文章编号：0253-374X(2019)04-0583-10 DOI:10.11908/j.issn.0253-374x.2019.04.019

融合频繁项集和潜在语义分析的股评论坛主题发现方法

张 涛[1,2]，翁康年[1]，顾小敏[1]，张玥杰[3,4]

(1. 上海财经大学 信息管理与工程学院，上海 200433；
2. 上海市金融信息技术研究重点实验室(上海财经大学)，上海 200433；
3. 复旦大学 计算机科学技术学院，上海 200433；4. 上海市智能信息处理重点实验室(复旦大学)，上海 200433)

摘要：针对股评论坛主题发现，提出基于频繁项集与潜在语义相结合的短文本聚类(STC_FL)框架.在基于知网的知识获取后得到概念向量空间，挖掘并筛选出重要频繁项集，然后采用统计和潜在语义相结合的方法进行重要频繁项集的自适应聚类.最后，提出TSC-SN(text soft classifying based on similarity threshold and non-overlapping)算法，通过参数调优策略选择和控制文本软聚类过程.股吧论坛数据实证分析发现：所提出的STC_FL框架和TSC-SN算法可充分挖掘文本潜在语义信息，并有效降低特征空间维度，最终实现对短文本的深层次信息挖掘和主题归类.

关键词：主题发现；股吧论坛；频繁项集；潜在语义分析；文本软聚类

中图分类号：TP391 **文献标志码**：A

Topic Discovery Method of Stock Bar Forum Based on Integration of Frequent Item-set and Latent Semantic Analysis

ZHANG Tao[1,2], *WENG Kangnian*[1], *GU Xiaomin*[1], *ZHANG Yuejie*[3,4]

(1. School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China; 2. Shanghai Key Laboratory of Financial Information Technology (Shanghai University of Finance and Economics), Shanghai 200433, China; 3. School of Computer Science, Fudan University, Shanghai 200433, China; 4. Shanghai Key Laboratory of Intelligent Information Processing (Fudan University), Shanghai 200433, China)

Abstract: To achieve more effective topic discovery of stock bar forum, this paper presents a framework with short text clustering based on frequent item-set and latent semantic (STC_FL). The important frequent item-sets are acquired with the concept vector space based on HowNet, and then a combination pattern of statistics and latent semantics is used to realize the self-adaptive clustering of important frequent item-sets. Finally, the algorithm of text soft classifying based on similarity threshold and non-overlapping (TSC-SN) is proposed. Text soft clustering is selected and controlled with parameter optimization. By taking the real stock bar forum data as a specific case of empirical analysis, it is shown that STC_FL framework and TSC-SN algorithm can fully exploit the latent semantic information of text and reduce the dimension of feature space, which realizes the deep information mining and topic classification of short texts.

Key words: topic discovery; stock bar forum; frequent item-set; latent semantic analysis; text soft clustering

我国证券市场的发展历史短，各项机制还不够健全，因此交易行为常常受到市场信息和传闻的影响.特别是2015年我国股票市场在52个交易日内呈现股灾式暴跌，整体跌幅高达40.31%，年内A股市场惊现17次千股跌停，这暴露出我国证券市场发展的不成熟和股民们的非理性投资决策行为，股市的频繁剧烈波动已超出传统金融学理论的解释范围.研究表明，投资者情绪可显著影响股票市场的表现，如何通过相关论坛股评信息的主题挖掘来度量投资者情绪对股市表现的影响，已成为金融领域的重要研究方向.

网络论坛积累了大量短文本，短文本携带着丰

收稿日期：2018-05-01
基金项目：国家自然科学基金(61572140)；上海市科委项目(17DZ1100504，16511104704)；教育部人文社会科学研究规划基金(19YJA630116)
第一作者：张 涛(1970—)，男，教授，博士生导师，工学博士，主要研究方向为数据挖掘、智能优化算法.
E-mail:taozhang@mail.shufe.edu.cn
通信作者：顾小敏(1992—)，女，硕士，主要研究方向为数据挖掘、智能优化算法. E-mail:gxm13694309801@163.com

富的用户信息,成为极具价值的新型信息资源[1].因此,从论坛的丰富信息中挖掘出用户真正关心的主题[2],不仅有助于管理层及时了解网络热点信息,还便于对网络舆情的监管[3-4].然而,网络论坛的文本数据具有低质、简短和冗余等问题,使得在基于现有向量空间模型的文本聚类方法处理时陷入高维稀疏、语义缺失的困境.对此,基于深度学习的方法效率较高,但需要依赖大量数据集进行训练,而实际应用中很难获取庞大的数据集.机器学习方法易于解释和理解,便于进行参数调整和模型改进,本文中提出的主题发现方法就是利用改进的机器学习算法进行短文本筛选和频繁项集的聚类.

选取新浪财经股吧论坛版块的评论作为数据集,利用基于频繁项集与潜在语义相结合的短文本聚类(STC_FL)框架和TSC-SN(text soft classifying based on similarity threshold and non-overlapping)算法对论坛数据进行深层次主题分析与挖掘,实现在线股评文本的自动聚类.

1 相关研究工作

一般从以下两个方面对投资者情绪进行考量:从隐性情绪指数的视角,选择公认可测变量来衡量;从显性指数的角度,通过实际调查来获取投资者的情绪[5].面向股评论坛的主题发现是通过对股评文本进行挖掘来获得潜在的主题和热点,然后分析用户发帖行为和情绪指标,并将其用于股市表现分析,以支持投资者的合理投资决策[6].

利用概率模型进行各类文本热点主题挖掘的方法已在信息处理领域得到广泛应用[7].常见的主题发现模型涵盖概率潜在语义索引(PLSI)模型、隐含狄利克雷分配主题(LDA)模型和潜在语义索引(LSI)模型等.其中,LDA模型最为经典,可用于从大量文档集中挖掘潜在的主题信息[8].Shams等[9]将共生关系作为先验领域知识应用到LDA模型中,自动从共生关系等方面的相关主题提取相关的先验知识,提高模型效果.Kim等[10]采用LDA模型,并结合基于变分期望最大化(EM)算法的学习模型参数推理算法,实现Twitter朋友和内容的推荐.Zhang等[11]提出基于群体LDA模型的受众检测方法,将图书模块和图书章节信息融入到模型中.李扬等[12]基于LDA模型将由文本提取的潜在主题用作分类特征,提出基于主题模型的阈值调整半监督文本情感分类模型.然而,基于概率模型的主题发现方法在训练过程中对语料依赖程度较高[13],应用于短文本数据效果不佳,主题中常出现高频重复词而无法直观看出主题,并且容易出现过拟合[14].

基于词频统计的主题挖掘方法也得到一定的关注与应用,最具代表性的是K-means算法.该算法在处理大规模数据时效率较高,不足之处在于初始聚簇中心容易选择不当而导致文本聚类结果为局部最优.针对该算法的不足,Laszlo等[15]利用遗传算法改进K-means算法对初始聚簇中心敏感的问题,尝试将该算法应用于高维数据聚类中.Sun等[16]引进Bradley和Fayyad的初始点迭代算法,提高了K-means算法聚类结果的准确性.然而,基于词频统计的主题发现方法是基于距离来度量文本之间和文本与聚簇类别间的相似度大小,而现实中文本特征项常常具有高维性.

基于频繁项集的热点主题挖掘方法的基本假设是:同一个主题聚簇中的文档集应共享更多的频繁项集,而不同主题聚簇间的文档集则共享较少的频繁项集.在此假设下按照频繁项集将文本划分至不同主题类别下[17].该方法得到了广泛的研究和应用.Chen等[18]提出了基于模糊频繁项集挖掘的层次文档聚类.Wang等[19]将频繁项集的概念用于数据库中的事务聚类和文本聚类,提出基于频繁项集的文本聚类算法.在应用中,学者们也对基于频繁项集的聚类算法不断改进.Zhang等[20]提出MC(maximum capturing)算法,利用文档所包含的频繁项集来度量文档间相似度,并将文档集划分至相似度高的聚簇中.Sethi等[21]提出混合频繁项集挖掘方法,通过对数据集进行垂直布局来解决迭代中数据集扫描的问题,提高算法效率.Djenouri等[22]提出频繁项集挖掘仿生方法,考虑频繁项集的递归性质,并引入粒子群优化算法.

基于频繁项集的方法从文本中挖掘频繁出现的词集合,可有效降低文本特征维度,又可对聚簇结果的聚类主题进行基本描述.然而,针对面向股评论坛中短文本比例较高的特殊情形,依然需要考虑以下三个问题:① 聚类过程中忽略文本所包含的潜在语义关系,造成语义缺失和不合理聚类;② 聚类中仍涉及初始聚簇中心选择与聚类数确定的问题;③ 采用的聚类算法仍属于文本硬聚类,仅将文本划分至唯一聚簇中.为解决这三个问题,有必要建立一种频繁项集和潜在语义的融合机制,有效结合两种方法的优势,以实现对短文本深层次信息的挖掘和主题归类.

2　面向股评论坛的主题发现新框架

为解决现有主题挖掘方法处理网络股评论坛中短文本数据所存在的困难,构建一种面向股评论坛主题发现的短文本聚类框架. 利用频繁项集与潜在语义相结合的 STC_FL 框架从在线股评抽取主题词,再使用 TSC-SN 算法基于主题词进行文本检索,从而实现特有的股评文本聚类,如图 1 所示. 知网(HowNet)是以揭示概念与概念之间和概念所具有的属性之间的关系为基本内容的常识知识库. 针对文本中所蕴含的潜在语义关系,引入知网作为背景知识库建立基于概念的向量空间,并在文本集相似度计算的基础上,采用基于统计和潜在语义相结合的度量模式. 通过较长频繁项集预估主题个数,以解决聚类结果数目的最优设定. 针对融合频繁项集与潜在语义关系的文本软聚类,在文本检索阶段对 TSC-SN 算法设置短文本与主题簇间相似度阈值与簇间非重叠度参数,灵活选择和控制文本与主题间的对应关系. 采用频繁项集和概念映射来降低向量空间维度,弥补基于向量空间的聚类所存在的语义缺失问题;融合频繁项集与潜在语义,有效降低特征空间维度的同时充分考虑潜在语义关系;在对主题词相关文本进行检索时控制短文本与主题簇间相似度阈值,同时引入簇间非重叠度概念,利用新型文本集划分策略实现文本软聚类.

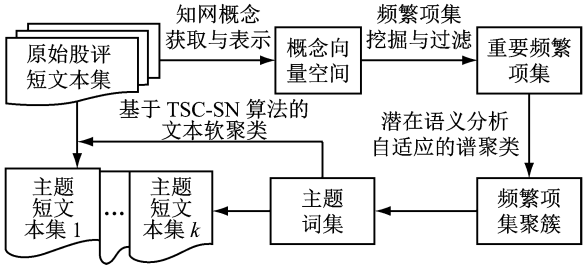

图 1　基于频繁项集和潜在语义的短文本聚类基本框架

Fig.1　Basic framework of short text clustering based on frequent item-sets and latent semantics

2.1　概念获取与表示

为了使具有潜在语义关系的词能够表达同一主题,引入语义知识源——知网作为背景知识库来加强语义间关联,在关键词向量空间中,将关键词映射至知识库中的某个概念,以概念来代替关键词特征项,在更高的概念层面上实现文本相似度度量,从而使同一主题的文本更容易聚集在一起.

2.1.1　词义消岐

当某个语义场与文本中的语境相符时,语义场中的词也有可能出现在文本中,可通过对比文本中的词和语义场中的词来实现语义消岐. 通过计算各语义场中词在文本中的重要程度来选取概念定义式(DEF),采用语义场密度进行度量,表现为语义场中词在文本中出现的频率之和. 对于一个多义词 w,其第 i 个 DEF 的语义场密度定义如下所示:

$$\rho(w_i)=\sum_{j=1}^{q_i}f(t_j)$$

式中:t_j 表示第 i 个语义场中第 j 个词;$f(t_j)$表示多义词 w 的语义场中第 j 个词在文本中出现的频率;q_i 为第 i 个语义场中所有词的个数. 语义场密度越大,语义场中的词对文本就越重要,针对词义消歧的 DEF 由下式确定:

$$Z_{\mathrm{DEF},w}=\max_i(\rho(w_i))$$

2.1.2　义原抽取

由知网的概念层次树特点可知,义原在概念树中的层次越深,所表达的含义就越具体,其描述能力就越强[23]. 可以认为,义原离概念树根节点越远,同时下位义原个数越少,该义原的描述能力就越强. 义原权值计算如下所示:

$$W(Z_{\mathrm{DEF},wj})=W_{\mathrm{tree}}\left(\log(d_{\mathrm{root},j}+a)+\frac{1}{m_j+b}+c\right)$$

式中:$W(Z_{\mathrm{DEF},wj})$为 DEF 中第 j 个义原 $Z_{\mathrm{DEF},wj}$ 的权值;W_{tree}为所在概念树的权重;$d_{\mathrm{root},j}$为义原 j 在概念树中的层次;m_j 为义原 j 的下位义原数;a、b、c 为控制权值 $W(Z_{\mathrm{DEF},wj})$取值的因子. 最终,义原的选取由下式确定:

$$Z_{\mathrm{DEF},w}\leftarrow\max_j(W(Z_{\mathrm{DEF},wj}))$$

2.1.3　概念向量空间构建

在对文本、关键词进行概念抽取后,即可构建基于概念的向量空间. 假设分词和预处理后的文本 $d=\{t_1,\ f_1(d),\ \cdots,\ t_i,\ f_i(d),\ \cdots,\ t_n,\ f_n(d)\}$,$t_i$ 表示文本 d 中的第 i 个关键词,$f_i(d)$表示文本 d 中 t_i 的词频,概念向量空间表示的生成算法如图 2 所示.

2.2　基于潜在语义分析的频繁项集聚类

针对所构建的概念向量空间,利用频繁模式增长(FP-growth)算法进行频繁项集挖掘,但得到的频繁项集存在冗余度高的问题. 为此,采用相似度过滤获取重要频繁项集. 首先剔除所有频繁项集的子集,然后对剩余频繁项集计算相似度. 将频繁项集相似度定义为 Jaccard 系数形式,如下所示:

$$\mathrm{Sim}(I_i,I_j)=J(I_i,I_j)=\frac{|\ I_i\cap I_j\ |}{|\ I_i\cup I_j\ |}$$

式中:I_i 表示频繁项集 i;$J(I_i,I_j)$表示 I_i 与 I_j 的

输入:文本 d 的关键词向量空间 $V_t(d)=(t_1, f_1(d), \cdots, t_i, f_i(d), \cdots, t_n, f_n(d))$,阈值为 θ

while $d\neq\varnothing$ 且 $i\leqslant n$

从 d 中依次取出关键词 t_i;

判断关键词 t_i 在知网中是否存在;

if t_i 为未登录词

if $f_i(d)<\theta$

去除;

else t_i 的概念 $z_i=\{t_i\}$,并将概念 z_i 和词频 $f_i(d)$ 加到概念向量空间 $V_c(d)$ 中;

else 查询知网,获取 t_i 的概念

if t_i 只有一个 DEF 定义

计算每个义原的权值 $W(Z_{\mathrm{DEF},wj})$,选择权值最大者作为 t_i 的概念 z_i,并统计 z_i 频率,将 z_i 加入至概念向量空间 $V_c(d)$ 中;

else 通过词义消歧选择 t_i 的语义场密度最大的 DEF,再选择其中权值最大的义原计算频率,作为 t_i 的向量加入至概念向量空间 $V_c(d)$ 中;

$i=i+1$;

end

return 文本 d 的概念向量空间 $V_c(d)=\{z_1, f_1(d), \cdots, z_i, f_i(d), \cdots, z_k, f_k(d)\}$

图 2 概念向量空间表示的生成算法

Fig.2 Generation algorithms for conceptual vector space representation

Jaccard 系数;$|I_i\cap I_j|$ 表示 I_i 与 I_j 的交集元素个数;$|I_i\cup I_j|$ 表示 I_i 与 I_j 的并集元素个数. 若频繁项集相似度大于设定值,则剔除,否则保留. 将每一频繁项集作为一个检索词串,从文本中查询出相关文本集合. 因此,两个频繁项集间的相似度计算就可由其相关文本集间相似度来替代,如下所示:

$$\begin{cases}\mathrm{Sim}_1(I_i,d_k)=\sum\limits_{j=1}^{g_i}W_jf_{jk}\\ \mathrm{Sim}(I_i,I_j)=\mathrm{Sim}(D_i,D_j)\end{cases}\tag{1}$$

式中:D_i 和 D_j 分别为包含频繁项集 I_i 和 I_j 的文本集;g_i 为频繁项集 I_i 中词的个数;W_j 为每个词的权重;f_{jk} 为词 t_j 在文本 d_k 中出现的次数. 设 ζ 为频繁项集与文本间最小相似度,当 $\mathrm{Sim}_1(I_i, d_k)\geqslant\zeta$ 时,将文本 d_k 划分至频繁项集 I_i 的相关文本集 D_i 中. 由此,即可得到频繁项集相似度较高的文本集.

2.2.1 文本潜在语义分析

潜在语义分析(LSA)是 Scott 等于 1990 年提出的一种索引与检索方法[7]. 基于该方法的表示过程为矩阵奇异值分解(SVD)与降维,具体步骤如下所示:

(1) 分析文档集,建立词-帖子矩阵. 假设帖子数量为 n,涵盖 m 个词,$\boldsymbol{X}_{m\times n}=(X_{ij})=(\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_n)$,$X_{ij}$ 表示词 i 在帖子 j 中出现的频数.

(2) 运用 SVD 将 $\boldsymbol{X}_{m\times n}$ 分解为三个矩阵的乘积,$\boldsymbol{X}_{m\times n}=\boldsymbol{USV}^{\mathrm{T}}$. 其中,$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 分别为 $m\times m$ 与 $n\times n$ 的正交矩阵,$\boldsymbol{S}$ 为对角矩阵,$\boldsymbol{S}$ 的非零对角元素 $\delta_i(i=1, 2, \cdots, r)$ 为矩阵 $\boldsymbol{X}_{m\times n}$ 的奇异值,r 为非零对角元素的个数.

(3) 对 SVD 后的矩阵进行降维,剔除较小奇异值. 计算得到原矩阵的相似矩阵 $\boldsymbol{X}'$,构建潜在语义空间,将文档向量与查询向量映射至一个子空间,该空间中来自文档矩阵的语义关系被保留,从而计算出帖子间的相似度.

2.2.2 文本语义相关度度量

为充分考虑自然语言中所蕴涵的语义问题,提出将语义和统计相结合的文本语义相关度度量方法. 在考察频繁项集相关的文本集间相关度时采用以下两种计算方式:基于 Jaccard 系数和基于 SVD 相似矩阵. 基于 Jaccard 系数和基于 SVD 相似矩阵计算式如下所示:

$$\begin{cases}\mathrm{Sim}_2(D_i,D_j)=\dfrac{|D_i\cap D_j|}{|D_i\cup D_j|}\\ \mathrm{Sim}_3(D_i,D_j)=\cos(\boldsymbol{c}_i,\boldsymbol{c}_j)=\\ \qquad\dfrac{x_{i1}x_{j1}+x_{i2}x_{j2}+\cdots+x_{iR}x_{jR}}{\sqrt{x_{i1}^2+x_{i2}^2+\cdots+x_{iR}^2}\sqrt{x_{j1}^2+x_{j2}^2+\cdots+x_{jR}^2}}\\ \mathrm{Sim}(I_i,I_j)\leftarrow\mathrm{Seqcom}(\mathrm{Sim}_2(D_i,D_j),\\ \qquad\mathrm{Sim}_3(D_i,D_j))\end{cases}\tag{2}$$

式中:$\boldsymbol{c}_i$ 为文本集 D_i 中所有文本合并生成的长向量;$x_{ir}(r=1,2,\cdots,R)$ 为 $\boldsymbol{c}_i$ 中的元素;$\mathrm{Sim}_2(D_i, D_j)$ 和 $\mathrm{Sim}_3(D_i, D_j)$ 分别为基于 Jaccard 系数和基于 SVD 相似矩阵的潜在语义分析所计算的文本集语义相关度;Seqcom(*, *)为最终文本集之间的语义相关度. 设 η 为文本集之间 Jaccard 系数最小语义相关度,ω 为文本集间的潜在语义最小相似度,则 Seqcom(*, *)计算按照以下策略进行:

步骤 1 计算度量文本集 D_i 和 D_j 间语义相关度的 Jaccard 系数. 若 $J(D_i, D_j)\geqslant\eta$,则 D_i 和 D_j 语义相关,否则执行步骤 2.

步骤 2 计算相关文本集 D_i 和 D_j 间的潜在语义相关度 $\cos(\boldsymbol{c}_i, \boldsymbol{c}_j)$,若 $\cos(\boldsymbol{c}_i, \boldsymbol{c}_j)\geqslant\omega$,则 D_i 和 D_j 语义相关,否则两者不相关.

2.2.3 基于潜在语义分析的聚类

字符较多的频繁项集表达完整且明确的主题,利用较长频繁项集进行聚类所得到聚类数可作为总

体频繁项集 V 的初始聚类数. 选取较长频繁项集集合 $I^*=\{v_i|v_i\in V,\ |v_i|>2\}$，设定初始簇 $C_1=\{v_1|v_1\in I^*\}$，初始簇集 $C=\{C_1\}$，初始簇数目 $K=1$，则对 $\forall v_i\in I^*$，依次比较 v_i 与当前所有簇 $C_k\in C$ 间的相似度. 对较长频繁项集聚类后将簇按大小排序，依次累计簇的元素个数，直至累计之和大于集合 I^* 长度的 80%为止，此时已累计簇的数量即为预估的聚类数 K. 为此，频繁项集与簇间的相似度计算如下所示：

$$\mathrm{Sim}_4(v_i,C_k)=\frac{1}{|C_k|}\sum_{j=1}^{|C_k|}\mathrm{Sim}(v_i,v_j)$$

对任一频繁项集 v_i 与簇 C_k 间的相似度，可利用 v_i 与 C_k 中所有频繁项集的平均相似度来计算.

重要频繁项集的聚类采用谱聚类算法. 设 $G=(V^*,E)$ 为带权无向图，两个顶点 o_i 与 o_j 间边权重 $W_{ij}^*=\mathrm{Sim}(I_i,I_j)$，所有顶点间的相似度构成相似矩阵 $\boldsymbol{Y}^*=(W_{ij}^*|o_i,o_j\in V^*)$，频繁项集聚类即为图 G 的 K 路分割问题，目标为各子图内相似度最大而子图间相似度最小. 由以上方法，构建基于频繁项集和潜在语义的聚类算法，如图 3 所示.

输入：重要频繁项集集合 $V'=\{v_i|v_i=I_i,\ i=1,2,\cdots,N\}$，用于挖掘频繁项集的文本集 $D^*=\{d_j|j=1,2,\cdots,M\}$，词权重集 $W=\{W_p|p=1,2,\cdots,P\}$，参数 η、ω、ζ 以及簇与频繁项集间最小相似度 γ

初始化：初始化每个频繁项集 v_i 的相关文本集 $D_i=\varnothing$，$\forall v_i\in V'$，$d_j\in D^*$，根据式(1)计算 $\mathrm{Sim}_1(v_i,d_j)$；

if $\mathrm{Sim}_1(v_i,d_j)\geqslant\zeta$

将 d_j 加入至 v_i 的相关文本集 D_i 中；

建立相似度矩阵 $\boldsymbol{X}^*$，元素 W_{ij} 由式(2)中的 $\mathrm{Sim}_2(D_i,D_j)$ 和 $\mathrm{Sim}_3(D_i,D_j)$ 比较得到；

if $\mathrm{Sim}_2(D_i,D_j)=J(D_i,D_j)\geqslant\eta$

$W_{ij}=\mathrm{Sim}_2(D_i,D_j)$；

else if $\mathrm{Sim}_3(D_i,D_j)=\cos(\boldsymbol{c}_1,\boldsymbol{c}_2)\geqslant\omega$

$W_{ij}=\mathrm{Sim}_3(D_i,D_j)$；

else $W_{ij}=\min\{\mathrm{Sim}_2(D_i,D_j),\mathrm{Sim}_3(D_i,D_j)\}$；

end

return 相似度矩阵 $\boldsymbol{X}^*$；

预估的聚类数 K

按照谱聚类算法对频繁项集进行聚类

图 3　基于频繁项集和潜在语义的聚类算法

Fig. 3　Clustering algorithm based on frequent itemsets and latent semantics

2.3　基于 TSC-SN 的文本软聚类

基于主题簇的主题词抽取，主要从词性、词频、词的簇内支持度以及词的簇间区分度综合考虑. 有关词 t_{ki} 的主题词分值计算式如下所示：

$$\mathrm{score}(t_{ki})=W(i)\ln(f_i)S_k(i)\lg\left(\frac{|I_{\mathrm{Ker}}|}{|I_i|}\frac{|c|}{|c_i|}+1\right)$$

式中：f_i 为词 t_{ki} 在高质量文本集中出现的频率；$S_k(i)$ 为簇 C_k 中包含词 t_{ki} 的频繁项集的个数；I_{Key} 为重要频繁项集集合；$|I_i|$ $(I_i\in I_{\mathrm{Key}})$ 为包含词 t_{ki} 的频繁项集个数；$|C_i|$ 为包含词 t_{ki} 的聚类数；$|C|$ 为总聚类数；$W(i)$ 为词 t_{ki} 的词性权重.

短文本聚类可看作在主题词基础上进行信息检索，寻找出与短文本 d_i $(d_i\in D)$ 相似度较大的聚簇 C_k $(C_k\in C)$，簇与短文本相似度度量依据式(1)计算. TSC-SN 算法允许同一文本划分至多个主题. 设文本与聚簇间的相似度阈值为 λ，簇间非重叠度参数 p_{nol} 的临界值为 δ. 主题词集 T_k 与短文本 d_i 间的相似度 $\mathrm{Sim}_1(T_k,d_i)>\lambda$ 时，将文本划分至相似度大于 λ 的若干个聚簇中，实现文本与主题间一对多的对应关联. p_{nol} 的计算式如下所示：

$$p_{\mathrm{nol}}=\frac{N}{\sum_{i=1}^{K'}|C_{ij}|}$$

式中：N 为文本总数；$|C_{ij}|$ 为初始簇 C_i 经过第 j 次文本划分后所包含的文本数；K' 为主题簇个数. 基于 TSC-SN 算法的文本软聚类算法的具体步骤如下所示：

步骤 1　计算短文本 $d_i\in D$ 与簇 $C_k\in C$ 的主题词 $T_k=\{t_{k1},t_{k2},\cdots,t_{ks}\}$ 间的相似度，将短文本 d_i 划分到相似度最大的簇，即 $\mathrm{argmax}(\mathrm{Sim}_1(T_k,d_i))$.

步骤 2　降低相似度阈值 θ，$\theta\in[0,1]$，可从 1 开始逐渐下调. 选定 θ 后将 $\mathrm{Sim}_1(T_k,d_i)>\theta$ 时的文本划分至相似度大于 θ 的若干簇中.

步骤 3　计算在选定 θ 下的 p_{nol}，若 $p_{\mathrm{nol}}\leqslant\delta$，则聚类结束.

步骤 4　重复步骤 2 和步骤 3，直至 $p_{\mathrm{nol}}\leqslant\delta$.

在对主题词相关的文本进行检索时，控制短文本与主题簇之间的 θ，不断降低 θ，计算每次降低后的总体文本 p_{nol}，直到满足 $p_{\mathrm{nol}}\leqslant\delta$ 为止. 由此，既可控制总体文本重叠度，又可实现文本软划分.

3　实验分析

实验数据来源于新浪财经股吧论坛，涵盖 2015 年 5 月至 2015 年 12 月期间与七个股市热点事件相关的 64 286 条评论数据，日均股评发帖量 262 条. 该期间内国内股市行情波动较大，经历比较明显的上

涨和下跌,并且引发股民热烈讨论,有利于论坛中多样化主题和热点的挖掘.基于在线股评数据,根据知网中所蕴含的概念上下位关系,知网中的义原共构成“事件树”、“实体树”、“专有名词树”、“属性树”、“次特征树”等九棵概念树.鉴于名词与动词更能体现文本的语义内涵,赋予“实体树”和“事件树”更高的权重,分别设置为1.00和0.25.“次特征树”中“领域”分支下的义原能加强文本的主题区分度,将其权重设置为0.15.“专有名词树”主要涵盖国家名称义原,但这些词本身已是不可再分的语义单位,因此这类义原不参与概念抽取,将其权重设为0.其他概念树中所包含的概念对文本类别区分的贡献都比较小,相应权重均设置为0.1.针对义原权值$W(Z_{\mathrm{DEF},wi})$计算中所涉及的三个参数a、b、c,分别设置为1.50、5.00和0.15.经过文本预处理后所得到的关键词数为46 382,特征空间的概念数为19 075,特征空间维度缩减58.9%,有效缓解概念向量空间表示中所存在的高维度问题.

3.1 重要参数设置

3.1.1 重要频繁项集数的参数分析

为通过频繁项集过滤策略获得比较完整与冗余性低的重要频繁项集集合,特别分析最小支持度min_sup和频繁项集间的Jaccard系数最大相似度α与重要频繁项集个数的关系,分别设置α的不同取值,观测每个取值下过滤后的重要频繁项集数与最小支持度min_sup之间的变化规律,如图4所示.

由图4可知,在α的不同设置中,过滤后的频繁项集占频繁项集总数的百分比均不超过20%,有利于提高频繁项集聚类的效率.为挖掘出更多的频繁项集,这里将min_sup设置较低,由此可得到大量包含主题信息的频繁项集,再通过过滤策略得到高质量的重要频繁项集.过滤策略的方法复杂度低,不会增加过多的时间消耗.α设置越高,过滤后的重要频繁项集所占百分比越高.当α取值为0.4与0.5时,重要频繁项集的百分比相差较小;当α取值为0.6时,重要频繁项集的百分比显著增大.这主要是因为基于FP-growth算法挖掘获取的频繁项集中包含大量3-项集.当α取值为0.4或0.5时,两个3-项集中若有两个重叠项,则被过滤掉;当α取值为0.6时,两个3-项集都会被保留.这说明α取值为0.6是不合理的,会造成大量冗余频繁项集未被过滤.另外,过滤后频繁项集的比例与min_sup成反比关系,这是因为min_sup越高就会产生越多的1-项集和2-项集,这些项集几乎是其他频繁项集的子集,很容易被过滤掉,使得重要频繁项集的比例降低.

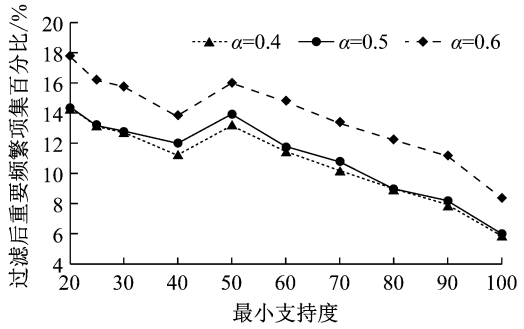

图4 过滤后频繁项集所占百分比与最小支持度的关系

Fig.4 Relationship between frequent item-sets proportion and minimum support degree after filtering

3.1.2 聚类数的参数分析

为进一步分析min_sup与α、频繁项集与频繁项集簇间最小相似度β对预估聚类数的影响,选取min_sup∈{20, 25, 30, 40, 50, 60}、α∈{0.4, 0.5, 0.6}以及β∈{0.2, 0.4, 0.6}时来预估聚类数,实验结果如表1所示.

表1 针对不同参数的预估聚类数比较

Tab.1 Comparison of predicted clustering numbers for different parameters

min_sup	频繁项集数	不同参数下预估聚类数								
		α=0.4			α=0.5			α=0.6		
		β=0.2	β=0.4	β=0.6	β=0.2	β=0.4	β=0.6	β=0.2	β=0.4	β=0.6
20	15 437	10	11	23	9	9	12	7	8	12
25	11 602	8	10	20	8	8	19	6	7	10
30	8 574	7	9	16	8	8	15	5	7	9
40	6 091	8	10	13	7	9	10	6	7	8
50	3 763	7	9	11	8	8	12	6	7	8
60	2 995	7	8	11	6	7	9	6	6	7

由表1可知,聚类数随着min_sup和α的增加而逐渐减小,主要因为min_sup增加时一些话题无法产生较长频繁项集,在预估聚类数时直接将其忽略.另外,当α增加时,新增加的频繁项集往往被分配到规模较大的前几个频繁项集簇中,而在估计聚类数时选择频繁项集累计总数占总频繁项集数80%以上的簇个数作为聚类数.因此,当更多频繁项集划入较大规模的簇中时,聚类数会减少.此外,β对预估

聚类数影响较大. 当 β 设置为 0.2 或 0.4 时,针对 α 和 min_sup 的不同设置,聚类数相近并且比较稳健. 当 β 设置为 0.6 时,原来比较相似的簇会被划分成更小的簇,聚类数也明显增多.

综合上述分析,考虑效率与准确性的平衡,设定 min_sup=25、α=0.6 以及 β=0.4.

3.2　文本软聚类性能评估

3.2.1　主题词提取

将名词、动词与形容词的权重分别设定为 1.00、0.25 和 0.15,按前文方法对主题词簇中每个词打分后,选择排序在前 τ 位的词为该簇主题词,这里设定 τ=4. 针对聚类数 K 不同设置的各事件主题词提取结果如表 2 所示.

由表 2 可知:当聚类数 K=7 时,股市暴跌这一事件分裂为两个子主题,一类讨论股市暴跌时国家是否会及时出台救市政策,另一类讨论暴跌所带来的恐慌情绪与投资者信心受挫,通过股吧论坛原文数据分析可发现,对于股市暴跌这一事件的讨论词区分度较大,一定程度上说明股市暴跌时投资者情绪波动较大,意见分歧明显;当聚类数 K=8 时,救市事件也被分裂为两个子主题,一类讨论国家出台相关救市政策及影响,另一类讨论为防止大盘崩盘央行紧急制定各种政策;当聚类数 K=6 时,这些分裂簇会消失,其他簇则几乎不变. 这说明本文所选取的聚类方法在主题抽取方面比较稳定且准确.

表 2　针对聚类数 K 不同设置的各事件主题词提取结果

Tab.2　Extraction results of event key words with different settings for cluster number K

编号	事件	主题词		
		K=6	K=7	K=8
1	大盘暴跌	暴跌、股灾、熊市、恐慌	巨震、救市、出手、见底、暴跌、股灾、恐慌、信号	沪指、救市、出手、见底、暴跌、股灾、恐慌、信号
2	投资者恐慌	行情、跌停、恐慌、暴跌	跌停、暴跌、行情、A 股	跌停、恐慌、杀跌、出逃
3	做空 A 股	做空、调查、严打、传闻	做空、杠杆、调查、打击	做空、恶意、庄家、调查
4	救市	救市、央行、护盘、发布	救市、央行、出台、护盘	救市、护盘、利好、政策、央行、稳定、紧急、崩盘
5	降准/降息	降息、双降、降准、利好	降息、降准、利好、双降	降息、利好、反弹、影响
6	蓝筹 ETF 申购	蓝筹、ETF、申购、汇金	蓝筹、ETF、1 200 亿、申购	蓝筹、ETF、申购、护盘
7	暂停 IPO	国务院、IPO、暂停、新股	国务院、IPO、暂停、A 股	国务院、IPO、暂停、A 股

注:ETF 为交易型开放式指数基金; IPO 为首次公开募股.

3.2.2　文本聚类

通过计算文本与频繁项集簇中主题词之间的相似度,将文本划分至相似度最高的主题词簇下,围绕 2015 年股市大幅下跌前后的评论数据进行文本聚类,部分聚类结果如图 5 所示.

author	title	cluster1	distance1	cluster2	distance2
鸿牛2014	降准降息预期骤然加大!	5	0.77015	5	0.77015
业水蕊	股市再现恐慌暴跌，投资者信心不足	1	0.76382	2	0.63817
财富哥	沪指为何再次暴跌百点 需高度警惕主力动向	1	0.92782	1	0.92782
金眼论市	暂停IPO救起血崩中的A股？	7	0.91967	7	0.91967
绿荫梧桐	救市出败笔再引被动抛压逾六成个股跌停	4	0.82513	4	0.82513
李清远	恐慌情绪释放的四大表象 耐心等待机会!	2	0.79955	2	0.79955
用户55721!	降息将准双降维稳市场　利好版块券商、地产、汽车	5	0.91637	5	0.91637
展源	期待更有力的实则性救市政策	4	0.80992	4	0.80992
游资会馆	央行双降超预期 下周直奔3500	5	0.76412	5	0.76412
羊竟瞻卧觉	四股力量恶意做空A股 救市能否见效？	3	0.81747	4	0.60808
股海光头52	究竟谁在做空A股	3	0.82976	3	0.82976
左洋平	为国家队救市改变策略叫好	4	0.81263	4	0.81263
以天为剑	再论暴跌的前因后果及救市的必要性	4	0.76795	1	0.62795
大侠猎庄1(	恐慌性杀跌现羊群效应，调整或一步到位!	2	0.79791	2	0.79791
趋势无敌手	沪指暴跌超4% 中字头成暴跌重灾区	1	0.95534	1	0.95534
K线狙击手	恐慌中酝酿月线级别的新一轮反攻	2	0.73591	2	0.73591
海娃8888	探讨股市为什么暴跌200多点：散户没有股指期货对冲工具	1	0.83651	1	0.83651
狼眼财经8(	股民关注重点：提防大C浪 周三或将开启暴跌模式	1	0.92381	1	0.92381
刘晋城微博	全面救市成功六大利好激活人气	4	0.81263	4	0.81263
赵楠-个股2	传汇金大举赎回上证180ETF，短期继续利空于大蓝筹	6	0.80373	6	0.80373
金瑜宝	六大利空致沪指跌逾4% 暴跌后抄底还是卖出	1	0.91983	1	0.91983
上海龙世斌	暂停IPO可能会引发一批实力股民资金出逃，需慎之又慎!	7	0.80373	7	0.80373
龍哥论市	海外做空A股，政府能见招拆招吗？	3	0.82914	3	0.82914

图 5　基于频繁项集的短文本聚类部分结果

Fig.5　Results of short text clustering based on frequent item-sets

首先根据 $\mathrm{argmax}(\mathrm{Sim}_1(T_k, d_i))$ 将短文本 d_i 划分至相似度最大的簇中，此时 $p_{\mathrm{nol}}=1$，对应图5中第一次聚类结果；若设定 $\delta=0.8$，则降低 θ（$\theta\in[0, 1]$）. 选定 $\theta=0.6$，将符合 $\mathrm{Sim}_1(T_k, d_i)>0.6$ 的文本划分至相似度大于0.6的若干簇中，对应图5中第二次聚类结果，此时再次计算 $\theta=0.6$ 下的 p_{nol}（0.916）. 因 $p_{\mathrm{nol}}>\delta$，需重复调低 θ 值，将文本进行软划分之后再计算 p_{nol}. 随着 θ 值增大，p_{nol} 呈现缓慢上升趋势，这是因为聚类文本长度较短，大部分仅表达一个主题，少数文本与多个主题簇之间相似度均较高. 有关 p_{nol} 随文本与 θ 变化情况，如图6所示.

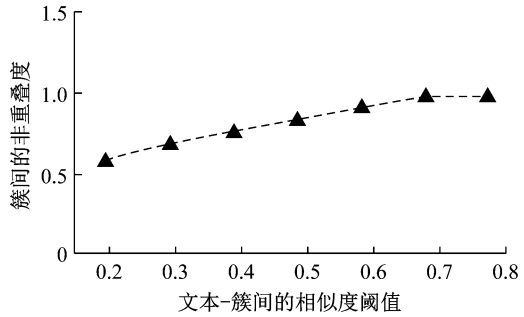

图6 聚类簇间非重叠度与文本-簇相似度阈值关系

Fig.6 Relationship between non-overlapping degree of clusters and text-cluster similarity threshold

通过重复对 θ 进行取值与文本软划分，发现将 θ 取值为0.4时所计算出的 $p_{\mathrm{nol}}=0.762$，满足终止条件 $p_{\mathrm{nol}}<\delta=0.8$.

3.3 整体性能对比分析

针对频繁项集聚类效果的评估，选择聚类后簇内平均紧密度 c 与簇间平均分离度 s 作为比较对象，计算式如下所示：

$$c=\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i=1}^{|C_k|}l(I_i,\boldsymbol{u}_k)$$

$$s=\frac{1}{K(K-1)}\sum_{i,j}l(\boldsymbol{u}_i,\boldsymbol{u}_j)$$

式中：$\boldsymbol{u}_k$ 为聚类簇 C_k 的中心向量；$\boldsymbol{u}_i$ 与 $\boldsymbol{u}_j$ 分别为不同聚类簇的中心向量；l 为欧氏距离. 高质量的聚类算法应具有低簇内紧密度和高簇间分离度. 整体性能评估采用涵盖准确率、召回率及 F 值，F 值为准确率与召回率的加权平均. 考虑到当 $\alpha\in\{0.4, 0.6\}$ 与 $\beta\in\{0.2, 0.4\}$ 时，所估计的聚类数集中分布在{6, 7, 8}，因此将聚类数 K 值设置为6、7、8.

3.3.1 频繁项集聚类性能对比分析

为验证基于知网获取概念向量空间 TSC-SN 算法的性能，选取基于关键词向量空间的 V_SC 谱聚类算法、V_K-means 算法、V_TSC-SN 算法进行比较. 因四种聚类算法并非都在欧氏空间进行聚类，无法直接比较算法的簇内平均紧密度 c 与簇间平均分离度 s，因而选择比值 c/s 作为评价指标. 四种算法的参数设置均相同，对比结果如表3所示.

表3 四种聚类算法的性能对比

Tab.3 Comparison of performance among four clustering algorithms

K	各算法 c/s 值			
	TSC-SN	V_SC	V_K-means	V_TSC-SN
6	0.107	0.158	1.416	0.147
7	0.079	0.112	1.209	0.105
8	0.081	0.094	1.031	0.093

由表3可知，针对不同聚类数，TSC-SN 算法和 V_TSC-SN 算法的 c/s 值小于 V_SC 与 V_K-means 算法，相比于基于欧氏空间的距离度量法，TSC-SN 算法的频繁项集聚类效果更优. TSC-SN 算法的 c/s 值也小于 V_TSC-SN 算法，说明基于知网获取概念向量空间的聚类结果优于基于关键词向量空间的聚类结果，验证了本文算法的有效性.

3.3.2 主题发现性能对比分析

为评估本文算法所获取的主题类别效果，计算出相应的最大 F 值，如表4所示.

表4 不同事件的文本聚类整体性能

Tab.4 Clustering performance with different events

K	不同事件最大 F 值							整体性能 F 值
	1	2	3	4	5	6	7	
6	0.932	0.815	0.925	0.783	0.869	0.754	0.851	0.847
7	0.931	0.874	0.926	0.797	0.866	0.756	0.853	0.857
8	0.927	0.893	0.901	0.739	0.843	0.742	0.829	0.839

由表4可知，在本文所提出的基于频繁项集和潜在语义相结合的论坛主题发现算法框架下，不同事件的最大 F 值整体上均较高. 当 K 为7时，大部分事件的最大 F 值优于 K 取6与8时的情况. 另外，K 为6与7时，不同事件的最大 F 值相差较小，因为“大盘暴跌”和“投资者恐慌”这两个主题经常同时出现，文本软划分时这两个主题簇重叠度较高.

为进一步验证本文算法在基于文本聚类的主题发现上的整体性能，选取基于关键词向量空间的 V_EM 算法、V_K-means 算法、V_TSC-SN 算法以及基于概念向量空间但未考虑潜在语义的 C_TSC-SN 算法进行比较，结果如图7所示.

由图7可知，TSC-SN 算法的整体性能最优，F 值最大. V_EM 和 V_K-means 算法的整体性能 F 值均低于其他三种算法. 这主要是因为大部分文本较短，从而造成向量空间的稀疏性，使得仅从欧氏距离度量相似度比较低效，由此得到聚类中心向量所

表达的主题不集中，聚类结果不理想. TSC-SN 算法与 V_TSC-SN 算法相比，前者略优于后者，两种算法效果优于 C_TSC-SN 算法，说明结合潜在语义进行相似度分析后所得到的主题簇更为全面.

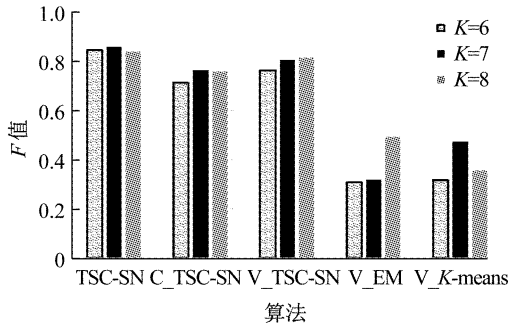

图 7　不同聚类算法 F 值对比

Fig. 7　Comparison of F-measure values among different clusting algorithms

3.3.3　时间性能对比分析

为验证 TSC-SN 算法的时间性能，选取基于概念向量空间的 C_K-means 算法、V_SC 算法进行比较，实验结果如图 8 所示.

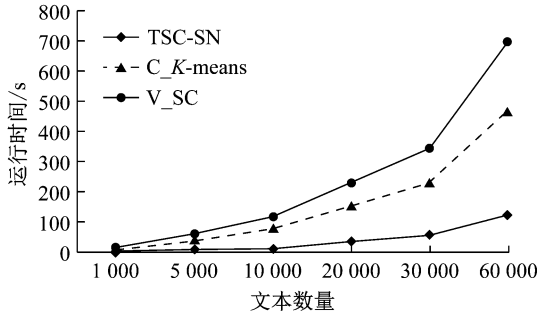

图 8　不同聚类算法时间性能对比

Fig. 8　Comparison of time performance among different clustering algorithms

由图 8 可知，TSC-SN 算法在时间性能上表现最优，并且随着文本数量的增加运行时间增加较为缓慢. 比较 TSC-SN 和 V_SC 的运行时间可见，基于知网获取概念向量空间后可有效缓解短文本高维度问题，降低算法运行时间.

4　结语

针对股评论坛主题发现问题，提出利用频繁项集和潜在语义相结合的框架从在线股评抽取主题词，使用 TSC-SN 算法基于主题词进行文本检索以实现文本软聚类，进而获取股评论坛相关文本的主题. 实验结果表明，该方法具有明显优势. 利用潜在语义信息与多层次聚类优化策略，是提高大规模短文本聚类效果以获取文本主题的有效方式. 未来研究将进一步拓展目前的整体框架与文本情感倾向性分析的融合，考虑短文本中修饰词、专有词项的词法层检测和语义层分析，充分利用短文本中的多样性信息，延伸更为深层次的主题发现与情感获取.

参考文献：

[1] 王仲远，程健鹏，王海勋，等. 短文本理解研究[J]. 计算机研究与发展，2016，53(2)：262.
WANG Zhongyuan，CHENG Jianpeng，WANG Haixun，*et al*. Short text understanding：a survey [J]. Journal of Computer Research and Development，2016，53(2)：262.

[2] ZHENG Y，MENG Z，XU C. A short-text oriented clustering method for hot topics extraction [J]. International Journal of Software Engineering & Knowledge Engineering，2015，25(3)：453.

[3] SRIRAM B，FUHRY D，DEMIR E，*et al*. Short text classification in Twitter to improve information filtering [C]// Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010). Geneva：ACM，2010：841-842.

[4] 蔡淑琴，张静，王旸，等. 基于中心化的微博热点发现方法[J]. 管理学报，2012，9(6)：864.
CAI Shuqin，ZHANG Jing，WANG Yang，*et al*. Micro-blogging hotspot discovery method based on centralization [J]. Chinese Journal of Management，2012，9(6)：864.

[5] 翟延冬，王康平，张东娜，等. 一种基于 WordNet 的短文本语义相似性算法[J]. 电子学报，2012，40(3)：617.
ZHAI Yandong，WANG Kangping，ZHANG Dongna，*et al*. An algorithm for semantic similarity of short text based on WordNet [J]. Acta Electronica Sinica，2012，40(3)：617.

[6] 杨震，王来涛，赖英旭. 基于改进语义距离的网络评论聚类研究[J]. 软件学报，2014，25(12)：2777.
YANG Zhen，WANG Laitao，LAI Yingxu. Online comment clustering based on an improved semantic distance [J]. Journal of Software，2014，25(12)：2777.

[7] BLEI D M，NG A Y，JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research Archive，2003，3：993.

[8] 徐戈，王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报，2011，34(8)：1423.
XU Ge，WANG Houfeng. The development of topic models in natural language processing [J]. Chinese Journal of Computers，2011，34(8)：1423.

[9] SHAMS M，BARAANI-DASTJERDI A. Enriched LDA (ELDA)：combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction [J]. Expert Systems with Applications，2017，80：136.

[10] KIM Y，SHIM K. TWILITE：a recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation [J]. Information Systems，2014，42(3)：59.

[11] ZHANG P，GU H，GARTRELL M，*et al*. Group-based latent

Dirichlet allocation (Group-LDA): effective audience detection for books in online social media [J]. Knowledge-Based Systems, 2016, 105:134.

[12] 李扬，孔雯婧，谢邦昌. 基于主题模型的半监督网络文本情感分类研究[J]. 数理统计与管理，2016(6)：961.
LI Yang, KONG Wenjing, XIE Bangchang. Study on semi-supervised sentiment classification of web context based on topic model [J]. Journal of Applied Statistics and Management, 2016(6): 961.

[13] ZOGHBI S, VULIC I, MOENS M F. Latent Dirichlet allocation for linking user-generated content and e-commerce data [J]. Information Sciences, 2016, 367/368: 573.

[14] 曹丽娜，唐锡晋. 基于主题模型的BBS话题演化趋势分析[J]. 管理科学学报，2014，17(11)：109.
CAO Lina, TANG Xijin. Trends of BBS topics based on dynamic topic model [J]. Journal of Management Sciences in China, 2014, 17(11): 109.

[15] LASZLO M, MUKHEJEE S. A genetic algorithm that exchanges neighboring centers for k-means clustering [J]. Pattern Recognition Letters, 2007, 28(6): 2359.

[16] SUN Y, ZHU Q, CHEN Z. An iterative initial-points refinement algorithm for categorical data clustering [J]. Pattern Recognition Letters, 2002, 23(7): 875.

[17] 彭敏，黄佳佳，朱佳晖，等. 基于频繁项集的海量短文本聚类与主题抽取[J]. 计算机研究与发展，2015，52(9)：1941.
PENG Min, HUANG Jiajia, ZHU Jiahui, *et al*. Mass of short texts clustering and topic extraction based on frequent item-sets [J]. Journal of Computer Research and Development, 2015, 52(9): 1941.

[18] CHEN C L, TSENG F S C, LIANG T. Mining fuzzy frequent item-sets for hierarchical document clustering [J]. Information Processing and Management, 2010, 46(2): 193.

[19] WANG K, XU C, LIU B. Clustering transactions using large items [C]//Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM 1999). Kansas City: ACM, 1999: 483-490.

[20] ZHANG W, YOSHIDA T, TANG X. Text clustering using frequent item-sets [J]. Knowledge-Based Systems, 2010, 23: 379.

[21] SETHI K K, RAMESH D. HFIM: a Spark-based hybrid frequent itemset mining algorithm for big data processing [J]. Journal of Supercomputing, 2017, 73: 1.

[22] DJENOURI Y, COMUZZI M. Combining Apriori heuristic and bio-inspired algorithms for solving the frequent itemsets mining problem [J]. Information Sciences, 2017, 420: 1.

[23] 刘青磊，顾小丰. 基于《知网》的词语相似度算法研究[J]. 中文信息学报，2010，24(6)：31.
LIU Qinglei, GU Xiaofeng. Study on HowNet-based word similarity algorithm [J]. Journal of Chinese Information Processing, 2010, 24(6): 31.

(上接第553页)

[6] 孟忠伟，张川，李路，等. DOC对柴油机颗粒物排放影响的试验研究[J]. 内燃机工程，2017，38(2)：67.
MENG Zhongwei, ZHANG Chuan, LI Lu, *et al*. Experimental investigation on the influence of DOC on diesel engine particle emission [J]. Chinese Internal Combustion Engine Engineering, 2017, 38(2): 67.

[7] LEFORT I, TSOLAKIS A. A thermally efficient DOC configuration to improve CO and THC conversion efficiency [R]. Detroit: SAE, 2013.

[8] SOUTHWARD B W L, BASSO S, PFEIFER M. On the development of low PGM content direct soot combustion catalysts for diesel particulate filters[R]. Detroit: SAE, 2010.

[9] VÄLIHEIKKI A, KÄRKKÄINEN M, HONKANEN M, *et al*. Deactivation of Pt/SiO_2-ZrO_2 diesel oxidation catalysts by sulphur, phosphorus and their combinations [J]. Applied Catalysis B: Environmental, 2017, 218: 409.

[10] KOLLI T, KANERVA T, HUUHTANEN M, *et al*. The activity of Pt/Al_2O_3 diesel oxidation catalyst after sulphur and calcium treatments[J]. Catalysis Today, 2010, 154(3): 303.

[11] SHARMA H, MHADESHWAR A. A detailed micro kinetic model for diesel engine emissions oxidation on platinum-based diesel oxidation catalysts (DOC) [J]. Applied Catalysis B: Environmental, 2012, 127:190.

[12] CHATTERJEE D, BURKHARDT T, RAPPE T, *et al*. Numerical simulation of DOC + DPF + SCR systems: DOC influence on SCR performance[J]. SAE International Journal of Fuels & Lubricants, 2009, 1(1):440.

[13] 楼狄明，张斌，谭丕强，等. 车用柴油机氧化催化转化器仿真模拟与试验分析[J]. 内燃机，2008(4):30.
LOU Diming, ZHANG Bin, TAN Piqiang, *et al*. Simulation and experimental analysis of diesel oxidation catalyst in the automobile diesel engine [J]. Internal Combustion Engines, 2008(4):30.

[14] 向立明，谭金龙. 车用柴油机氧化催化转化器仿真分析[J]. 内燃机，2014(2):48.
XIANG Liming, TAN Jinlong. Simulation analysis of diesel oxidation catalyst in the automobile diesel engine[J]. Internal Combustion Engines, 2014(2):48.

[15] HERREROS J M, TSOLAKIS A. Reduction of low temperature engine pollutants by understanding the exhaust species interactions in a diesel oxidation catalyst [J]. Environmental Science & Technology, 2014, 48(4):2361.

[16] KATARE S R, PATTERSON J E, LAING P M. Aged DOC is a net consumer of NO_2: analyses of vehicle, engine-dynamometer and reactor data[R]. Detroit: SAE, 2007.

[17] AL-HARBI M, HAYES R, VOTSMEIER M, *et al*. Competitive NO, CO and hydrocarbon oxidation reactions over a diesel oxidation catalyst [J]. Canadian Journal of Chemical Engineering, 2012, 90(6):1527.