基于改进权重的贝叶斯推理和 TFIDF 算法文本主题词提取研究

邵晓根[1],鞠训光[1],胡局新[1],马忠伟[2]

(1. 徐州工程学院信电工程学院,江苏 徐州 221111)
(2. 湘潭大学信息工程学院,湖南 湘潭 411105)

[**摘要**] 本文针对中文文本主题词提取的 TFIDF 算法不足进行了改进,综合考虑关键词在文本中出现的频率及位置权重,设计了贝叶斯推理和 TFIDF 主题词提取混合算法,并基于候选词排序位置进行了正向、逆向和中间向前后的提取测试,结果表明,本算法比单纯 TFIDF 算法正向提取平均准确率提高了 6.2%.

[**关键词**] 贝叶斯推理,位置权重,主题词提取,TFIDF 算法
[**中图分类号**]TP391;TP301　[**文献标志码**]A　[**文章编号**]1001-4616(2014)01-0057-04

Research of Text Subject Extraction Based on Improved Weight for Bayesian Reasoning and TFIDF Algorithm

Shao Xiaogen[1], Ju Xunguang[1], Hu Juxin[1], Ma Zhongwei[2]

(1. Department of Information and Electrical Engineering, Xuzhou Institute of Technology, Xuzhou 221111, China)
(2. College of Information Engineering, Xiangtan University, Xiangtan 411105, China)

Abstract: The shortcoming of the TFIDF algorithm is improved for Chinese text topic word extraction. This paper considers the keywords appearing frequency, position weight in the text, the hybrid algorithm of Bayesian Reasoning and TFIDF was designed to extracte topic words, and the topic words was extracted from forward, reverse and middle based on sorting position of the candidate words. The results was higher average accuracy than the simple TFIDF by 6.2%.

Key words: Bayesian reasoning, position weight, topic words extraction, TFIDF algorithm

汉语文本主题词提取是自然语言理解处理的基础. 主题词常应用于文摘、索引、分类、聚类和检索查重等方面. 目前主题词的提取方法主要有基于词典、基于规则和基于统计的提取方法,三类方法各有优缺点. 基于统计的提取方法最为流行,已经取得了较好的研究及应用[1,2].

基于统计的方法是通过构造评估函数,对特征集合中的每个特征进行评估,并对每个特征打分,这样每个词语都获得一个评估值,又称为权值. 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. 显然,决定文本特征提取效果的主要问题是评估函数的选取及质量.

基于统计的特征提取方法目前已有的算法有:由 Salton 在 1988 年提出的 TF-IDF、词频方法、互信息、信息增益、交叉熵和主成分分析法等方法.

上述几种评价函数都是试图通过概率找出特征与主题词之间的联系,信息增益的定义过于复杂;互信息的效果要好于交叉熵,这是因为互信息是对不同的主题类分别抽取特征词,而交叉熵与特征在全部主题类内的分布有关,是对全部主题类来抽取特征词. 这些方法,在英文特征提取方面都有各自的优势,但用于中文文本效率不高. 主要存在 2 个方面的原因:(1)特征提取的计算量太大,效率太低;(2)经过特征提取后生成的特征向量维数太高,而且不能直接计算出特征向量中各个特征词的权重.

所以,本文综合考虑候选词语位置和频率信息的改进权重 TFIDF 方法[1],并将其应用于市级科技项

收稿日期:2013-07-13.
基金项目:科技部国家中小企业创新基金项目(11C26213204533)、徐州市科技计划项目(XF11C052).
通讯联系人:鞠训光,博士,副教授,研究方向:智能计算、数据挖掘、云计算. E-mail:375768447@qq.com

目主要研究内容的主题词提取.由于单纯应用TFIDF算法提取的主题词准确率较低,考虑申报者对科技项目主要研究内容的写作方法和技巧,表征其核心研究内容的主题词位置因人而异,受关键词语在文本中出现的位置影响较大,采用将每一个待检测关键词作为贝叶斯先验概率[2-4],利用贝叶斯推理求得每一个待检测关键词的后验概率,以提高主题词的提取准确率.

1　改进TFIDF算法

1.1　权值函数设计

考虑主题词词频和位置两个因素,权值函数构造如下:

$$\mathrm{weight}_i=\alpha\times\mathrm{fre}_i+\mathrm{loc}_i, \tag{1}$$

其中,weight_i为候选词i的权重;fre_i为词频权重因子;loc_i表示位置权重因子;α为调节因子.

1.2　算法各因子计算方法

(1)词频因子

$$\mathrm{fre}_i=\frac{f_i}{1+f_i}, \tag{2}$$

其中,f_i为候选词i在项目主要研究内容中出现的频次.

(2)位置因子

位置因子根据文献[4,5]中采用的候选词在文本中第一次出现的位置确定其权重值,同时考虑便于计算,测试结果表明取值为以下值时,主题词提取准确率较好:文本前1/3的候选词权重设定为1.5,中间权重设定为0.5,后1/3的权重设定为1.0.

(3)调节因子

经过试验发现,词语之间的词频因子影响不是很强,影响其关联关系较强的是位置因子,测试时采用文献[5]中的LMS法则计算α.

2　贝叶斯推理和TFIDF混合算法

2.1　混合算法思想及流程

单独应用TFIDF算法求得的主题词准确率较低,所以考虑将每一个待检测候选词作为贝叶斯先验概率求解,再利用贝叶斯推理求得每一个待检测关键词的后验概率,计算的结果能够在很大程度上提高主题词的提取准确率.算法流程如图1.

图1　混合算法流程

Fig.1　Hybrid algorithm flow

2.2　贝叶斯统计优化算法原理

首先,建立中文关键特征提取的特征搜索模型,设为$Z(x,y)$,根据语义特征关键模型,可以提取中文文本中待搜索的关键特征[6].

建立中文特征的搜索模型之后,需根据建立的阀值完成主题词的智能抽取,阀值判断方法如下[4]:

$$\begin{cases}(Z(x,y)+T)/\log T\geqslant 0.5, & \text{提取 } T,\\ (Z(x,y)+T)/\log T<0.5, & \text{不提取 } T.\end{cases} \tag{3}$$

把待检测的候选词关键特征进行空间映射,形成集合$T=\{T_1,T_2,\cdots,T_s\}$,其中T_i表示第i个中文语义关键特征信息,s为待检测候选词特征集合.

根据候选词特征权值计算结果,进行如下判断:

$$E(l)=\sin x/\sum_{j=2}^{s}\log T. \tag{4}$$

应用贝叶斯弥补TFIDF算法的不足,本算法计算相似特征的方法如下:

$$S=\sqrt{(\boldsymbol{t}_1-\boldsymbol{p}_1)^2w_1+(\boldsymbol{t}_2-\boldsymbol{p}_2)^2w_2+\cdots+(\boldsymbol{t}_i-\boldsymbol{p}_i)^2w_i},$$

其中,$\boldsymbol{t}_i$ 为 TFIDF 算法计算得到的样本特征向量;$\boldsymbol{p}_i$ 为文本特征库的特征向量;w_i 为 TFIDF 算法计算得到的权值. 但是,这些特征的相似度没有形成一定的关联,存在着偏差. 因此,采用贝叶斯决策思想统计文本特征之间的关联程度,考虑将某些文本特征的权值距离适当偏移,然后将其变化的因素进行概率的相似度计算. 不妨假设,候选词文本的特征量设为 I_1 和 I_2,计算贝叶斯统计下的相似度计算公式为:

$$S(I_1,I_2)=P(\Delta\in\Omega_i)=P(\Omega_i|\Delta), \tag{5}$$

其中,Ω_i 是文本特征中的一种变化程度的度量方式,表示候选词的变化量,也称同类变化量,Ω_E 表示候选词特征异类的文本信息变化量[3,4],也称异类变化量. $P(\Omega_i|\Delta)$ 表示同类后验概率,可通过 $P(\Delta|\Omega_i)$ 和 $P(\Delta|\Omega_E)$ 计算得到[3,4]. 如此,可很好地表达不同类别下的特征关联程度,即文本的变化关联程度. 融合贝叶斯算法后,计算公式为:

$$S(I_1,I_2)=\frac{P(\Delta|\Omega_i)P(\Omega_i)}{P(\Delta|\Omega_i)P(\Omega_i)+P(\Delta|\Omega_E)P(\Omega_E)}. \tag{6}$$

从式中可知,文本的特征提取识别问题就转化成了计算相关概率的问题. 即只需求出特征文本空间的特征变化程度 Δ 属于 Ω_i 或者属于 Ω_E 的概率,就可识别出是否为主题词. 当 $P(\Omega_i|\Delta)>P(\Omega_E|\Delta)$ 或 $S(I_1,I_{2i})>1/2$ 时[7,8],就可以判定出提取的词语为该文本的核心关键词语即主题词.

2.3 改进 TFIDF 算法流程

通过贝叶斯统计的方法对 TFIDF 算法进行改进,改进后的算法描述如下:

首先,假定一个主题词的特征集合为 H,对 H 中的特征向量进行贝叶斯统计,计算相关的分布概率. 利用计算得到的贝叶斯统计结果,一个相应的特征向量可以用一个文档特征向量来表示. 因此,在提取一个随机文本的特征主题词的时候,可利用 k 个不同的贝叶斯统计结果[9],完成特征向量的统计. 统计结果可以定义为:

$$H=\{S_1,w_1,S_2,w_2,\cdots,S_n,w_n\}, \tag{7}$$

其中:

$$\begin{aligned}&S_k=(t_{k1},w_{k1},t_{k2},w_{k2},\cdots,t_{kn},w_{kn})\in V(d),\\&w_i=\max(w_{i1},w_{i2},\cdots,w_{in}).\end{aligned} \tag{8}$$

在对候选词进行识别时,同一次输入的文本属于相同的主题. 算法实现如下:

(1)输入需要提取的特征文本信息,利用 TFIDF 公式计算并生成文本特征向量 $\boldsymbol{V}$.

(2)运用贝叶斯统计原理计算某一候选主题词特征值权值 $\boldsymbol{V}=(v_1,w_1,v_2,w_2,\cdots,v_n,w_n)$,并进行相关的概率与相似度计算.

(3)若相似概率极大值大于通过先验知识确定的阈值[10],可以把其加入到候选集合中;否则认为 V 不是一个可以提取的主题词,加入到其他的非主题词特征空间中.

(4)对 H 中的词进行退化处理,若 $t\in H$,t 的权值修改更新为 $w=0.9w$[4],并删除权值低于阈值的候选评语.

(5)按正向、逆向和中间向前后分别计算位置及权值,根据权值进行候选主题词排序,利用贝叶斯相似概率统计的方法,重新计算权值,即 $\text{weight}_i=\alpha\times\text{fre}_i+\text{loc}_i$.

(6)输出 H 中权值最高的几个词作为科技项目主要研究内容的主题词.

3 实验结果

考虑科技项目主要研究内容的写作手法及技巧,其核心研究内容受关键词语在文本中出现的位置影响较大,所以,主题词抽取测试时设计了基于候选词排序的位置,进行了正向、逆向和中间向前后的方法分别进行提取.

实验测试时随机从历史语料库中选取 20 项徐州市科研项目申报信息的主要研究内容,测试中候选词集的生成参见文献[11,12]. 主题词提取准确度综合测试如下:

通过对文本“地震自然灾害预测预报”的逆向顺序权重进行提取的主题词的权值计算结果见表 1.

通过对文本“地震自然灾害预测预报”的中间向两边提取,提取的主题词的权值结果见表 2.

表1　逆向提取权重值计算结果

Table 1　Weight results of reverse extraction

主题词	词频因子	位置因子	权值
传感系统	0.986	1.500	3.472
微电子机械系统	0.875	1.500	3.250
技术	0.745	1.000	2.490
部件	0.667	1.000	2.334
系统	0.750	0.500	2.000
加工	0.488	0.500	1.476

表2　中间向前后权重计算结果

Table 2　Weight results of forward and backward from middle

主题词	词频因子	位置因子	权值
传感器	0.986	1.500	3.472
传感系统	0.875	1.000	2.750
微电子机械系统	0.875	1.000	2.750
系统	0.750	0.500	2.000
部件	0.667	0.500	1.834
地震	0.366	0.500	1.232

分析:表1的逆向权重和表2的中间向前后方向的权重计算表明,将提取出不同的主题词语,即预示着主题词的提取准确率将提高.

根据候选词排序位置,经正向、反向、中间向前后综合抽取测试,主题词抽取结果与单纯正向提取及单纯TFIDF算法[1,2]抽取结果相比,总体准确度平均提高了约6.2%.结果如表3所示.

表3　综合考虑后的测试结果

Table 3　Test results of complex

科研文本	准确度		
	文本正向提取/%	文本正逆向提取/%	文本中间向两边提取/%
地震自然灾害预测预报与应急处理研究	72	82	84
徐州地区农村饮用水安全现状调查研究	95	95	95
徐州生态环境对城市化建设承载能力	73	84	86
新型强力纳米焊割气技术开发	75	78	80
南刺五加悬浮细胞培养及植株再生研究	83	83	83
草莓果产地综合保鲜技术研究	83	85	86
多功能智能矿灯瓦斯报警器	66	76	77
YH型调换绳装置	75	77	79
JW-27X射线荧光能谱仪	82	82	84
高性能水性船舶防腐涂料	80	82	84
徐州社会主义新农村体育建设发展	76	84	87
工程机械培训平台建设	65	75	79
微生物有机肥研究	83	86	88
全自动大面积太阳能系统工程	75	84	86
甲烷传感器调校装置	73	75	75
细胞生理精密显微操作器	80	83	83
新型节能型门窗开发	80	85	88
煤矿工作面液压支架自动控制系统	76	79	81
数字化太阳能微波治疗仪	68	73	75
盐酸右美托咪定及注射液	77	77	81
综合	76.9	81.3	83.1

4　结语

实验结果表明改进的算法对中文主题词提取有很大改善,弥补了算法只考虑候选词频信息对权重的不完全贡献,改进后的算法综合考虑候选词的词频、位置以及候选词排序位置方向等多因素特征,提高了主题词提取的准确度.

[参考文献]

[1]　施聪莺,徐朝军,杨晓江. TFIDF算法研究综述[J]. 计算机应用,2009,6(29):167-170.

[2]　刘兴林,彭宏,马千里. 基于增量词集频率的文本主题词提取算法研究[J]. 计算机应用研究,2010,27(9):3 237-3 238.

(下转第65页)

完成路径规划,达到目标. 本文的单目视觉控制的路径规划系统,可以结合云计算平台应用,以便克服数据准确性差、视频容量大等弱点,使得高速并行计算与海量存储成为可能. 这样该系统除了视频采集设备等无需另外安装高速计算和大规模存储设备. 展望未来,人工智能将普及应用,不再神秘. 智能机器人的整体是一个大家族,可以相互联络沟通.

[参考文献]

[1] 仝敬辰. 移动机器人基于三维激光测距与单目视觉的室内场景认知[D]. 大连:大连理工大学计算机科学与技术学院,2010.

[2] Leonard J J, Durrant-Whyte H F. Mobile robot localization by tracking geometric beacons[J]. IEEE Transactions on Robotics and Automation, 1991, 7(3): 376-382.

[3] Sangwoo M, Unghui L, David H S. Study on real-time obstacle avoidance for unmanned ground vehicles[C]//International Conference on Control, Automation and Systems. Taipei: IEEE, 2010.

[4] DeSouza G N, Kak A C. Vision for mobile robot navigation a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(2): 237-26.

[5] 姚靖靖,邱于兵,敖俊宇. 移动机器人避障路径规划改进人工势场法[J]. 科学技术与工程,2011,13(11):1 671-1 851.

[6] 于红斌,李孝安. 基于栅格法的机器人快速路径规划[J]. 微电子学与计算机,2005,22(6):98-100.

[7] 禹建丽,程思雅,Kroumov V. 一种移动机器人三维路径规划算法[J]. 中原工学院学报,2008,19(2):37-40.

[8] 厉茂海,洪炳熔. 移动机器人同时定位和地图创建的一种新方法[J]. 南京理工大学学报:自然科学版,2006,3(30):302-305.

[责任编辑:黄　敏]

(上接第 60 页)

[3] 饶丽丽,刘雄辉,张东站. 基于特征相关的改进加权朴素贝叶斯分类算法[J]. 厦门大学学报:自然科学版,2012,51(4):682-685.

[4] 刘林. 基于词语权重改进的朴素贝叶斯分类算法的研究与应用[D]. 广州:中山大学软件学院,2009.

[5] 管瑞霞,陆蓓. TFLD:一种中文文本关键词自动提取方法[J]. 机电工程,2010,27(9):123-126.

[6] 李艳美,张卓奎. 基于贝叶斯网络的数据挖掘方法[J]. 计算机仿真,2008,25(2):117-119.

[7] Sarah Petersen, Mari Ostendorf. Assessing the reading level of web pages[C]//Proceedings of Interspeech (poster). Pittsburgh, 2006: 833-836. .

[8] Christopher D Manning, Prabhakar Raghavan, Hinrich Schutze. Introduction to Information Retrieval[M]. Cambridge: Cambridge University Press, 2008: 96-100.

[9] Harry Zhang, Shengli Sheng. Learning weighted naive bayes with accurate ranking[C]//Fourth IEEE International Conference on Data Mining(ICDM'04). Brighton, 2004. DOI: 10. 1109/ICDM. 2004. 10030

[10] 卫洁,石洪波,冀素琴. 基于 Hadoop 的分布式朴素贝叶斯文本分类[J]. 计算机系统应用,2012,212:210-212.

[11] 胡局新,鞠训光. 自学习分词算法在科研项目查重系统中的应用[J]. 科技通报,2013,29(6):14-19.

[12] 胡局新,鞠训光. 基于贝叶斯推理和 TFIDF 算法的中文关键词智能抽取[J]. 微电子学与计算机,2012,29(9):197-200.

[责任编辑:丁　蓉]