

结合时序和语义的中文微博话题检测与跟踪方法

陈铁明, 王小号, 庞卫巍, 江韵

(浙江工业大学计算机科学与技术学院, 浙江 杭州 310023)

摘要: 微博文本具有短小快捷、主题多变等特点, 社交话题检测与跟踪研究面临新的挑战。结合微博的话题时序性和短文本语义相似度等特点, 提出了基于微博聚类的话题检测与跟踪系统方法。首先, 通过定义微博文本的时序频繁词集, 给出面向热点话题的特征词选择方法; 然后, 根据时序频繁特征词集, 利用最大频繁项集获得微博初始聚类; 针对初始簇间存在文本重叠情况, 提出基于短文本扩展语义隶属度的簇间重叠消减算法, 获得完全分离的初始簇; 最后, 根据簇语义相似度矩阵, 给出凝聚式话题聚类方法。通过新浪微博完成实验测试, 表明所提方法可用于中文微博热点话题检测与跟踪。

关键词: 微博文本; 频繁词集; 特征选择; 聚类; 话题检测; 时序; 语义

中图分类号: TP301

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2016.00048

Time series and semantics-based chinese microblog topic detection and tracking method

CHEN Tie-ming, WANG Xiao-hao, PANG Wei-wei, JIANG Jie

(College of Computer Science & Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: As a widely used tool in social networks, microblog is definitely with short document, quick broadcasting and topic changeable, which results in big challenging for social topic detection and tracking. A new systematic framework for micro-blog topic detection and tracking was proposed based on the microblog clustering using temporal trend and semantic similarity. Firstly, a feature words selection method for hot topics was presented by defining the temporal frequent words set. Secondly, an initially clustering was conducted depending on the selected temporal frequent words set. As far as the overlaps between initial clusters concerned, an effective overlap elimination algorithm was proposed, by introducing the extended short document semantic membership, to separate any possible overlapped initial clusters. Finally, an aggregated topic clustering method was employed using the cluster semantic similarity matrix. The experiments were at last done on some real-world dataset from Sina microblog. It show that the method for chinese microblog topic detection and tracking can obtain excellent performance and results.

Key words: microblog text, frequent words, feature selection, clustering, topic detection, time series, semantics

1 引言

话题检测与追踪 (TDT, topic detection and

tracking) 是指在没有人工干预的情况下自动判别新闻数据流的主题, 以应对日益严重的信息过载问题^[1], 其主要任务是从媒体信息流中抽取并监

收稿日期: 2016-03-17; 修回日期: 2016-04-27。通信作者: 陈铁明, tmchen@zjut.edu.cn

基金项目: 国家自然科学基金资助项目 (No.U1509214); 浙江省自然科学基金资助项目 (No.LY16F020035)

Foundation Items: The National Natural Science Foundation of China (No.U1509214), The Natural Science Foundation of Zhejiang Province (No.LY16F020035)

控不同的新闻报道,在没有话题先验知识的前提下,检测并组织出信息流中预先未知的话题^[2]。话题检测的研究自1996年由美国国防高级研究规划署(DARPA)和国家标准技术局(NIST)联合发起以来,已获得了较广泛的响应^[3-5],已有学者将Single-Pass^[6]、增量 k -means^[7]、层次聚类^[8]、关键词元统计^[9]等不同类型的聚类算法应用到话题检测中,并在一定程度上提高了话题检测效果^[10,11]。

由于微博文本具有短小快捷、主题多变等特点,以及其非结构化文本带来的向量高维性和稀疏性等问题,传统的新闻话题检测研究也面临着新的挑战:1) 基于经典Bag of Words思想构造的文本表示模型VSM会导致“高维诅咒”问题,需要研究面向微博文本的新型特征选择方法;2) 短文本的特性稀疏性,会导致传统基于空间距离的相似度计算方法无法有效地衡量2个短文本之间的相似度,需要研究新的文本相似度计算方法;3) 传统聚类算法通常需要预设初始簇或终止簇的数量,但这种先验知识在文本聚类的话题识别中往往是未知的,若设置不合适的数量参数,则会导致不理想的聚类结果,因此,需要根据微博文本集的内部性质合理地设置话题初始簇和终止簇的数量。

本文围绕中文微博社交话题检测面临的新问题,研究提出一种考虑频繁特征词时序特性和微博短文本语义相似度的聚类新方法,并设计实现基于聚类结果的微博话题检测与跟踪方法。

2 相关工作

2.1 微博话题检测系统方法研究

微博作为一种新兴网络媒体,具有信息增长快、内容实时性强、用语随意化等一系列特点,国外学者率先在微博话题检测领域取得了初步的研究成果。文献[12]将Twitter用户视为网络传感节点,采用基于关键词的贝叶斯决策方法设计开发了一套基于Twitter的实时地震监控原型系统,并取得了80%以上的检测率;文献[13]提出了一种在Twitter上收集、分组、排序和追踪突发新闻的方法,将相似度较高的tweets归为一组并视为一个新闻,再根据每组新闻话题所含词汇之间的关联度和流行度对这些新闻话题进行排序,最终得到爆炸性新闻;文献[14]提出了基于社会关系

评价和时序的Twitter热点话题检测方法,将tweets抽取为一系列词语的集合,通过词语的生命周期模型挖掘出突发词语,通过用户社会网络关系挖掘特定用户的影响力,并以此计算tweets的重要性,最后,通过创建基本话题表获得热点话题。随着新浪微博等主流平台的快速流行,国内学者也开始了面向中文微博话题检测的研究工作。文献[15]根据微博数据量大、信息破碎等特点,提出了基于隐主题模型的微博新闻话题检测方法。

2.2 面向微博短文本的聚类算法研究

文本内容聚类仍然是微博话题检测的核心,但微博短文本的特有属性使传统的聚类算法无法获得较好的应用效果,因此,微博短文本聚类方法的研究至关重要。下面主要介绍有较大发展潜力的2个方向^[16]。

1) 基于扩展语义信息

Gabrilovich等^[17]发现借助诸如维基百科等外部资源可以扩充短文本的特征,提高短文本之间的相似度。Baghel等^[18]通过引入WordNet将频繁词集转换为频繁概念集,然后,提出基于频繁概念集的文本聚类算法。Zelikovitz等^[19]通过创建潜在语义索引来挖掘词语之间的语义关系,提高了短文本处理的效果。

2) 基于词间共现或顺序关系

Beil等^[20]最早提出基于频繁项集的聚类FTC(frequent term-based clustering)算法。FTC算法使用频繁词集来表示簇,采用贪婪式的启发策略,频繁词集选择的次序会影响最终的聚类结果。Li等^[21]提出基于频繁词序列的文本聚类方法CFWS,算法通过挖掘频繁词序列构造初始簇,然后,使用 k -mismatch方法来合并初始簇并得到聚类结果。Fung等^[22]则针对FTC算法的缺点,提出基于频繁项集的层次聚类(FIHC, frequent itemset-based hierarchical clustering)算法。

由于微博短文本特性,若直接将FIHC算法应用到微博聚类,主要将面临如下2个问题:①普通的频繁项集仅表明某些特征项的共现关系频繁,无法保证共现关系可以代表文本集的隐含话题,影响话题簇的正确构造与划分;②由于微博内容简短、特征稀少,同一话题的某些微博可能语义相似但表述方式不同,从而被错误地划分,

最终影响聚类结果。

3 本文方法

3.1 方法框架

为解决微博短文本聚类所面临的问题，本文采用 FIHC 算法^[22]“先建簇后消重再凝聚”的思想，提出一种结合时序频繁和语义聚类的新方法（TS-FIHC, time and semantics FIHC），聚类的主要过程如图 1 所示。

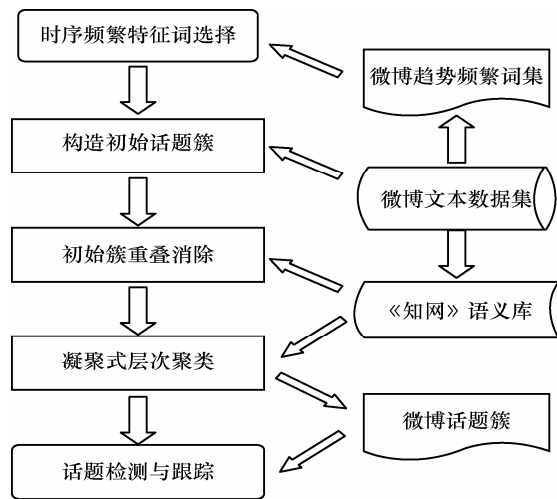


图 1 TS-FIHC 微博话题聚类检测的基本过程

话题通常具有时间属性，微博的热点话题更具有时序性，即在某个时间点上具有突发性，发展趋势极度不平衡。因此，首先根据时间滑动窗口定义频繁词集的时序趋势度，提出一种基于时序词频的微博文本特征选择，并采用选择的时序趋势频繁词集划分微博的初始话题簇；为了更精准地消除初始簇间的文本重叠，采用知网的中文语义相似度模型，根据最大语义隶属度原则分离各个初始的话题簇；最后，通过定义簇间语义相似度矩阵，完成微博话题簇的凝聚式层次聚类，根据参考优化得到最终的话题簇，并实现话题检测与跟踪。

3.2 基于时序词频的微博文本特征选择

微博文本有异于普通文本，它的时序特性表明所有的微博都是沿时间轴展开，具有先后顺序，若对时间轴设置一个时间窗口，则可将所有的微博分散到不同的时间窗口中，得到基于时间滑动窗口的微博集合。因此，可从时序角度考虑微博词汇的特征选择。

定义 1 趋势基数。

将某个特征词 T_j 在第 i 个时间滑动窗口内的趋势基数 BT_{ji} 定义为该词汇在之前连续的 k 个时间滑动窗口内出现频率 TF 的平均值（这里的 k 定义为时间窗口参数，且当 $i \leq k$ 时，取 $k=i-1$ ）

$$BT_{ji} = \frac{\sum_{s=1}^k TF_{(i-s)}}{k} \quad (1)$$

定义 2 趋势增长率。

记某个特征词 T_j 的趋势基数为 BT_{ji} ，在第 i 个时间滑动窗口内出现的频率为 TF_i ，定义 T_j 在第 i 个时间滑动窗口内的趋势增长率为

$$GT_{ji} = \frac{TF_i}{BT_{ji}} = \frac{k \cdot TF_i}{\sum_{s=1}^k TF_{(i-s)}} \quad (2)$$

定义 3 时序趋势度。

特征词的话题趋势与它的趋势增长率及趋势基数成正比，因此定义微博话题频繁词集时序趋势度的计算公式为

$$TT_{ji} = GT_{ji} \cdot \ln BT_{ji} \quad (3)$$

下面验证话题趋势特征词汇选择与一般频繁特征词汇选择的效果。鉴于当前一般的微博话题热度以天为计时单位，且社交网络的新闻热度一般不超过一周，因此，本文对随机抽取的新浪微博进行统计，根据时间滑动窗口设置的天数（不超过 7）不同，给出基于时序频繁词集和一般频繁词集的特征词汇选择结果的差异比例曲线。图 2 显示基于时序词频的微博特征选择结果与一般基于频繁词集的方法相比，在特征数量较少时均具有明显的差异，体现了基于时序词频的微博特征选择对区分微博热点话题的适用性，时间滑动窗口以 3 天为宜。

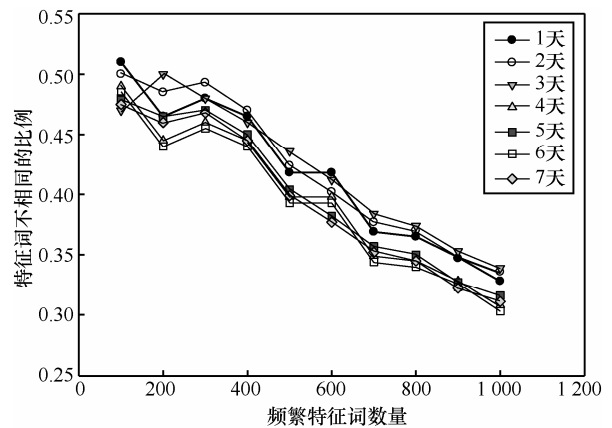


图 2 不同滑动窗口大小的时序频繁特征词集与一般频繁特征词集的差异

3.3 基于时序频繁词集的微博初始聚类

定义 4 对文本集 D 中某个项集 X , 若项集 X 在 D 中出现的次数大于一个预设的比例, 则称 X 是文本集 D 上的频繁项集, 这个预设比例称作最小支持度 s 。

定义 5 对文本集 D 的某个趋势词集 W , 若 W 在 D 中的支持度 $s(W) \geq \min_s$, 则称趋势词集 W 是文本集 D 上的频繁趋势词集, $\min_s$ 为全局最小支持度。

本文采用关联分析中广泛使用的频繁集挖掘算法 Apriori 来计算挖掘频繁趋势词集, 执行如下步骤: 1) 扫描文本集 D , 利用词频趋势度统计候选项集出现的次数, 收集满足最小支持度 $\min_s$ 设定的项集, 记为频繁项集; 2) 利用产生的频繁 k -项集构造强关联规则, 利用频繁 k -项集构造候选 $(k+1)$ -项集, 反复迭代直至候选 $(k+1)$ -项集为空。

频繁趋势词集可较好地描述微博中隐含的话题信息, 本文利用频繁趋势词集划分构造初始簇, 即将包含某频繁趋势词集的微博都划分为一个簇, 得到基于频繁趋势词集的初始簇。

定义 6 若一个时序词汇特征项集在初始簇所含的微博集中出现的比例超过一个预设的最小比例, 就称此时序特征项集是簇时序频繁项集, 这个预设的最小比例被记作簇最小支持度 θ 。

由于簇时序频繁项集是指那些在聚类簇中出现频率较高的核心词汇, 这些核心词汇在一定层面上代表了这个聚类簇所描述话题的隐含语义, 可以通过抽取各个初始簇的簇时序频繁项集来代表这个初始簇的话题语义信息。

3.4 基于语义隶属的簇间重叠消减算法

由于话题检测的目标是将每条微博归属到一个话题簇中, 因此还需要设计初始簇间重叠部分的消减算法, 结合微博短文本的特点, 将重叠的微博最终分配给最合适的初始簇。

定义 7 若微博 doc_j 被分配到初始簇 C_i 中, 则称微博 doc_j 支持簇 C_i 。

定义 8 记 D_i 和 D_j 是支持簇 C_i 和 C_j 的所有微博的集合, 且 $D_i \cap D_j \neq \emptyset$, 即簇 C_i 和 C_j 间存在共享的微博, 则称簇 C_i 和 C_j 存在簇间重叠。进一步地, 记簇间重叠的微博集合为 C^D , 这里的 $D = \{C_i, C_j\}$, $C^D = D_i \cap D_j$ 。

定义 9 微博话题语义隶属度。

本文将微博 doc_j 对初始簇 C_i 的话题语义隶属度函数定义为

$$Score(doc_j \rightarrow C_i) = \frac{\sum_{l=1}^n \max_{k=1,2,\dots,m} \{sim(f_{ik}, t_{jl})\}}{n} \quad (4)$$

其中, 簇频繁 1-项集 $\{f_{i1}, f_{i2}, \dots, f_{im}\}$ 表示初始簇 C_i 的话题特征项, $\{t_{j1}, t_{j2}, \dots, t_{jn}\}$ 表示初始簇 C_i 中微博文本 doc_j 的特征项; $sim(f_{ik}, t_{jl})$ 簇特征项 f_{ik} 和文本特征项 t_{jl} 在《知网》中定义的语义相似度^[23], n 为微博文本 doc_j 特征项的数目, m 为簇特征项的数目。当然, 这里可通过设置合理的簇最小簇支持度 θ 来控制初始簇的话题特征的数量。

基于微博语义隶属度的初始簇重叠消减算法的具体描述如下。

输入: 带有重叠的初始簇 $C_1, C_2, \dots, C_n$

输出: 重叠消减的初始簇 $C'_1, C'_2, \dots, C'_n$

1) 记 m 个重叠的微博文本初始簇集合为 $C^{D_1}, C^{D_2}, \dots, C^{D_n}$, $D_i \subset \{C_1, C_2, \dots, C_n\}$

2) 初始化一个二维数组向量 $clusterScoreHash := \{\}$

3) FOR each i from 1 to n

4) FOR each weibo doc_j in C^{D_i}

5) FOR each cluster $C_k \subset C^{D_i}$

6) $currScore = Score(doc_j \rightarrow C_k)$

7) IF $doc_j \notin clusterScoreHash$

8) add $\langle doc_j, currScore \rangle$ to $clusterScoreHash$

9) ELSEIF $currScore \geq currScore$ of $doc_j \in clusterScoreHash$

10) update $\langle doc_j, currScore \rangle$ to $clusterScoreHash$

11) ELSE

12) delete doc_j from C_k

13) ENDIF

14) $C'_k = C_k$

15) ENDFOR

16) ENDFOR

17) ENDFOR

该算法的复杂度为 $O(n)$, 即仅需对所有重叠初始簇中的微博文本扫描一遍即可消减所有初始簇之间的重叠, 最后, 删除初始簇分离后大小为

0 的空簇, 就可得到非空的候选话题簇。

3.5 基于语义相似度的凝聚式话题聚类算法

通过初始簇间重叠消减可得到微博聚类话题检测的候选话题簇, 但有时这些话题簇都可归属于某一个大话题, 因此, 有必要再对候选话题簇进行凝聚式层次聚类, 合并话题簇以缩减大话题数目, 为用户提供更聚焦的微博热点话题。

为合并候选话题簇, 首先需度量 2 个候选话题簇之间的相似性。由于候选话题簇由大量微博文本组成, 为保证凝聚式层次聚类的效率, 应避免让候选话题簇中所有微博文本参与相似性度量的计算, 因此, 本文选择候选话题簇中的主要频繁特征词集构成簇的特征向量, 用该特征向量表示候选话题簇。

定义 10 簇特征向量。

针对候选话题簇 CT_i , 挖掘出 CT_i 的簇频繁 1-项集, 即构成该簇的簇特征向量, 记为 $\vec{t}_{CT_i} = (t_{i1}, t_{i2}, \dots, t_{in})$ 。

定义 11 簇相似度矩阵。

记 2 个不同候选话题簇 CT_i 和 CT_j 的簇特征向量分别为 $\vec{t}_{CT_i} = (t_{i1}, t_{i2}, \dots, t_{in})$ 和 $\vec{t}_{CT_j} = (t_{j1}, t_{j2}, \dots, t_{jm})$, 则 CT_i 和 CT_j 的特征项构成的簇语义相似度矩阵按表 1 的方式定义。

表 1 话题簇 CT_i 和 CT_j 的簇语义相似度矩阵

	t_{i1}	t_{i2}	$\dots$	t_{in-1}	t_{in}
t_{j1}	$\text{sim}(t_{j1}, t_{i1})$	$\text{sim}(t_{j1}, t_{i2})$	$\dots$	$\text{sim}(t_{j1}, t_{in-1})$	$\text{sim}(t_{j1}, t_{in})$
t_{j2}	$\text{sim}(t_{j2}, t_{i1})$	$\text{sim}(t_{j2}, t_{i2})$	$\dots$	$\text{sim}(t_{j2}, t_{in-1})$	$\text{sim}(t_{j2}, t_{in})$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
t_{jm-1}	$\text{sim}(t_{jm-1}, t_{i1})$	$\text{sim}(t_{jm-1}, t_{i2})$	$\dots$	$\text{sim}(t_{jm-1}, t_{in-1})$	$\text{sim}(t_{jm-1}, t_{in})$
t_{jm}	$\text{sim}(t_{jm}, t_{i1})$	$\text{sim}(t_{jm}, t_{i2})$	$\dots$	$\text{sim}(t_{jm}, t_{in-1})$	$\text{sim}(t_{jm}, t_{in})$

定义 12 话题簇语义相似度。

为避免过多非关键特征词对簇间语义相似度的噪音, 仅选取相似度矩阵中语义相似度最大的 k 组特征项对候选话题簇之间的相似度进行计算, 记为 $\{\text{sim}(t_i, t_j)_1, \text{sim}(t_i, t_j)_2, \dots, \text{sim}(t_i, t_j)_k\}$, 并将候选话题簇之间的语义相似度定义为

$$\text{sim}(CT_i, CT_j) = \frac{\sum_{l=1}^k \text{sim}(t_i, t_j)_l}{k} \quad (5)$$

基于候选话题簇的语义相似度, 采用 λ 表示 2 个簇合并时设定的簇语义相似度的最小阈值, μ 表示话题簇凝聚后期望得到的最少簇数目, 给出

话题簇凝聚式层次聚类方法的操作步骤如下。

1) 抽取各个候选话题簇的特征向量, 计算候选话题簇的语义相似度。

2) 构建候选话题簇的语义相似度矩阵, 由簇相似度的定义可知 $\text{sim}(CT_i, CT_j) = \text{sim}(CT_j, CT_i)$, 即该相似度矩阵为一个对称矩阵。

3) 从相似度矩阵中选择最大的簇间相似度, 记为 $\max\{\text{sim}(CT_i, CT_j)\}$, 若 $\max\{\text{sim}(CT_i, CT_j)\} \leq \lambda$, 执行 6); 否则, 执行 4)。

4) 由于 $\max\{\text{sim}(CT_i, CT_j)\} > \lambda$, CT_i 和 CT_j 之间的相似性较大, 故将 CT_i 和 CT_j 2 个簇合并, 形成一个新的簇 CT'_i , 删除原来的 CT_j , 并重新计算簇特征向量, 更新语义相似度矩阵。

5) 若簇间语义相似度矩阵的行数或列数小于等于预设的最小簇数目 μ , 执行 6); 否则, 聚类尚未结束, 重新回到 3)。

6) 凝聚式层次聚类结束, 得到最终的话题聚类簇。

4 实验结果与分析

4.1 数据采集及其预处理

由于目前尚没有公开的中文微博话题检测 Benchmark 语料集, 本文自主采集了新浪微博数据展开实验分析。鉴于话题具有群体属性, 聚焦了浙江工业大学微群里的 9 015 位新浪微博用户, 通过官方提供的 Open API 定向抓取分析了这些用户在 2012 年 7 月 20 日至 2012 年 8 月 30 日期间发布的所有微博, 总计 495 026 条。

尽管 Open API 抓取的微博是较完整的结构化数据, 在数据分析前仍需对微博文本进行一系列的预处理, 主要包括中文分词及词性标注、停用词过滤、词性选择和特征选择等步骤。本文采用中科院的 ICTCLAS (institute of computing technology, chinese lexical analysis system) 系统, 通过分词算法标注的词性, 选择话题描述性较强的名词、动词或形容词作为文本特征, 直接降低文本词汇特征的维度。

最后, 在采用本文提出的微博话题检测方法之前, 还需进一步执行如 3.2 节所述的基于时序词频的微博特征词汇选择。

4.2 微博聚类实验及其效果分析

为了评估聚类算法的效果与性能, 本文对抓

取的微博进行筛选和话题人工标注,得到手工标注的10个话题类别,总计13 356条微博,话题标注情况如表2所示。

表2 手工标注的10个话题类别分布情况

序号	标记簇	簇大小
1	{保钓, 钓鱼岛}	1 438
2	{京东, 电商}	1 427
3	{海葵}	2 463
4	{流星雨}	157
5	{爆料哥, 周克华}	680
6	{伦敦, 奥运}	2 000
7	{小米}	528
8	{星座, 运势}	3 115
9	{叶诗文}	799
10	{羽毛球}	749

不失一般性,本文采用纯度和 F -Measure 值2个指标客观评价聚类效果。一般来说,聚类结果的纯度越大,聚类的效果越好;聚类结果的 F -Measure 值越大,聚类的效果就越好。

1) 聚类算法的参数分析

簇最小支持度 θ 的大小直接影响初始簇在特征提取阶段所得的簇特征数量,并进一步影响簇间重叠部分的微博对初始簇的话题语义隶属度计算,最终影响初始簇间重叠的分离效果。

为了分析参数 θ 的选择效果,首先将人工标注的10个类别的簇随机分为2组,各含5个人工标注簇,分别记为“#PartI”和“#PartII”(随机获得10组不同的#PartI和#PartII);选择不同的簇最小支持度 θ ,不同 θ 对聚类结果 F -Measure 平均值的影响如图3所示。由测试结果知,当簇最小支持度 θ 取0.5~0.6时,可获得较好的聚类效果。

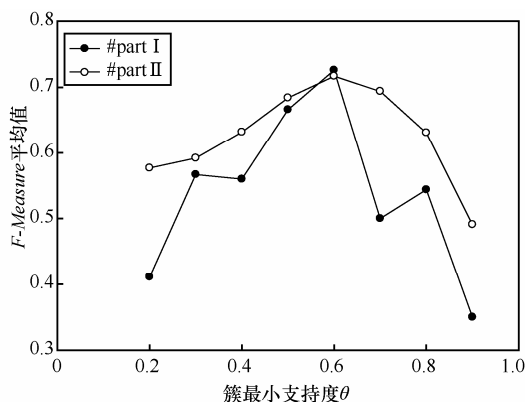


图3 不同簇最小支持度 θ 对 F 值的影响

在凝聚式层次聚类中, λ 为候选话题簇间相似度的最小阈值,当所有的簇间相似度都小于 λ 时,话题合并终止;由实验结果可知,当 λ 取0.6~0.7时可获得最好的聚类效果,如图4所示。

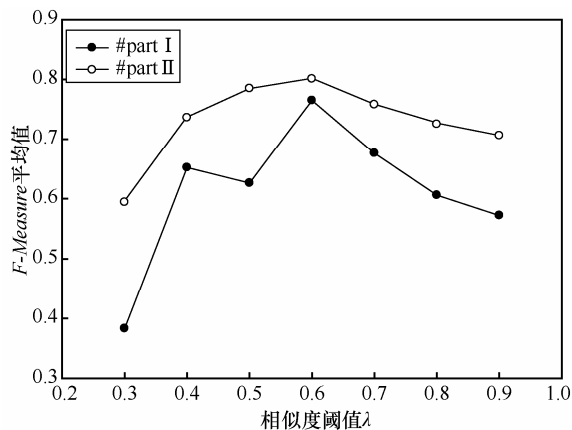


图4 话题簇相似度最小阈值 λ 对 F 值的影响

2) TS-FIHC 与 FIHC 的效果比较

分别抽取人工标注话题中的2个、4个、6个、8个和10个标注话题共5组话题数据作为测试基准,比较本文提出的TS-FIHC算法和FIHC算法聚类的纯度和 F -Measure 值,结果分别如图5和图6所示。其中,TS-FIHC算法采用TS-FIHC- k 代表频繁 k -项集($k=1,2,3,4$)。

由实验结果可知,由于改进的TS-FIHC方法考虑了文本语义,使初始簇的分离和候选话题簇的合并更加合理,因而获得了比FIHC更好的纯度和 F -Measure 值。另一方面,所有TS-FIHC-1的聚类效果也优于TS-FIHC- k ,同时,TS-FIHC-1算法可避免对频繁 k -项集的挖掘,大大降低了算法开销。

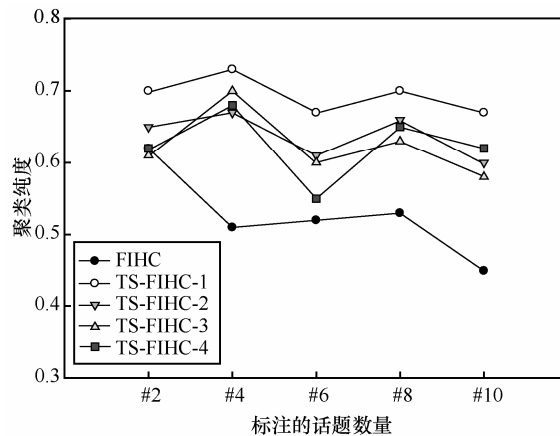


图5 不同话题数量的聚类纯度比较

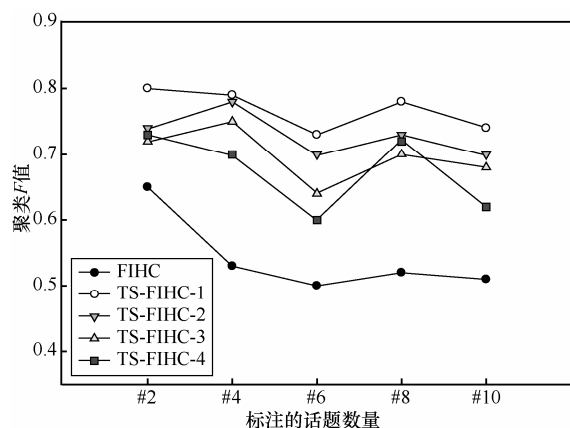


图6 不同话题数量的聚类F值比较

为了进一步比较本文提出的 TS-FIHC 与 FIHC 在微博增量处理环境下的聚类效果，图 7 给出了在处理 10 000~100 000 条不同微博数量时，2 种聚类方法的 F 值效果趋势。分析结果表明，TS-FIHC 随着微博样本数量的增加，聚类效果有明显的增强，而 FIHC 则基本保持不变。

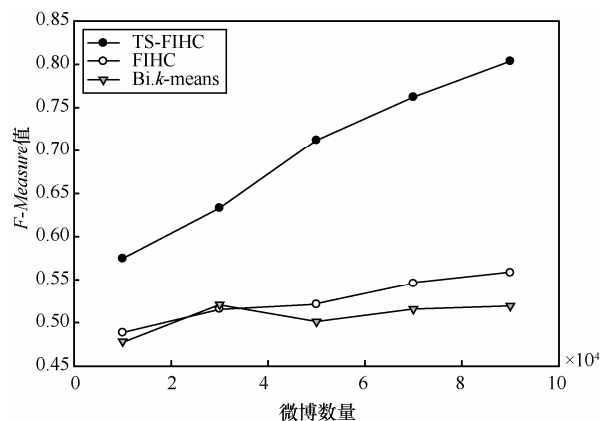


图7 不同微博文本数量的F-Measure值比较

3) TS-FIHC 与 FIHC 的性能比较

下面实验分析 TS-FIHC_k 较 FIHC 的性能优势。继续使用抽取人工标注话题中的 2 个、4 个、6 个、8 个和 10 个标注话题共 5 组话题数据作为测试基准，分别计算 TS-FIHC_k 和 FIHC 执行聚类算法的时间消耗 (Intel Core i3, 3.2 GHz, 2 GB RAM)，结果如图 8 所示。由实验结果知，与 FIHC 算法相比，TS-FIHC₁ 由于减少了初始簇的数目，具有明显的性能提升。

为了进一步测试 TS-FIHC 聚类性能对微博样本处理数量的扩展性，图 9 给出了在处理 10 000~100 000 条不同微博文本时，TS-FIHC 与各种聚类算法的性能扩展性比较。实验结果表明，TS-FIHC

具有较高的对样本增量运算的适应性，适合大规模海量微博数据的分析。

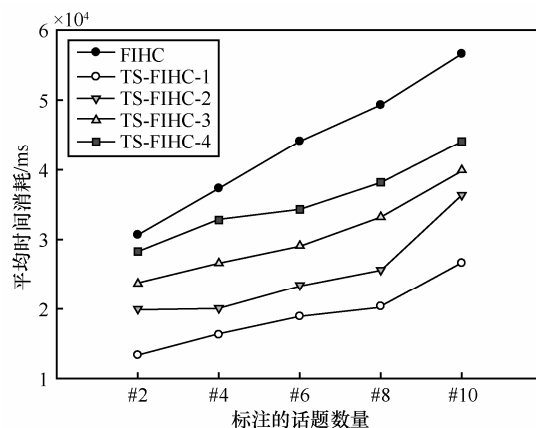
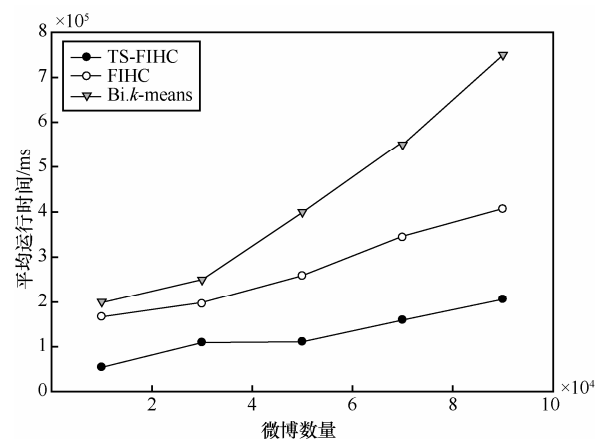
图8 TS-FIHC_k与FIHC的运行性能比较

图9 不同聚类方法的运行性能扩展性比较

4.3 基于聚类的微博话题检测与跟踪

为了评估聚类方法对微博热点话题的检测效果，本文手工标记了 2012 年 8 月 15 日的热门话题作为参照对象，采用 TS-FIHC 方法检测出话题簇，并抽取话题簇的描述特征。将最终聚类得到的 Top-10 话题簇与标注的热门话题进行对比，其结果如表 3 所示。

通过比较表 3 中标注话题和 Top-10 话题簇，可以看出，本文方法 TS-FIHC 有效检测出指定时间段的微博热点话题，并且从话题簇中抽取的簇特征基本上描述了话题簇的主要内容，使话题簇具有较好的可理解性。

下面进一步验证话题跟踪的效果。例如，图 10 给出了 2012 年 7 月 20 日至 2012 年 8 月 20 日 1 个月期间，{伦敦奥运}、{周克华}、{钓鱼岛}3 个热门话题的发展趋势，从关键的时间点可见，

话题曲线的发展趋势与社会事件的实际情况基本相符。

表3 2012年8月15日的话题检测评估

序号	人工标注话题 (7个)	聚类结果 Top-10
1	电商大战	{京东, 苏宁, 价格}
2	鸡蛋灌饼	{垃圾, 杭州, 里脊, 生菜, 口感, 鸡蛋, 滋味, 经典, 小吃, 外皮}
3	钓鱼岛	{保钓, 钓鱼岛, 香港, 日本, 人士}
4	向日葵	{偶像, 好友, 先锋, 盛宴, 大麦, 门票, 上海...}
5	高温	{星座, 双鱼, 射手, 处女}
6	小米手机	{杂志, 免费, 书店}
7	周克华	{美国, 中国}
8		{今年, 向日葵, 土壤, 积累, 微生物}
9		{图片, 手机, 小米, 发布会}
10		{高温, 杭州, 气温, 重庆}

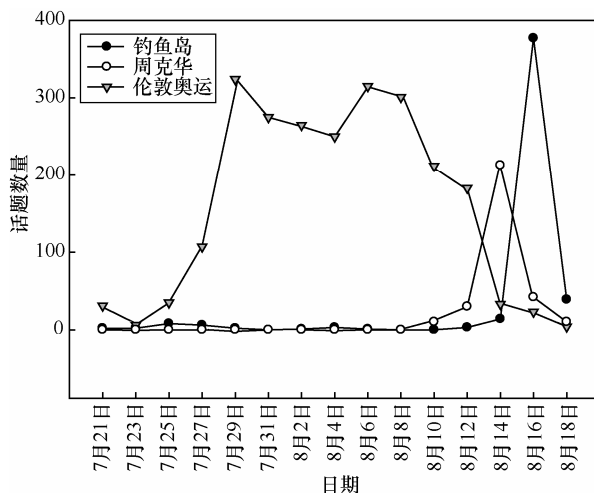


图10 热门话题的发展演变趋势

5 结束语

本文针对中文微博内容开展热点话题检测研究，将话题检测问题归结为短文本聚类问题，提出了一种系统的解决方案：1) 利用微博话题的时序特性，提出面向微博频繁特征词的话题趋势性相关度量；2) 针对初始簇间的微博重叠问题，提出一种基于语义隶属度划分的初始簇重叠消减算法，克服微博短文本语义表达简短及形式多样化等引起的话题二义性问题；3) 通过定义初始簇间的语义相似度，给出一种面向微博话题的凝聚式层次聚类方法，可实现话题检测与跟踪；4) 通过真实的微博数据分析实验，验证本文方法的有效性。

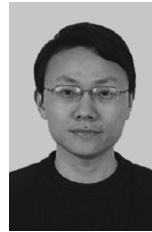
由于微博热点话题的独特性，话题检测仍面临诸多问题值得进一步深究：1) 由于《知网》语义库词汇量有限，词汇间的相似度不能随话题的不同而动态变化，因此，可进一步挖掘统计微博词汇的共现情况，研究基于词汇共现状况的动态语义相似度计算方法；2) 并非所有的微博都仅讨论一个话题，有些微博可能将多个话题串联，即一条微博可能同时属于多个话题，因此，值得进一步研究一些软聚类的方法来扩展微博话题检测；3) 通过聚类获得的话题簇通常还隐含较多潜在信息，但这些隐含的信息不易被直接发现，因此，可进一步研究可视化方法来展现话题簇之间的相互联系，直观辅助话题检测。

参考文献:

- [1] ALLAN J. Topic detection and tracking: event-based information organization[M]. Kluwer Academic Publisher, 2002.
- [2] NIST. The 2003 topic detection and tracking task definition and evaluation plan[EB/OL]. <http://www.nist.gov/speech/tests.tdt/tdt2003/evalplan.html>.
- [3] ALLAN J, CARBONELL J, DODINGTON G, et al. Topic detection and tracking pilot study: final report[C]//The Darpa Broadcast News Transcription and Understanding Workshop. c2000:194-218.
- [4] WAYNE C. Multilingual topic detection and tracking: successful research enabled by corpora and evaluation[C]//The Language Resources and Evaluation Conference. c2000:1487-1494.
- [5] 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, et al. The study of topic detection based on algorithm of division and multilevel clustering with multistrategy optimization[J]. Journal of Chinese Information Processing, 2006, 20(1):29-36.
- [6] 贾自艳, 何清, 张俊海, 等. 一种基于动态进化模型的事件探测和追踪算法[J]. 计算机研究与发展, 2004, 41(7): 1273-1280.
- JIA Z Y, HE Q, ZHANG J H, et al. A new event detection and tracking algorithm based on dynamic evolution model[J]. Journal of Computer Research and Development, 2004, 41(7):1273-1280.
- [7] YAMRON J P, KNECHT S, MULBREGT P V. Dragon's tracking and detection systems for the tdt2000 evaluation[C]//Topic Detection and Tracking Workshop. c2000: 75-80.
- [8] DAI X Y, CHEN Q C, WANG X L, et al. Online topic detection and tracking of financial news based on hierarchical clustering[C]//2010 International Conference on Machine Learning and Cybernetics. c2010: 3341-3346.
- [9] 张阔, 李涓子, 吴刚, 等. 基于关键词元的话题内事件检测[J]. 计算机研究与发展, 2009, 46(2): 245-252.
- ZHANG K, LI J Z, WU G, et al. Word committee based event identification[J]. Journal of Computer Research and Development, 2009, 46(2):245-252.
- [10] 洪宇, 仓玉, 姚建民, 等. 话题跟踪中静态和动态话题模型的核

- 捕捉衰减[J]. 软件学报, 2012, 23(5):1100-1119
- HONG Y, CANG Y, YAO J M, et al. Descending kernel track of static and dynamic topic models in topic tracking[J]. Journal of Software, 2012, 23(5):1100-1119.
- [11] 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.
- ZHANG X M, LI Z J, CHAO W H. Research of automatic topic detection based on incremental clustering[J]. Journal of Software, 2012, 23(6): 1578-1587.
- [12] SAKAKI T, OKAZAKI M, MATSUO Y. Earthquake shakes twitter user: real-time event detection by social sensors[C]//The 19th International Conference on World Wide Web. c2010: 851-861.
- [13] PHUVIPADAWAT S, MURATA T. Breaking news detection and tracking in twitter[C]//2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology(WI-IAT). c2010:120-123.
- [14] CATALDI M, CARO L D, SCHIFANELLA C. Emerging topic detection on twitter based on temporal and social terms evaluation[C]//The 10th International Workshop on Multimedia Data Mining. c2010: 1-10.
- [15] 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[J]. 模式识别与人工智能, 2012, 3: 382-387.
- LU R, XIANG L, LIU M R, et al. Extracting news topics from microblogs based on hidden topics analysis and text clustering[J]. Pattern Recognition and Artificial Intelligence, 2012, 3:382-387.
- [16] 王永恒. 海量短语信息挖掘技术的研究和实现[D]. 长沙: 国防科学技术大学. 2006.
- WANG Y H. Research and implementation of information mining on massive short messages[D]. Changsha: National University of Defense Technology. 2006.
- [17] GABRILOVICH E. Feature generation for textual information retrieval using world knowledge[J]. ACM SIGIR Forum, 2007, 41(2): 123.
- [18] BAGHEL R, DHIR R. Text document clustering based on frequent concepts[C]//2010 1st International Conference on Parallel, Distributed and Grid Computing (PDGC). c2010: 366-371.
- [19] ZELIKOVITZ S. Transductive LSI for short text classification problems[C]//The 17th International FLAIRS Conference. c2004.
- [20] BEIL F, ESTER M, XU X. Frequent term-based text clustering[C]//The 8th ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. c2002: 436-442.
- [21] LI Y J, CHUNG S M, HOLT J D. Text document clustering based on frequent word meaning sequences[C]//Data & Knowledge Engineering . c2008: 381-404.
- [22] FUNG B C M, WANG K, ESTER M. Hierarchical document clustering using frequent itemsets[C]//The Siam International Conference on Data Mining, San Francisco. c2003.
- [23] 许云, 樊孝忠, 张锋. 基于《知网》的语义相似度计算[J]. 北京理工大学学报, 2005, 25(5): 411-414.
- XU Y, FAN X Z, ZHANG F. Semantic relevancy computing based on hownet[J]. Transactions of Beijing Institute of Technology, 2005, 25(5):411-414.

作者简介:



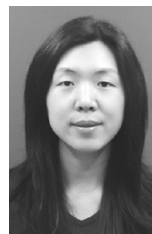
陈铁明 (1978-), 男, 浙江诸暨人, 博士, 浙江工业大学教授, 主要研究方向为网络与信息安全。



王小号 (1981-), 男, 浙江新昌人, 浙江工业大学讲师, 主要研究方向为信息安全。



庞卫巍 (1989-), 男, 浙江绍兴人, 浙江工业大学硕士生, 主要研究方向为网络安全与本文挖掘。



江颜 (1972-), 女, 浙江平湖, 博士, 浙江工业大学副教授, 主要研究方向为网络信息安全。