文章编号:1004—5570(2015)06 -0106 -04

基于 TF -IDF 的网页新闻分类的研究与应用

李春梅

(安徽新华学院 信息工程学院,安徽 合肥 230088)

摘要: 文本分类作为处理和组织大量文本数据的关键技术,为用户准确、快速查找所需信息提供依据。通过 TF-IDF 算法计算文本词汇的词频,并根据词频排序选择特征项,再用 Simhash 和余弦相似度算法计算文本之间的相似度,最后采用准确率和召回率为评价标准,根据评价结果分析两种算法的优劣。

关 键 词: 文本分类; TF-IDF; Simhash; 余弦相似度

中图分类号: TP391.6 **文献标识码:** A

DOI:10.16614/j.cnki.issn1004-5570.2015.06.022

Research and application of TF-IDF classification based news website

LI Chunmei

(Department of Information Engineering, Anhui Xinhua University, Hefei, Anhui 230088, China)

Abstract: With the rapid development of the internet and the increasing popularity, how to quickly and accurately find the information from the large amount of information is the challenge of information science and technology. Text classification is an important technology of the process and organize large amounts of text data for users to accurately and quickly find the information to provide a basis. In this paper, TF-IDF algorithm calculate text vocabulary word frequency, and select features based on word frequency sorting items, and then, select the similarity Simhash and cosine similarity algorithm calculate the similarity between the text. Finally, the precision and recall rate of evaluation criteria to analyze the merits of the two algorithms based on evaluation results.

Key words: text categorization; TF-IDF; simhash; cosine similarity

0 引言

文本分类作为处理文本数据的关键技术,可以解决信息杂乱现象的问题。而且作为信息过滤、检索、搜索引擎、文本数据库、数字图书馆等领域的技术基础,文本分类技术的应用前景非常广泛[1-3]。

通过对目前文本分类技术进行研究,以新闻网页的分类为背景,研究基于 TF-IDF 的新闻分类方法。并在 TF-IDF 的研究基础上对采用余弦相似度和 Simhash 算法计算相似度的分类方法进行对比。

1 相关工作

1.1 文本分类

词匹配法[4,5]是最早被提出的文本分类方法。目前统计学习方法已成为文本分类领域绝对的主

收稿日期:2015 -06 -07
基金项目:安徽省高等学校自然科学研究项目(KJ2015A309)
作者简介:李春梅(1978 -),女,讲师,研究方向:数据挖掘,E - mail:437385999@qq.com.

流。统计分类算法将样本数据成功转化为向量表示,计算机通过对规则的学习进行分类。常用的分类算法[6]有:遗传算法,朴素贝叶斯,最大熵,线性最小平方拟合,支持向量机,KNN,决策树,Rocchio,神经网络等。

1.2　TF-IDF 算法

TF-IDF[7,8]是一种统计方法,用来评估一个字词对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,但同时会与它在语料库中出现的频率成反比。

在一份给定的文本中,词频(term frequency,TF)指的是一个词在该文件中出现的次数。逆向文件频率(inverse document frequency,IDF)[9,10]是对词重要性的度量。一个词的 IDF,可以由文件总数除以包含该词的文件数目,再将得到的商取对数得到。若某一词语在一个文件内部的词频较高,而在整个文件集中的词频较低,则这个词语的权重较大。因此,TF-IDF 可以过滤常词语,保留文档中可以突出文档特征的词语。

2　基于 TF-IDF 的网页新闻分类方法

2.1　问题描述

基于 TF-IDF 的网页新闻分类主要采用 IF-IDF 算法进行词频的计算,再根据词频的统计结果选择词频最高的词集合做特征项,通过余弦相似度和 Simhash 算法计算新闻文本之间的相似度。

2.2　算法步骤

输入:经过分词的文本

输出:每个词的 TF-IDF 值,取前 TOP-N 个词作为特征词。

Setp1:统计每个词在文本中出现的次数,计算词的 TF 值。

Setp2:统计每一个词在多少个文本中出现,计算词的 IDF 值。

Setp3:计算每一个词的 TF* IDF 值。

其中,TF 值的计算方法如公式(1)所示:

$$tf_{i,j} = \frac{n_{i,j}}{\sum T_j} \tag{1}$$

$tf_{i,j}$ 是 j 这篇文档中 i 词语的词频,$n_{i,j}$ 是 i 在 j 文档中出现的次数,T_j 代表 j 文档中的每一个词语。

IDF 值的计算方法如公式(2)所示:

$$idf_i = \log \frac{|D|}{\sum t_i \in d + 1} \tag{2}$$

idf_i 代表 i 这个词汇的逆向文件频率,D 表示文档的个数,t_i 表示 i 词汇,d 表示文档集。

TF-IDF 值的计算方法如公式(3)所示:

$$tf - idf_i = tf_{i,j} * idf_i \tag{3}$$

2.3　文本相似度计算

通过 TF-IDF 的计算,获取到每篇文档的前 Top-N 个特征词,将这些特征词构建成一个一维向量,每一篇文档都用一个一维向量表示,两个向量之间使用余弦相似度计算出来的值小于等于 1,即为两个文本之间的相似度。值越大,两个文本越相似。余弦相似度[11,12]的计算如公式(4)所示:

$$sim(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \overline{R}_u)(R_{u,j} - \overline{R}_U)}{\sqrt{\sum_{u \in U}(R_{u,i} - \overline{R}_U)^2}\sqrt{\sum_{u \in U}(R_{u,j} - \overline{R}_u)^2}} \tag{4}$$

$sim(i,j)$ 代表 i 文本和 j 文本的相似度,$R_{u,i}$ 代表 i 文本中的 u 特征词的权重。

但由于一篇文章的特征词较多而导致整个一维向量特征过多,使得计算的代价太大,对于万亿级别的网页的搜索引擎来说是不可接受的,所以也可以采用 Simhash 算法[13,14]进行相似度计算,该算法的主要思想是,将高维的特征向量映射成一个 F-bit 的指纹(fingerprint),通过比较两篇文章 F-bit 指纹的海明距离来计算两篇文本的相似性。

3　实验与分析

实验采用数据堂提供的新浪新闻数据集。数据集包含了 2 000 个网页新闻文本。

3.1　实验步骤

根据之前的分析和选择的数据集进行文本分类实验,实验步骤如下:

Step 1:利用结巴分词软件对文本进行分词;

Step 2:去除停用词。如:的,得,地等停用词对计算 TF-IDF 产生噪音;

Step 3:通过公式(1)(2)(3)计算每个词的 TF-IDF 值;

Step 4:取 TF-IDF 前 TOP-N 的词,用余弦相似度和 Simhash 算法进行对比实验,验证两种方法的使用场景。

3.2 评价标准

实验评价标准主要采用准确率(P)和召回率(R)。准确率和召回率同为检索系统中的两大基本指标。准确率,又称"精度""正确率"。是符合检索条件的文档数量与总文档数量的比率。

准确率的计算公式(5)所示: $P = \frac{d}{T}$ (5)

d 为相似的文档数目,T 是数据集中文档总数。

召回率(Recall)衡量的是检索系统的查全率。是符合检索条件的文档数量和数据集中该类别文档集中文档业务量的比率。

召回率的计算公式(6)所示: $R = \frac{d}{D}$ (6)

d 为相似的文档数目,D 是该分类中的文档数目。

3.3 实验和分析

实验取 TF-IDF 前 TOP-N,N 取(50,40,30,20,10)进行。通过实验来验证余弦相似度算法和 Simhash 算法在文本分类中应用的差异。

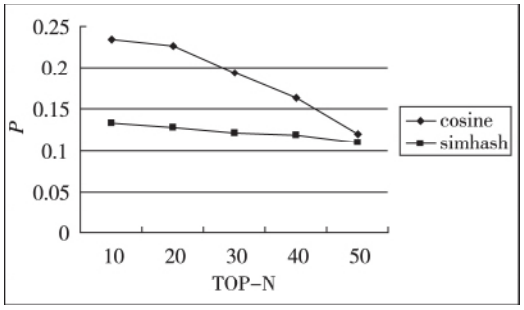

图1 正确率曲线

Fig. 1 The curve of correct ratio

从图1中可以看出,在提取特征词较少时,余弦相似度的准确率高,但随着特征词数量的增加,余弦相似度与 Simhash 的准确率比较接近。

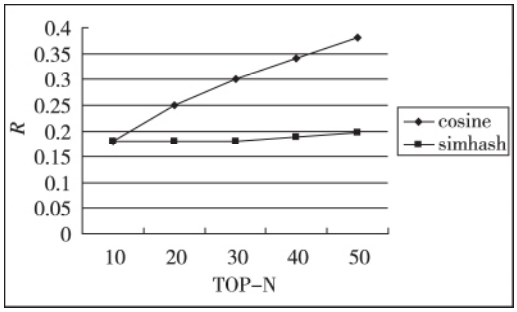

图2 召回率曲线

Fig. 2 The curve of recall ratio

从图2中可以看出,在提取特征词较少时,余弦相似度的召回率与 Simhash 的召回率比较接近。但随着特征词数量的增加,余弦相似度的召回率在上升,但 Simhash 方法的召回率变化平稳,故与特征词的选取关系不大。

从实验结果中可以看出,余弦相似度算法更适合于新闻文本的分类。

4 结果与讨论

通过 TF-IDF 计算新闻文本中的词频,并在计算 TF-IDF 的基础上,采用余弦相似度和 Simhash 方法进行对比实验,得出以下结论:

1) 在选取特征词较少时,余弦相似度比 Simhash 准确率高,但2种方法召回率接近;

2) 随着特征词增加时,余弦相似度与 Simhash 准确率接近,但余弦的召回率高于 Simhash 方法;

3) 基于 TF-IDF 的余弦相似度算法更适用于网页新闻分类。

经分析,产生以上结论的原因主要有:

1) Simhash 受特征词选取数量影响较小;

2) 余弦相似度比 Simhash 方法更适合于小文本量的稀疏阵列。

参考文献:

[1] 李荣陆. 文本分类及其相关技术研究[D/OL]. 上海:复旦大学,2005[2015-03-04]. http://wenku.baidu.com/link?url=KoqtdpdA38OWRyLQysYdad2evajjluW2DTMdJ3BR_7TItS9ytXY3H9uPFu9ZNsmTfkKmmLTSTz4ItoAlx36Bwoa6FbeX0g0utW7DURwESrq.

[2] 孙强,李建华,李生红. 基于 Python 的文本分类系统开发研究[J]. 计算机应用与软件,2011,28(3):13-14.

[3] 齐鹏,李隐峰,宋玉伟. 基于 Python 的 Web 数据采集技术[J]. 电子科技,2012,25(11):118-120.

[4] 黄昌宁,赵海. 中文分词十年回顾[J]. 中文信息学报,2007,21(3):8-19.

[5] 王勇. 中文文本分类特征选择和特征加权方法研究[D/OL]. 重庆:重庆大学,2012:1-2[2015-05-08]. http://cdmd.cnki.com.cn/Article/CDMD-10611-1012049124.htm.

[6] 余苗,杨瑞娟,程伟,等. 基于 TF-IDF 分类算法的雷达情报分发技术[J]. 计算机工程与设计,2012,33(5):1822-1826.

[7] 王园,龚尚福. 基于二次 TF* IDF 的互信息文本特征选择算法研究[J]. 计算机应用与软件,2011,28(4):129-131.

[8] 黄承慧,印鉴,侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法 [J]. 计算机学报,2011,34(5):856-864.
[9] 徐建民,王金花,马伟瑜. 利用本体关联度改进的 TF-IDF 特征词提取方法 [J]. 情报科学,2011,29(2):279-283.
[10] 罗欣,夏德麟,晏蒲柳. 基于词频差异的特征选取及改进的 TF-IDF 公式 [J]. 计算机应用,2005,25(9):2031-2033.
[11] 吴桂玲. 基于欧氏距离和余弦相似度特征选择的入侵检测模型 [J]. 中小企业管理与科技,2010(2):231-232.
[12] 张祖平,徐昕,龙军,等. 文本相似性度量中参数相关性与优化配置研究 [J]. 小型微型计算机系统,2011,32(5):983-988.
[13] 马成前,毛许光. 网页查重算法 Shingling 和 Simhash 研究 [J]. 计算机与数字工程,2009,37(1): 15-17.
[13] 张春霞,郝天永. 汉语自动分词的研究现状与困难 [J]. 系统仿真学报,2005,17(1):138-143,147.
[14] 董博,郑庆华,宋凯磊,等. 基于多 SimHash 指纹的近似文本检测 [J]. 小型微型计算机系统,2011,32(11):2152-2157.

(上接第 96 页)

[2] 李琨. 水平受荷嵌岩桩的静载试验研究及有限元分析 [D/OL]. 太原: 太原理工大学,2009: 6 [2015-10. 21]. http://www.cnki.net/KCMS/detail/detail.aspx?QueryID=3&CurRec=1&recid=&filename=2010047379.nh&dbname=CMFD2011&dbcode=CMFD&pr=&urlid=&yx=&uid=WEEvREcwSlJHSldRa1FiNk5TcnlDaUNiQ0NqR1JLNnEvaWt4NE5CdHFuNE1hMG44V2lpWTliTmkrcnFGVmNEejdnPT0=MYM9A4hF_YAuvQ5obgVAqNKPCYcEjKensW4IQMovwHtwkF4VYPoHbKxJw!!&v=MzAzMDM2SHJPOEdkTExwcEViUElSOGVYMUx1eFlTN0RoMVQzcVRyV00xRnJDVVJMK2VaK1J1RkNqaFVMektWMTI=
[3] 王建华,陈锦剑,柯学. 水平荷载下大直径嵌岩桩的承载特性研究 [J]. 岩土工程学报,2007,29(8): 1194-1198.
[4] 劳伟康,周立运,王钊,大直径柔性钢管嵌岩桩水平承载力试验与理论分析 [J]. 岩石力学与工程学报,2004,23(10):1770-1777.
[5] 贵州省地方志编纂委员会. 贵州省志. 科学技术志 [M]. 贵阳: 贵州人民出版社,1992: 37-38.
[6] 陈明祥. 弹塑性力学 [M]. 北京: 科学出版社,2007: 355-356.
[7] 钱家欢,殷宗泽. 土工原理与计算(第二版) [M]. 北京: 中国水利水电出版社,1996: 60-64.
[8] 谭峰屹,汪稔,赵丽. 柔性桩复合地基承载力数值计算 [J]. 岩土力学,2011,32(1):288-292.
[9] 赖永标,胡仁喜,黄书珍. ANSYS11. 0 土木工程有限元分析典型范例 [M]. 北京: 电子工业出版社,2007: 40.
[10] 何周. 基于 ANSYS 的结构内力计算方法 [J]. 科技信息,2011(31):5,36.
[11] 陈魁. 应用概率统计 [M]. 北京: 清华大学出版社,2000: 262-275.