

关联挖掘下的海量文本信息深入挖掘实现

彭其华

(四川理工学院 计算机学院, 四川 自贡 643000)

摘 要: 研究基于关联度挖掘的海量网络文本挖掘方法;随着计算机和网络技术的快速发展,网络上的文本呈现海量增长的趋势,传统的网络文本挖掘方法采用基于特征提取的方法实现,能够实现小数据量下的文本挖掘,但是在信息量的快速增长下,传统方法已经不能适应;提出一种基于关联度挖掘的海量网络文本挖掘方法,首先采用特征提取的方法对海量文本进行初步的分类和特征识别,然后采用关联度挖掘的方法对各个文本特征之间的关联度进行计算处理,根据关联度的大小最终实现文本挖掘,由于关联度可以很好的体现特征文本之间的相互关系;最后采用一组随机的网络热门词汇进行测试实验,结果显示,算法能够很好适应海量文本下的挖掘实现,具有很好的应用价值。

关键词: 关联度挖掘;海量文本;特征提取

中图分类号: TP37

文献标识码: A

文章编号: 1000-7180(2013)10-0157-04

Text Mining Method of Massive Network Based on Correlation Mining

PENG Qi-hua

(School of Computer Science, Sichuan University of Science and Engineering, Zigong 643000, China)

Abstract: The text mining method of massive network based on correlation mining was research on. With the rapid development of computer and network technology, the text rendering of network grew fast, the traditional network-based text mining method extracted feature from text to achieve text mining, but with the rapid growth in the amount of information, the traditional methods cannot meet the need of development. So a text mining method of massive network based on correlation mining was proposed, the feature was extracted with mass text to finish initial classification and characteristics identification, and then the method of mining correlation between the characteristics of the various texts correlation was used to do calculate the coefficient of correlation, according to the coefficient of correlation, the text was divided into several types, so the correlation can reflect the relationship between the characteristics of the text well. Finally, a team of random words were used to test the ability of the algorithm, and the result shows that the algorithm can adapt to massive text excavation well with good application value.

Key words: correlation mining; massive text; feature extraction

1 引言

关联挖掘算法是一种近年来逐渐广泛得到应用的控制算法,它具有关联特性好,挖掘深度深的优点,所以广泛应用在各种场合.海量网络文本挖掘中的挖掘方法,是一种基于特征提取和关联的文本识别算法,最重要的部分是对关联样本中各个样本的关联特性进行计算.所以关联挖掘算法可以很好的

在海量网络文本挖掘中来^[1-5].

为了解决大信息量下的海量网络文本挖掘处理瓶颈,提出一种基于关联挖掘算法的文本分类识别方法,对各个文本之间和各个文本之内的词语、短语、句之间的相关特征进行提取,采用关联挖掘算法对文本之间的关联性进行计算,对文本之间和内部的特征进行识别分类关联,这样更加细致的描述文本与文本之间的内在关联度.本文最后采用一组随机的网络词汇进行系统测试.

2 网络文本特征提取方法

海量网络文本挖掘的实现方法是根据文本特征,对特征进行提取的基础上,采用特征融合处理的方法从海量的网络数据文本中提取文本归类识别^[6-7].

较为成熟的文本特征提取方法是对不同的词语实现不同的加权,以表示本词语在文档中出现的重要性.

在加权实现方法中,最容易实现的加权方式为布尔加权方式,即当一个词语在文档中出现时,加权为 1,反之,加权为 0. 定义为^[6]

$$w_{ij} = \begin{cases} 1 & f_{ij} \geq 1 \\ 0 & f_{ij} \leq 1 \end{cases} \quad (1)$$

其中:

f_{ij} — 词汇 i 在文档 d_j 中出现的频率;

w_{ij} — 词汇的加权结果.

权重反映了文本数据出现的频率,同时直接反映文本数据的重要性,是一种基于信息理论的权重计算方法,基于熵权重的文本挖掘方法定义为^[8]

$$\begin{cases} w_{ij} = \log(f_{ij} + 1, 0) * \\ \left(1 + \log \frac{1}{N} \sum_{k=1}^N \left[\frac{f_{ik}}{n_i} \log\left(\frac{f_{ik}}{n_i}\right)\right]\right) \end{cases} \quad (2)$$

其中:

f_{ij} — 词汇 i 出现的频率;

n_i — 出现研究特征的次数.

f_{ik} — 词语 i 在文档 k 中出现的频率;

N — 文档数目;

在进行文本数据挖掘的识别过程中,需要采用数字化的归一化方法对其进行基于计算机的处理.

通过归一化处理,可以很好的对文档中关键词进行分类的度量,表征各个关键词的重要性质,从而实现识别关键字^[8].

在传统挖掘方法的基础上,做自适应权重修改,形成如下定义:

$$x(L_{ac}) = 1 - \frac{L_{ac}}{\max(L_{ac}) + l} \quad (3)$$

为了表征文本个数与最大相关系数之间的相互联系,定义为

$$\max(L_{ac}) = \log_2 k \quad (4)$$

其中:

$\max(L_{ac})$ — 特征提取后,关联特征项,各个特征类信息熵的最大值;

k — 类别数;

l — 常数系数.

所以,结合特征项的类间和类内信息熵分布加权方法定义为

$$\begin{cases} W_{ik}(d) = \frac{IDF_1}{IDF_{const}} \times a(L_{ac}) \\ IDF_1 = \text{coeff}_1 + tf_{ik}(d) \times \log\left(\frac{N}{n_k} + 0.01\right) \\ IDF_{const} = \text{coeff}_{const} + \\ \sqrt{\sum_{i=1}^n (tf_{ik}(d))^2 \times [\log\left(\frac{N}{n_k} + 0.01\right)]^2} \end{cases} \quad (5)$$

其中:

coeff_1 — IDF_1 的固定系数;

coeff_{const} — IDF_{const} 的固定系数.

在式(7)中,通过变化加权时采用的固定系数 coeff_1 和 coeff_{const} ,可以对 IDF_1 和 coeff_{const} 的值进行响应的改进,通过不断调整,达到较好的分类效果.

实际挖掘实现时需要从以下几方面实现:

(1) 训练科目编辑功能 根据训练科目的要求,把文本以参数的形式表现出来,包括文本特征参数、文本数量参数、文本性质参数、文本类别参数和文本容量参数等. 训练计划是指根据文本的特点,在一定的挖掘环境下,采取优化技术措施对抗海量环境以具体的形式表现出来. 模拟训练系统的关键,训练计划可以任意设计,人为增加训练的强度和难度,从而达到更好的训练层次效果^[9-10].

(2) 结合文本挖掘性能,以海量为仿真对象,采用深度关联挖掘技术,实时对文本数据进行挖掘,适应训练科目的计划要求.

(3) 根据训练计划模拟文本挖掘,通过优化调整技术措施进行优化训练. 优化调整技术措施要与文本挖掘的期望性能相结合,充分体现优化模拟挖掘的真实性.

(4) 考核评估根据被挖掘数据的效果信息等,记录操作步骤,计算挖掘结果中的变化率等特征参数,对其整个操作过程进行定性或定量的评价. 考核评估的关键是建立合理的评估体系,对优化训练的结果进行科学的评估,从而使训练更具有针对性.

3 基于关联挖掘的网络文本深挖实现方法

关联挖掘是一种数据处理的新挖掘方法,采用数据挖掘的方法,对数据的关联度进行衡量考虑的基础上,对文本特征进行挖掘.

文本挖掘时,首先假设文本待挖掘区域为,该区域为需要进行文本挖掘的区域,采用挖掘处理的方法,对该区域进行挖掘分析,把它视为一张图像,一个区域,该区域图像为 $p \times q$ 像素.在待挖掘区域 W 中,假设需要设置 Q 同等性能指标的文本,各个文本在区域 W 中的坐标参数假设为 (x_i, y_i) ,各个文本由于性能参数相同,故其挖掘半径是相同的,这里假设为 r ,因此,最终能有效实现全区域挖掘的半径为 $R = 2r$,相对应的每个文本坐标 (x_i, y_i) 的圆心可表示为 $W_i = \{x_i, y_i, r\}$,其中 W_i 的半径表示为 r ,为挖掘半径.区域内所需的目标像素点假设为: (x, y) ,根据欧式距离的定义,得到该目标点与源节点之间的距离为

$$h(w_i, k) = \sqrt{(x_i - x)^2 + (y_i - y)^2} \quad (6)$$

假设 $K(r_i)$ 为区域中的像素点 (x, y) 能被节点 W_i 的挖掘的概率.该概率具有一定的规律性,可用如下公式表示:

$$K_{upg}(x, y, w_i) = \begin{cases} 1 & d(w_i, k) \leq r - r_u \\ u\left(\frac{-\alpha_1 \theta \beta_1}{\theta_2 \beta_2}\right) + \alpha_2 & r - r_u < d(w_i, k) < r + r_u \\ 0 & \text{else} \end{cases} \quad (7)$$

式中, $r_u (0 < r_u < r)$ 表示挖掘能够实现通信传输和监测的有效性能参数, $\alpha_1, \alpha_2, \beta_1, \beta_2$, 为对应于文本检测系数,该系数与其特征相关的, θ_i 为输入参数变量,表示式为

$$\theta_i = r_u - r + (w_i, k) \quad (8)$$

在整个挖掘中,通过多个文本节点,实现对区域的协同挖掘,非此即彼,协同进行,从而提高了对区域的最优挖掘,提高了挖掘效率,协同信息挖掘的检测概率可表示为

$$K_{upg}(W_{pg}) = 1 - \prod_{w_i \in W_{pg}} (1 - K_{upg}(x, y, w_i)) \quad (9)$$

在挖掘系统中,设文本中待挖掘区域为 W ,同样为 $p \times q$ 像素,同样,如上所述,各个节点集合的协同信息挖掘的检测概率为 $K_{upg}(W_{pg})$,通过该概率,可以得到区域内的各个相点是否为信息挖掘,由此,得到信息待挖掘区域为 W 的节点集合的区域挖掘率为

$$Q_w(w) = \frac{\sum K_{upg}(W_{pg})}{p \times q} \quad (10)$$

在利用上述方法重构的文本集合中,文本 x_j 到除 x_i 本身外文本 x_i 的距离小于 r 的 x_j 的文本类似度定义为

$$Q = \sum_{j \neq i} H(r - \|x_i - x_j\|) \quad (11)$$

此处设定关联函数概念,定义为把所有可能在距离上比给定的距离 r 小的点相对于占有总的点对数的比例称为关联函数,表示为

$$C_N(r) = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i+1}^N H(r - \|x_i - x_j\|) \quad (12)$$

式中,分子为 2 为了排除重复计数,用范数表示两文本之间的类似度.可得到两文本类似度为两个矢量的最大分差量,为

$$\|x_i - x_j\| = \max_{1 \leq k \leq m} |x_{i-(k-1)\tau} - x_{j-(k-1)\tau}| \quad (13)$$

对于文本类似不大于 r 的向量,可有称为具有关联性的矢量.在此,假设一维文本数据序列为 n ,则相空间重构中的向量点个数为 $N = n - (m-1)\tau$.计算相点中有关联的相点对数,占一切可能的 $N(N-1)/2$ 种配对的比例称为关联积分,表示为

$$C_m(r) = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i+1}^N H(r - \|x_i - x_j\|) \quad (14)$$

4 系统仿真与结果分析

4.1 仿真环境描述

为了测试基于关联挖掘的海量网络文本挖掘算法与传统文本挖掘算法的性能比较,构建仿真参数见表 1.采用随机网络热门词汇,对所有的文本数据进行分类统计和处理.

表 1 仿真参数描述

热门词汇	出现文本数量	出现词汇数量
BBS	9	363
斑竹	34	154
菜鸟	15	419
灌水	27	284
纯净水	7	185
刷屏	34	351
楼主	18	273
沙发	43	222
大虾	42	347
板凳	29	310
顶	24	397
抓狂	44	478
路过	41	261
火星帖	32	440
教授	40	86
犬科	33	489
大学	17	135
超强	14	126
汗	17	437
马甲	26	368

4.2 结果分析

从表 1 可以看出,各种文本文档的分类基本均衡,每个类别的文本分类系数分布如图 1 所示。

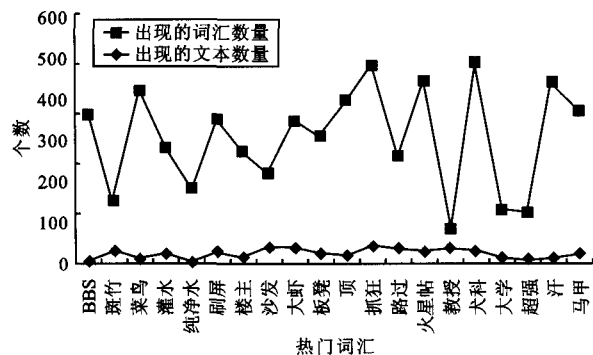


图 1 热门词汇分布

从图 1 可以看出,热门词汇分类基本均衡,类内方差小。

在表 1 构建的实验参数环境下进行系统性能测试。

采用传统文本挖掘方法对文本进行挖掘的结果如图 2 所示。

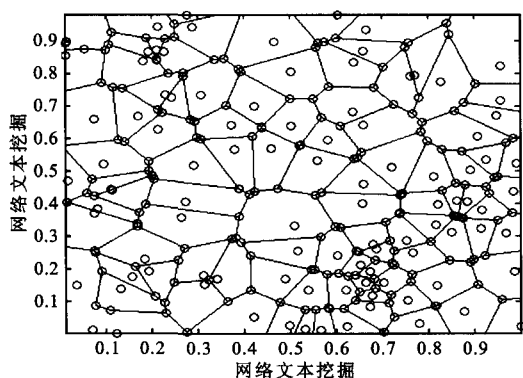


图 2 传统文本挖掘方法效果

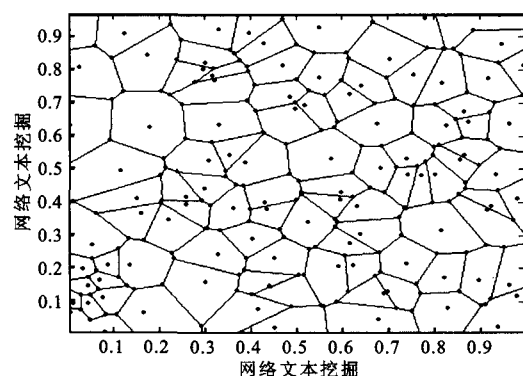


图 3 基于关联挖掘的挖掘结果

从图 2 和图 3 的比较可以看出,基于关联特性挖掘的海量网络文本分类挖掘方法能够对文本词汇进行更加详细的挖掘分类,随着文本词汇特征数目

的增加,文本分类所需要实现的计算量也是飞速增长的。

算法收敛性是算法稳定的关键,本文的改进算法不但能够提取更多的网络词汇特征而且算法收敛性强,算法收敛特性的比较曲线如图 4 所示。

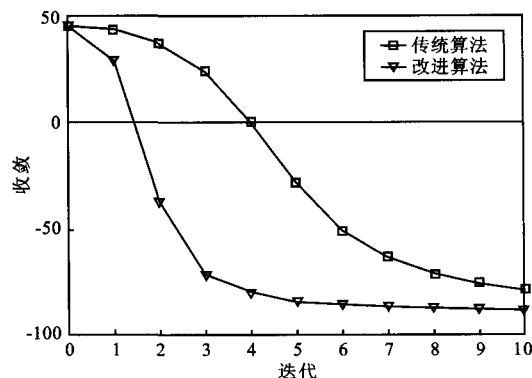


图 4 传统算法与改进算法的收敛特性比较图

从图 4 可以看出,算法收敛后逐渐稳定到一个较小的误差。

5 结束语

本文研究了一种基于关联度挖掘的海量网络文本挖掘方法。为了适应计算机和网络信息技术快速发展下文本呈现海量增长下传统挖掘方法不能很好实现挖掘的缺陷,研究关联度挖掘的实现方法,首先采用特征提取的方法对海量文本进行初步的分类和特征识别,然后采用关联度挖掘的方法对各个文本特征之间的关联度进行计算处理,根据关联度的大小最终实现文本挖掘,由于关联度可以很好的体现特征文本之间的相互关系;最后采用一组随机的网络热门词汇进行测试实验,结果显示,算法能够很好适应海量文本下的挖掘实现,具有很好的应用价值。从本文的实验结果可以得出:

- (1) 基于关联度挖掘的网络海量文本挖掘方法能够很好的应对大文本量下的挖掘实现;
- (2) 基于关联挖掘的算法收敛性强于普通的算法;
- (3) 基于关联挖掘的算法能够对文本进行更好的文本间关联度计算和特征分析,从而最终实现网络海量文本数据的挖掘。

参考文献:

- [1] 乔良. 文本挖掘技术研究及其在信息检索中的应用[J]. 软件导刊, 2009, 8(4): 160-161.
- [2] 李素建, 王厚峰, 俞士汶, 等. 关键词自动标引的最大熵模型应用研究[J]. 计算机学报, 2004, 27(9): 1192-1197.

(下转第 164 页)

WEB 页面资源的充分复制。

表 2 响应时间记录表

序号	文本转换率(K)	文本正确率(L)	文本字数	转换耗时/ms
1	100	100	150	3
2	100	100	225	3
3	100	100	380	3
4	100	100	513	4
5	100	100	680	5
6	100	100	877	5
7	100	100	920	5
8	100	100	1 030	5
9	100	100	1 716	6
10	100	100	2 309	6
11	100	100	5 765	7
12	100	100	6 780	7
13	99.99	100	8 670	8
14	99.99	100	18 629	8

5 语 束 语

本文分析了现有的 WAP 转码方案,设计了一种基于 XML 的中间转换模型,给出相关算法和实现程序,实现从 HTML 页面转换到 WML 页面。最后给出了实例,对页面转换进行了测试,并记录相关的算法运行时间,对模型性能进行分析,证明了本模型转换时的准确性和具有更高的效率。针对其它格式的转码方案是下一步研究的方向。

参考文献:

- [1] 李青山,陈平. 一种基于内容的 HTML 到 XML 转换策略[J]. 计算机工程与应用,2001,37(9):30-32.
- [2] 张仁良. 基于自动生成技术的杉达学院 WAP 网站的设计与实现[D]. 上海:上海交通大学,2009.
- [3] 温泉. WAP 页面转换的实现技术研究[D]. 江苏:河海大学,2004.
- [4] 钱程,阳小兰. HTML 到 XML 转换研究[J]. 计算机与现代化,2011(8):39-41.
- [5] 林碧英,李四达. HTML 页面到 WAP 页面转码器的研究[J]. 中国电力教育(S3),2006,:21-22.
- [6] 李佳,面向移动平台的网络信息获取技术的研究与实现[D]. 哈尔滨:哈尔滨工业大学,2006.
- [7] 董树明. 半结构化 Web 信息抽取技术及其应用研究[D]. 江苏:东南大学,2004.
- [8] 葛冬梅. 基于内容个性化的 HTML 页面转换研究[D]. 黑龙江:哈尔滨工程大学,2008.
- [9] 刘杰,束博. 一种高效的 HTML/XHTML 至 WML 的转换方法[J]. 北京工商大学学报:自然科学版,2006,24(6):45-48.
- [10] 刘洪,贺深,黄河燕. WAP 页面转换代理系统原理及实现[J]. 计算机工程与应用,2002(4):177-179.
- [11] 马黎,基于信息剪裁的 HTML2WML 智能转换算法[J]. 微型电脑应用,2009(25)1:1-4.

作者简介:

李超建 男,(1978—),硕士,讲师.研究方向为计算机网络、嵌入式技术.

朱晓姝 女,(1973—),硕士研究生导师,教授.研究方向为网络计算、数据挖掘.

(上接第 160 页)

- [3] 李伟光,薛龙飞,母晓峰. 智能多媒体教室无线控制研究[J]. 科技通报,2013,29(4):220-222.
- [4] Hu WM, Xie D, Tan TN, et al. A hierarchical self-organizing approach for learning the patterns of motion trajectories[J]. Chinese Journal of Computers, 2003, 26(4):417-426.
- [5] Zelnik—Manor L, Irani M. Event—based analysis of video[C]// Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2001: 123-130.
- [6] 许晓昕,李安贵. 一种基于的网络聊天关键词提取算法[J]. 计算机技术与发展,2006,16(3):122-123.
- [7] Zhou Yanao, Tang Jianbo, Wang Jiaqin. An improved feature selection algorithm based on information entropy[C]//Proceedings of the 26th Chinese Control Con-

ference CCC. [s. l.]:IEEE,2007:312—315.

- [8] 徐燕,李锦涛,王斌,等. 基于区别分类能力的高性能特征选择方法[J]. 软件学报,2008,19(1):82-89.
- [9] 巩知乐,张德贤,胡明明. 一种改进的支持向量机的文本分类算法[J]. 计算机仿真,2009,26(7):165-168.
- [10] Zhang D, Gatica-Perez D, Bengio S, et al. Semi-supervised adapted HMMs for unusual event detection [C]// Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2005:611-618.

作者简介:

彭其华 男,(1969—),硕士研究生,讲师.研究方向为计算机软件开发、电子商务.