

基于贝叶斯推理和 TFIDF 算法的中文关键词智能抽取

胡局新, 鞠训光

(徐州工程学院 信电工程学院, 江苏 徐州 221008)

摘 要: 针对传统的 TFIDF 中文关键词智能抽取模型中, 遇到冗余、或者动态性较强的词汇时, 词汇挖掘效果不好的问题, 提出一种基于贝叶斯推理和 TFIDF 算法的中文关键词智能抽取方法. 利用贝叶斯统计原理对文本信息进行概率化的统计, 运用贝叶斯决策理论对 TFIDF 算法进行优化, 克服传统的 TFIDF 算法存在着缺陷. 实验结果表明, 优化后的 TFIDF 算法在进行中文关键词智能抽取中, 抽取的准确性大幅提高.

关键词: TFIDF; 文本挖掘; 贝叶斯统计; 关键词提取

中图分类号: TP393.08

文献标识码: A

文章编号: 1000-7180(2012)09-0197-04

Based on the Bayesian Reasoning and TFIDF Algorithm of Chinese Keywords Intelligent Extraction

HU Ju-xin, JU Xun-guang

(Department of Information and Electrical Engineering, Xuzhou Institute of Technology, Xuzhou 221008, China)

Abstract: In traditional TFIDF Chinese keywords intelligent extraction model, meet the redundant, or dynamic words, the words mining result is bad. Proposed based on the bayesian reasoning and TFIDF algorithm of Chinese keywords intelligent extraction method. Using bayesian statistical principle to text information of probability statistics. Using bayesian decision-making theory to optimize TFIDF algorithm. Overcome traditional TFIDF algorithm of defects. The experimental results show that the optimal algorithm for Chinese keywords TFIDF in intelligent extraction, the accuracy of the extraction greatly increased.

Key words: TFIDF; text mining; Bayesian statistical; key words extraction

1 引言

对于关键文本中的关键字数据挖掘是近年来较为流行的技术, 引起了很多相关学者的重视. 与一般的文本挖掘不同, 传统的数据挖掘一般都是根据较为规范的文本信息模型, 可以通过相关的聚类算法对较为结构化的数据进行分类, 使其结构化程度较高^[1]. 但是现实中的文本信息往往呈现非结构化或半结构化的特征, 由此转化的关联特征会形成一个海量数据的环境. 所以, 文本挖掘技术首要解决的问题就是如何合理地、便利地表示出这些文本特征, 准确的确定待挖掘文本中的关键词特征^[2].

对于一般的中度文本文章(一般大于 200 个词)可以利用 TFIDF 算法来进行文本中的关键词的识别, 取得了较好的识别效果, 也得到了广泛的应用^[3]. 但是, 随着网络信息个性化程度的不断增加, 文本中的信息种类也越来越多. 数据的个性化、突发性变化较强. 传统的 TFIDF 算法应用到这些环境下的时候, 工作效率较低. 本文中在 TFIDF 的基础上提出一种基于贝叶斯推理和 TFIDF 算法的中文关键词智能抽取方法. 这种算法通过贝叶斯决策思想划分文本特征, 对一些非结构化的文本特征进行有效的检测识别, 得到的挖掘效果也更为准确^[4].

2 基于贝叶斯推理和 TFIDF 算法的中文关键词

2.1 TFIDF 算法

传统的 TFIDF 算法是一个对文本分类较为流行的算法,其是运用数学上的特征向量表示相关的文本特征,通过把文本特征建立一个高维度的集合.对集合中的每个文本特征项 t ,计算相应的特征权值,关键词被分解成一个类似于 $\langle t, w \rangle$ 的集合.假设特征项 $(t_1, t_2, \dots, t_n)$ 代表文档内容的特征项,可以用一个高纬度的坐标系表示^[7].计算出的文本特征权值 $(w_1, w_2, \dots, w_n)$.这些值分别对应着不同维度上的坐标.假设文档中的语义特征 d 都可映射成此空间上的一个特征向量 $V(d) = (t_1, w_1, t_2, w_2, \dots, t_n, w_n)$.在这类方法中,每一个特征关键词都被匹配了一个 Term 对应的权重,权值 w 代表 Term 在文档关键词识别时的重要程度.这一权值也反映了关键词在文本中出现的次数.如果出现的次数越多,表明文本的重要程度越大,也就是越关键,即 TFIDF 公式:

$$W_j = TF_j \times IDF$$

其中, W_j 表示文档中特征词的权重; TF_j 表示关键词出现的次数; IDF 是关键词在整个搜索区域中出现的频率, IDF 就是 DF 的倒数, $IDF = \log(|D|/DF_j)$, D 是可识别的关键词数量.这种方法能够突出关键词在文档中的出现频率,用出现次数的方法体现关键词的重要意义,效果比较明显.

2.2 贝叶斯统计优化

由于传统的方法存在一定的弊端,贝叶斯统计决策能够很好的弥补这种弊端,计算文本相似特征的方法如下:

$$S = \sqrt{(t_1 - p_1)^2 w_1 + (t_2 - p_2)^2 w_2 + \dots + (t_{14} - p_{14})^2 w_{14}}$$

其中, t_i 为 TFIDF 算法计算得到的样本特征向量; 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.但是由于此公式并没有反映出特征文本之间出现的次数.例如:文本库中的相关文本特征在某些程度上存在一定的相似度.但是,这些相似度没有形成一定的关联,存在着偏差.因此,采用贝叶斯决策的思想统计文本特征之间的关联程度,考虑将一些文本特征的权值距离适当偏移,再把包含了这种变化的因素考虑到概率的相似度计算中.设关键词文本的特征量设为 I_1 和 I_2 ,而 TF_j 表示特征文本关键词中与标准关键词文本的特征差异化程度.根据这些变量,计算贝叶斯统计下的相似度的计

算公式如下:

$$S(I_1, I_2) = P(\Delta \in \Omega_i) = P(\Omega_i | \Delta)$$

其中, Ω_i 是文本特征中的一种变化程度的度量方式,表示关键词的变化量,也称同类变化量, Ω_E 表示关键词特征在不同于本类的文本信息中的变化量,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.这样,通过贝叶斯统计的方法能够很好的计算出不同类别下的特征关联程度,已经文本中的变化的关联程度.加入贝叶斯决策后,文本相似度的计算公式为:

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

其中, $P(\Omega_E)$ 表示文本的一种存在条件,例如是网络文本还是聊天文本等等.统计出文本特征在所有分类特征中出现的相关概率等等.于是,文本的特征提取识别问题就转化成了对概率的有效判断问题.只需要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,就可以进行中文关键词智能抽取.当 $P(\Omega_i | \Delta) > P(\Omega_E | \Delta)$ 或 $S(I_1, I_2) > 1/2$ 时,就可以判定为中文关键词智能抽取成功.

2.3 特征向量空间的降维运算

由于在文本特征检测的过程中,可能存在着一定的误差.假设这个误差为 Δ .而 Δ 由多维空间的坐标 $\Delta x, \Delta y$ 组成,是一个二维的向量.为了保证文本中关键词抽取过程中的简洁性,需要对文本中的特征词汇进行降维处理.降低特征词的维度可以选择主成分分析的方法,通过把关键词特征向量映射到相应的低纬度空间,保证特征最大化,就可以降低计算的复杂程度.

如果 n 维特征文本向量为 $x = (x_1, \dots, x_n)^T$,用 $\bar{x} = E[x]$,相关矩阵 $R_x = E[xx^T]$,协方差矩阵可以用 $C_x = E[(x - \bar{x})(x - \bar{x})^T]$ 来表示.降维的过程可以表示为文本关键特征变换到一个低纬度的特征空间,设为 $y = (y_1, \dots, y_n)^T$.计算方法如下:

$$y = U^T x = (u_1, u_2, \dots, u_n)^T x = \begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_n^T \end{pmatrix} x$$

其中, $y_i = u_i^T x$, $i = 1, 2, \dots, n$.

特征向量用表示:

$$x = (U^T)^{-1}y = U_y y = U^T x = (U^T)^{-1}y$$

$$= U_y = (u_1, u_2, \dots, u_n) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \sum_{i=1}^n y_i u_i$$

选择一个映射后的特征集合 $\{y_1, \dots, y_n\}$ 作为关键字 x 的估计,其他的用 b_i 代替,则估计式为:

$$\hat{x} = \sum_{i=1}^m y_i u_i + \sum_{i=m+1}^n b_i u_i$$

误差:

$$\epsilon^2(m) = E[(x - \hat{x})^T (x - \hat{x})]$$

$$= \sum_{i=m+1}^n E[(y_i - b_i)^2]$$

这种方法能够较好的保存原有的关键词汇的信息,在映射过程中最大程度的保证特征不发生退化,保存大量的原有文本特征信息。

2.4 改进后 TFIDF 算法流程

通过贝叶斯统计的方法对 TFIDF 算法进行改进,改进后的算法流程如下:

首先利用一个关键词的特征集合 H ,对 H 中的特征向量进行贝叶斯统计,计算相关的概括分布.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,在提取随机文本特征关键词的时候,可利用 k 个不同的贝叶斯统计结果,完成特征向量的统计.统计结果可以定义为:

$$H = \{S_1, w_1, S_2, w_2, \dots, S_n, w_n\}$$

其中:

$$S_k = (t_{k1}, w_{k1}, t_{k2}, w_{k2}, \dots, t_{kn}, w_{kn}) \in V(d)$$

$$w_i = \max(w_{i1}, w_{i2}, \dots, w_{in})$$

算法实现如下:

(1) 对计算得到的关键词特征信息进行文本信息初始化,利用 TFIDF 公式通过特征映射的方法生成文本特征向量 V .

(2) 运用提出的贝叶斯统计原理计算 $V = (v_1, w_1, v_2, w_2, \dots, v_n, w_n)$ 相关的概率分布与相似度计算.其中:

$$fsim_k = \sum_{i=1}^n \sum_{j=1}^n (v_i * t_j)$$

如果, $v_i * t_j = 1$, $v_i = t_{j0}$, 并且 $v_i \neq t_j$, 那么 $v_i \in V$, $t_j \in S_k$, $w_j \in S_k$.

(3) 若 $fsim$ 计算得到的极大值结果超过了先验知识确定规定判断阈值,可以对其进行第一次分类,分配到相关集合中.否则,认为 V 不是一个可以提取的关键特征词汇,对不是关键特征的词汇也可

以加入到相应的特征空间中,加以保存.

(4) 如果计算结果 H 中的词,特征发生了较为明显的衰退,也急速 $t \in H, t \in V$, t 的权值 $w = w \times 0.9$,可以把其看做是干扰特征,加以删除.

(5) 根据贝叶斯统计与算法计算出的权值对关键词集合中的元素进行排序,并利用贝叶斯相似概率统计的方法,重新计算权值,即 $w_{ki} = 0.5w_{ki} + 0.5w_i$.

(6) 输出的判断特征中,权值最高的几个词作为本次文本输入的关键词.

3 实验结果及对比

为了验证基于贝叶斯推理和 TFIDF 算法的中文关键词智能抽取中的效果,对一个采集外部多样性信息的文本信息进行中文关键词智能抽取,为了营造实验环境,保证节点检测的高效性,对文本信息不做明确要求,以保证其随机性结构特征.

本文选取的文本信息集合中包含 1300 个可能与关键词特征类似的信息,其中有部分信息是无效的,可以通过对算法改进前后,对中文关键词智能抽取的准确定位统计,通过不断的增加文本关键信息的数量,不断的统计监测结果的准确率,对两种不同的方法进行统计,获取的实验数据如图 1 所示.

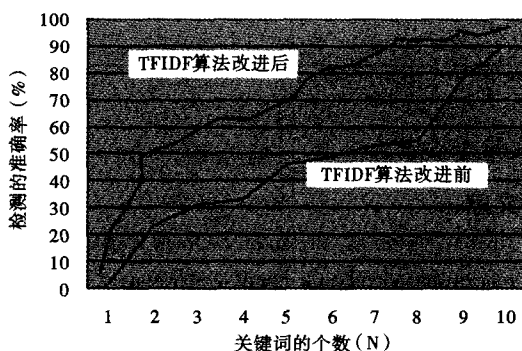


图1 中文关键词智能抽取检测曲线示意图

在图1中,通过对加入的关键词干扰信息等级不同,随着待检测关键词数量的不断增加,不同方法检测准确率也随之增加.本文算法在相同节点数量的前提下,检测的结果都优于传统的方法.为了进一步说明情况,对平均的检测准确率进行统计,实验数据如表1所示.

本文方法运用贝叶斯统计原理,对非规则文本特征中的中文关键词进行智能抽取,克服了传统方法对非结构化的文本挖掘效果不好的弊端,因此本文算法得到的结果优于传统算法.

表 1 两种检测方式准确率

算法	关键词抽取 样本数目	漏检率 (%)	检测准确率
传统 TFIDF 算法	200	23.1%	53%
加入贝叶斯统计 的 TFIDF 算法	200	16%	67%

4 结束语

本文提出一种基于贝叶斯推理和 TFIDF 算法的中文关键词智能抽取方法. 利用贝叶斯统计原理对文本信息进行概率化的统计, 运用贝叶斯决策理论对 TFIDF 算法进行优化, 克服传统的 TFIDF 算法存在着缺陷. 实验结果表明, 优化后的 TFIDF 算法在进行中文关键词智能抽取中, 抽取的准确性大幅提高. 实验证明, 本文提出的检测方法, 取得了很好的效果, 有较大的使用价值和应用前景.

参考文献:

- [1] Brendan Kitts, David Freed, Martin Vrieze. Cross-sell: A Fast Promotion Tunable Customer-Item Recommendation Method Based on Conditionally Independent Probabilities[C]// Proceedings of KDD, [S. L.]: Based on ACM Press, 2000: 437-446.

- [2] Mladenec D. Personal web watcher: design and implementation [M]. Department for Intelligent System, J. Stefan Institute, 2000.
- [3] Susan Gauch, Jason Chaffee. Ontology-based personalized search and browsing[J]. Web Intelligence and Agent Systems 2003, 1(3): 219-234.
- [4] Konstan JA, Miller B N Maltz, Herlocker D, et al. Grouplens: applying collaborative filtering to usenet news[J]. Communications of the ACM, 1997: 165-169.
- [5] YK Woon, WK Ng, EP Lim. Online and incremental mining of separately—grouped Web access Logs[C]// Proceedings of the 3rd International Conference on Web Information Systems Engineering . singapore: IEEE, 2002 :1-10.
- [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 (02): 117-119
- [7] 陈作聪 孙树峰 基于混沌算法的无线安全认证技术研究[J]. 微电子学与计算机, 2007(12): 115-117.

作者简介:

胡局新 男, (1979—), 硕士, 讲师. 研究方向为数据库技术、智能计算.

鞠训光 男, (1966—), 博士, 副教授. 研究方向为智能计算、数据挖掘.

(上接第 196 页)

参考文献:

- [1] Juva I, Vaton S, Virtamo J. Quick traffic matrix estimation based on link count covariance[C]// proceedings of the IEEE Intl Conf on Icc, Istanbul: IEEE Communication Society, 2006: 603-608.
- [2] 程光, 龚俭. 大规模网络流量宏观行为周期性分析研究[J]. 小型微型计算机系统, 2003, 24(6): 992-994.
- [3] 洪飞, 吴志美. 基于小波的多尺度网络流量预测模型[J]. 计算机学报, 2006, 29(1): 166-170.
- [4] 曹建华, 刘渊, 戴悦. 一种基于灰色神经网络的网络流量预测模型[J]. 计算机工程与应用, 2008, 44(5): 155-157.
- [5] 冯海亮, 陈涤, 林青家, 等. 一种基于神经网络的网络流量组合预测模型[J]. 计算机应用, 2006, 26(9): 2206-2208.

- [6] 李海燕, 李咚. 改进的神经网络在网络流量预测中的应用研究[J]. 计算机仿真, 2011, 28(9): 132-135.
- [7] 饶云华, 曹阳, 杨艳. 自相似网络通信量的多尺度预测研究[J]. 计算机工程与应用, 2005, 41(28): 26-28.
- [8] 田妮莉, 喻莉. 一种基于小波变换和 FIR 神经网络的广域网网络流量预测模型[J]. 电子与信息学报, 2008, 30(10): 2499-2502.
- [9] 蔡晓丽, 宁慧. 基于最大熵算法网络流量预测模型研究[J]. 计算机仿真, 2011, 28(10): 107-110.
- [10] 李捷, 侯秀红, 韩志杰. 基于卡尔曼滤波和小波的网络流量预测算法研究[J]. 电子与信息学报, 2007, 29(3): 725-725.

作者简介:

段谟意 男, (1964—), 本科, 副教授. 研究方向为计算机网络.