IISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 2,Mar. 2015,pp. 464－468
DOI:10. 16337/j. 1004-9037. 2015. 02. 026

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于短文本信息流的热点话题检测

宗　慧　刘金岭

(淮阴工学院计算机工程学院,淮安,223003)

摘　要:短文本信息流在传递公开信息时携带了丰富且具有极大价值的信息资源。根据短文本信息流特点,利用训练数据集中的信息熵来构建决策树检测模型进行热点话题检测,该方法先是计算出各话题类别的平均信息量和每个特征词对于短文本信息流进行划分的信息增益率,再通过选择具有最大信息增益率的特征词进行测试,完成自上而下的决策树建树过程,最后利用叶子结点的类型确定热点话题。在真实短信文本信息流上实验表明,该方法具有明显的检测稳定性和较高的数据处理效率。

关键词:短文本;信息流;热点话题;决策树

中图分类号:TP391　　**文献标志码**:A

Hot Topic Detection Based on Short Text Information Flow

Zong Hui, Liu Jinling

(College of Computer Engineering, Huaiyin Institute of Technology, Huai'an, 223003, China)

Abstract: Potential information with high value are carried by short text information flow in transmission. A model of decision tree for hot topic is established with the information entropy of training data set, according to the characteristics of short text information flow. The average amount of information of each topic categories and the information gain ratio of each characteristic word for distinguishing short text information flow are computed in the first step by the above algorithm of decision tree. Then, the characteristic word with maximum information gain ratio is selected for the job of test, while the top-down construction process of the decision tree is accomplished. Finally, the hot topic is determined according to the leaf node type. The experiment result on real short text information flow shows that the proposed algorithm is more stable and faster than others.

Key words: short text; information flow; hot topic; decision tree

引　　言

短文本信息流存在于当今广泛使用的手机短信、互联网即时通信、论坛、博客和微博等系统中,一个短文本信息流通常涉及多个话题。话题检测任务旨在根据短文本信息讨论的话题以及信息间的对话关系,将信息分检到多个队列,每个队列是一段主题明确的话题。以话题组织的短文本信息数据,比原始的按时间顺序组织的短文本信息流更便于内容管理和进一步的挖掘。然而,人工在海量的文本信息流

基金项目:国家级星火计划"农村民生建设信息反馈平台建设"(2011GA690190)资助项目。
收稿日期:2014-06-05;**修订日期**:2014-06-30

中抽取话题是一项费时费力、甚至不可能完成的艰巨任务。另一方面,短文本信息流中富含一些社会热点问题、社会突发事件信息等。热点话题的及时发现和分析对于行业调研、信息安全、舆情预测、金融证券等领域有广阔的应用前景,同时它也将带来比较有价值的应用服务。因此,对热点话题发现自动化和智能化的研究得到了广泛关注并成为重点研究领域之一。

短文本信息流话题检测是根据话题以及信息间的对话关系,将每一个短文本分检到话题簇中。由于短文本信息流具有如下特点[1]:(1) 短文本信息长度短、信息量少,以词为维度的向量空间模型呈现出高维稀疏的特点;(2) 短文本信息流中广泛存在的谐音词和简写词,短信文本信息流中常常会出现网络用语、口语化短语和一些变形字等,如网络用语"FB"意指"腐败",聊天常用语"小资"意指"单身女",论坛常用语"潜水"意指"呆在聊天室里不说话"等。变形字如垃圾短信中用"发 piao"表达"发票",用"收费交底"表达"收费较低"等,呈现出奇异性特点;(3) 按时间顺序组织的短文本信息流更便于内容管理和进一步的挖掘,此时,在短文本信息流中不同话题的短文本随时间变化而交错地长时间出现,即使相邻的短文本也可能隶属于不同的话题。

1　相关研究

美国国防高级研究计划局(Defense advanced research projects agency,DARPA)在 1996 年第一次提出了针对新闻流数据的话题动态性挖掘研究,即话题检测与追踪(Mopic demecmion and mracking,MDM),并建立了 MDM 的语料库,其主要任务是:话题分割与识别;话题题检与跟踪;主题识别。迄今为止,研究者共采用了 3 种方法:基于词典的抽取方法、基于规则的抽取方法和基于统计机器学习的抽取方法。本文采用机器学习方法处理短文本信息流中多话题检测,利用流量检测问题的核心工作主要包括两个方面:(1)构建适当的短文本信息流属性向量;(2)选择适当的机器学习方法。相关研究有:张振亚等[2]在文本信息的检索中引入了用户焦点的个性化推荐,该方法可以有效地减轻用户因从包含大量无关信息的查询结果中筛选感兴趣信息而产生的负担,较好地满足了用户对文本信息检索的时间要求。但是它没有考虑到被检测的特征词对于不同的话题具有不同的重要程度,从而使得特征词属于话题类别的边界变得模糊。Shen 等[3]使用两个查询的预余弦相似度作为判断依据,但由于存在词汇不匹配问题,所以没有利用搜索引擎返回理想的结果。Ozmutlu 等[4-5]使用多元线性回归搜索模式和 Monte-Carlo 模拟算法进行新话题识别,虽然在搜索意图的连续性、时效性上取得了较好的效果,但是在识别搜索意图发生改变时效果较差。Roughan [6]利用贝叶斯定理对属性集合和类型变量之间概率关系进行数据流检测的方法(Naïve Bayes,NB),该方法假定文本向量的特征词相互独立且遵循高斯分布,通过计算训练集上各类特征词的统计值来获取高斯分布的各项参数值。然而在实际的文本数据流检测问题中,这些条件的假设通常难以满足。使用 NB 方法来处理流量检测问题时,检测准确率只有 65%左右[7]。Moore[8]采用基于核估计技术的朴素贝叶斯(Naïve Bayes using kernel density estimation,NBK)方法来进行数据流检测,从实验结果来看,检测准确率从 65%提高到 95%左右。

在该类问题研究的已有算法中主要使用到了时间间隔、查询相似度以及查询模式等特征,利用这些特征进行话题检测时,出现了词汇不匹配问题,如"电脑"和"计算机"。也有时会出现两个词义非常接近的词汇,可能属于两个不同的话题,如"电脑不能运行的原因"和"汽车不能行驶的原因"。本文利用短文本信息流的特点,给出了基于决策树的方法(Short text decision tree classification mothd,ST_DTC),该方法利用训练数据集中的信息熵来构建检测模型,从而有效地解决了话题边界的划分,同时也解决了词汇不匹配问题。这种利用检测模型的简单查找来完成未知短文本信息流样本的检测,具有良好的检测稳定性。

2　基于短文本信息流的热点话题检测

决策树方法是以实例为基础的归纳学习算法,它从一个无次序、无规则的实例集合中归纳出一组采

用树形结构表示的检测规则,在检测、预测、规则提取等领域得到广泛应用。本文利用决策树对短文本信息流进行了热点话题检测,先是通过训练集合的学习,形成决策树检测模型;然后利用生成的决策树模型对类型未知的样本进行检测。在使用决策树模型对类型未知样本进行检测时,从根节点开始逐步对该样本的属性进行测试,并沿着相应的分支向下行走,直至到达某个叶节点,此时叶节点所代表的类型即为该样本的类型。本文提出的算法框架如算法 1 所示。

算法 1 短文本信息流热点话题分析的 ST_DTC 决策树算法

输入 短文本信息流 ST_F;

输出 话题检测集合 S_ST:

(1)利用文献[9]将 ST_F 表示成特征向量集的形式

$$ST\_F_V=\{ST_V \mid (w_1,t_1;w_2,t_2;\cdots\cdots;w_n,t_n)\} \tag{1}$$

式中:w_i 是短文本 ST_V 的特征词;t_i 是 w_i 在短信文本 ST_V 中的权值,其含义是 w_i 在短文本中的重要程度。这是因为去除那些对分类结果无影响的词,比如"的"、"在"这一类词,可以降低文本向量的维数,以提高运算的效率[10]。

(2)假设在短文本信息流 ST_F 中需要检测 k 个不同的话题 $C_1,C_2,\cdots\cdots,C_k$,则 ST_F_V 对话题类别的平均信息量

$$H(ST\_F_V)=-\sum_{p=1}^{k}P(C_p)\log_2 P(C_p) \tag{2}$$

式中:$P(C_p)=|C_p|/|ST\_F_V|$ $(1\leqslant p\leqslant k)$

(3)对于如果特征词 w_i 隶属于 m 个不同的话题,则可将 ST_F 划分为 m 个子集:$S_1,S_2,\cdots\cdots,Sm$。进而将话题 $C_1,C_2,\cdots\cdots,C_k$ 分为 $k\times m$ 个小子集 $C_{iq}=\{ST_V \mid w_i\in C_{iq},ST_V\in C_{iq}\subseteq ST\_F_V,1\leqslant i\leqslant n,\ 1\leqslant q\leqslant m\}$。

(4)利用特征词 w_i 进行划分后,ST_F_V 对于话题检测的平均信息量为

$$H(ST\_F_V/w_i)=-\sum_{q=1}^{m}P(C_q)[-\sum_{j=1}^{k}P(C_{jq})\log_2 P(C_{jq})] \tag{3}$$

其中,$P(C_q)=\sum_{j=1}^{k}|C_{jq}|/|ST\_F_V|$,$P(C_{jq})=|C_{jq}|/|ST\_F_V|$,那么利用 w_i 对 ST_F_V 划分的信息增益量 $f_G(ST\_F_V,w_i)$,有

$$f_G(ST\_F_V,w_i)=H(ST\_F_V)-H(ST\_F_V/w_i) \tag{4}$$

式(4)表明 $f_G(ST\_F_V,w_i)$ 等于使用 w_i 对 ST_F_V 进行划分前后不确定性下降的程度。

(5)使用特征词 w_i 对 ST_F_V 进行划分的信息增益率为

$$f_{GR}(ST\_F_V,w_i)=\frac{f_G(ST\_F_V,w_i)}{f_{sp}(ST\_F_V,w_i)} \tag{5}$$

其中

$$f_{sp}(ST\_F_V,w_i)=-\sum_{l=1}^{m}(|S_l|/|ST\_F_V|)\log_2(|S_l|/|S|) \tag{6}$$

(6)通过选择具有最大信息增益率的特征词作为测试,ST_DTC 决策树方法自上而下地完成决策树的建树过程。为了去除噪声点和孤立点[11]引起的分支异常,ST_DTC 决策树方法利用训练数据集中剩余的样本,对生成的初始决策树进行了剪枝,进而得到最终的 ST_DTC 决策树,此时,叶节点所代表的话题类型即为该样本的类型。

(7)在话题集合 S_ST 中输出大类别的热点话题[11]。

3 实验及结果分析

为了验证相关结论,作者从江苏某短信运营商截取 2012 年 2 月 1 日 0 点整到 4 月 30 日 24 点 0 分

时间段的近 9 万 4 千条手机短信文本集合进行了人工标注,抽取出了如载有化学品的船在江阴段沉船、江苏沿江部分城市出现市民抢购矿泉水、元宵节祝福等 142 个热点话题。为了问题的简化,实验前根据文献[12]提前将样本集通过分词、特征提取和降维等预处理为短信文本向量集,得到 18 000 条的短信文本向量集 ST_F_V。

3.1　话题检测效率比较

为了测试 3 种算法效率的可靠性,在短信文本向量集 ST_F_V 上进行了 5 次仿真实验,其时间值取 3 次实验结果的均值。而每次实验都是随机抽取了 ST_F_V 的 30%作为训练样本集,将剩余的 70%作为测试数据集。结果如表 1 所示。

表 1　NB,NBK 和 ST_DTC 算法运行时间比较

Table 1　Time consuming of NB,NBK and ST_DTC　　s

时间	NB	NBK	ST_DTC
训练时间	13.87	14.63	93.26
测试时间	183.47	265.82	58.34

实验结果可以看出,ST_DTC 算法构建决策树的过程比较复杂,即花费的时间比较长,但是在对未检测的短信文本检测时的处理速度上明显地快于其他两种方法,这是因为 ST_DTC 算法在短信文本信息流的检测中只需根据特征词进行自上而下地简单比较,处理起来相对容易。而 NB,NBK 对未检测的短信文本进行检测时需要计算出每条短信属于每个话题类别的概率,然后比较出最大概率值,这个计算过程相对复杂,因此所耗费的时间较长。

3.2　准确率比较

下面实验是对 ST_DTC,NB 和 NBK 算法进行准确率比较。假设类别 i 中包含 n_i 条短信文本,类别 j 中包含 n_j 条短信文本,类别 j 中隶属于类别 i 的短信文本条数记为 n_{ij},则检测准确率 $P(i,j)$的定义如下[13]

$$P(i,j)=\frac{n_{ij}}{n_j} \tag{7}$$

实验结果如图 1 所示。横轴表示随机抽取的样本数量,纵轴表示与真是话题检测的准确率。

从图 1 可以看出,在上述 3 种机器学习方法检测结果中,由于 ST_DTC 算法在样本预测中部依赖于短文本信息流样本的分布,可以有效地避免样本分布变化的影响,因此其检测稳定性较好,从而保证了整体检测准确率随着短信文本训练集合的逐步增大而保持相对稳定的增加,而 NB 和 NBK 方法的检测

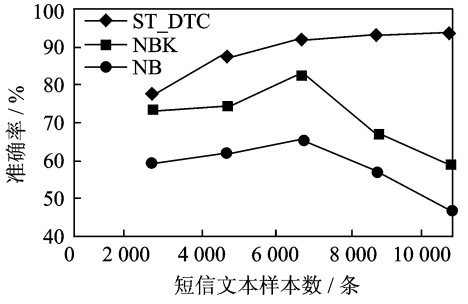

图 1　ST_DTC,NB 和 NBK 算法准确率比较

Fig. 1　Accuracy comparison of ST_DTC,NB and NBK

准确率不仅没有随着短信文本训练集的增大而提高,相反却随着短信文本信息训练集的增大而出现显著下降。这不仅是由于 NBK 方法的局部最优性所带来的抖动,更主要的是因为训练短信文本信息集和测试短信文本信息集中各类样本的分布存在较大差异,而基于贝叶斯定理的 NB 和 NBK 方法都是假设先验概率保持不变。当这一假设条件无法满足时,NB 和 NBK 方法也就随之失效。

4 结束语

从短文本信息流中挖掘热点话题是目前研究的热点之一。现有的热点话题提取技术主要对基于文本相似度聚类方法的改进,不能很好地反应短文本信息流特征的稀疏性、奇异性和动态性,本文先是引入 ST_DTC 决策树方法来处理短文本信息流的检测问题,该方法利用训练短文本集中的信息熵来构建检测模型,并通过检测模型的简单查找来完成未知文本信息流样本的检测,其性能较好。

参考文献:

[1] 刘金岭,倪晓红,王新功. 手机短信文本信息流的自动文摘生成[J]. 现代图书情报技术,2013,29(2):43-49.
Liu Jinling, Ni Xiaohong, Wang Xingong. Automatic abstracting generating based on mobile short message text information flow[J]. New Technology of Library and Information Service, 2013, 29(2): 43-49.

[2] 张振亚,陈恩红,王进,等. RealCC 在文本信息检索的个性化推荐中的应用研究[J]. 数据采集与处理,2004,19(3):338-342.
Zhang Zhenya, Chen Enhong, Wang Jin, et al. Enabling personalization recommendation with RealCC for text information retrieval based on user-focus[J]. Journal of Data Acquisition and Processing, 2004, 19(3): 338-342.

[3] Shen X, Tan B, Zhai C. Implicit user modeling for personalized search[C]//Proceedings of the Conference on Information and Knowledge Management. Bremen, Germany: [s. n.], 2005: 824-831.

[4] Ozmutlu S, Ozmutlu H C, Spink A. Automatic new topic identification in search engine transaction logs using multiple linear regression[C]//Proceedings of the 4lst Hawaii International Conference on System Sciences. Hawaii, USA: [s. n.], 2008: 140-148.

[5] Ozmutlu S, Ozmutlu H C, Buyuk B. Using Monte-Carlo simulation for automatic new topic identification of seach engine transaction logs[C]//Proceedings of the 2007 Winter Simulation Conference. Washington, USA: [s. n.], 2007: 2306-2314.

[6] Roughan M, Sen S, Spatscheck O, et al. Class-of-service mapping for QoS: A statistical signature-based approach to IP traffic classification[C]//Proc of the ACM SIGCOMM Internet Measurement Conf. Taormina: [s. n.], 2004: 135-148.

[7] Zuev D, Moore A W. Traffic classification using a statistical approach[C]//Proc of the PAM 2005. LNCS 3431, Heidelberg: Springer-Verlag, 2005: 321-324.

[8] Moore A W, Zuev D. Internet traffic classification using Bayesian analysis techniques[C]//Proc of the 2005 ACM SIGMETRICS Int'l Conf on Measurement and Modeling of Computer Systems. Banff: [s. n.], 2005: 50-60.

[9] 刘金岭. 基于语义的高质量中文短信文本聚类算法[J]. 计算机工程,2009,35(10):201-202.
Liu Jinling. High quality algorithm for chinese short messages text clustering based on semantic[J]. Computer Engineering, 2009, 35(10): 201-202.

[10] 史岳鹏,朱颢东. 基于类别相关性和优化的 ID3 特征选择[J]. 数据采集与处理,2011,26(2):231-234.
Shi Yuepeng, Zhu Haodong. Feature selection based on category correlation and improved ID3[J]. Journal of Data Acquisition and Processing, 2011, 26(2): 231-234.

[11] 刘金岭,王新功. 基于中文短信文本聚类的热点事件发现[J]. 情报杂志,2013,32(2):30-33.
Liu Jinling, Wang Xingong. Hot events detection based on Chinese SMS text clustering[J]. Journal of Intelligence, 2013, 32(2): 30-33.

[12] 刘金岭. 基于降维的短信文本语义分类及主题提取[J]. 计算机工程与应用,2010,46(23):159-161
Liu Jinling. Dimensionality reduction of short message text classification and thematic extraction of semantic[J]. Computer Engineering and Applications, 2010, 46(23): 159-161

[13] 黄九鸣,吴泉源,刘春阳,等. 短文本信息流的无监督会话抽取技术[J]. 软件学报,2012,23(4):735-747.
Huang Jiuming, Wu Quanyuan, Liu Chunyang, et al. Unsupervised conversation extraction in short text message streams [J]. Journal of Software, 2012, 23(4): 735-747.

作者简介:**宗慧**(1980-),女,讲师,研究方向:中文信息处理,E-mail:43122990@qq.com;**刘金岭**(1958-),男,教授,研究方向:中文信息处理。