基于词共现关系和粗糙集的微博话题检测方法①

兰 天[1,2], 郭躬德[1,2]

[1](福建师范大学 数学与计算机科学学院, 福州 350007)
[2](福建师范大学 网络安全与密码技术福建省重点实验室, 福州 350007)

摘 要: 为解决传统词共现方法在微博中检测话题时计算复杂度大、查全率不高、查准率低的情况, 提出一种基于粗糙集原理的改进词共现算法(RSCW). 通过词共现关系形成词共现矩阵, 并由共现矩阵找出极大完全子图作为话题簇中心, 最后由粗糙集原理找出每个话题的关键词集合. 在 NLPIR 微博内容语料库和实时获取的微博数据集上的实验结果表明, 该方法能够有效地从大规模微博信息中检测突发新闻, 提高突发新闻的识别率.

关键词: 微博; 词共现图; 粗糙集; 话题检测

News Topic Detection on Chinese Microblog Based on Rough Set and Word Co-Occurrence

LAN Tian[1,2], GUO Gong-De[1,2]

[1](School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, China)
[2](Network Security and Cryptography key laboratory of Fujian province, Fujian Normal University, Fuzhou 350007, China)

Abstract: Traditional word co-occurrence detection methods in microblog news encounter the problems of high computational complexity, high time consuming, low recall rate and low precision. An improved algorithm of word co-occurrence detection based on rough set is proposed in this paper aiming at solving these problems. It builds a word co-occurrence matrix through word co-occurrence relation, and finds out the maximum complete subgraph as topic cluster center via co-occurrence matrix, finally identifies each topic keyword set using the rough set theory. The experimental results carried out on the microblog content corpus of NLPIR and the real-time collection of microblog data set verify that this method can effectively detect news topic from the massive microblog information and realize the news topic tracking.

Key words: microblog; word co-occurrence graph; rough set; topic detection

1 引言

互联网的兴起、移动数据网络的飞速发展等因素, 使网民对于社交网络平台的使用率得到了极大的提高. 微博作为一种分享和交流的平台, 集实时性、及时性、开放性、共享性等特点, 使突发话题的传播比在其他传统媒体传播的速度要快、范围要广, 并成为最主流的社交网络平台[1]. 根据新浪公布的微博用户的最新数据, 截止 2014 年 12 月底, 该微博的注册用户已超 5 亿, 日活跃用户为 4629 万. 微博消息的传播已成为平民大众网络信息传播的主要途径, 微博每日信息的发布和转载量已过亿, 其中每秒信息量的峰值达到 7000 条之多. 同时新浪微博还作为许多政府部门、公共安全机构、非政府组织等发布重要申明和传播重大信息的平台[2].

随着 Twitter 等微博的流行, 话题检测与跟踪(Topic Detection and Tracking, TDT)作为文本挖掘的一个方向, 在微博上的运用越来越遭到人们的重视. 在国外, Twitter 被研究用于辅助各种突发事件的应对, 例如重大火警、交通路况、自然灾害等, 已经取得了一些进展[3]. Takeshi[4]等提出基于 Twitter 的实时监控地震系统, 在实际应用过程中检测到了 80%以上的地震发生, 时效性超过了当地的地震告警机构. 在国内, 新浪微博也邀请了政府组织、媒体机构、名人名流等地加入, 使得重大信息的发布都变得尤其迅速. 据新

① 基金项目:国家自然科学基金(61070062,61175123);福建高校产学合作科技重大项目(2010H6007)
收稿时间:2015-09-21;收到修改稿时间:2015-10-30 [doi:10.15888/j.cnki.csa.005155]

浪微博获悉: “2015 年除夕当天, 新浪微博日活跃用户首次突破 1 亿, 比去年上涨 46%, 除夕当天抢微博红包的总次数超过 1.01 亿次, 据统计, 2 月 18 日 0 点至春晚结束, 有 3470 万微博网友参与春晚互动, 讨论春晚的微博达到4505万条, 相关内容的总互动量达6941万, 相关话题总阅读量达 41.5 亿”[5]. 可见对于微博上信息的研究具有很大的价值.

传统的微博话题检测方法识别率不够高, 并且由于微博热门话题的生命周期较长, 容易导致突发话题容易被以往话题给覆盖, 因此不能及时发现. 并且由于中文的一词多义性, 导致不同话题存在相同的关键词而混淆. 文本旨在解决以上问题, 提出基于词共现关系和粗糙集的微博话题检测方法.

本文的其余部分安排如下: 在第二节中, 介绍了近年来对微博话题研究的相关工作; 在第三节中, 介绍传统的基于词共现关系的微博检测算法; 在第四节中, 详细介绍了本研究中提出的改进算法; 在第四节中, 对于实验数据进行介绍和分析; 最后在第五节中, 总结了本次研究工作以及未来需要进行的研究.

2 相关工作

近年来, 许多学者在微博话题的检测方面进行了大量研究. Zheng Feiran[3]在实验中分析了时间窗收集数据参数的选择, 并使用增量式聚类的方法来进行主题检测. Zhou Jinhua[6]把衰减的词共现图方法与潜在语义分析方法相结合并应用于多文档文摘, 得到了较好的主题抽取效果. Zhao Wenqing[7]提出一种基于词共现图的方法来识别微博中的新闻话题, 有效地解决了传统的话题检测算法在短文本稀疏数据处理上的不足. Bai Qiuchan[8]提出了基于关联规则词共现的文本主题聚类算法, 提高了聚类效率和准确性. Song Shien[9]提出了一种组合词上下文与传统词共现的方法, 并考虑到词的倾向性, 提高了观点词抽取的性能. Fang Ran[10]提出了一种情感内容加权的话题检测方法, 通过加大含有负面情感的短文本的权重,来提高聚类时对话题的查全率. Ge Bing 等[11]基于知网的词汇语义相似度计算方法研究,提出了改进的词汇相似度计算方法. Zhu Zhengyu 等[12]进一步改进了基于 《知网》 的词汇语义相似度计算方法, 使得利用语义相似度来解决语义相似问题. Cui Zhengyan[13]利用《知网》进行语义拓展, 并采用knn算法进行分类, 提高了准确率和召回率. Jin Chunxia[14]利用动态向量计算中文短文本的内容相似度, 提高了短文本聚类效果. Shi Jianhong[15]针对高维、稀疏的中文微博数据,利用 LDA 进行隐主题建模, 并采用频繁项集进行挖掘, 较好地实现数据降维和话题发现. Liu Zhiming[16]实验中表明, 采用SVM和IG,以及TF-IDF 作为特征项权重,三者结合对微博的情感分类效果最好. Qiu Yunfei[17]等针对传统话题检测方法不能很好处理微博中用语不规范、随意性强、指代不明以及存在大量网络用语的问题,提出了一种基于潜在狄利克雷分配(LDA)模型的主题树检测方法,提突发话题的识别率.

虽然已经有许多学者对微博突发新闻的检测进行了研究, 并提出了许多算法, 但还是存在着一些问题: 首先, 微博数据量的庞大, 运算耗时长; 其次, 不同主题的相似话题较难区分; 再次, 热点话题并非都是突发话题, 对重复性话题的辨识度不高. 因此本文就以上问题提出了相应的解决办法, 首先增加时间窗之间计算的独立性, 以降低计算量; 其次, 利用词共现矩阵将词汇分解成不同的话题完全子图, 实现一词多义的话题区分; 再次, 粗糙集中邻域的划分, 确保话题簇边界的划分, 减少话题间的相互覆盖的影响.

3 微博新闻话题检测相关常用方法

3.1 文本预处理

通常每收集一个单位时间窗的微博信息, 都要进行一次预处理, 以方便进行主题检测. 在预处理的过程中包括: 文本分词、停用词过滤、核心词提取等.

汉语中词是最小、能独立活动、有意义的语言成分[10], 因此也是进行文本处理的第一个关键步骤. 本文章中采用ICTCLAS[19]分词系统, 它是一种中文文本分词工具, 有词性标注、命名实体识别、用户词典等功能, 并且在微博分词中具有比较好的效果. 停用词过滤将使用频率较大但没有实际意义的字进行剔除, 例如: “的”、“了”; 同时剔除微博中特定的格式标签, 如“@用户”.

每一则有价值的微博话题, 都表达了某一事物的某一状态, 因此我们将微博的构成看成: “核心词+修饰词+无关词”. 核心词表示话题的主体, 主要由名词或部分动词构成; 修饰词则表达话题的一个状态或者行为, 主要为形容词和部分动词构成; 核心词和修饰词之外的则为无关词, 它们在信息中并没有起到实际

意义, 去除他们不影响信息表达的内容. 例如原微博“啊啊. 你在东北？？转: @haosong2004 冬天的棉袄穿上了, 立马就显暖和”, 分词并去除无关词后得到“东北 冬天 棉袄 暖和”, 处理结果同样能表达出原来的内容.

3.2 词共现聚类方法

词间的共现现象是指两个词之间形成的统计关系. 在文献 8 中指出, 词汇频繁共同出现在相同的上下文中, 就可以认为这个词汇组合是比较稳定的, 它们表达了某个潜在的主题信息[8]. 因此, 可以根据时间窗内的词共现情况, 找到该时间窗内的新闻主题, 而词共现形成的聚类, 正好是同一新闻主题下主题词的聚类.

微博信息不同于传统话题检测面向的对象, 它具有很强的时序性. “微博信息收集时间窗化的方法是一种有效地将连续的时间离散化的方法. 在主题词检测时需要控制好时间窗口的长度, 时间窗较小时, 容易受到噪声数据干扰, 查准率和查全率都较低; 时间窗较大时, 虽然选出的主题词较精确, 但是由于粒度较大, 有的话题被漏掉. ”[3]引入当前时间窗口中某词的频率 F:

$$F=\frac{F_i}{F_{\max}} \tag{1}$$

在式(1)中, F_i 表示该词在当前时间窗内出现的频次, F_{max} 表示当前时间窗内最高词频.

由于微博信息收集是一种增量式的数据, 因此通常引入增长系数 G_{ij} 来表示词 i 在某一时间窗格 j 的词频增长速度, 定义为当前窗格中该词的频率除以之前 K 个窗格的频率平均值:

$$G_{ij}=\frac{F_{ij}}{\overline{F_i}}=\frac{F_{ij}\times K}{\sum_{u}^{K}F_{iu}} \tag{2}$$

在式(2)中, F_{iu} 表示主题词 i 在 u 时间窗内的出现频率, K 是回顾窗的大小, 即当前时间窗之前的 K 个时间窗组成.

文献[7]提出, 构造一个符合权值来评价一个词是主题词的程度:

$$\omega_{ij}=\log G_{ij}+\alpha\log\frac{F_{ij}}{F_{\max}} \tag{3}$$

在式(3)中可以发现, 一个词在当前时间窗内出现的频率越高, 在过去出现的频率越小, 那么该词为主题词的程度就越高. 其中 a 的取值范围在 1 到 1.5 之间最好.

文献[6-9]均引入了增长系数 G_{ij}, 而该系数的计算必然增加时间窗之间计算的复杂度. 为了增加各个时间窗口之间计算的独立性, 最大程度地减少计算量, 我们在研究中去除了增长系数, 且该处理对结果的影响较小.

词共现模型[19]是基于统计的自然语言处理领域的重要模型之一. 在找出主题词列表之后, 再根据主题词在微博话题中的共现程度进行聚类, 即在几个主题词同时出现在同一话题中的概率越高, 说明这几个词是共同描述一个主题. 在文献[19]中给出了下面两个定义:

定义 1. 词汇 w_x 相对于词汇 w_y 的相对共现度 $R(w_x|w_y)$则定义为

$$R(w_x\mid w_y)=\frac{f(w_xw_y)}{f(w_y)} \tag{4}$$

在式(4)中, $f(w_xw_y)$为单位时间窗内词 w_x 和词 w_y 在同条微博中出现的次数, $f(w_y)$为词 w_y 在单位时间窗内出现的次数. 可知, 通常 $R(w_x|w_y)$不等于 $R(w_y|w_x)$.

定义 2. 词汇 w_x 与词汇 w_y 之间共现度 $C(w_x|w_y)$则定义为

$$C(w_x,w_y)=\frac{R(w_x\mid w_y)+R(w_y\mid w_x)}{2} \tag{5}$$

可知有 $C(w_x,w_y)=C(w_y,w_x)$.

在计算获得主题词表中, 各主题之间的共现度, 便可以根据主题词之间达到一定的共现度阈值进行聚类.

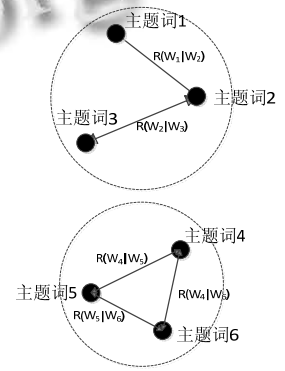

图 1 多连通图形成簇

文献[7]中提出, 基于词共现原理识别微博新闻话题的基本步骤:

(1)主题词共现图中点集 N_S 的生成. 微博文本预处理后得到主题词表, 将其中的主题词作为词共现图

G 的点集;

(2)对词共现图中的点集连边. 根据点集 N_S 中两个词之间的贡献度是否超过一定的阈值进行连边;

(3)微博新闻话题的确定. 根据点之间的连边形成多个连通区域, 每个连通区域构成一个簇(例如图 1 中的两个簇), 每个簇与一个热点话题对应;

(4)微博新闻话题表示. 如果一个词汇与越多的词汇形成共现词组合, 则具有较为积极的主题意义, 它很可能是某个主题的领域词汇. 根据一定阈值挑选出领域词汇作为该主题的核心词. 利用式(6)来计算每个簇中主题词的信息量大小, 其表示对簇集的贡献程度大小.

$$G(w_i)=\sum_{(w_i,w_j)\in E(G)} C(w_i,w_j) \tag{6}$$

式(6)中, $E(G)$是图 G 中的边集; 通过对主题词 w_i 的信息量 $G(w_i)$进行排序, 选出 K 个对话题簇贡献较大的主题词, 作为该热心话题的核心词.

3.3 基于最大 Code 码的极大完全子图算法

文献[21]提出基于最大 Code 码的极大完全子图算法 FMCSG, 在计算时通过 Code 码来进行子矩阵的剪枝, 极大提高运算量, 在词共现矩阵这一高维数据情况下有较好的效果.

定义 3. 设图 G 的邻接矩阵 X 为:

$$X(v_1,v_2,...,v_n)=\begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_n \end{array}\begin{bmatrix} 0 & x_{1,2} & x_{1,3} & \cdots & x_{1,n} \\ x_{2,1} & 0 & x_{2,3} & \cdots & x_{2,n} \\ x_{3,1} & x_{3,2} & 0 & \cdots & x_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & x_{n,3} & \cdots & 0 \end{bmatrix}$$

称 $code(X)=x_{1,2}x_{1,3}...x_{1,n}x_{2,3}...x_{n-2,n}x_{n-1,n}$ 为邻接矩阵的 code 码.

定义 4. 设 $H=(V',E')$ 是图 $G=(V,E)$ 的顶点个数为 k 的极大完全子图. 如果 G 的邻接矩阵 X 的左上角 k 阶子矩阵与 H 的邻接矩阵相同, 称 code(X)为图 G 的极大 code 码(Max-Code).

定义 5. 两个 k 阶邻接矩阵进行矩阵连接生成的 k+1 阶矩阵所对应的 code 码称为 k+1 阶候选码, 记他们的集合为 Ccode(k+1).

定义6. 给定图 $G=(V,E),|V|=n$. 称图 $G'=(V',E')$ 为图 G 的逆导出子图, 其中 V'是这样一种点的集合: 设 v_0 是 G 中顶点度最大的点(若此点不唯一, 任选其一), $V'\leftarrow\{v_0\}$, 且将 v_0 称为 V'中的核心点. 对于 $\forall v_i\in V$, 如果 v_i 和 v_0 之间有边相连, $V'\leftarrow\{v_i\}\cup V'$; E'是这样的集合: $\forall v_i,v_j\in V'$, 如果 $(x_i,x_j)\in E$, 则 $(x_i,x_j)\in E'$.

定义7. 给定图 $G=(V,E)$, $G'=(V',E')$ 是 G 的逆导出子图. 称 $G_c'=(V'',E'')$ 为图 G 的逆导出补图, 其中 $V''=(V-V')\cup\{v_i\mid\forall v_i\in V',v_j\in V-V'$, 有 $(v_i,v_j)\in E$, E'' 是这样的集合: $\forall v_i,v_j\in V''$, 如果 $(v_i,v_j)\in E$, 则 $(v_i,v_j)\in E''$.

定理1. 设 G' 和 G'' 分别是图 G 的逆导出子图和逆导出补图, 则有 $G=G'\cap G''$.

定理 2. 一个图可以由若干个图的逆导出子图构成.

定义8. 设 F_2 是图 G 的逆导出子图中的核心点 v_0 与其他各点 $v_i(0<i\le n)$ 两两组合构成的顶点序列的集合, $F_2=(v_0v_1,v_0v_2,...,v_0v_m)$, 二元关系 $\equiv_{head}$, 称为 Head 关系, 规定 $X\equiv_{head}Y$, 当且仅当 $Head(X)=Head(Y)$ 且 $X,Y\in F_2$.

定理 3. Head 关系为等价关系.

定义 9. 设图 G 的逆导出子图为 G', 其核心点为 v_0, $\sum$ 为 G 极大完全子图的顶点序列, 当等价类中有 $X\in\sum$ 且 $Head(X)=v_0$, 用 $\varphi_c(v_0)$ 表示等价类.

算法　FMCSG

输入: 图 G;

输出: Nest; //Nest 代表极大完全子图的顶点序列集, 初始值为空;

(1)　$G_0\leftarrow G$;

(2)　求出图 G_0 的逆导出子图 G'与逆导出补图 G'_c;

(3)　If G'_c 为空, 退出;

(4)　$k\leftarrow1$;

将 G'的核心点 v_1 对应的一阶矩阵分别与 G'中其他顶点对应的一阶矩阵相连接, 生成二阶矩阵;

(5)　for k=2 to n do // n 表示图的顶点个数;

将 G'中含有极大子 Code 码的 k 阶矩阵中的顶点序列放入 $\varphi_c(v_1)$;

由 k 阶矩阵两两相连接, 生成 k+1 阶矩阵;

生成候选 code 码, 找到极大子 code 码;

$\varphi_c(v_1)\leftarrow$极大子 code 码对应的顶点序列;

(6)　如果 $\varphi_c(v_1)$ 中某一顶点序列中的所有顶点均出现在另一顶点序列中, 则删除该顶点序列, 直到 $\varphi_c(v_1)$ 中不存在这样的顶点序列为止;

(7) $G_0 \leftarrow G'_c$, 转(2);

设图 G 的结点数为 n, 可以证明算法 FMCSG 的时间复杂度为:

$$T(n) = T(n-1) * T(n-1) + O(n^2).$$

4 基于词共现关系和粗糙集的微博话题检测方法

4.1 基于相关度矩阵建立话题极大完全子图

在传统的词同现图中, 通常通过主题词聚类形成的簇来确立新闻, 在这里已经隐含了一个假设, 即每个主题词只属于一个话题. 这一假设这显然限制了大部分话题的表达, 由于多个话题可能具有相同的主题词, 按照这一假设必然造成不同话题的混杂, 在共现图中错误地聚类成一个话题簇. 例如 NLPIR 提供的微博语料库中, 2012 年 2 月份 28 天以来的微博内容进行检测, 我们发现“城管”相关的微博始终占据着热门新闻头条. 专门对含有“城管”的微博进行提取, 按照传统的词共现算法, 有趣的是, 我们在每一个时间窗几乎都只能得出一条“城管—执法-小贩”, 似乎城管永远都是负面话题.

我们在这里通过建立词相关度矩阵, 采用文献 21 提出的 FMCSG 算法, 以词共现关系矩阵作为图的邻接矩阵, 构建最大完全子图, 使同一个核心词和在多个话题簇中出现, 以解决一词多义的问题. 采用以上方法, 我们可以得出一条“城管-江苏-研究生”, 微博内容大意为: 江苏常州城管一线有 12 名硕士. 该微博内容具有正面意义. 可以看出, 该话题在传统算法中被大数据淹没了, 仅按照传统的对簇集贡献度的大小筛选核心词, 就容易造成话题间的相互覆盖或者混杂.

4.2 改进粗糙集理论的核心词簇

粗糙集理论是一种处理不确定性问题的一种软计算方法. 粗糙集理论主要工作在于对空间的划分. 一个对象 a 是否属于集合 X 需根据现有知识判断, 可分为三种情况: (1)对象 a 肯定属于集合 X; (2)对象 a 不可能属于集合 X; (3)对象 a 可能属于也不可能不属于集合 X. 因此在空间上划分为下邻域、上邻域和外界. 从经典粗糙集理论对样本点分布进行描述:

设训练集样本 S=(X1,X2,…,Xn)(其中 n 表示训练集类别数量), v 表示待测样本点, Xi 表示训练集中第 i 类, $\underline{R}(X_i)$ 表示第 i 类的上邻域, $\overline{R}(X_i)$ 表示第 i 类的下邻域, 则

① 若 $\exists v \in \underline{R}(X_i)$, 则 $v \in \overline{R}(X_i)$;

② 若 $\exists v \in \underline{R}(X_i)$, 则 $v \notin \overline{R}(X_j)(j \neq i)$.

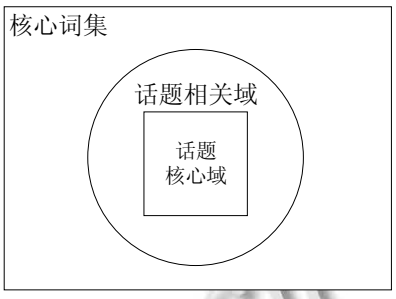

图 2 话题粗糙词集示意图

由图 2 所示, 根据粗糙集理论对空间划分的思想, 可以建立话题核心词集合. 相对粗糙集理论, 这里我们对话题集合的空间划分和对象的知识进行了重新的定义. 根据微博中核心词间的相关度矩阵, 确定了 n 个极大完全子图, 每一个极大完全子图作为一个话题的代表词集, 这里在空间中命名为话题核心域. 同样在空间中划分出话题相关域和话题外界. 改进后, 粗糙集中点之间的距离度量不再是简单的与中心点的距离, 而是要满足与区域内的点两两之间的距离都要小于阈值, 这里表现为共现度.

定义 10. 话题集 $T = (t_1, t_2, ..., t_m)$, m 为话题总数, t_i 为第 i 个话题.

定义 11. 核心词集 $W = (w_1, w_2, ..., w_n)$, n 为核心词总数, w_i 为第 i 个核心词.

定义 12. 相关度矩阵 $R_{n\times n} = \{r_{i,j} \mid 1 \le i, j \le n\}$, $r_{i,j}$ 为 w_i 和 w_j 的相关度.

定义 13. 话题核心域 $\underline{\mathrm{RT}}(\mathrm{t_i})$ 表示第 i 个话题的核心词词集, 其中 $\forall w_i, w_j \in \underline{\mathrm{RT}}(\mathrm{t_i})$, 满足 $\mathrm{r_{i,j}} > 0$.

定义 14. 话题相关域 $\overline{\mathrm{RT}}(\mathrm{t_i})$ 表示第 i 个话题的相关词词集, 则 $\forall w_i, w_j \in \overline{RT}(\mathrm{t_i})$, 给定一个错误容忍度 ε, 满足 $\frac{n}{N} > \varepsilon$, N 表示话题总词对的个数, n 表示相关词对的个数.

定义 15. 两个话题核心域的词集共现度矩阵进行连接形成 k 阶矩阵 X, 当满足以下条件时可以进行合并, 视为同一话题:

$\frac{num_1}{num_1 + num_2} > \delta$, 这里参数 δ 取 0.83.

矩阵 X 中 code 码全为 1 的数量记为 num_1, 0 的数量记为 num_2.

改进后的粗糙集内数据之间的联系更加紧密, 由核心域中的词汇作为话题类别的判定, 使结果更加严谨, 极大的提高准确性, 克服了微博话题词汇的稀疏性问题; 而话题相关域中的词汇作为该话题下相关词汇的补充, 使得话题的查全率得到保证. 因此改进后的粗糙集能很好地运用于微博话题的检测, 可行性较高.

4.3 话题检测及话题更新

由于过去时间窗口内产生的热门话题可能带来持续的热度, 导致当前时间窗口内可能产生与过去相同话题, 这些话题并不是我们要检测的目标, 因此需要与过去相邻时间窗内的话题进行比较.

首先需要比较话题核心域中的词汇, 若话题核心域中词汇的满足定义 15 中合并条件时, 表示该话题与过去的话题相同, 不作为突发话题; 若不满足合并条件时, 则表明该话题为突发话题. 在这里对于同一话题下的核心词汇发生变化的情况便进行了简单的处理, 变化较大时, 作为新的突发话题处理; 变化较小时引起的社会效应较小, 可以忽略, 因此作为过去的热门话题.

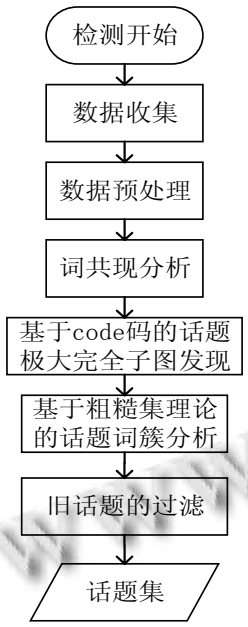

图 3 话题检测流程示意图

图 3 所示即为基于粗糙集原理的改进词共现算法(RSCW). 首先通过词共现分析, 获取核心词之间的相关度; 然后根据词之间的相关度建立共现矩阵, 通过矩阵获得所有极大完全子图; 最后根据极大完全子图作为粗糙集中的核心域, 通过话题词集的扩展, 加入话题相关词, 即可获得每一个话题下的词集.

通过该方法, 可以容易地解决词汇在不同话题中一词多义的表达情况, 并且通过极大完全子图的划分, 缩短了计算量; 通过粗糙集理论, 无需考虑集合外的词汇, 进一步缩短了算法的计算量, 提高了话题获取精度.

5 实验结果及分析

5.1 实验数据说明

本实验使用两个语料库内容, 第一个为 NLPIR[18] 微博内容语料库-23 万条中 2011 年 12 月份的微博数据, 其中已剔除研究无关数据(包含空数据, 图片微博, 视频微博), 有用数据共计 81596 条. 图 4 中反应了 12 月份中每日的微博信息收集量, 平均每日的微博数量为 2720 条; 由图 5 和图 6 中 24 小时的微博发布量我们可以发现, 不同时间段内的微博发布量是有很大差异的, 这是由于人们的作息习惯导致, 在睡眠时间的微博发布量明显下降. 传统时间窗口数据为固定时间长度的信息量收集, 因此会造成各个时间窗内数据量起伏大, 检测效率低, 检测精度缺失. 在研究中对时间窗口的时间长度进行弹性设置, 我们将时间窗口长度以搜集的数据量为单位, 设为周平均 2 小时的数据量 206 条. 图 7 中反应了 12 月份微博与核心词数量的分布图, 可见大多微博的核心词数量在 6 个左右. 图中核心词数量较少的部分, 占据了较大部分的微博数量, 它们并没有表达出实际的话题意义, 此类信息可视为无用信息.

第二个语料库为通过爬虫获取的话题榜中 2015 年 7 月 3 号到当月 21 号以来微博的收集, 得到 28155 条微博数据, 含有 40 个话题. NLPIR 微博内容语料库提供的话题信息数据量大, 但是较为稀疏, 该语料库记为 Set1; 通过爬虫获取的语料库数据量小, 但是话题信息较为紧凑, 该语料库记为 Set2. 将算法通过两个不同特点的语料库进行比较, 以体现它的效果.

5.2 评价指标

话题正确率 P1: 时间窗内检测出的正确话题簇数作为分子, 时间窗内检测出的话题簇数作为分母;

话题召回率 P2: 时间窗内检测出的正确话题数作为分子, 时间窗内的总话题数作为分母;

话题重复率 P3: 时间窗内检测出的重复话题簇数作为分子, 时间窗内检测出的正确话题数作为分母;

话题覆盖率 P4: 时间窗内检测出的含有多个话题

的簇数作为分子, 时间窗内检测出的正确话题数作为分母;

话题正确率表现算法的查准率; 话题召回率表现算法的查全率; 话题重复率表现算法找出的话题集之间表征的话题的重复情况, 该值越低说明算法性能越好; 话题覆盖率表现一个话题集中涵盖多个话题的情况, 该值越低说明算法性能越好.

同时在实验中使用 6 种不同数据量进行实验, 来考察算法对不同数据量的执行能力.

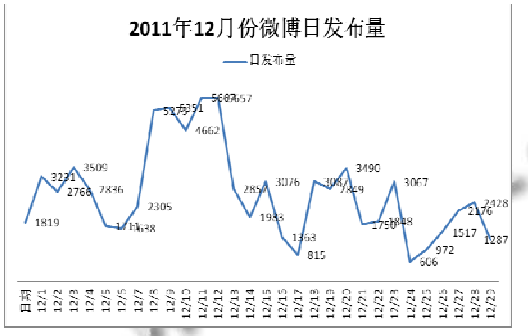

图 4 2011 年 12 月份微博日发布量

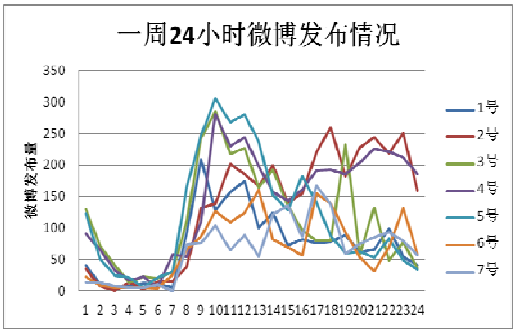

图 5 24 小时微博发布量走势图

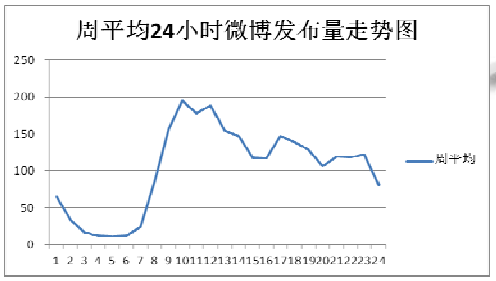

图 6 周平均微博 24 小时发布量走势图

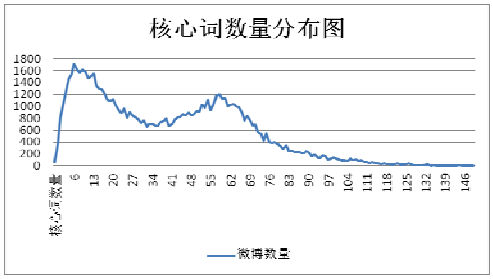

图 7 核心词数量分布图

5.3 实验结果分析

本实验对比算法采用传统的词共现模型算法 7(记为 CW-model), 并通过调整时间窗口内数据的倍数(从 1 倍到 6 倍)以观察在不同数据量大小下的检测效果.

表 1 语料库 Set1 下 2 种算法在不同数据量下的 4 种精度情况

	语料库一					
P1	x1	x2	x3	x4	x5	x6
CW	63.27%	67.35%	71.43%	65.31%	61.22%	57.14%
RSCW	69.39%	75.51%	79.59%	73.47%	73.47%	67.35%
P2	x1	x2	x3	x4	x5	x6
CW	48.98%	53.06%	53.06%	53.06%	53.06%	53.06%
RSCW	61.22%	65.31%	71.43%	65.31%	61.22%	63.27%
P3	x1	x2	x3	x4	x5	x6
CW	5.41%	8.11%	9.00%	10.38%	10.96%	11.50%
RSCW	5.41%	10.81%	12.73%	13.51%	15.59%	17.62%
P4	x1	x2	x3	x4	x5	x6
CW	5.08%	6.78%	6.78%	8.47%	8.47%	8.47%
RSCW	0.00%	1.69%	3.39%	4.24%	3.87%	4.16%

表 2 语料库 Set2 下 2 种算法在不同数据量下的 4 种精度情况

	语料库二					
P1	x1	x2	x3	x4	x5	x6
CW	72.00%	71.85%	71.72%	70.00%	68.39%	66.88%
RSCW	84.00%	88.89%	93.10%	93.33%	90.32%	87.50%
P2	x1	x2	x3	x4	x5	x6
CW	72.50%	75.00%	77.50%	75.00%	72.50%	67.50%
RSCW	75.00%	80.00%	85.00%	87.50%	90.00%	87.50%
P3	x1	x2	x3	x4	x5	x6
CW	7.69%	7.14%	0.00%	0.00%	0.00%	0.00%
RSCW	19.05%	16.67%	14.81%	14.29%	14.29%	14.29%
P4	x1	x2	x3	x4	x5	x6
CW	13.08%	18.57%	16.67%	23.33%	23.33%	23.33%
RSCW	0.00%	0.00%	7.41%	7.14%	7.14%	7.14%

1)不同数据量下, 算法的效能

由表 1 和表 2 可以看出 RSCW 在 5 倍数据的情况下精度最好, 并且两种算法的精度都是随着数据量变大而变大, 超过一定阈值后变小. 这种变化正是由于如果数据量继续增大, 将导致词与词之间的相关度趋于平和, 即任意两个词之间都能建立起联系, 并且所有词对的相关度都倾向于均值.

2)不同语料库下, 算法的效能

2 种算法在语料库 Set2 的效能比 Set1 的更好, 这是由于语料稀疏情况下, 非话题下的词汇相关度会造成误导的影响. 由于语料稀疏, 导致相同话题下的词对相关度不高, 使得集合包含非话题的词对.

3)不同精度下, 算法的效能

由 2 种算法在 4 种精度的效能上可以看出, 传统

的词共现算法(CW)的话题覆盖率最高, 重复率最低, 这是因为 CW 算法处理一词多义的情况最差, 当数据越稀疏, 词对之间的关联度越低时表现得越明显; RSCW 算法的精度、准确率、覆盖率效能表现得都最好, 但是重复率表现得较差, 这是由于 RSCW 对话题的划分较为严谨, 导致由于关联度不高的词对被分开, 使话题称为多个自话题簇, 但是这对话题检测的影响并不是很大.

由以上总结可看出, RSCW 算法的性能是具有一定优越性的.

6 未来工作及总结

本文通过在词共现度基础上建立相关度矩阵, 采用极大完全子图处理词汇多义现象, 并通过粗糙集理论对话题集合进行构建, 有效地检测出微博中的突发话题. 在未来的工作中, 还需要处理好非话题微博的识别, 防止对算法造成的影响, 并且由于微博话题下的词汇会根据热点发生变化, 因此要研究话题集合中词汇的自适应变化情况.

参考文献

1 Kwak H, Lee C, Park H, et al. What is Twitter, a social net-work or a news media? Proc. of the 19th International Conference on World Wide Web. 2010.

2 Zhao JJ, Wu WL, Zhang XL, Qiang Y, Liu T, Wu LD. A short-term trend prediction model of topic over Sina Weibo dataset. Journal of Combinatorial Optimization, 2014, 28(3): 613−625.

3 郑斐然,苗夺谦,张志飞,高灿.一种中文微博新闻话题检测的方法.计算机科学,2012,1:138−141.

4 Sakaki T, Okazaki M, Matsno Y. Earthquake shakes Twitter users: Real-time event detection by social sensors. Proc. of the 19th International Conference on World Wide Web. 2010.

5 新浪.今年除夕微博日活跃用户首次突破 1 亿.http://news.xinhuanet.com/tech/2015-02/25/c_1114430665.htm

6 周进华,刘贵全.基于衰减词共现图的多文档摘要研究.小型微型计算机系统,2009,1:173−177.

7 赵文清,侯小可.基于词共现图的中文微博新闻话题识别.智能系统学报,2012,5:444−449.

8 白秋产,金春霞,章慧,周海岩.词共现文本主题聚类算法.计算机工程与科学,2013,7:164−168.

9 宋施恩,樊兴华.基于词共现和词上下文的领域观点词抽取方法.计算机工程与设计,2013,11:4012−4015.

10 方然,苗夺谦,张志飞.一种基于情感的中文微博话题检测方法.智能系统学报,2013,3:208−213.

11 葛斌,李芳芳,郭丝路,汤大权.基于知网的词汇语义相似度计算方法研究.计算机应用研究,2010,9:3329−3333.

12 朱征宇,孙俊华.改进的基于《知网》的词汇语义相似度计算.计算机应用,2013,8:2276−2279,2288.

13 崔争艳.基于语义的微博短信息分类.现代计算机(专业版),2010,08:18−20,24.

14 金春霞,周海岩.动态向量的中文短文本聚类.计算机工程与应用,2011,33:156−158.

15 史剑虹,陈兴蜀,王文贤.基于隐主题分析的中文微博话题发现.计算机应用研究,2014,3:700−704.

16 刘志明,刘鲁.基于机器学习的中文微博情感分类实证研究.计算机工程与应用,2012,1:1−4.

17 邱云飞,郭弥纶,邵良杉.基于主题树的微博突发话题检测.计算机应用,2014,8:2332−2335.

18 张华平.NLPIR 微博内容语料库—23 万条.自然语言处理与信息检索共享平台.http://www.nlpir.org/.

19 张华平.ICTCLAS2015 版本.自然语言处理与信息检索共享平台.http://ictclas.nlpir.org/.

20 耿焕同,蔡庆生,于琨,赵鹏.一种基于词共现图的文档主题词自动抽取方法.南京大学学报(自然科学版),2006,2:156−162.

21 郭平,康艳荣,史晓晨.基于最大 Code 码的极大完全子图算法.计算机科学,2006,2:188−190,200.