◎数据库、数据挖掘、机器学习◎

基于隐含语义分析的微博话题发现方法

马雯雯[1],魏文晗[1],邓一贵[1,2]
MA Wenwen[1], WEI Wenhan[1], DEGN Yigui[1,2]

1.重庆大学 计算机学院,重庆 400044
2.重庆大学 信息与网络管理中心,重庆 400044
1.School of Computer Science, Chongqing University, Chongqing 400044, China
2.Center of Information and Network, Chongqing University, Chongqing 400044, China

MA Wenwen, WEI Wenhan, DEGN Yigui. Micro-blog topic detection method based on Latent Semantic Analysis. Computer Engineering and Applications, 2014, 50(1):96-100.

Abstract: As the large popularity of micro-blog and awareness continues to improve, hot topic of micro-blog detecting has become the current research focuses. For short texts, there exist high-dimension, sparse, synonymy and polysemy problems for Vector Space Model(VSM) text presentation, making it difficult to measure the similarity of the texts accurately. This paper presents a two-stage cluster based on Latent Semantic Analysis(LSA) topic detection approach. Firstly, the concept of hot topic is introduced to select micro-blogs with certain attention, using LSA to model the dataset. Then CURE algorithm of hierarchical clustering is employed to determine the initial centers. Finally, the hot topic clustering results are obtained through *K*-means clustering. Experimental results on real micro-blog dataset verify the validity of the method.

Key words: Latent Semantic Analysis(LSA); Vector Space Model(VSM); topic detection; micro-blog; two-stage clustering

摘 要:随着微博的大量普及和关注度的不断提高,微博热点话题发现已成为当前研究热点。针对于短文本、向量空间模型(VSM)文本表示方法存在高维度、稀疏,以及同义多义问题,导致难以准确度量文本相似度,提出一种基于隐含语义分析的两阶段聚类话题发现方法。引入话题热度的概念来选取具有一定关注度的微博文本,用隐含语义分析(LSA)对数据集进行建模;用层次聚类的CURE算法确定初始类中心;用*K*-means聚类得到热点话题的聚类结果。真实微博数据集的实验结果验证了该方法的有效性。

关键词:隐含语义分析;向量空间模型;话题发现;微博;两阶段聚类
文献标志码:A **中图分类号:**TP393 doi:10.3778/j.issn.1002-8331.1203-0250

1 引言

随着互联网技术的发展及其应用的迅猛增长,继Web2.0技术之后微博客(简称“微博”)的应用受到了越来越多网民和机构的关注。微博具有内容简单易懂,发布快捷及时,传播速度快,信息来源渠道广泛,内容时效性强等特点。作为一种新兴的传播载体,微博已成为民众表达舆情的重要窗口[1],对话题发现、信息安全等领域的舆情监督具有重要作用。另外,基于社交网络的话题发现数据并不局限于文本信息,还可以利用非文本信息,如评论数等。这些新特点使面向社交网络的话题发现研究得到了更多重视。

微博数据主要由普通用户产生,无论是用词、形式还是具体内容的质量都参差不齐,给话题发现造成很大困难。尽管话题发现研究已经开展多年,但是由于互联

基金项目:重庆市自然科学基金(No.cstc2011jjA40023)。
作者简介:马雯雯(1986—),女,硕士,主要研究方向:计算机网络与信息安全;魏文晗(1986—),男,硕士,主要研究方向:信息安全;邓一贵(1971—),男,博士,高级工程师,主要研究方向:计算机网络与信息安全,移动代理。E-mail:ma-wen1024@163.com
收稿日期:2012-03-13 **修回日期:**2012-06-05 **文章编号:**1002-8331(2014)01-0096-05
CNKI**网络优先出版:**2012-07-16, http://www.cnki.net/kcms/detail/11.2127.TP.20120716.1500.029.html

网数据来源的多样性和特征抽取的不确定性,目前话题发现研究主要集中在新闻类数据上,针对社交网络(含微博)话题检测的研究相对较少。

国内外一些学者就著名的"Twitter"英文微博数据进行了相关研究[2-5]。文献[3,5]利用文本生成模型LDA在大规模Twitter数据集上挖掘话题;杨冠超[4]提出一个迭代式的语义分析和话题热度预测模型—Topic Rank,通过时间片划分和话题的关键词集合两个概念计算话题影响力。实际上,中文和英文在形式和表达上存在很大差异:英文结构严谨、表意明确;中文结构较灵活、语义丰富但意象相对模糊。再加上中西文化上的差异,使得基于Twitter数据的研究成果并不适用于中文微博。针对中文微博进行话题发现的研究相对较少。郑斐然[6]采用向量空间模型在线检测中文微博消息中的关键字,通过聚类方法来搜寻新闻话题。常用的基于关键字的向量空间模型(Vector Space Model,VSM)[7]将文本嵌入到正交向量空间以便于数学处理,却忽略了中文的"同义"、"多义"及高维向量问题,因此其发现话题的准确率较低,且高维的向量空间使得计算过程耗费时间较长[6]。另外,VSM基于任意词项间是独立的这一假设,这不符合实际的语言环境。

传统的话题发现方法主要基于关键词匹配来发现频度较高的话题,未考虑话题的语义相关性。隐含语义分析能挖掘出文字背后潜在的语义信息,其奇异值分解过程能在保留大部分语义信息的同时,降低向量空间的维度,减少计算量。为兼顾文本的语义信息,更精确地发现热点话题,本文引入隐含语义分析(Latent Semantic Analysis,LSA)的方法对中文微博数据建模,通过一个两阶段的聚类策略发现近期网络上较受关注的话题。本文的研究内容和话题检测与跟踪(Topic Detection Tracking,TDT)领域的研究十分相似,不同的是TDT研究多采用TREC会议提供的TDT语料[8],并不能完全反映网络上舆情发展的真实情况,而本文抓取真实的中文微博语料开展研究,更具实用价值。

2 理论基础

2.1 LSA基本思想

传统向量空间模型反映的是简单的词频和分布关系。微博文本词条和形式的多样性在一定程度上掩盖了其实际的语义信息。隐含语义分析在向量空间模型的基础上处理词条关系,试图跨越对自然语言的理解,运用统计的方法来发现词语使用过程中潜在的语义结构,用概念取代关键词,从而削减了词语和文档间的语义模糊度,在一定程度上缓解了向量空间模型中同义词、多义词的影响,提高了话题发现的精度。

隐含语义分析(Latent Semantic Analysis,LSA)模型最早由Dumans等人[9]提出,其基本思想是将原始的向量空间通过奇异值分解投影到低维的正交矩阵,从而转换到潜在的语义空间。不同于VSM,该模型建立在文本中的词语之间有紧密联系的假设基础上,用一个$m \times n$维(m为文档集中特征词个数,n为文档集包含的文档数)的特征矩阵$\boldsymbol{A}$描述文本中词项的共现性。即:

$$\boldsymbol{A}=[a_{ij}]_{m\times n} \tag{1}$$

通过对$\boldsymbol{A}$进行奇异值分解,取前k个最大的奇异值及其对应的奇异矢量构成一个新矩阵$\boldsymbol{A}_k$,来近似表示原词条-文档矩阵$\boldsymbol{A}$。

2.2 奇异值分解(SVD)

文本表示成词条矩阵后,通过奇异值分解计算矩阵$\boldsymbol{A}$的近似矩阵$\boldsymbol{A}_k(k \ll \min(m,n))$。抽取词的概念到概念空间,形成最小的表述文档的概念集合。例如,在A文档中出现的"书"和在B文档中出现的"报告"、"手册"、"指南"等会被认为是同一个概念。经奇异值分解后,矩阵$\boldsymbol{A}$可以表示为3个矩阵的乘积:

$$\boldsymbol{A}=\boldsymbol{U}_{m\times m}\boldsymbol{S}_{m\times n}\boldsymbol{V}^{\mathrm{T}}_{n\times n} \tag{2}$$

其中,$\boldsymbol{U}$和$\boldsymbol{V}$分别是与矩阵$\boldsymbol{A}$对应的左、右奇异向量矩阵,对角矩阵$\boldsymbol{S}$的元素为$\boldsymbol{A}$的奇异值序列:$\sigma_1,\sigma_2,\cdots,\sigma_r$,且$\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > 0$。分别取$\boldsymbol{U}$、$\boldsymbol{V}$的前$k$列$\boldsymbol{U}_k$、$\boldsymbol{V}_k$,构成$\boldsymbol{A}$的$k$秩近似矩阵$\boldsymbol{A}_k$:

$$\boldsymbol{A}_k=\boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}^{\mathrm{T}}_k \tag{3}$$

式(3)中,$\boldsymbol{U}_k$和$\boldsymbol{V}_k$列向量为正交向量,行向量分别作为词向量和文本向量。$\boldsymbol{A}_k$近似表示矩阵$\boldsymbol{A}$,降维因子k值的选取直接关系到语义空间模型的效率:k值过小会丢失一些有用信息,k值过大则导致运算量增加。考虑到计算响应速度和存储空间限制,k值一般在100~300[10-11]之间。

奇异值分解和取k秩近似矩阵有两个优点:(1)将词条和文本映射到同一个k维的语义空间内,削减了原词条矩阵的"噪声",突出了词和文本之间的语义关系。(2)在保留大部分有效信息的同时大大降低了空间维度,可有效提高聚类速度。

2.3 聚类方法

聚类方法可分为:划分法、层次法、基于密度的方法、基于模型的方法和基于网格的方法等。有些聚类算法只擅长处理球状分布或相近大小的簇,另一些则对孤立点比较敏感。基于层次聚类的CURE算法解决了上述两方面的问题:用多个代表点取代传统的单个点或质心来表示簇,并通过收缩因子α调节代表点收缩或向簇中心移动,使之可适应非球形或大小不均的簇,且准确度高。基于划分的K-means具有算法简单,收敛速度快等优点,但它对初始聚类中心数K(不同于上文降维因子k)非常敏感,目前一般采用$\sqrt{n}$作为K的最大值。

综合CURE算法高准确率和K-means算法高效率的特点,本文采用CURE和K-means结合的两阶段聚类

策略,来提高大规模微博文本话题发现的准确率和效率。该方法先用CURE算法进行初步聚类,得到*K*-means算法的输入参数:聚类个数及其对应的初始类中心,从而缓解*K*-means初始聚类中心的随机性和先验性导致聚类结果波动的问题。

好的聚类结果应使得簇内数据点"紧密",而簇间则尽可能"分散"。文献[12]定义了一个指标Q(见式(4)),将每个簇看做一个"大数据点",该指标不依赖于具体算法,实验验证效果较好。因此,用$Q(c)$的定义选择CURE算法聚类中心:设$\|\boldsymbol{X}-\boldsymbol{Y}\|$是点$\boldsymbol{X}$和$\boldsymbol{Y}$之间的欧式距离,给定数据集$D$的一个划分$C^k=(C_1, C_2, \cdots, C_k)$,$|C_k|$表示簇中元素个数,有

$$Q(c)=\frac{Scat(C^k)}{Sep(C^k)}=\frac{\sum_{i=1}^{k}\sum_{X,Y\in C_i}\|\boldsymbol{X}-\boldsymbol{Y}\|^2}{\sum_{i=1}^{k}\left(\sum_{j=1,j\neq i}^{k}\frac{1}{|C_i||C_j|}\sum_{X\in C_i,Y\in C_j}\|\boldsymbol{X}-\boldsymbol{Y}\|^2\right)} \tag{4}$$

其中,*Scat*为簇内任意两个数据点之间距离的平方和;*Sep*为簇间平均距离。由前述定义可知,$Q(c)$值越小,表示聚类数的选择越合理。

3　基于隐含语义模型的微博话题发现方法

3.1　基本思想和处理流程

为克服传统VSM进行词项匹配方法的不足,本文采用隐含语义分析建模的方法,先用"热度"的定义初步选择某时间段内关注度较高的微博;然后通过奇异值分解将词项和文档映射到潜在语义空间,挖掘每个短文本的隐含语义信息;最后,采用CURE和*K*-means算法相结合的两阶段聚类策略,发现网络上的热门话题。方法处理流程如图1。

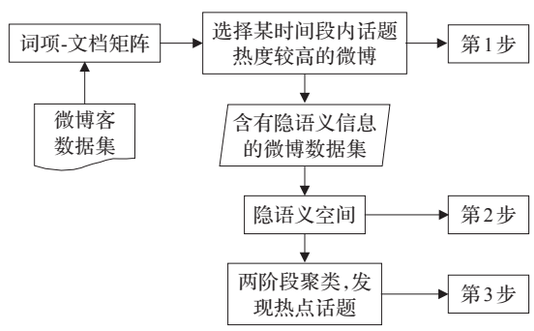

图1　处理流程图

3.2　方法描述

(1)筛选关注度高的微博。统计每条微博文本的转发数*N_rel*和评论数*N_com*。在选择某时间段内关注度较高的微博时,需要一个指标用于选择有一定关注度的话题。观察分析后发现,当某一话题引起网民关注后,其转发数和评论数会逐渐上升,从而在较短时间内传播开来,这在一定程度上反应了该话题的热度。故有:

定义1(热度)微博b的热度H_T为该微博的转发数*N_rel*与评论数*N_com*的加权和,即

$$H_T=\alpha N\_relb_i+\beta N\_comb_i \tag{5}$$

其中N_relb_i或$N_comb_i=\begin{cases}0, & N<10^2\\ 0.3, & 10^2\leqslant N<10^3\\ 0.5, & N\geqslant 10^3\end{cases}$。式(5)中,$\alpha$和$\beta$为调节因子,且$\alpha+\beta=1$,$N$为微博文本数,每个话题的热度$H_T\in[0,1]$。多次实验发现,$\alpha=0.4$,$\beta=0.6$时筛选的微博与人工判断的热点话题最接近。另外,只选择$H_T\geqslant 0.3$的微博做后续处理,即认为N_relb_i和N_comb_i为0或很小的微博不具备热点话题的条件。热度的定义提供了初步判断一个话题是否为热点话题的宏观标准,与微博内容本身无关。下述的隐含语义分析则从微观角度分析文字背后隐含的信息,再通过两阶段文本聚类算法来发现热点话题。

(2)构造隐含语义空间。对第一步筛选出的微博用ICTCLAS[13]进行中文分词、去停用词处理。对分词后的文本建索引得到71 514×18 456维的词项-文档矩阵$\boldsymbol{A}$(式(1)),其中a_{ij}表示第i个词在第j篇文档中的权重。由于微博文本短小且数目大,单个文本中出现的词条非常有限,因此,$\boldsymbol{A}$一般为稀疏矩阵。$\boldsymbol{A}$中特征词条权重有多种不同的计算方法[14],本文a_{ij}的计算采用目前最常用且效果最理想的TF·IDF法,计算方法如下:

$$a_{ij}=\frac{tf_{ij}\times \mathrm{lb}(N/n_i+0.01)}{\sqrt{\sum_{j=1}^{t}(tf_{ij}\times \mathrm{lb}(N/n_i+0.01))^2}} \tag{6}$$

其中,tf_{ij}表示第j个文档中第i个词出现的频度,N为文本集的总文本数,n_i为含有词条i的文本个数。

用式(6)计算a_{ij},对$\boldsymbol{A}$矩阵用麻省理工大学开发的SVDLIBC[15]进行SVD处理,得到矩阵$\boldsymbol{A}$的近似矩阵$\boldsymbol{A}_k$,其中$\boldsymbol{U}$和$\boldsymbol{V}$的前k列$\boldsymbol{U}_k$、$\boldsymbol{V}_k$分别作为词向量和文档向量,这就是隐含语义分析。在此基础上再进行下一步处理。

(3)两阶段聚类算法流程如下:

① $S=RandomS(D)$　//从数据集D中随机抽取样本S。

②将样本S划分成等大的n份,簇c_i、c_j中任意两个数据项p、p'以最小平均距离$dist(c_i,c_j)=\frac{1}{n_in_j}\sum_{p\in c_i}\sum_{p\in c_j}|p-p'|$进行局部聚类。

③通过随机取样剔除孤立点。

④新簇代表点$w.rep=p+\alpha\times(w.mean-p)$(收缩因子$\alpha=0.5$)。

⑤依据公式(4)分别计算每个层次上簇集合的$Q(c)$

表1　12天(同时段)内网站上热点话题与本文结果比较

新浪微博热点话题 (微博条数)	实验所得结果对应的微博条目
1月电影抢先看(159 249 348)	我参与江苏卫视 发起的投票 如果你是荔枝春晚的导演,你会邀请谁?
肖艳琴的遗书(8 050 161)	向这些人们致敬,春节快乐
今天你买到票了吗?(7 873 006)	全国列车今起全部实行实名制 实名验票最快8秒
CBA京粤大战(6 063 180)	南京枪杀储户案嫌犯8年致7死
2012我希望的(4 580 456)	今天,两个数字要公布,一龙门飞甲截至今晚,票房突破五亿大关二……
微博账号以一敌百(1 476 917)	看完十三钗后无处宣泄内心的压抑
你会悄悄关注谁(309 560)	突发新闻:蒙牛集团官网被黑,黑客质问蒙牛:你有良心吗?
全国列车都实名制了(91 966)	小三逼死原配门的当事人肖艳琴死而复生现身,讲述事情原貌..

值,抽取其中使得目标函数 $Q(c)$ 值最小的层次。

⑥ *LabRecord*(*D*)　//用相应的簇标签标记每个数据样本。

⑦对每一个类别 C 所有样本求其平均值,得到相应类中心。

⑧依据④和⑤得到的 K 值和聚类中心,执行 K-means算法,得到热点话题的聚类结果,结束。

上述③中,剔除孤立点分为两个阶段:(1)聚类过程中,若一个簇增长过慢(以点的个数作为阈值),则去除它;(2)聚类快结束时,剔除非常小的簇。

4　实验及分析

(1)数据集

目前还没有通用的中文微博数据集,本文通过新浪微博提供的API抓取了2 135个用户从2011年12月26日到2012年1月6日共12天,发表的所有微博数据。数据清洗后,选取长度为4个字符以上的微博文本共18 456条,每条微博平均长度为30个字符。

(2)实验结果及分析

实验1　将数据集随机分为两等份,分别记为D_set1和D_set2,预处理后进行CURE聚类,图2给出 Q 值和聚类数之间的关系。由图可知,Q 值取0.17时,聚类结果趋于稳定,K 值约等于30。然后进行 K-means聚类。

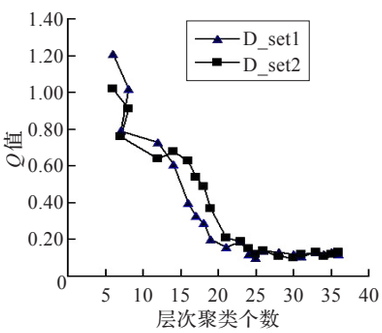

图2　Q 值和聚类数的关系

为了更直观地展现实验结果,列出同时段内微博网站推荐的"热点话题",与部分实验结果作对比。见表1(第1列是某时段内网站推荐的"热点话题",取前8个;第2列是利用本文方法得出的最"热门"话题所对应微博条目的前8个)。

实验2　为了比较传统VSM和相似矩阵经LSA处理后的效果,对D_set1进行手工标注,将标注出的第 i 个主题定义为 T_i,其中主题 T_i 的聚类准确率[16]定义为:

$$precision(T_i)=\max_{c_j\in c}\frac{n_{ij}}{n_j} \tag{7}$$

图3中的准确率为各个主题准确率的加权平均值。从图中可以看出,经LSA处理后,相似矩阵的生成时间和算法时间复杂度大大降低。k(LSA处理后的维度)值取200时,聚类准确率最高。在数据规模上,规模较小时,LSA和VSM的聚类准确率差别不太明显,当数据规模达到4 000之后,传统VSM(未经LSA处理)的数据因噪声影响,聚类准确率下降明显,而经LSA处理之后的数据,由于考虑到了字面背后隐含的语义信息,且在预处理阶段对微博的"热度"先做了筛选,在准确率和速度上都表现出了较明显的优势。

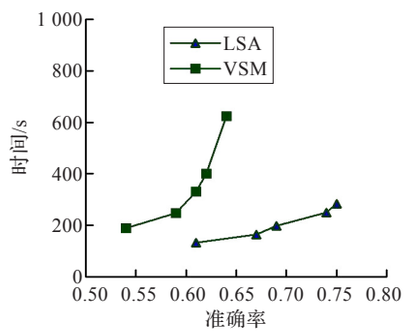

图3　聚类准确率与时间的关系

表1可以看出,第1列和第2列的热点话题不尽相同。本文在考虑微博评论数和转发数的同时,分析了微博内容隐含的语义信息,因而,实验所得微博的话题更接近该时间段内人们讨论最多的内容。相对而言,网站上的热点话题排名并不能全面反映大众真正关心的话题,主要是因为排名仅仅基于评论数和转发数或关键字匹配,而相当一部分评论是与主题无关的广告或链接,还有部分微博是人为发起的投票,对真正的热点话题结果造成了干扰。本文利用隐含语义分析的方法避免了语言形式多样性等带来的负面影响,得到的热点话题更接近实际状况。

5 结语

随着社交网络的兴起,面向社交网络,尤其是微博、论坛的话题发现具有越来越重要的现实意义。选取传播速度最快的微博作为发现话题的对象,具有很强的实用性。本文利用隐含语义分析的方法,解决了传统向量空间模型中高维和同义、多义问题。采用层次聚类CURE与*K*-means算法相结合的两阶段聚类策略,提高了话题发现的效率和准确率。同时也应该注意到,微博内容的权威性与用户角色有较强的相关性,本文对数据集的处理过程并未考虑这一特征;另外,利用本文方法得到的热点话题还不能完整地表述事件的时间等要素。因此,下一步工作主要集中在:细分用户角色并评估其发布的微博内容的权威性,进一步提高话题发现的准确率;如何更完整地展示话题的内容也是未来的目标。

参考文献:

[1] 李心妍,刘俐俐.浅析微博中的"微舆情"[J].新闻世界,2011(7):111-112.

[2] Lee Chunghong, Chien Tzanfeng, Yang Hsinchang.An automatic topic ranking approach for event detection on microblogging messages[C]//Proceedings of 2011 IEEE Internation Conference on Systems, Man, and Cybernetics (SMC),2011:1358-1363.

[3] 张晨逸,孙建伶,丁逸群.基于MB-LDA模型的微博主题挖掘[J].计算机研究与发展,2011,48(10):1795-1802.

[4] 杨冠超.微博客.热点话题发现策略研究[D].浙江:浙江大学,2011:20-22.

[5] 路荣,项亮,刘明荣,等.基于隐主题分析和文本聚类的微博客新闻话题发现研究[C]//第六届全国信息检索学术会议论文集,2010.

[6] 郑斐然,苗夺谦,张志飞,等.一种中文微博新闻话题检测的方法[J].计算机科学,2012(1):138-141.

[7] Raghavan V V, Wong S K M.A critical analysis of vector space model for information retrieval[J].Journal of the American Society for information Science, 1986, 37(5): 279-287.

[8] Connel M, Feng A, Kumaran G, et al.Umass at TDT 2004[C]//Proc of TDT 2004, 2004.

[9] Deerwesster S, Dumais S T, Fuvnas G W.Indexing by latent semantic analysis[J].Journal of the American Society for Information Sciens, 1990, 41(6): 391-407.

[10] 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.

[11] Wei C, Yang C, Lin C.A latent semantic indexing-based approach to multilingual document clustering[J].Decision Support Systems, 2008, 45(3): 606-620.

[12] Sun Jingtao, Zhang Qiuyu, Yuan Zhanting.Calculation of latent semantic weight based on fuzzy membership[C]// Proceedings of ISNN 2008, 2008: 91-99.

[13] 张华平.基于多层隐马尔科夫模型的中文词法分析[C]//第4届ACL会议暨第二届SIGHAN研讨会,札幌,日本,2003:63-70.

[14] Debole F, Sebastiani F.Supervised term weighting for automated text categorization[C]//Proceedings of the 18th ACM Symposium on Applied Computing, Melbourne, US, 2003: 784-788.

[15] Gale L D.A sequential algorithm for training tex classifiers[C]//Proceedings of ACM SIGIR Conference, 1994.

[16] 赵世奇,刘挺,李生.一种基于主题的文本聚类方法[J].中文信息学报,2007,21(2):58-62.

(上接73页)

[5] Mosso F, Tebaldi M, Torroba R, et al.Double random phase encoding method using a key code generated by affine transformation[J].International Journal for Light and Electron Optics, 2011, 122(6): 529-534.

[6] 鹏翔,位恒政,张鹏.光学信息安全导论[M].北京:科学出版社,2008:64-95.

[7] 刘建东,付秀丽.基于耦合帐篷映像的时空混沌单向Hash函数构造[J].通信学报,2007,28(6):30-38.

[8] Ge Xin, Liu Fenlin, Lu Bin, et al.An image encryption algorithm based on spatiotemporal chaos in DCT domain[C]// Proceedings of the 2nd IEEE International Conference on Information Management and Engineering(ICIME), 2010: 267-270.

[9] Yin Ruming, Yuan Jian, Yang Qiuhua, et al.A stream cipher based on discretized spatiotemporal chaotic system[C]// Proceedings of the 1st International Conference on Information Science and Engineering, 2009: 1613-1616.

[10] Wang Yong, Luo Longyan, Xie Qing, et al.A fast stream cipher based on spatiotemporal chaos[C]//Proceedings of the International Symposium on Information Engineering and Electronic Commerce, 2009: 418-422.

[11] He Jun, Zheng Jun, Li Zhibin, et al.Color image cryptography using multiple one-dimensional chaotic maps and OCML[C]// Proceedings of the International Symposium on Information Engineering and Electronic Commerce, 2009: 85-89.

[12] He Bo, Zhang Fang, Luo Longyan, et al.An image encryption algorithm based on spatiotemporal chaos[C]//Proceedings of the 2nd International Congress on Image and Signal Processing, 2009: 1-5.