改进的话题检测和跟踪算法研究

肖 红,许少华
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:话题检测可以及时发现互联网舆情热点和突发性事件,并可对话题进行持续跟踪,从而实时掌握舆情事件动向。文中提出了一种基于聚类的改进话题检测和跟踪算法。首先,对文本的特征向量进行改进,增加了基于句子主干的主干向量。然后对每个检测到的话题提取两个中心向量,一个是基本中心向量,另一个是基于主干向量提炼的主干中心向量。在此基础上再通过计算每个文本与中心向量之间的距离进行聚类分析,保证话题中各个文本之间的内聚性。同时基于主题词抽取,在主题词的基础上计算话题之间的主题相关性,有效地实现了子话题检测功能,从而提高了话题检测和跟踪的准确性。通过对 10 大网站 5 个频道超过两周数据量的测试,结果表明此方法在一定程度上提高了话题检测和跟踪的正确率,并具有一定的适应性和推广性。

关键词:话题检测和跟踪;聚类算法;特征向量;网络舆情
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2014)09-0084-05
doi:10.3969/j.issn.1673-629X.2014.09.019

Improved Algorithm Study on Topic Detection and Tracking

XIAO Hong, XU Shao-hua
(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: The topic detection can detect hot Internet public opinion and emergencies, and can carry out the continuous tracking of the topic, which can get a real-time grasp of public opinion trends. Propose an improved algorithm for detecting and tracking based on topic clustering in this paper. First, to improve the feature vectors of document, increase the backbone vectors based on sentence trunk. Then two center vectors are extracted from each detected topic, in which one is the basic center vector and another is the main center vector. On this basis, by calculating the distance between the document vector and the corresponding center vector, the cluster analysis is performed to ensure the cohesion of each document for the same topic. Meanwhile, based on keyword extraction, the theme correlation between different topics is calculated to improve the accuracy of topic detection and tracking. Taking the top 10 sites 5 channel data for more than two weeks as the test data, the experimental results show that this method improves the accuracy of topic detection and tracking to some extent, and has certain adaptability and generalization.

Key words: topic detection and tracking; cluster algorithm; feature vector; network public opinion

0 引 言

网络舆情监测是一个包含众多关键技术的综合性课题,其中话题检测和追踪是舆情系统较为核心的一个应用功能。通过话题检测可以分析每天的舆情热点,并在此基础上对某个话题进行跟踪,从而分析整个话题,即舆情事件的起因、发展至消亡的全过程。随着互联网技术的发展和普及,网络媒体每天发布的新闻资讯、博客、论坛帖子呈指数性增长,微博更是随时随刻更新海量的碎片化信息。因此如何快速、准确地从海量信息中发现热点话题并对其进行有效的跟踪是舆情监测系统研究的重点。

话题检测与跟踪(Topic Detection and Tracking, TDT)主要是利用聚类技术对海量的网络数据进行聚类分析,将讨论和报道相同话题的新闻资讯、论坛帖子、博客内容聚合到一个统一的分类中,建立一个热点话题,并在后续新增的数据中进行增量聚类,对已经存在的话题进行追踪分析。其中核心的算法就是聚类算法,聚类有较为成熟的算法模型,其算法也是多种多

收稿日期:2013-11-04 **修回日期**:2014-02-09 **网络出版时间**:2014-07-17
基金项目:国家自然科学基金资助项目(61170132)
作者简介:肖 红(1979-),女,黑龙江林甸人,硕士,副教授,通讯作者,研究方向为人工智能与图像处理。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140717.1233.045.html

样,有 SVM、KNN、贝叶斯等。但话题检测不等于聚类,一个热点话题可能包括多个子话题,是一个多层次聚类,需要结合行业应用特点进行改进和参数调整,从而更好地适应网络舆情应用环境。

文中提出了一种改进的话题检测和跟踪算法,是在增量聚类方法的基础上,引入句法分析,分析文章中各个句子的主干,在句子主干的基础上抽取 VSM 向量,作为控制话题发散的特征向量。同时为了加快聚类的速度,引入一个中心向量的概念,降低向量运算的时间复杂度,对于增量检测的文章只需要与中心向量进行相似性分析即可,并控制其与中心向量的距离,保证话题的内聚性。最后还需要对话题进行合并处理,根据话题的主题进行子话题检测和合并。实验结果表明,该话题检测方法相比同类方法具有更高的准确率和效率,并且保证了话题的内聚性,进而对话题进行精确的跟踪分析。

1　相关工作

目前话题检测的算法模型主要有两种,即统计模型和语义模型。

基于统计的方法是一种传统且成熟的算法,它主要是在文本切词的基础上对文本中出现的各个词进行词频统计,再按词频高低选取一定数量的特征词作为文本的特征表示,虽然在计算特征词权重时进行了综合评判,但核心仍然是统计词频。通过两个文本的高频特征词集合之间的匹配,计算出共有特征词数量,在此基础上通过权重计算文本之间的相关性程度。由于完全基于机械的概率统计,所以统计模型无法有效区分同一话题下的不同事件。

语义模型利用自然语言处理技术,对文本进行语法分析和语义识别,在此基础上提取语义主题模型,再通过分析两个文本主题在语义上的相关性来计算。语义模型较统计模型来说正确率大大提高,但程序实现复杂,并且在语言书写比较自由的网络环境中纯粹的语义模型算法适应性较差。

近年来,随着网络舆情系统应用的深入,话题检测和跟踪得到了很多专家学者的持续研究,并取得了较好的应用效果。但是大多数研究者都是基于内容的统计模型,在统计模型的基础上建立向量矩阵,再通过矩阵分析发现话题,并对话题进行增量聚类。如文献[1]就是一种基于增量型聚类的自动话题检测研究,它在聚类过程中动态调整话题的特征向量,虽然正确率提高了,但是调整特征向量以后需要对已聚类的文章再次进行相似性计算,整个计算的效率不高。文献[2]也是基于统计模型,用 VSM 向量进行聚类分析,然后再二次聚类,对聚类进行合并和拆分,再得到最终的结果。

另外也有一些学者基于语义模型进行了一些研究和实践,如文献[3-6]都采用了语义模型,但都是将语义模型和统计模型相结合,实践证明语义模型和统计模型相结合,并加以一定的算法改进策略会得到较好的应用效果。

2　改进的话题检测和跟踪算法

文中所提出的改进算法,是先通过统计模型得到各文本的特征向量,然后再利用句法分析得到各句的句子成分,并对句法树进行剪裁抽取各个句子的主干,再对句子主干中的词汇进行统计分析得到一个主干词特征向量。这两个向量分别称作基本向量和主干向量,再利用这两个向量进行双向量的聚类分析。按基本向量和主干向量分别聚类,再对这两个聚类结果求交集,交集部分就是所检测到的新话题。

而对于已有话题的增量检测,为了加快速度对话题对应的各文本特征向量进行优化和提炼,从而形成一个中心向量,新增加的文章先提取基本向量和主干向量,再利用两个向量分别与两个中心向量进行相似性计算,根据计算结果再进一步确定具体所属的话题。

通常情况下,互联网上的一个舆情事件(话题)包含多个子话题,光从向量相似度分析的结果上进行判断,很有可能会将属于同一话题的两个子话题分别归结为两个独立的话题,但是在舆情系统应用中应该将这两个话题归结到一个共同的父话题中,这样才能对话题进行持续跟踪分析。

2.1　特征向量抽取

目前对于文本的特征表示,业界使用最广泛的文本表示模型是向量空间模型[7-9],在 VSM 向量中每篇文章都被表示成一个向量:$\boldsymbol{D}_i = (T_1:W_1;T_2:W_2;\cdots;T_n:W_n)$,其中 T_i 表示特征词,W_i 表示 T_i 的权重[10-11]。对于每一个特征的权重采用成熟的 TF-IDF 模型来计算,TF (Term Frequency)表示特征词在文档中出现的频率,IDF 表示逆文档频率[12-13]。TF×IDF 值在信息检索领域广泛使用,用来衡量一个词语所包含的信息量,它是文中关键词抽取工作的基准度量之一,采用如下公式:

$$\mathrm{TF}\times\mathrm{IDF} = \mathrm{tf}(t)\times\log(N/\mathrm{df}(t)+0.01) \quad (1)$$

其中,t 是候选词汇;$\mathrm{tf}(t)$ 是 t 在测试文献中的词频;N 是训练样本数目;$\mathrm{df}(t)$ 是 t 在整个训练集中出现的文献数目。

为了增强话题内文章的内聚性,也就是提高聚类的正确性,系统采用双向量,一个是基本向量,另一个是主干向量。基本向量的抽取完全根据文本内容的切词结果进行统计词频,再筛选出若干特征词作为文本

的特征向量。

文中的研究创新性地应用了双向量策略,并且是在句法分析的基础上得到文章句子主干向量作为文章的特征,大大减少了噪音数据对于聚类准确率的影响。

2.2 基于句法分析的主干向量抽取

主干向量的抽取和普通向量的抽取方法一致,只不过需要对文本进行预处理。文中采用语义分析模型,对文本中的所有句子进行句法分析,在语法树的基础上对分析得到的句法树进行裁剪,得到句子主干。再以句子主干为基础抽取文本的特征向量。其中句法分析方法采用文献[3]中所提及的方法,将中心词驱动模型和结构上下文模型有效结合在一起进行句法分析,最终提取句子主干。其算法流程如图 1 所示。

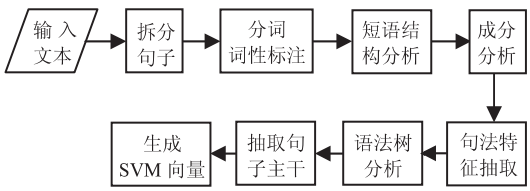

图 1 主干向量抽取算法流程

2.3 中心向量提炼

文中的中心向量是指代表话题特征的一个 VSM 向量,这个中心向量不是具体某个文本的特征向量,而是经过计算从话题所有文本 VSM 中提炼出来的。前面提到系统中有两种向量,即基本向量和主干向量,同样中心向量也有两个,一个是基本中心向量,另一个是主干中心向量。

在提炼中心向量时首先对话题中所有的向量特征(词汇)进行统计,按出现频次排序,选取频次大于一定阈值的特征词组成新的向量,这个向量就是代表话题的中心向量,每个特征词的权重计算是在各向量中权重取平均值的基础上乘以一个系数,其计算公式如下:

$$W_{0i} = \frac{1}{M}\sum_{i=1}^{m} w_i \times (1 + \log(\mathrm{tf}_{ij} + 1) \times \log(\frac{N+1}{\mathrm{df}_i})) \tag{2}$$

其中,M 为当前话题中所有文档数;N 为所有话题中包括的所有文档数;W_{0i} 表示中心向量中某一特征的最终权重值;tf_{ij} 是特征词 t_i 在话题 j 中的频次;df_i 为文本频次。

2.4 子话题检测

由于 VSM 向量的特征维比较多,而且每一维都会影响计算出的向量距离。而同一话题随着时间的推移会出现众多子话题,如果纯粹用向量运算的方法去计算,很难将它们归结为同一个话题。因此需要用其他的特征来检测是否属于同一话题。文中主要采用主题检测方法来实现,也就是用一个或者几个主题词来刻画文章主题。主题词的提取方法众多,很多学者都是基于互信息与贝叶斯网络进行主题词抽取,还有一些学者是通过词聚类[2,4]方法来进行抽取。文中采用基于同现关系的计算方法,具体参考文献[14]。

主题词抽取旨在找出一些对主题贡献大,具有较强主题表现力的非高频词汇。对于每一个检测出的新话题对所包含的所有文章进行挖掘分析,先计算各个词的权重,然后根据权重筛选得到最终的主题词。在计算词的权重时充分利用句法分析的结果,将句子主干词作为一个权重因子,其权重计算公式如下:

$$W(w_i) = \mathrm{TF} \times \mathrm{IDF} \times f_{\mathrm{p}}(w_i) f_{\mathrm{l}}(w_i)(1 + \frac{F(w_i)}{F(w_i)+1}) \tag{3}$$

式中,TF×IDF 是词频信息;$F(w_i)$ 代表词汇 w_i 有固定共现关系的词汇个数,$1 + \frac{F(w_i)}{F(w_i)+1}$ 是词汇 w_i 的主题表现力因子;$f_{\mathrm{p}}(w_i)$ 是位置权重因子,如果是文章标题中出现的词则值为 1,如果是非标题中出现则其值为 0.5;$f_{\mathrm{l}}(w_i)$ 是词干权重因子,如果 w_i 是句子主干词则权重为 1,非主干词则权重为 0.5。

计算出各个词的词频以后再根据权重排序,筛选出前 n 个词作为话题的主题词。在进行话题合并时计算话题之间主题词的相关度即可,主题词相关度用向量之间的距离计算公式,但需要优先考虑主题词重合度,将其作为一个权重因子,其具体的计算公式如下:

$$\mathrm{Sim}(T_i, T_j) = \mathrm{Sim}(X, Y) \times (1 + \frac{R(i,j)}{N_i + N_j}) \tag{4}$$

式中,$\mathrm{Sim}(X,Y)$ 是话题 T_i 和 T_j 所对应主题词向量的距离;$R(i,j)$ 代表两个话题中词汇的重合度;N_i 和 N_j 分别是两个话题的主题词个数。

2.5 算法实现

舆情系统话题检测和跟踪是按天为单位进行处理。文中首先从抓取的新闻数据中提取当天的数据,然后对每篇新闻进行分析挖掘,分别生成基本向量和主干向量。如果当天已经有话题生成则先对数据进行增量分类,此时只需要计算待分类文本向量与话题中心向量之间的距离,向量之间的距离用余弦夹角公式来计算,其公式如下:

$$\mathrm{Sim}(X,Y) = \frac{XY}{|X||Y|} = \frac{\sum_{i=1}^{n}(x_i y_i)}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \tag{5}$$

每一个话题用一个七元组来表示,其结构如下:

(1)话题 ID;

(2)话题标题;

(3)基本中心向量;

(4)主干中心向量;

(5)文章个数;

(6)创建时间;

(7)更新时间。

每一篇新闻用基本向量和主干向量分别进行一次分类,分类时先通过向量相似性在已有的向量列表中进行相似性计算,根据相似度阈值筛选出最相似的话题作为分类结果,此时得到两个分类结果集 $C_1(T_1, T_2, \cdots, T_n)$ 和 $C_2(T_1', T_2', \cdots, T_n')$,对这两个集合根据话题 ID 求交集,并按相似度从大到小排列,选择相似度最大的话题作为最终的分类结果,此时将这篇新闻归类到此话题当中。

遍历完一遍以后,再对剩下的数据用 KNN 方法进行聚类。此时以剩余数据为集合,在这些数据集上相互之间分别用两个向量进行相似性计算,再对两个相似度值求平均值得到最终的相似度值。根据相互之间的相似性关系划分成一个个的话题组,对于超过两篇以上文章且相似度大于阈值的话题作为新发现话题存入话题列表。同时对新发现的话题进行两种操作,首先根据新话题所包含的文章,分别提取各文章相应的向量,对其进行加权合并,得到话题的中心向量。中心向量各特征词的权重取平均值即可。然后用中心向量到最近一周的历史话题中检测,根据向量间的夹角计算出是否是某一个话题的延续。如果找不到相关话题,则根据话题主题进行子话题检测,如果检测到,则将根据检测到的同级子话题生成父话题,并建立它们之间的关系,从而对话题进行跟踪和演化分析。其详细的检测跟踪流程如图 2 所示。

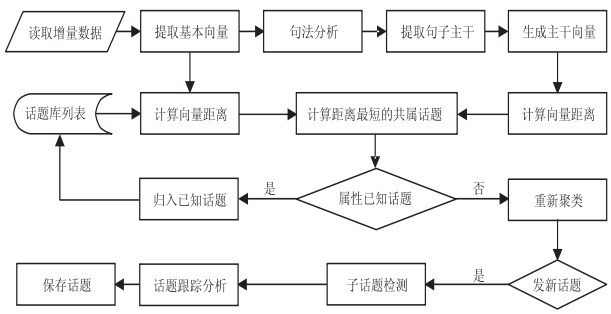

图 2　话题检测跟踪算法流程

3　实验结果

因为一个网络舆情事件或者话题其平均的生命周期大约是两周,所以在文中的实验中,以两周内的网络新闻报道为测试数据进行话题检测和跟踪测试,主要抓取了新浪、搜狐、网易、腾讯、新华网、人民网、环球网等 10 家主流媒体的国内、国际、时政、社会、财经 5 个频道的数据,共计 8 万多条数据。在测试数据样本集上进行了两组实验,第一组实验对文中的算法和普通的基于聚类的算法进行对比测试,第二组实验主要是对文中的算法本身在不同的参数设置和策略下进行测试。

3.1　第一组测试

第一组测试主要是对比测试传统算法与文中算法之间的差异性,话题检测相似度阈值和话题间跟踪阈值统一设定为 40 和 30,其测试结果见表 1 和表 2。

表 1　话题检测实验结果

算法	话题总数	正确数	正确率/%
普通算法(KNN)	249	137	55.02
文中改进算法	112	102	91.07

表 2　话题跟踪实验结果

算法	测试周期内延续性话题总数	测试周期内延续性话题正确数	正确率/%	话题持续时间/天	
				平均	最长
普通算法(KNN)	102	43	42.15	12	>14
文中改进算法	64	57	89.06	7	12

3.2　第二组测试

从第一组实验中可以看出文中提出的算法明显提高了话题检测的正确率,但策略和参数比较多,不同的参数和策略对算法正确率和运算速度会有一定的影响。为了找到最佳的参数设置,笔者主要从相似度阈值、聚类种子选取等方面进行了一系列有针对性的测试,其测试结果分别如表 3 至表 5 所示。

表 3　不同相似度阈值下话题检测实验结果

相似度阈值	话题总数	正确数	正确率/%	话题识别率
30	123	99	80.48	高
40	112	102	91.07	较高
50	81	80	98.76	低

表 4　聚类种子对话题检测结果影响的实验结果

聚类种子选取	话题总数	正确数	正确率/%	话题识别率
随机	125	98	78.4	低
全局最优选取	112	102	91.07	高

表 5　不同阈值下的话题跟踪实验结果

相似度阈值	延续性话题总数	延续性话题正确数	正确率/%	话题持续时间/天	
				平均	最长
20	53	46	86.79	>14	>14
30	64	57	91	7	13
40	32	30	93.75	6	10

综合上面的实验结果,文中提出的改进算法有效提高了话题检测和跟踪的正确率,而且在聚类时如果采取一定的策略选择最优化的种子文本则正确率会进一步提高,同时在话题检测和跟踪阶段相似度阈值分别选择 40 和 30 时此算法的效果最佳。

4 结束语

话题检测和跟踪是网络舆情系统最重要的应用之一,几乎每个舆情系统都会应用话题检测和跟踪分析网络舆情热点和突发事件。由于互联网每天发布的信息量大,如何快速准确地发现热点话题和突发性事件是近几年来舆情应用研究的热点。大多数研究者的共识就是应用统计模型,在自动聚类的基础上进行话题检测,将计算正确率的提高归结于聚类算法的改进上,很少关注聚类时的策略和方法的改进。但实际上策略和方法上的改进在一定应用范围内比算法改进更有效。文中就是运用语法分析抽取文中所有句子的主干,并根据主干提取 VSM 向量。基于双向量的话题检测大大提高了检测的正确率,在网络舆情热点监测方面得到了较好的验证和应用。下一步的工作将继续深入研究特征向量的提取方法和中心向量提炼策略,在语义分析的基础上进一步提高话题检测的正确率,从而得到更好的应用效果。

参考文献:

[1] 张小明,李舟军,巢文涵. 基于增量型聚类的自动话题检测研究[J]. 软件学报,2012,23(6):1578-1587.

[2] 王振宇,吴泽衡,唐远华. 基于多向量和二次聚类的话题检测[J]. 计算机工程与设计,2012,33(8):3214-3218.

[3] 米海涛,熊德意,刘 群. 中文词法分析与句法分析融合策略研究[J]. 中文信息学报,2008,22(2):10-17.

[4] 马 彬,洪 宇,陆剑江,等. 基于线索树双层聚类的微博话题检测[J]. 中文信息学报,2012,26(6):121-128.

[5] 洪 宇. 基于语义结构和时序特征的话题检测与跟踪技术研究[D]. 哈尔滨:哈尔滨工业大学,2009.

[6] 常 鹏,马 辉. 高效的短文本主题词抽取方法[J]. 计算机工程与应用,2011,47(20):126-128.

[7] Hofmann T. Probabilisitic latent semantic analysis[C]//Proc of UAI99. [s. l.]:[s. n.],1999.

[8] Li Hang, Yamanishi K. Topic analysis using a finite mixture model[J]. Information Processing and Management, 2003, 39(4):521-541.

[9] Landauer T, Foltz P, Laham D. Introduction to latent semantic analysis[J]. Discourse Processes, 1998, 25:259-284.

[10] 唐籍涛,李 飞,郭昌松. 网络舆情监控中新词识别问题的研究[J]. 计算机技术与发展,2012,22(1):119-121.

[11] 林鸿飞,高 天,姚天顺. 中文文本的可视化表示[J]. 东北大学学报:自然科学版,2000,21(5):501-504.

[12] 金 珠,林鸿飞,赵 晶. 基于 HowNet 的话题跟踪及倾向性分类研究[J]. 情报学报,2005,24(5):555-561.

[13] 赵 华,赵铁军,张 姝,等. 基于内容分析的话题检测研究[J]. 哈尔滨工业大学学报,2006,38(10):1740-1743.

[14] Hotho A, Stumme G. Conceptual clustering of text clusters [C]//Proceedings of FGML workshop. Piscataway, NJ, USA: IEEE, 2002:1-9.

(上接第 83 页)

室环境下协助医生进行手术以提高手术效率,减轻医护人员劳动强度。医生可以在手术过程中随时使用自己的手势、身体姿态及语音作为输入元素来实现病例资料、医学影像和化验报告的切换;同时可以实现医学影像的放大缩小、翻页、旋转等功能。

参考文献:

[1] 余 涛. Kinect 应用开发实战[M]. 北京:机械工业出版社,2012.

[2] 董士海. 人机交互的进展及面临的挑战[J]. 计算机辅助设计与图形学学报,2004,16(1):1-13.

[3] Hernandez-Lopez J J, Quintanilla-Olvera A L, Lopez-Ramirez J L, et al. Detecting objects using color and depth segmentation with Kinect sensor[J]. Procedia Technology, 2012,3:196-204.

[4] 李 斌,吴国斌. Kinect 引领人机交互变革[J]. 程序员,2011(9):100-103.

[5] 邓 瑞,周玲玲,应忍冬. 基于 Kinect 深度信息的手势提取与识别研究[J]. 计算机应用研究,2013,30(4):1263-1265.

[6] 何 贝,王贵锦,林行刚. 结合 Kinect 深度图的快速视频抠图算法[J]. 清华大学学报:自然科学版,2012,52(4):561-565.

[7] 吕开阳,叶华茂,李晓光,等. Kinect 体感技术在动物外科实验教学中的应用及展望[J]. 中国医学教育技术,2012,26(2):171-173.

[8] Kuehn T. The Kinect sensor platform[J]. Advance in Media Technology,2011(6):2192-2198.

[9] Gallo L, Minutolo A, de Pietro G. A user interface for VR-ready 3D medical imaging by off-the-shelf input devices[J]. Computers in Biology and Medicine,2010,40(3):350-358.

[10] 陈一民,张云华. 基于手势识别的机器人人机交互技术研究[J]. 机器人,2009,31(4):351-356.

[11] 郝颖明,朱 枫. 外科手术计算机辅助导航技术[J]. 生物医学工程杂志,2004,21(2):306-310.

[12] Chang Y J, Chen S F, Huang J D. A Kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities[J]. Res Dev Disabil, 2011, 32(6): 2566-2570.

[13] 罗 元,谢 彧,张 毅. 基于 Kinect 传感器的智能轮椅手势控制系统的设计与实现[J]. 机器人,2012,34(1):110-113.

[14] Michael N, Fink J, Kumar V. Cooperative manipulation and transportation with aerial robots[J]. Autonomous Robots, 2011,30(1):73-86.

改进的话题检测和跟踪算法研究

作者： 肖红， 许少华， XIAO Hong， XU Shao-hua
作者单位： 东北石油大学 计算机与信息技术学院,黑龙江 大庆,163318
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(9)