doi:10.3969/j.issn.1000-436x.2013.z1.034

# 基于有意义串聚类的微博热点话题发现方法

贺敏[1,2]，王丽宏[2]，杜攀[1]，张瑾[1]，程学旗[1]
（1. 中国科学院 计算技术研究所，北京 100080；2. 国家计算机网络应急技术处理协调中心，北京 100029）

**摘 要**：针对微博数据特征稀疏、内容碎片化的特点，提出一种基于有意义串聚类的热点话题发现方法。结合重复串计算、上下文邻接分析和语言规则过滤多种策略，提取能够表达独立完整语义的有意义串，并将微博数据建模在相对较小的有意义串空间，通过聚类产生候选话题，根据热度排序发现热点话题。微博数据实验结果表明，该方法在一定程度上实现对微博高维稀疏空间的降维，对于微博空间的热点话题发现有效可行。
**关键词**：热点话题；微博；有意义串；特征聚类
**中图分类号**：TP391 **文献标识码**：A **文章编号**：1000-436X(2013)Z1-0256-07

# Microblog hot topic detection method based on meaningful string clustering

HE Min[1,2], WANG Li-hong[2], DU Pan[1], ZHANG Jin[1], CHENG Xue-qi[1]
(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China;
2.National Computer network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China)

**Abstract:** Aiming at the properties of sparse feature, content fragmentation for microblog data, a hot topic detection method was proposed based on meaningful string clustering. The multiple strategies including repeated string detection, context analysis and language rule filtering were combined to extract meaningful strings. Candidate topics were generated by clustering with distribution of meaningful strings in documents. The hot topics were detected according to hotness sorting for candidate topics. As is shown from the experiment results on microblog data, the method achieves good effect in solving the problem of data sparseness. It is effective and feasible to hot topic detection for microblog.
**Key words:** hot topic; microblog; meaningful string; feature clustering

## 1 引言

微博是近年来兴起的 Web2.0 新媒体。用户可以通过手机、即时通信工具、Email、Web 等媒介在个人微博上发布 140 字以内的文本信息及图片、影音等多媒体内容，展现个人最新动态，实时分享身边信息。微博用户数量大，信息传播速度快，已经成为信息产生、发展、传播的重要平台。

微博平台上每天产生的信息数量庞大，据统计，新浪微博 2012 年 11 月日均发微博量约 1.366 亿条，平均每分钟约 94 907 条。微博在为用户带来新鲜及时且丰富繁杂信息的同时，也带来了严重的信息过载和信息碎片问题。而面向微博数据的话题发现技术，能够从话题粒度上重新组织微博数据，成为解决上述问题的关键技术之一。及时、准确的发现热点话题，能够帮助个人了解社会热点和重要资讯，辅助国家发现网络舆情事件和舆论趋势，在舆情监控、信息安全等领域具有重要的现实意义。

但微博数据具有内容短小、数据巨大、信息零碎、用语不规范等不同于传统新闻文档的显著特点，这些新特点为面向微博的热点话题发现技术带来了新的挑战。

1) 数据高维稀疏导致内容关系难以准确计算。传统特征向量文本表示方法中，通常以词作为特

收稿日期：2013-07-05
基金项目：国家科技支撑基金资助项目 (2012BAH46B01)；国家自然科学基金资助项目(61170230)
**Foundation Items:** The National Science and Technology Support Project (2012BAH46B01);The National Natural Science Foundation of China(61170230)

征，并使用 TF-IDF 方法来衡量每个特征(即向量每一维) 的权重。但是对于微博来说，它的文本内容非常短，同一个词出现在不同短文本中的概率会远小于长文本，这种数据的稀疏性，使得传统文本表示方法很难准确计算文本间的相似度。

2) 不完整的信息碎片对语言上下文感知产生严重依赖。每个用户随时都可以发表微博，信息具有原创性和时效性的同时，也表现出草根性和随意性，用词口语化、不规范现象严重，简称、缩略语大量存在，这些碎片信息需要依赖语言上下文来辅助分析。

3) 新词涌现导致词空间的动态变化。随着网络事件的事态发展，微博空间不断涌现出大量的新词，用传统的静态词典中的词语来表示文本，将会遗漏部分关键特征。

本文针对上述挑战，提出将有意义串作为微博的基本表达单元，通过有意义串聚类来发现热点话题。有意义串是结合频繁字符串发现、上下文语境分析和语言规则过滤方法从微博中分析出的完整独立的粗粒度语义单元，可能包含新词、命名实体，以及比词粒度更大的词组或者短语，其语义更加明确、完整，且与传统词空间相比，有意义串空间维度相对较低。因此，有意义串既能辅助发现微博中频繁出现的新词、网络流行语、暗语等不规范用语，又能在一定程度上缓解传统词空间的高维稀疏性问题。

## 2　相关工作

传统的话题检测一般采用文本聚类方法来实现，其中，文本表示通常使用向量空间模型(VSM)，文本相似度计算多采用欧式距离和余弦夹角距离，聚类方法通常有中心向量法、层次聚类法、*K*-Means 算法、Single-Pass 算法等。目前关于话题检测的研究主要集中在文本特征的选择和聚类方法的适应性改进等方面。在特征选择方面，自然语言处理(NLP)技术被用于辅助统计策略发现特征[1~4]，Kumaran 将报道描述成 3 种向量空间, 分别为全集特征向量、仅包含命名实体（NE）的特征向量和排除命名实体的特征向量[1]，对比了 3 种向量空间模型对话题检测的影响, 并验证 NE 极大地促进了话题之间的区分。在聚类算法改进方面，Papka 等首先提出 SinglePass 的聚类思想[5]，是一种效果较好的在线话题检测方法；雷震等提出了一种用于事件检测的增量 *K* 均值聚类算法[6]，使用密度函数法进行聚类中心初始化；骆卫华等提出了针对事件特点的单粒度话题识别方法，采用了自底向上的层次凝聚式聚类方法[7]。这些方法在新闻报道领域取得了较好的效果，但是对于特征非常稀疏的微博信息，这些方法的适应性还有待评估。

近年来，也出现了一些针对微博文本方面的研究，主要思路有 2 种，一种是通过特征扩展来解决特征稀疏的问题，如 Sharifi 等将 Twitter 文本分类到预定义的话题类别时，通过抽取作者 profile 文件和文本记录中的领域相关的特征集合，对 Twitter 文本的特征进行扩充[8]。Liu 等试图借助 HowNet 实现特征扩展，但也带来了一些噪音信息，导致处理效果提高不明显[9]；另一种思路是提高关键特征提取准确率，Lee 等结合滑动窗口技术提出了一种名为 BursT 特征权重算法，该算法能够在动态环境抽取重要特征，与递增 TF*IDF 进行比较，具有明显的优势[10]。Du 等提出了另一种新的关键词权重计算方法，词的权重计算主要考虑到用户的权威性、受众数量、回复数量及关键词集等因素[11]。

特征扩展的方法虽然对缓解特征稀疏有帮助，但是又引入了新的特征，增加了特征维度，也会带来一些噪音，不能从根本上解决特征稀疏的问题。关键特征提取时，往往综合考虑除文本外的多种因素，加大了信息要素提取的难度和计算复杂度。本文方法的实现过程只利用了文本信息，在降低特征空间维度的同时，具有简洁、高效的优势。

## 3　基于有意义串聚类的微博热点话题发现

### 3.1　微博有意义串

一条微博信息内容短小，包含的语义成分也比较少。通过微博信息中的几个关键字符串，就几乎涵盖了一条微博信息的主要内容。例如，下面微博内容：“美国总统大选的选民投票日将于当地时间 6 日举行，美国总统奥巴马和共和党总统候选人罗姆尼终迎来最后“对决”。虽然近半美国选民已心有所属，但两人仍不遗余力拉票至最后一刻。由于今年选情异常，究竟谁能在未来 4 年入住白宫仍难以预料”。

此条微博信息的关键字符串是“美国总统大选”、“奥巴马”和“罗姆尼”。

这些关键字符串或者是有意义的短语，或者是未登录的新词和命名实体，都具有语义完整性，但突破了词典中词语的界限，可能是多个词语的组

合。这些字符串包含具体语义，是灵活独立的语言单元，能在多种不同语境中使用，称为“有意义串”，本文将有意义串作为表示微博关键信息的新特征。下面给出这种新特征的定性描述。1) 语义上：表意完整单一、所指明确，在意义上有一定的完整性和专指性；2) 结构上：内部结构比较稳定，具有一定的凝固性；3) 语用上：是一个独立的语言单元，使用环境灵活多样；4) 统计上：在大规模真实文本中具有一定的流通度，并非临时性的组合，可重用性强，具有统计意义。

将有意义串作为新特征来表示微博文本，与通常的词语特征相比，主要有以下 3 个方面的优势。

1) 能够降低特征维度，缓解特征稀疏问题。如有意义串“黄斑病变性失明”是一个特征，而用词语来表示将划分为“黄斑”、“病变”、“性”“失明”4 个特征。

2) 有意义串中的新词和命名实体经常是表征一个话题的关键词，对于提高话题检测准确率有重要贡献。如新词“国五条”、“禽流感”本身就是话题。

3) 有意义串中的短语或词组作为文本特征项，与词语特征项相比，引入了一定的词序，表达了更强的语义概念，提高了特征项的完整独立性。如短语“袭击巴格达”表达的语义为“袭击”的对象是“巴格达”，而仅仅通过“袭击”和“巴格达”2 个词语不能完整表达出这层意思，有可能理解是“巴格达”参与了“袭击”。

有意义串作为独立使用的语言单元，内部结合度比较高，而上下文环境灵活多变，为了描述这种上下文语境灵活程度，Feng 等提出了 Accessor Variety 的概念[12]，本文在其基础上，对语言单位进行了扩展，定义字符串的邻接种类如下。

**定义 1** 邻接种类(variety of neighbour)：真实文本中，与字符串 $S$ 左边或者右边相邻的元素的集合，分为左邻接集合和右邻接集合。元素可以是字、词等不同粒度的语言单位，一个句子的开始和结尾也分别记为一个元素。左邻接集合中和右邻接集合中元素的数目，称为左邻接种类 $VN_{\mathrm{L}}$ 和右邻接种类 $VN_{\mathrm{R}}$。

例如，下面是微博信息中的几个真实例子。

1) 钟南山透露禽流感病毒尚未明显变异。

2) 广东的防控禽流感形势趋缓。

3) 有 7 人感染禽流感事件。

4) 发现一宗禽流感病毒疑似病例。

“禽流感”是近年来出现的新词，在词典中并未收录。如果将词作为邻接分析的粒度，字符串“禽流感”的左邻接种类 $VN_{\mathrm{L}}$=4，右邻接种类 $VN_{\mathrm{R}}$=3。字符串“禽流”的 $VN_{\mathrm{L}}$=4，$VN_{\mathrm{R}}$=1。当一个串的左邻接种类和右邻接种类同时满足一定条件时，该串才具有语用灵活性，可能成为有意义串。当字符串的左、右邻接种类一个较大，另一个为 1 时，该串的语义不完整，是其他有意义串的子串。因此，邻接种类体现了字符串的上下文语言环境，是字符串语义完整性的一种度量。

### 3.2 有意义串提取算法

准确提取微博中的有意义串新特征，是高效发现热点话题的基础。语义完整性是有意义串的基本属性，也是判断一个串是否有意义的的首要特性。但是，借助计算机来理解语义比较困难，而有意义串在结构、语用、统计上还有一些其他特性，这些性质能够表示成计算机可计算的形式。本文从统计、语用、结构 3 个方面来综合分析，辅助判断一个串的语义完整性，提取大规模微博数据中的有意义串。首先通过重复串发现寻找候选特征，然后由上下文邻接类别判断特征的语用灵活程度，最后利用语言规律过滤频繁垃圾特征，获取微博信息中有意义的字符串。算法如图 1 所示。

```
Input：Microblog Corpus
SegMent the corpus with ICTCLAS;
Find Repeated Strings of the corpus;
For each Repeated String i{
    Comput the min VN;
    if(min VN<T_VN) continue;
    if (isStopWord(first word)or isStopWord(last word)) continue;
    if(isFuctionWord(middle word)) continue;
    put the Repeated String into Meaningful String Set;
}
Output：Meanningful String Set
```

图 1 有意义串提取算法

微博信息虽然数量巨大，但很多都是全文转发或者稍加修改后转发，信息近似重复率很高。具有统计意义的微博新特征应该包含在重复串[13]中，是重复串集合的子集，因此发现重复串是提取频繁有意义串的第一个步骤。本文采用中国科学院计算技术研究所的 ICTCLAS 分词系统对微博信息分词，

然后使用龚才春提出的低频剪枝方法[13]来发现重复串，其中重复串的长度需要限制在一定范围内。

发现重复串后，将对重复串进行上下文邻接分析。本文以词为基本粒度单位，计算每个串的左邻接类别 $VN_L$ 和右邻接类别 $VN_R$，选取左邻接类别和右邻接类别中的较小值记为 min$VN$。当重复串的 min$VN$ 小于阈值 $T_{VN}$ 时，认为该串语用不灵活，不是有意义串。

经过语用分析后，有一些重复串的 min$VN$ 值也比较大，语用环境很灵活，但是语义不够明确，本文将进一步根据语言规律过滤无意义垃圾字符串。通过观察，垃圾字符串具有以下 2 种特征之一。

1) 串首或者串尾包含停用词，如“又不能”、“可被”、“奥巴马在”。

2) 串中间含有副词、助词、连词等虚词，如“做的事情”“行走被车上”。

在汉语语言中，副词、助词、介词等虚词只起连接修饰作用，属于封闭集合。因此，本文从汉语语言规律出发，人工整理停用词集合和虚词集合，通过首尾停用词过滤和串中虚词过滤 2 种规则来进一步提纯有意义串。

### 3.3　热点话题发现算法

微博热点话题是指以微博为传播媒介，被一定人群广泛、持续关注，并能够反映网络舆论状况的信息集合。热点话题发现包含 2 个层次的含义，首先是话题发现，然后是热点话题筛选。本文将有意义串作为新特征，通过特征聚类发现话题，以话题相关文档数量来反映话题的热度，进而筛选发现热点话题。下面是微博热点话题发现涉及的基本概念。

特征(feature)：表示微博文本信息的基本单位，可以是字、词、有意义串或者 n-gram。

文档(DoC)：一个文档就是一条微博信息。

话题(topic)：围绕特定事件或活动展开的描述和讨论，话题的内容体现在一组相关文档中。

一个话题可以通过若干关键特征来表示，这些关键特征在该主题相关的文档中频繁共现，由此推断，经常在相同文档中共现的特征应该是同一话题相关的特征。通过特征在文档中的共现情况对特征进行聚类，得到围绕同一话题的特征簇。通常的方法以文档作为聚类对象，聚类结果为文档簇；本文以 3.2 节方法提取的新特征作为聚类对象，聚类结果为特征簇，即话题。新特征聚类与文档聚类相比，需要重新设计文档选择、特征表示、权重计算 3 个方面的方法。

微博空间文档数量巨大，大量信息可能与公众参与的热点话题无关，所以在特征聚类之前，需要进行文档选择，删除与热点话题无关的文档。这样能够显著降低文档空间维度，进而降低计算复杂度。下面引入文档特征数的概念来选择文档。

**定义 2**　文档特征数(VF, variety of featrue)：给定特征集合 $F=\{f_1, f_2, f_3\cdots\}$和文档集合 $D=\{\mathrm{DoC}_1, \mathrm{DoC}_2, \mathrm{DoC}_3\cdots\}$，文档 $\mathrm{DoC}_i$ 的特征数 $VF$ 是指 $\mathrm{DoC}_i$ 中出现且包含在集合 $F$ 中的非重复特征的数目。

根据 3.1 节有意义串的描述，包含较少特征的文档游离于话题之外，包含较多特征的文档才可能构成话题。因此，文档选择可通过文档特征数量的统计来确定。当一个文档的 $VF$ 值低于阈值 $T_{VF}$ 时，包含特征较少，被列入热点话题无关范围。

文档聚类采用空间向量模型，将文档表示为特征空间上的向量。本文对特征进行聚类时，采用特征的文档表示模型，也就是根据特征在文档中的分布情况，将特征表示为文档空间上的向量。具体表示形式如下：

$$\mathrm{DoC}ID_1\,\mathrm{DoC}ID_2\cdots\mathrm{DoC}ID_n$$

$$\mathrm{Feature}\langle \mathrm{weight}_1,\ \mathrm{weight}_2,\ \cdots\mathrm{weight}_n\rangle$$

其中，Feature 是一个特征，DoC$ID$ 表示文本编号，表征文档空间的 $n$ 个维度，weight 是特征在每个维度上的权重。

新特征在文档中的权重计算方法可以采用布尔值和 TF 值 2 种，布尔值表示特征在文档中出现与否，取值 0 和 1；TF 值反映了特征在文档中出现的频繁程度，归一化计算公式为：$w_{i,j}=\dfrac{TF_{i,j}}{\sum_{j=0}^{N-1}TF_{i,j}}$，其中，$w_{i,j}$ 表示特征 $i$ 在文档 $j$ 上的权重，$TF_{i,j}$ 是特征 $i$ 在文档 $j$ 中的频次，$N$ 是文档总数。

聚类算法方面，Steinbach M[14]对文本领域常用的 3 类聚类算法（凝聚式聚类，$k$-means 聚类以及 Bisecting $K$-means 聚类)进行了实验比较，结果表明 Bisecting $K$-means 的性能优于其余二者。本文也采用 Bisecting $K$-means 算法进行聚类。

聚类产生候选话题后，根据每个话题的相关文档数量来对话题进行热度排序，获得最热的 TOPN 个热点话题。热点话题发现的算法代码如图 2 所示。

```
Input：Microblog Coupus
Extract meaningful strings of the corpus;
for each document i{
    Compute the VF;
    if (VF<T_VF) delete document from the corpus;
}
for each meaningful string feature j{
    Represent the feature with document vector;
    Compute the weight on every document dimension ;
}
    Cluster the features with Bisecting K-means;
    for each topic k{
    compute the the number of relevant domument;
}
Rank topics with the number of relevant document;
Output:Hot Topics
```

图 2　热点话题发现算法

## 4　实验及结果分析

### 4.1　实验数据及评价标准

中文微博的研究还处于起步阶段，目前尚无公认的语料集和标注结果。本文通过互联网采集新浪微博 2012 年 10 月 30 日的微博信息 28 705 条。通过人工标注产生数据集中最热的 50 个话题，作为评价实验结果的标准。

实验采用 TOPN 的准确率 P@N 来评价算法有效性，准确率 $P$ 计算方法如下：

$$P=\frac{|C_S\bigcap C_R|}{|C_R|} \tag{1}$$

其中，$C_S$ 表示标注的 TOPN 话题集合，$C_R$ 表示实验产生的 TOPN 话题集合。

### 4.2　实验结果

本文设计了 3 组实验来验证热点话题发现的有效性及主要参数对实验结果的影响，实验 1 是本文方法与传统话题发现方法的对比实验，实验 2 调整有意义串提取过程中最重要的参数邻接类别，来观察对特征数量及话题结果的影响，实验 3 选取了特征权重计算时的 2 种不同方法，对比分析权重计算方法对话题结果的影响。

**实验 1**　与传统方法对比实验。

本实验也实现了传统方法：用词作为特征表示文档，采用 *K*-means 聚类算法来检测话题[15]，按照话题文档数排序产生热点话题。本实验中 $T_{VN}$ 取 2，权重计算方法采用 BOOL 值。实验结果如表 1 所示。

**表 1**　传统方法及本文方法实验结果对比

| 方法 | *P*@10 | *P*@20 | *P*@30 | *P*@40 | *P*@50 |
|---|---|---|---|---|---|
| 传统方法 | 60.00% | 70.00% | 76.67% | 75.00% | 74.00% |
| 本文方法 | 100.00% | 95.00% | 93.33% | 92.50% | 90.00% |

从表 1 中明显看出：本文方法在微博热点话题发现方面效果优于传统方法。传统方法由于特征稀疏，发现的热点话题质量不高；本文方法符合微博数据的特点，采用新特征来表示文本，较好地缓解了特征稀疏的问题，准确地描述了话题的关键信息。因此本文方法在微博数据的热点话题发现实验中取得较好效果。

在 $N$ 取 10 时，本文方法的准确率最高，而传统方法的准确率最低，这是因为传统方法发现的话题噪音较多，如“尖角 星星 微言 大道理”，而且排序比较靠前，在最热的 10 个话题中噪音最大；而本文方法发现的话题大部分是语义明确的具体话题，如“浙江温岭 虐童女幼师　颜艳红 补充调查 司法鉴定”，最热的话题与一般话题在文档数量上区分度很大，全部排到前面。

**实验 2**　邻接类别阈值变化对话题结果的影响。

在提取新特征的过程中，最重要的一个环节是关于邻接类别的语用分析，本实验通过调整阈值最小邻接类别 $T_{VN}$，观察对新特征数量、质量的影响，以及对热点话题结果的影响。实验中特征权重计算采用布尔值，结果如表 2 所示。

**表 2**　参数 $T_{VN}$ 变化实验结果对比

| $T_{VN}$ | 特征数 | TOP10 | | TOP20 | |
|---|---|---|---|---|---|
| | | 准确话题数 | 准确率 *P* | 准确话题数 | 准确率 *P* |
| 2 | 9 045 | 10 | 100% | 19 | 95% |
| 3 | 2 795 | 10 | 100% | 18 | 90% |
| 4 | 1 952 | 8 | 80% | 11 | 55% |

从表 2 可以看出，随着阈值 $T_{VN}$ 增大，对串的语用灵活性要求提高，产生的新特征数量逐渐减少，这与语言规律一致。当 $T_{VN}$ 取 2 和 3 时，热点话题的准确率较高，但是当 $T_{VN}$ 取 4 时，热点话题的准确率急剧下降。产生这一现象与微博信息的特点有关，因为微博发布灵活，词语变形情况比较普遍，

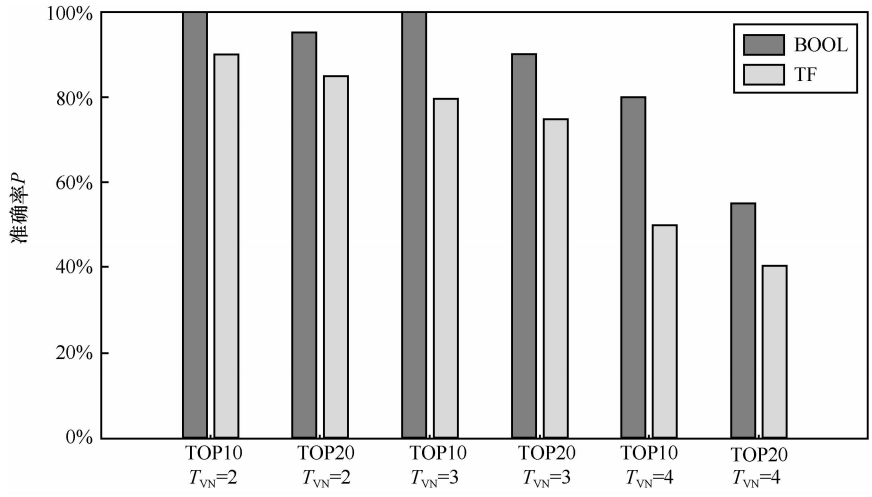



图 3　BOOL 和 TF 权重计算方法方法验结果对比

这导致表达相同含义的话题特征多样化，而且不会在整个文档空间频繁出现，呈现出空间局部性的特点。$T_{VN}$ 取值由 2 增大到 3 时，虽然特征数量大幅减少，但是大部分减少的是“希望改变”、“筒子们注意”这样的偶然碎片串，对于产生热点话题的关键特征影响不大。但是，当 $T_{VN}$ 增大到 4 时，有一部分关键频繁特征，如“罗姆尼当选”“食品添加剂”等，也被删除，因此对热点话题结果影响较大。

**实验 3**　特征权重计算方法对热点话题结果的影响。

分别采用 BOOL 方法和 TF 方法计算权重，热点话题准确率分布如图 3 所示。从图中看出，无论 $T_{VN}$ 取值为多少，BOOL 方法的结果都明显优于 TF 方法。这一结论与传统的文本聚类结果不同。这是因为长文本信息中，关键特征在一个文档中会频繁出现，频次发挥重要作用，而对于微博短文本信息，信息内容短小，特征分布非常稀疏，特征出现与否更为重要。统计发现，在 $T_{VN}$ 取值为 2 产生的特征中，一个文档中只出现一次的特征比例占到 82%，大于 1 次的特征比例仅占 18%。

下面结合话题实例，分析 BOOL 方法和 TF 方法的差异。在 $T_{VN}$=4 时，BOOL 方法发现了“深圳地铁故障”，而在 TF 方法中没有发现。“深圳地铁”这个词在部分文档中的分布情况如表 3 所示。表 3 所列文档都是与话题“深圳地铁故障”相关的，在 TF 方法中，只有文档 1 212、3 200、6 321 权重较高，聚类时“深圳地铁”与 3 个文档中的其他词聚成团的可能性大，所以这 3 个高频文档发挥的作用被放大，其他文档的作用被弱化。而在 BOOL 方法中，“深圳地铁”在各文档中的权重平均，与 8 个文档中的其他特征聚成团的可能性比较均匀。由此看出，微博信息中，某个话题的关键特征在文档中出现与否，直接决定了该文档是否与话题相关，而关键特征在文档中的出现频次作用较小。所以，BOOL 权重计算方法更适合于微博空间。

**表 3**　“深圳地铁”权重分布

| DoCID | *TF* | $W_{TF}$ | $W_{BOOL}$ | DoCID | *TF* | $W_{TF}$ | $W_{BOOL}$ |
|---|---|---|---|---|---|---|---|
| 587 | 1 | 0.062 5 | 0.1250 | 3 200 | 3 | 0.187 5 | 0.125 0 |
| 738 | 1 | 0.062 5 | 0.1250 | 3 841 | 1 | 0.062 5 | 0.125 0 |
| 1 212 | 3 | 0.187 5 | 0.1250 | 6 302 | 1 | 0.062 5 | 0.125 0 |
| 1 884 | 1 | 0.062 5 | 0.1250 | 6 321 | 5 | 0.312 5 | 0.125 0 |

## 5　结束语

本文提出了一种基于有意义串聚类的热点话题发现方法。方法首先针对微博数据高维稀疏、信息碎片化、新词不断涌现等特点，结合频繁字符串发现、语境上下文分析和语言规则过滤方法从微博中分析出独立的语义单元——有意义串。然后在微博的有意义串空间上聚类产生候选话题，进而通过热点排序实现微博热点话题的发现。有意义串既能帮助发现微博中频繁出现的新词、网络流行语、暗语等不规范用语，又能在一定程度上缓解传统词空间的高维稀疏性问题。实验表明，基于有意义串聚类的热点话题发现方法与传统方法相比，在准确率上有了明显的提升。

基于有意义串聚类的微博热点话题发现方法有效提升了热点话题的准确率，但未来仍需在如下

2 个方向上进行探索：1)微博富特征的应用，通过充分利用好友关系、链接关系、转发关系等丰富复杂的关联关系进一步提升热点话题发现的准确率。2)海量微博数据上的热点话题发现，通过设计分布式并行化的话题发现算法，应对大规模微博数据的计算问题，提高热点话题发现的效率。

## 参考文献：


[1] KUMARAN G., ALLAN J. Tex t classification and named entities for new event detection[A] . Proceedings of 27th ACM SIGIR Conference on Research and Development in Information Retrieval[C]. 2004, 297-304.

[2] ALLAN J, JIN H, RAJMAN M. Topic-based novelty detection[A]. Proceedings of the Johns Hopkins Summer Workshop[C].1999.

[3] YANG Y, CARBONELL J, JIN C. Topic-conditioned novelty detection[A]. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. 2002, 688-693.

[4] LAM W, MENG H, WONG K. Using contextual analysis for news event detection[J].International Journal on Intelligent Systems, 2001, (4):525-546.

[5] PAPKA R, ALLAN J. On-line new event detection using single pass clustering.Technical[R]. UM-CS-1998-021. University of Massachusetts.1998.

[6] 雷震, 吴玲达, 雷蕾等. 初始化类中心的增量 *K* 军执法及其在新闻事件探测的应用[J]. 情报学报. 2006, 25(3):289-295.
LEI Z, WU L D, LEI L.Incremental *k*-means method based on initlalisation of cluster centers and its application in news event detection[J]. Journal of The China Society For Scientific and Technical Information, 2006, 25(3):289-295

[7] 骆卫华, 于满泉, 许洪波等. 基于多策略优化的分治多层聚类算法的话题发现研究[J]. 中文信息学报. 2006, 20(1):29-36.
LUO W H,YU M Q,XU H B.The study of topic detection based on algorithm of division and multi-level clustering with multi-strategy optimization[J]. Journal of Chinese Information Processing, 2006, 20(1):29-36

[8] SHARIFI B, HUTTON M, KALITA J. Summarizing microblogs with topic models[A]. Proceeding of 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics[C]. 2010. 685-688.

[9] LIU Z, YU W, CHEN W. Short text feature selection and classification for microblog mining[A].Proceedings of International Conference on Computational Intelligence and Software Engineering[C]. 2010. 1-4.

[10] LEE C, WU C, CHIEN T. Burst: a dynamic term weighting scheme for mining microblogging messages[A]. Proceedings of 8th International Symposium on Neural Networks. 2011.

[11] DU Y Y. HE Y X. Miscoblog bursty topic detection based on userrelationship[A]. Proceeding of 6th IEEE Information Technology and Artificial Intelligence Conference[C]. 2011. 260-263.

[12] FENG H D. CHEN K. DENG X T. Accessor variety criteria for chinese word extraction[J]. Computer Linguistics, 2004, 30(1).

[13] 龚才春, 贺敏, 陈海强等. 大规模语料的频繁模式发现算法[J]. 通信学报.2007, 28(12):161-166.
GONG C C,HE M, CHEN H Q. Frequent-pattern discovering algorithm for large-scale corpus[J]. Journal on Communications, 2006, 20(1):29-36

[14] STEINBACH M, KARYPIS G., KUMAR V. A comparison of document clustering techniques[A]. Workshop on Text Mining[C]. 2000.

[15] YANG S, CHENG X, CHEN Y. Detect events on noisy textual datasets[A]. Proceedings of the 12th International Asia-Pacific Web Conference[C]. 2010.

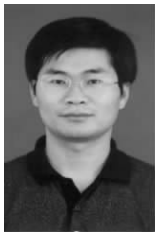

## 作者简介：


**贺敏**（1981-），女，山西忻州人，中国科学院计算技术研究所博士生，主要研究方向网络信息安全、舆情分析、自然语言处理等。

**王丽宏**（1967-），女，辽宁沈阳人，国家计算机网络应急技术处理协调中心副总工程师、研究员，主要研究方向为网络信息安全、舆情分析等。

**杜攀**（1981-），男，河南南阳人，中国科学院计算技术研究所助理研究员，主要研究方向为文本挖掘、信息检索、机器学习等。

**张瑾**（1978-），男，湖北应城人，中国科学院计算技术研究所高级工程师，主要研究方向为舆情分析、自然语言处理、话题分析、分布式处理等。



**程学旗**（1971-），男，安徽安庆人，中国科学院计算技术研究所研究员、博士生导师，主要研究方向为信息检索、文本挖掘、社会计算等。