基于字符串相似性聚类的网络短文本舆情热点发现技术

杨 震，段立娟，赖英旭
（北京工业大学 计算机学院，北京 100124）

摘 要：将每个短文本文档看成一个由文字、数字和标点构成的字符串，并基于字符串自身的特性直接计算其相似性，在此基础上进行短文本层次化聚类，进而发现网络舆情热点. 由于这种方法免去特征提取和文本表示过程，在一定程度上避免了传统方法在短文本表示时特征向量稀疏的不足，有效解决了短文本内容聚类问题. 实验结果表明，本文提出方法有效.

关键词：舆情分析；短文本处理；层次聚类
中图分类号：TP 393 **文献标志码**：A **文章编号**：0254－0037(2010)05－0669－05

互联网络信息爆炸、信息泛滥、信息污染、信息扰民、信息惑众等问题的日益严重极大影响普通用户对互联网信息正常、合理的使用. 更为严重的是，一些不法分子开始利用网络传播虚假和非法广告，散布谣言蛊惑人心，扰乱国家经济和社会秩序；敌对势力更是利用网络传播害国言论，制造事端，教唆动乱，严重地威胁着国家的稳定和安全. 信息安全重心已转向应用和数据安全，基于内容对互联网信息传播和利用进行监管（即舆情监控）的国家和社会需求越来越强烈，成为学术界和产业界广泛关注的一个热点[1].

在需求的推动之下，众多研究者利用模式识别、人工智能、知识发现为代表的智能技术对网络信息进行内容分析、语义挖掘，进而实施有效的信息过滤、话题发现以及趋势预测. 但需要指出的是，现有的技术实现距离需求期望仍有差距，解决互联网舆情预警问题的关键技术，特别是网络话题的发现技术还亟待提高，互联网内容安全形式不容乐观. 一方面，针对普通网络信息（长文本信息）舆情态势分析及舆情预警关键技术的研究已经大规模地展开，并取得了一定的研究成果. 总体来说，针对普通网络信息（长文本信息）的内容识别与过滤技术已经迈入实用阶段. 在文本表达方面，Salton 的向量空间模型和基于 Markov 过程的 n-gram 模型提供了有效的文本描述数学模型. 在文本特征选择方面，提出了基于词频/倒文档频度（TF/IDF）、信息增益（IG）、CHI、互信息（MI）等统计量的专门特征选择方法，同时，还将主成分分析、线性鉴别分析和奇异值分解的方法引入文本特征选择，衍生出了潜在语义索引（LSI）的重要概念. 在文本聚类/分类方面，贝叶斯分类器、支撑向量机（SVM）、神经网络、自组织映射（SOM）、k 近邻、k 均值、决策树、关联规则、向量相似度量以及分类器集成等模型得到了广泛应用.

然而另一方面，针对以即时消息、在线聊天记录、BBS 标题、手机短消息、微博客、博客评论、新闻评论等为代表的短文本信息舆情态势分析及舆情预警关键技术的研究力度不够，而恰恰是这一部分内容更能反映真实的网络舆情. 但是由于短文本独特的语言特征（稀疏性、实时性、不规范性等）[2]，使得一些针对长文本的内容处理方法性能劣化，甚至不可用. 因此，针对短文本自身特点，研究符合其特性的文本表达和特征选择方法，实现短文本的正确聚类成为了一个迫切的现实要求.

基于此，本文面向网上短文本信息舆情分析需求，基于字符串相似性研究短文本信息的聚类方法，以期解决短文本话题发现、传播及动态演变的特征分析等关键问题.

收稿日期：2009-12-10.
基金项目：国家“九七三”计划资助项目（2007CB311100）；北京市自然科学基金资助项目（4102012，4102013）；北京市教育委员会科技发展计划面上资助项目（KM200810005030）；北京工业大学青年科学基金资助项目.
作者简介：杨 震（1979—），男，贵州六盘水人，讲师.

1 网络短文本信息舆情分析系统架构

网络短文本信息舆情分析系统架构如图 1 所示. 首先系统对接收到的网络短文本信息进行数据接收和解码,把元数据送入元数据缓存,同时将其输入垃圾信息过滤器处理,将与舆情分析无关的短文本(包括 SP 定制信息、无意义信息、格式信息及其他无需进行内容监控和舆情预警的信息)判断为垃圾信息放入垃圾信箱,对有用信息内容进行话题发现,并对其传播和演变规律进行分析. 系统根据用户反馈,对分类器进行更新和重建,逐渐逼近实际应用的使用需求.

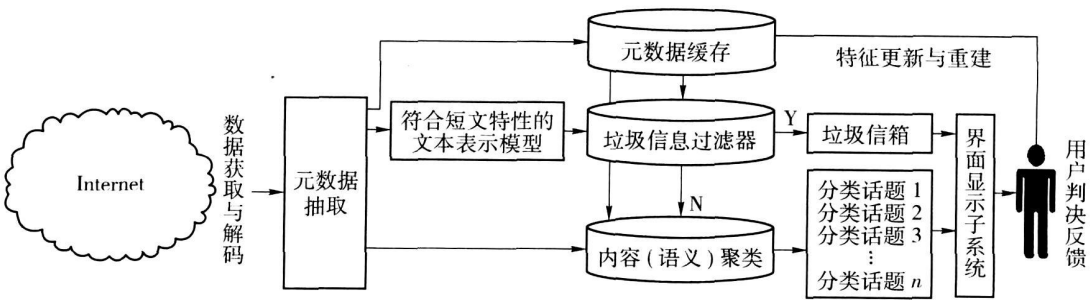

图 1 网络短文本信息舆情分析演示系统框图

Fig. 1 Flowchart of online public opinion hotspot detection

在实现有用信息(舆情分析相关信息)和垃圾信息(舆情分析无关信息)分离之后,需要对有用信息的聚类方法进行研究. 短文本作为全新的文本媒体对象,具有其自身特点(稀疏性、实时性、不规范性等),使得传统的聚类分析方法在短文本表示这个层次上遇到了极大的困难. 传统的文本表示模型,包括布尔模型、概率模型、向量空间模型都无法良好地表示,总会遇到特征向量稀疏性的问题,最终使得短文本的聚类变为简单层次上“词重现”一级的短文本聚集.

毫无疑问,对短文本间相似性的准确表达及正确度量将会对短文本聚类处理带来很大帮助,而传统的文本表示和特征提取方法会损失许多重要的信息, 如特征的顺序、上下文等特征,因而无法准确表达短文本间的相似性,进而使得聚类性能劣化甚至不可用.

因此,如何基于短文本自身的特性确定其相似性成为本文重要的研究内容. 本研究把每个短文本文档看成一个由文字、数字和标点构成的字符串,并基于字符串自身的特性计算其相似性,在此基础上进行短文本聚类,进而发现网络舆情热点. 由于这种方法免去了特征提取和文本表示过程,在一定程度上能够避免特征向量稀疏性的问题.

2 基于字符串相似性短文本聚类的热点发现

短文本作为全新的文本媒体对象,具有独特的语言特性. 为了避免由于特征向量稀释性导致短文本聚类蜕化为简单层次上“词重现”一级的短文本聚集,迫使研究者考虑能否跳过特征提取和文本表示环节,基于短文本的特性计算相似性. 通过将每个短文本文档看成一个由文字、数字和标点构成的字符串,那么可以借助比较 2 个字符串共同包含的子串个数和连续程度来衡量 2 个字符串的相似程度. 当然共同的子串越多, 2 个短文本文档就越相似. 这样一来,基于字符串相似性聚类的网络短文本舆情热点发现过程即可按照以下步骤处理:

步骤 1 预处理步骤. 对于采集的短文本 $M_i, i=1,2,3,\cdots,k$ 进行整理和清洗. 将输入的短文本信息转换为统一编码,去除乱码等噪声信息. 并按采集时间、上下文信息以及正文信息导入数据库.

步骤 2 基于字符串相似性计算各个短文本之间的相似程度. 假设字符串 A, B 间的相似性可表示为 $D(A,B)$,即以通过比较 2 个字符串共同包含的子串个数和连续程度来衡量 2 个字符串的相似程度,寻找

短文本 M_i, M_j 的最佳匹配.

步骤 3　基于短文本 M_i, M_j 之间的归一化相似度进行层次聚类(hierarchical clustering)[3]分析. 层次聚类法是一种高效的聚类算法,其基本思想是根据所定义的个体间相似度,从相似性最高的个体开始,向初始化空网络中添加新个体. 过程终止后,此时该网络的组成就被认为是划分为了若干簇. 层次聚类方法可分为凝聚的层次聚类和分裂的层次聚类.

步骤 4　利用层次聚类可视化的特点,对话题间的联系进行直观的度量,发现话题,进而对其传播及动态演变的特征进行分析.

其中,字符串相似测度 $D(.)$ 以及用以确定聚类数目的评价指标是本文接下来需要解决的重要问题.

2.1　基于编辑距离的字符串相似性计算

假设短信 M_i, M_j 分别由 m 和 n 个字符组成,分别由 $\{C_1^i, C_2^i, C_3^i, \cdots, C_m^i\}$ 和 $\{C_1^j, C_2^j, C_3^j, \cdots, C_n^j\}$ 表示. 那么短文本 M_i, M_j 之间的相似度就可由其包含字符串之间的相似度计算而来. 利用 Hungarian 算法去发现 M_i, M_j 和之间的最大匹配. 设 M_i 在 M_j 中的最大匹配是 $\{C_{j_1}^i, C_{j_2}^i, C_{j_3}^i, \cdots, C_{j_m}^i\}$, $j_k \in \{1,2,3,\cdots,n\}$, $k=1,2,3,\cdots,m$. M_j 在 M_i 中的最大匹配是 $\{C_{j_1}^i, C_{j_2}^i, C_{j_3}^i, \cdots, C_{j_n}^i\}$, $j_k \in \{1,2,3,\cdots,m\}$, $k=1,2,3,\cdots,n$. 那么基于最大匹配,短文本 M_i, M_j 之间的相似度定义为其间的编辑距离(Levenshtein 距离)[4-5]:

$$D(A,B) = \text{Levenshtein}(M_i, M_j). \tag{1}$$

在这样的定义下,$D(\cdot)$ 越小,说明字符串越相似.

2.2　层次化聚类数目选择方法

在层次化聚类分析中,如何选择恰当的聚类个数是一个非常复杂而又必须面对的问题. 尽管众多研究者进行了广泛的研究,提出了各种聚类有效性指标,包括信息熵、V_{wsj} 指标、Gap statistic、IGP、Scat/Sep 指标等[6],但如何确定数据的聚类个数仍然是一个富有挑战性的问题,一般来说只能通过试错法(trial-and-error)迭代确定.

实际上,一个好的聚类结果应该使得簇内的数据点之间是尽可能"紧凑"的,而簇间的数据点之间是尽可能"分离"的. 这样一来,一个可行的聚类个数选择依据可以定义如下:

$$Q = \frac{\text{簇内平均相似度}}{\text{簇间平均相似度}}. \tag{2}$$

可以对聚类簇的几何拓扑结构预先假定,或者不做任何限制,在此基础上度量平均相似度[6-7],本文使用基于简单的点对(pair-wise)相似性的度量方法. 假设待处理短文本集为 S,假定其可能被划分为 k 簇,即 $S=\{S_1, S_2, \cdots, S_k\}$,其中用 $|S_k|$ 表示簇中元素的个数,那么

$$Q = \frac{\frac{1}{k}\sum_{i=1}^{k}\sum_{A,B\in S_i}\frac{1}{|S_i|^2}D(A,B)}{\frac{1}{k^2}\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{A\in S_i, B\in S_j}\frac{1}{|S_i|\cdot|S_j|}D(A,B)} \tag{3}$$

依据前述定义,显然 Q 值越小说明聚类所选择的数目越合理.

3　实验结果

实验采用 SMS 短信库[8]作为评测语料库,这里我们使用了其中一个标注后的子集(共 4 486 条短信). 为了简化问题并且考虑到人工标注的可行性,将其标注为 5 个类别:日常生活、工作相关、非法和虚假信息、系统群发(非手写短信)和其他短信. 使用这样的分类体系是基于以下的考虑:

1) 这样的分类简单易行,且概念明确,易于标注实现;

2) 这样的分类体系虽然比较简单粗略,但其体系结构容易扩展,能为进一步的研究打下坚实的基础;

3) 这个分类体系也涵盖了一些研究热点所需要关注的短信类别.

在这里主要考察和网络异常发现相关的第 3 类(非法和虚假信息)和第 4 类(系统群发/非手写)短信. 图 2 给出了层次聚类数目和 Q 值之间的关系. 从图中可知,如果设定 $Q=0.1$ 为阈值,那么当聚类数目为 14 和 10 时,第 3 类和第 4 类短信聚类性能开始稳定下来,聚类结果趋于合理. 从聚类结果来看,首先,形式上相似的短信被聚为一类,如下所示(出于隐私保护的目的,我们去除了短信中的敏感信息):

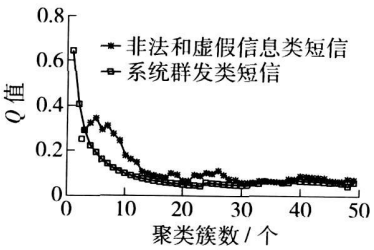

图 2 层次聚类簇数目和 Q 值间关系图

Fig. 2 The Q yielded by different number of clusters

短信簇 1

短信 1:哗哗哗! 铜蛋被砸开了,金花四溅!! 定睛一看,你得到了 27 分,目前你的总分为 27

短信 2:哗哗哗! 铜蛋被砸开了,金花四溅!! 定睛一看,你得到了 29 分,目前你的总分为 55

短信 3:哗哗哗! 银蛋被砸开了,金花四溅!! 定睛一看,你得到了 26 分,目前你的总分为 53

其次,被聚为一类的短信在内容上体现出一致性,如下所示(出于隐私保护的目的,我们去除了短信中的敏感信息):

短信簇 2

短信 1:不了解那你还把你电话给我

短信 2:你不接我电话那我也不接你电话了

短信 3:打电话怎么不接我给你往家里打电话的了

短信 4:你话费还剩多少了? 晚上给我打电话不?

这样就比较容易从同一个聚类簇中得到其共同主题,进而可以发现短信中舆情热点. 实验结果说明了本文提出方法的有效性. 由于篇幅所限,其他类似实验结果不再赘述.

4 结束语

以短信息为代表的网络舆情分析相关研究已经引起了人们的重视. 但由于短文本独特的语言特性,使得一些针对长文本的内容表示和聚类分析方法性能劣化,甚至不可用. 对短文本间相似性的准确表达及正确度量将会对短文本聚类处理带来很大的帮助,本文通过把每个短文本文档看成一个由文字、数字和标点构成的字符串,并基于字符串自身的特性计算其相似性,在此基础上进行短文本层次化聚类,进而发现网络舆情热点. 但是字符串相似性的计算方法无一例外均具有较高的计算复杂性,而且这样的相似性也仅局限在衡量某种程度的形式相似性上. 如何从更高的层次,即语义的角度来衡量其相似性成为下一步研究的重要方向.

参考文献:

[1] 中国信息产业商会信息安全产业分会. 中国信息安全产业发展白皮书(2005—2010)[EB/OL]. [2005-03-11]. http://www.itsec.gov.cn/webportal/document/baipishu.doc.

[2] 龚才春. 短本语言计算的关键技术研究[D]. 北京:中国科学院研究生院计算技术研究所,2008.
GONG Cai-Chun. Research on short text language computing[D]. Beijing: Institute of Computing Technology, Chinese Academy of Sciences, 2008. (in Chinese)

[3] SCOTT J. Social network analysis: a handbook[M]. 2nd Edition. London: Sage, 2000: 123-145.

[4] 车万翔,刘挺,秦兵,等. 基于改进编辑距离的中文相似句子检索[J]. 高技术通讯,2004,14(7):15-20.
CHE Wan-xiang, LIU Ting, QIN Bing, et al. Similar Chinese sentence retrieval based on improved edit-distance[J]. High Technology Letters, 2004, 14(7): 15-20. (in Chinese)

[5] 杨震,范科峰,雷建军,等. 基于语义的文本流形研究[J]. 电子学报,2009,37(3):557-561.
YANG Zhen, FAN Ke-feng, LEI Jian-jun, et al. Text manifold based on semantic analysis[J]. Acta Electronica Sinica,

2009，37(3)：557-561.（in Chinese）

[6] 陈黎飞，姜青山，王声瑞. 基于层次划分的最佳聚类数确定方法[J]. 软件学报，2008，19(1)：62-72.
CHEN Li-fei，JIANG Qing-shan，WANG Sheng-rui. A hierarchical method for determining the number of clusters[J]. Journal of Software，2008，19(1)：62-72.（in Chinese）

[7] BOUGUESSA M，WANG S，SUN H. An objective approach to cluster validation[J]. Pattern Recognition Letters，2006，27(13)：1419-1430.

[8] 马旭，徐蔚然，郭军，等. SMS-2008 标注中文短信息库[J]. 中文信息学报，2009，23(4)：22-26.
MA Xu，XU Wei-ran，GUO Jun，et al. SMS-2008：an annotated Chinese short messages corpus[J]. Journal of Chinese Information Processing，2009，23(4)：22-26.（in Chinese）

Online Public Opinion Hotspot Detection and Analysis Based on Short Text Clustering Using String Distance

YANG Zhen，DUAN Li-juan，LAI Ying-xu

（College of Computer Science，Beijing University of Technology，Beijing 100124，China）

Abstract：The unique language characteristic of short texts has made the performance of traditional natural language processing methods degradation，or even unavailable. Exact representation and calculation of the similarity between short texts are great helpful to content based clustering. That this paper treated each short text as a composition of characters，numbers and punctuation，and a similarity measure based on string similarity was proposed. Then a public opinion hotspot detection and analysis system based on short text hierarchical clustering was built. This method calculated the similarity directly which skipped the feature extraction and representation processing of short text，to a certain extent，and avoided using the sparse feature vectors. Experimental results show the effectiveness of the proposed method.

Key words：public opinion analysis；short text processing；hierarchical clustering

（责任编辑　张　蕾）