对特征集合中的每个特征进行评估，并对每个特征打分，使每个词语都获得一个评估值，即权值；

[0021] 然后将所有特征按权值大小排序；

[0022] 提取预定数目的最优特征作为提取结果的特征子集。

[0023] 所述特征词的权重计算通过改进的 TFIDF 算法完成，该改进的 TFIDF 的算法中加入一个可变常量，来对选取的特征词的权重进行调整，剔除干扰特征性在内间的影响，达到

为选取的特征词赋予更加合适的权重，提高文本分类的精确度。

[0024] 所述改进的 TFIDF 的算法的具体内容为：

[0025] IDF = logn×log(N/(n+k)+0.01)，其中 n ∈ N^+，求 n+k ≠ 0；

[0026] 其中，总文档文本数为 N，包含特征词条 t_i 的文档数为 n，k 为任意参数，该 k 为上述可变常量，对选取的特征词 t_i 的权重进行调整，在该公式中，当含特征词条 t_i 的文档数为 n 逐渐增大时，特征词 t_i 的文档区分能力逐渐增强，当 n 达到某一值时，特征词 t_i 的文档区分能力应随着 n 的增大而逐渐减少，在 IDF 公式中，IDF 先增后减，且 n 趋向于 1 和 n 趋向于 N 时，IDF 都趋向于 0。

[0027] 所述改进的 TFIDF 的算法中还增加一个类内离散度的新的权值来观察所选特征词在类内的分布情况，该类内离散度 CD 的计算公式如下：

[0028] $$CD=\frac{\sqrt{\frac{\sum_{j=1}^{m}(tf_{ij}-\overline{tf})^2}{m-1}}}{\overline{tf}}$$

[0029] 其中，$\overline{tf_{ij}}=\frac{1}{n}\sum_{j=1}^{n}tf_{ij}$；m 为类内总的文档数，$tf_{ij}$表示特征词 t_i在第 j 篇中出现的次数；$\overline{tf}$是特征词 t_i在类内各个文档中出现的次数的平均值；当类内的离散度 CD 取 1 或接近于 1 的值时，表示特征词只在少数的文档中出现，其分类能力差；当类内离散度取 0 或接近于 0 的值时，表示特征词在类内文档中每篇文档的 TF 值相等或大致相等，其分类能力好。

[0030] 本发明的一种文本分类的方法，具有以下优点：

[0031] 本发明提出的一种文本分类的方法，通实验结果表明，改进的 TFIDF 算法的精确度要高于传统的 TFIDF 算法，而且具有很强的适应性，能满足大部分不同文本的分类要求，有利于文本分类，实用性强，易于推广。

附图说明

[0032] 附图 1 为本发明的实现流程图。

[0033] 附图 2 为本发明中改进后的 TFIDF 算法流程图。

具体实施方式

[0034] 下面结合附图和具体实施例对本发明作进一步说明。

[0035] 本发明提供一种文本分类的方法，该方法中涉及到的名词解释如下：

[0036] TFIDF：TF-IDF 是一种统计方法，用以评估字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

[0037] 特征词：用户在使用搜索引擎时输入的、能够最大程度概括用户所要查找的信息内容的字或者词，是信息的概括化和集中化。一般在搜索引擎优化 SEO 行业谈到的特征词，往往是指网页的核心和主要内容。

[0038] 权重：权重是一个相对的概念，针对某一指标而言。某一指标的权重是指该指标在整体评价中的相对重要程度。权重是要从若干评价指标中分出轻重来，一组评价指标体系相对应的权重组成了权重体系。

[0039] 如附图 1、图 2 所示，其具体实现过程为，

[0040] 首先获取网络中的文本；

[0041] 对文本进行预处理，提取特征词，对网络中的文本进行分词，然后去除停用词；

[0042] 计算出网络文本中各特征词的权重，并将文本用向量模型进行表示。

[0043] 所述特征词的选取过程为：

[0044] 构造评估函数，对特征集合中的每个特征进行评估，并对每个特征打分，使每个词语都获得一个评估值，即权值；

[0045] 然后将所有特征按权值大小排序；

[0046] 提取预定数目的最优特征作为提取结果的特征子集。

[0047] 所述特征词的权重计算通过改进的 TFIDF 算法完成，该改进的 TFIDF 的算法主要解决问题有三个：1) IDF 没有考虑到特征词在内间的分布信息。2) TFIDF 没有考虑特征词不完全分类的情况。3) TFIDF 没有考虑特征词在类内的分布信息。故加入一个可变常量，来对选取的特征词的权重进行调整，剔除干扰特征性在内间的影响，达到为选取的特征词赋予更加合适的权重，提高文本分类的精确度。

[0048] 针对 IDF 没有考虑到特征项在内间的分布信息，我们对 IDF 公式进行了修改，增加了那些在一个类频繁出现的特征项的权重，减小了那些在均匀分布在不同类间的特征项的权重。同时针对 TFIDF 没有考虑到特征项不完全分类的情况，加强 TDIDF 公式对不同文档的适应性，我们引入了训练集和添加了参数 K，根据不同的文档类型调整参数 K 的大小。改进的 IDF 算法为：

[0049] $IDF = \log n \times \log(N/(n+k)+0.01)$，其中 $n \in N^{\wedge}+, n+k \neq 0$；

[0050] 其中，总文档文本数为 N，包含特征词条 t_i 的文档数为 n，k 为任意参数，当含特征词条 t_i 的文档数为 n 非常小，且趋向于 1 时，说明特征词 t_i 的文档区分能力很差，应具有很小的权重，在 IDF 公式中，当 n 趋向于 1 时，IDF 趋向于 0，正好满足；当含特征词条 t_i 的文档数为 n 非常大，且趋向于 N 时，说明特征词 t_i 的文档区分能力很差，应具有很小的权重，在 IDF 公式中，当 n 趋向于 N 时，IDF 趋向于 0，正好满足；当含特征词条 t_i 的文档数为 n 逐渐增大时，特征词 t_i 的文档区分能力应逐渐增强，当 n 达到莫一值时，特征词 t_i 的文档区分能力应随着 n 的增大而逐渐减少，在 IDF 公式中，IDF 先增后减，且 n 趋向于 1 和 n 趋向于 N 时，IDF 都趋向于 0，也正好满足要求。对不同类型的文档进行分类时，相同的特征词 t_i 应该具有不同的权重，因此我们加入一个可变常量 k，对选取的特征词 t_i 的权重进行调整，通过训练集求出最合适的 k 值，达到为选取的特征词 t_i 赋予更加合适的权重，从而提高文本分类的精确度。

[0051] 针对 IDF 没有考虑到特征项在类内的分布信息，所述改进的 TFIDF 的算法中还增加一个类内离散度的新的权值来观察所选特征词在类内的分布情况，该类内离散度 CD 的计算公式如下：

[0052] $$CD=\frac{\sqrt{\sum_{j=1}^{m}(tf_{ij}-\overline{tf_{ij}})^2}}{\overline{tf}}$$

[0053] 其中，$\overline{tf_{ij}}=\frac{1}{n}\sum_{j=1}^{n}tf_{ij}$；m 为类内总的文档数，$tf_{ij}$表示特征词 t_i在第 j 篇中出现的次数；$\overline{tf}$是特征词 t_i在类内各个文档中出现的次数的平均值；当类内的离散度 CD 取 1 或接近

于 1 的值时，表示特征词只在少数的文档中出现，其分类能力差；当类内离散度取 0 或接近于 0 的值时，表示特征词在类内文档中每篇文档的 TF 值相等或大致相等，其分类能力好。

[0054] 当对不同类型的文档进行分类时，相同的特征词 t_i 应该具有不同的权重，因此我们加入一个可变常量 k，对选取的特征词 t_i 的权重进行调整，通过训练集求出最合适的 k 值，达到为选取的特征词 t_i 赋予更加合适的权重，从而提高文本分类的精确度。

[0055] 本发明针对 IDF 没有考虑到特征项在类内的分布信息，增加类内离散度 CD。使得同样是集中分布于某一类别的不同特征项，类内分布相对均匀的特征项的权重应该比分布不均匀的要高。

[0056] 上述具体实施方式仅是本发明的具体个案，本发明的专利保护范围包括但不限于上述具体实施方式，任何符合本发明的一种文本分类的方法的权利要求书的且任何所述技术领域的普通技术人员对其所做的适当变化或替换，皆应落入本发明的专利保护范围。

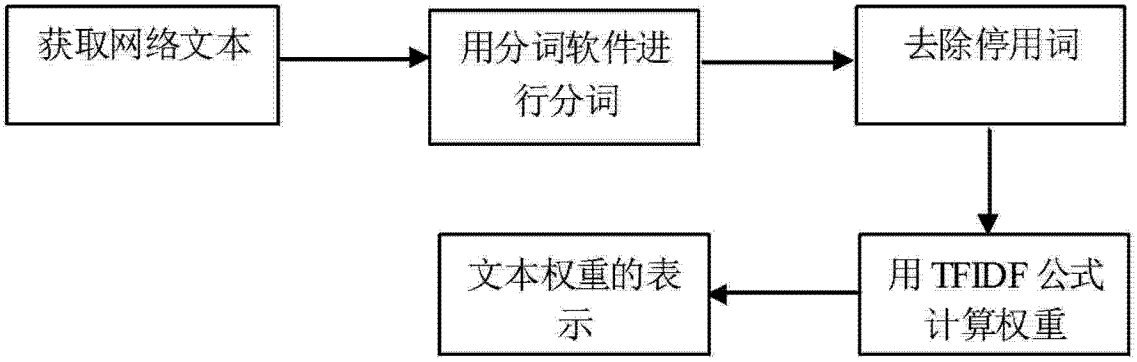

获取网络文本
用分词软件进行分词
去除停用词
用 TFIDF 公式计算权重
文本权重的表示

图 1

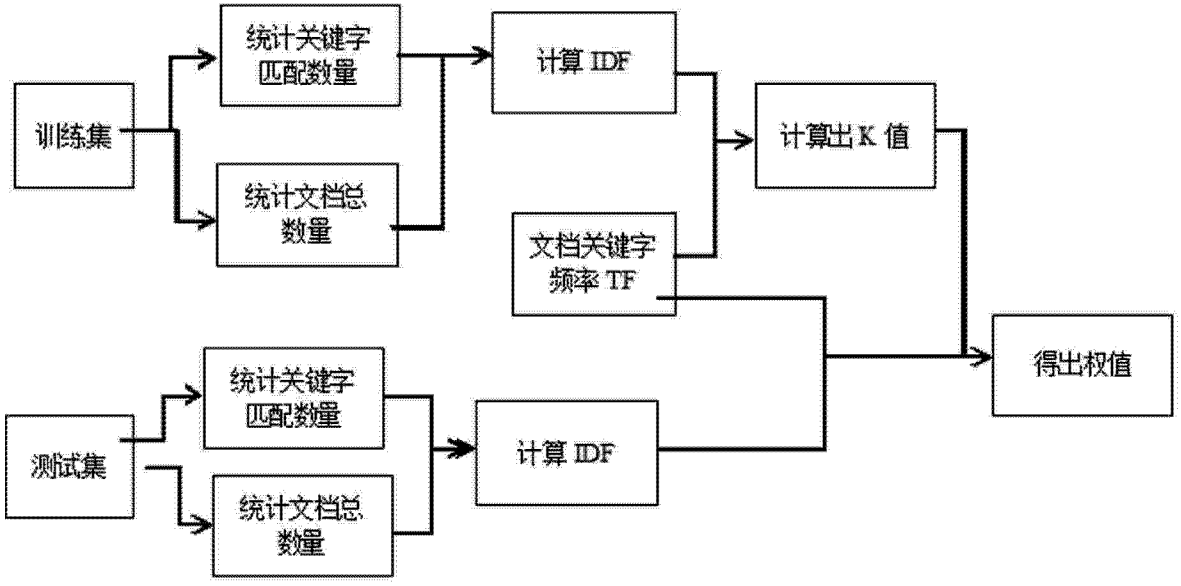

训练集
统计关键字匹配数量
统计文档总数量
计算 IDF
计算出 K 值
文档关键字频率 TF
得出权值
测试集
统计关键字匹配数量
统计文档总数量
计算 IDF

图 2