基于隐含语义分析的在线新闻话题发现方法

武高敏[1],张宇晨[1],韩京宇[1,2]
(1. 南京邮电大学 计算机学院,江苏 南京 210003;
2. 东南大学 计算机网络和信息集成教育部重点实验室,江苏 南京 211189)

摘 要:互联网的飞速发展和海量数据的不断增长,使得如何快速、有效地识别当前新闻热点信息成为迫切需求。在线新闻话题发现已成为当前研究热点。对于在线环境下的新闻文本特征表示,传统向量空间模型随着数据的增长向量维度不断增长,使得数据稀疏和同名异议问题愈加明显,导致文本相似度难以准确度量。使用基于特征加权的隐含语义分析将高维、稀疏的词-文档矩阵映射到隐藏的 k 维语义空间,充分挖掘词、文档之间的语义信息,以提高同主题文档间的语义相似度,克服在线环境下文本稀疏性和同名异议问题。此外,对于不断增长的大规模新闻数据,传统聚类算法存在时间复杂度过高或者输入依赖等问题,难以快速、有效地得到理想结果。基于新闻报道在时间上的顺序性和相关性,提出改进的 Single-pass 在线增量聚类算法检测话题类,并引入话题热度值的概念来筛选当前关注度较高的热点话题。实验结果表明,该方法能够有效提高话题检测的准确率,实现基于真实新闻数据集的在线话题捕捉。

关键词:话题发现;向量空间模型;隐含语义分析;文本聚类;奇异值分解
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2016)09-0001-07
doi:10.3969/j.issn.1673-629X.2016.09.001

Online News Topics Extraction Based on Latent Semantic Analysis

WU Gao-min[1],ZHANG Yu-chen[1],HAN Jing-yu[1,2]
(1. School of Computer Science and Technology,Nanjing University of Posts and Telecommunications,Nanjing 210003,China;
2. Key Laboratory of Computer Network and Information Integration of MOE,Southeast University,Nanjing 211189,China)

Abstract:With the rapid development of the Internet and the continuous increasing of massive data,how to identify the current news topic quickly and effectively is becoming an urgent demand,and online hot news topic detection has become an hot area of research. For online news stream,the degree of traditional Vector Space Model (VSM) will grow with the increasing of data,resulting in obvious problem of data sparsity and synonymy,which makes it difficult to quickly and accurately calculate the similarity of texts. The latent semantic analysis based on weighted features is used to map the sparse matrix with high-dimension of words and documents to the hidden k-dimension semantic space,making full use of the semantic information between words and documents to improve the semantic similarity between the same subject documents,overcoming the problems of text sparsity and synonymy in Internet. In addition,traditional clustering algorithm exists the problem of high time complexity and input dependency for increasing massive news data,which is difficult to get the expected result quickly and efficiently. A Single-pass online clustering algorithm is used to detect the topic clusters based on succession and correlation in time for news,and the concept of topic heat is introduced to screen the public attention of news topics. Experiment shows that the method proposed can effectively improve the accuracy of the detection of topics.

Key words:topic detection;vector space model;latent semantic analysis;text clustering;singular value decomposition

1 概 述

随着互联网技术的飞速发展,网络信息呈现爆炸式增长态势。据《第 35 次中国互联网络发展状况统计报告》,截至 2014 年 12 月,国内网页数量已达 1 899

收稿日期:2015-12-01 **修回日期**:2016-03-08 **网络出版时间**:2016-08-01
基金项目:国家自然科学基金重点项目(61003040,61100135,61302157)
作者简介:武高敏(1990-),女,硕士研究生,CCF 会员,研究方向为数据管理、知识库;韩京宇,教授,博士,CCF 会员,研究方向为数据管理、知识库。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160801.0909.070.html

亿,年增长 26.6%。其中,静态网页数量为 1 127 亿,占网页总量的 59.36%;动态网页数量为 772 亿,占网页总量的 40.64%。如何避免一些重要信息被海量数据淹没而从中快速、有效地获取当前网络中的热点,正成为新闻媒体、行政部门、企业单位等所关注的焦点。

在线话题检测[1](Online Topic Detection,OTD)是话题检测与追踪[2](Topic Detection and Tracking,TDT)的一个重要研究课题,关注于对在线环境下实时到达的新闻报道流在没有任何先验知识的情况下,从中识别出新的话题,或是识别已有话题的后续报道。目前,有关 OTD 的研究对于文本特征的表示多采用传统基于词的向量空间模型(VSM),以独立的词作为处理对象,假设词特征的统计独立性,根据每个特征词在文档集合中的分布状况赋予该词相应的权重,建立原始文档集合的词-文档矩阵。然而,在线环境下这种简单基于词的 VSM 文本特征表示法会随着数据的不断增长使得向量的维度越来越高,带来较大的时间和空间开销,同时使得数据稀疏性和同名异议问题越来越明显。OTD 研究的另一方面关注于不同场景下聚类算法的选择。传统基于划分的聚类算法和基于层次的聚类算法在离线话题检测研究中效果较为突出,但面对在线环境这两者或存在输入依赖问题,或难以很好地满足时间上的实时性需求,影响了在线话题发现的精度和效率。

考虑到新闻报道的特点,文中提出一种加权文本特征抽取方法构建 VSM 模型,并以此为基础利用隐含语义分析对原始词-文档矩阵进行主题建模,充分挖掘词、文档之间的语义信息,有效解决在线环境下传统 VSM 文本特征表示法因向量维度的不断增长导致的数据稀疏及同义词问题。此外,结合在线新闻数据流的时间性特点,同一话题相关的新闻报道往往聚集在一定的时间段内,话题存在着一定的生存周期。文中在信息采集过程中只选取当天时间范围内的新闻报道,并基于报道的时间属性进行排序,利用改进的 Single-pass 算法实现话题检测,克服经典 Single-pass 算法的顺序依赖性问题,并实现话题簇的周期更新机制。大多数的 TDT 研究均采用由 TREC 会议提供的 TDT[3] 语料,对于 OTD 研究而言,该语料无法真实、有效地反应在线环境下的新闻舆情状况。因此,文中基于实时抓取得到的真实新闻报道开展在线话题检测的研究,更具有现实意义。

2　相关工作

目前,国内外有关话题检测的研究工作主要集中在两方面:一是文本特征的表示,涉及特征的选择和权重的衡量;二是聚类算法的选择,主要考虑时空效率和结果的有效性。

对于文本特征的表示,大多数研究[4-7]以词为特征,利用 TF-IDF 模型衡量词的权重进行特征选择,从而将每个文本表示成一个基于词的特征向量,形成 VSM 模型[8]。再利用聚类技术,基于文档的 VSM 模型,将描述同一新闻事件的网页聚合到同一个类中表示话题。然而,这种以词为特征的向量表示方法对于在线环境下的大量新闻网页数据存在以下缺点:

(1)随着新数据的不断到来,向量维度将不断增长,至少上万维,而每个新闻网页的词数在 1 500 左右,使得数据稀疏性较大;

(2)基于独立词特征的 VSM 模型忽视了中文词在不同语境下的同名异议问题,抛开了文本潜在的语义信息,影响文本相似度计算的效果。

为挖掘文本潜藏的语义信息,克服传统 VSM 因高维度带来的数据稀疏问题,Deerwester[9] 等提出隐含语义分析(Latent Semantic Analysis,LSA)模型对文本进行主题建模,将文本映射到 k 维语义空间。k 维语义空间相对于传统 VSM 不仅向量维度大大减少,避免了数据稀疏性问题,而且更好地揭示了词、文档之间的语义信息。因此,LSA 在信息检索及自然语言处理领域应用广泛。Reisinger 等[10] 利用 LSA 获取词在文档集合中的语义信息,并基于此度量词之间的相似度进行词聚类。Yih 等[11-12] 基于词典对同义词、反义词加上正负极性,然后利用 LSA 消除同义词、反义词被映射到近似的词向量空间的问题。Valle-Lisboa 等[13] 利用 LSA 对文档进行语义分析,挖掘文档潜藏的语义结构,然后基于文档的语义信息对相似文档进行聚类。

在聚类算法的选择上,Gao 等[14] 利用报道内容的时间和地点信息度量文档之间的相似度,基于组平均距离的凝聚层次聚类算法对有关自然灾害的大规模新闻报道进行话题检测,在离线环境中虽然效果较好,但层次一旦确定就不能更改。对于在线话题发现,当新的文档到来时该算法必须重新计算当前整个文档集,无法满足实时话题检测的时间需求。李胜东等[15-16] 利用基于划分的 K - means 聚类算法实现话题检测,在离线文档集合的应用中效率较高,但对于在线网络话题捕捉,难以事先确定待划分类簇数目 k,初始聚类中心选择的随机性使得算法不能保证聚类结果是最优解,而且 K - means 算法本身对噪声和离群点数据较为敏感。因此,利用 K - means 算法实现在线新闻流数据的话题检测存在着一定的局限性。马雯雯等[17] 利用 LSA 对微博数据集进行主题建模,将文本映射到 k 维语义空间,然后采用层次聚类和 K-均值聚类相结合的聚类方法实现话题发现。该聚类算法的组合虽能够缓解 K - means 初始聚类中心的随机性和先验性导

致聚类结果波动的问题,但对层次聚类算法的选择依赖性较强,聚类结果的不确定性较为明显。周刚等[18]基于组合相似度计算策略,利用增量型聚类算法 Single-pass 进行微博数据的话题检测。作为典型的增量聚类算法,Single-pass 以其原理简单、计算速度快的优点常用于在线话题的发现,然而该算法易受输入顺序影响,对相同的输入文档集,聚类结果会因为输入顺序的不同而不同。税仪冬等[19]为解决 Single-pass 的顺序敏感性问题,在聚类阶段引入"代"的概念,对文档不再是一次一篇的输入,而是按批次添加,并且在每一批文档到来时先进行初步聚类,然后再将初步聚类结果与已有话题类簇进行 Single-pass 聚类,一定程度上缓解了 Single-pass 本身的缺点,但初步聚类算法的选择对于整个聚类结果的影响较大,很容易因为初步聚类的结果影响到最终的聚类效果。

3　基于隐含语义分析的在线新闻话题发现

为克服传统 VSM 文本特征表示法存在的不足,文中首先利用隐含语义分析对文本进行主题建模,然后基于新闻报道的时间属性利用改进的 Single-pass 算法实现实时新闻话题捕捉,避免传统 Single-pass 算法的顺序依赖问题。最后,基于新闻话题的群众参与度、来源渠道的多样性及相关报道数综合评判话题簇的热度值,筛选出当前关注度较高的多个话题。方法的整体流程如图 1 所示。

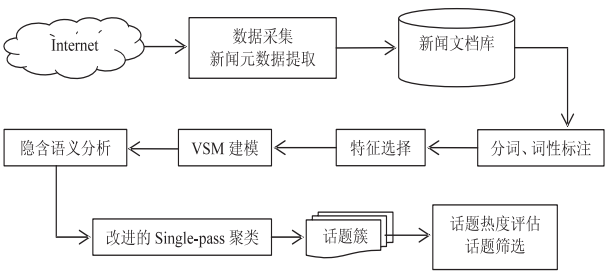

图 1　基于隐含语义分析的在线新闻话题发现方法处理流程

3.1　数据采集与预处理

(1)数据采集。

文中基于开源的 Scrapy 爬虫框架,对网易、腾讯、新浪三个新闻门户网站下国内、国际、社会、娱乐四大主题模块实现基于特定域名规则和 URL 正则过滤的在线舆情采集。其中国内、国际、社会三大模块基于同样的域名,具体域名规则见表 1,各网站新闻链接的 URL 正则表达式见表 2。

(2)数据预处理。

结合各个门户网站新闻网页结构,利用正则表达式匹配提取新闻元数据,包括:发表时间、标题、正文、新闻链接、来源网站、新闻评论链接、新闻评论数、新闻 ID、主题类型。基于抽取得到的新闻元数据,对新闻正文内容进行降噪处理,去除诸如"＊＊＊报道"、"编辑"、"＊＊＊电"等一些主题无关信息,以及微博、微信的推广信息。

表 1　各门户网站不同主题模块的域名规则

模块	网易	腾讯	新浪
社会、国际、国内	news. 163. com	news. qq. com	news. sina. com. cn
娱乐	ent. 163. com	ent. qq. com	ent. sina. com. cn

表 2　各门户网站网页 URL 正则表达式

网站	正则表达式
网易	(http://(?:news|war|ent)\.163\.com)/(\d{2})/(\d{4})/\d+/(\w+)\.html
腾讯	(.*)/a/(\d{8})/(\d+)\.htm
新浪	(http://(?:\w+\.)*(?:news\.){0,1}sina\.com\.cn)/(?:.*/){0,2}(\d{4}-\d{2}-\d{2})/(?:\d{4}|doc-\w+)(\d{6,8})\.(?:s){0,1}html

利用中科院分词系统 NLPIR[20]对新闻网页正文内容进行分词,基于停用词表去掉其中的停用词并提取实词。停用词表从哈工大停用词表和百度停用词表中整理而得,共包括 2 316 个停用词,实体词根据分词词性仅提取名词和动词。

3.2　文本特征化表示

(1)VSM 建模。

TF-IDF 模型是特征权重衡量的常用方法,考虑到一个关键词代表文档的能力不仅受词频、文档频率的影响,还有可能受位置信息等多种因素的影响。对于新闻报道,标题往往是对核心内容的简要概括。因此,标题中关键词的重要程度不容忽视。文中将基于关键词的位置信息实现特征加权的 TF-IDF 值计算。

定义 1:假设 D 为新闻文档集合,D 中包含的特征词数为 m,文档数为 n,$m \times n$ 的矩阵 $\boldsymbol{A}$ 记为文档集 D 的词-文档矩阵,$\boldsymbol{A}$ 中每一行代表一个词特征,每一列代表一个文档特征。

定义 2:根据定义 1,假设 $\text{weight\_}\{i,j\}$ 为矩阵 $\boldsymbol{A}$ 中 A_{ij} 的值,表示词 i 在文档 j 中出现的统计度量:

$$\text{weight\_}\{i,j\} = \text{tf\_}\{i,j\} * \log2(\text{ND}/(\text{df\_}\{i\}+1))$$

其中,$\text{tf\_}\{i,j\}$ 表示第 i 个词在文档 j 中出现的频次;ND 表示文档集合 D 中的文档数;$\text{df\_}\{i\}$ 表示特征词 i 的文档频度,即包含特征词 i 的文档数。

基于定义 2 得到的是特征词的初始特征权重,此时并未考虑关键词是否曾在新闻标题中出现,因此需要基于关键词的位置信息对其权重进行修正。

定义 3:考虑特征词是否在标题中,若在则对初始 TF-IDF 值乘上加权因子 ε ($\varepsilon>1$) 提高其权重,否则不予修正。特征加权后的 TF-IDF 值计算方法为:

$$\text{weight\_}(i,j)=\begin{cases}\text{tf\_}\{i,j\}*\log2(\text{ND}/(\text{df\_}\{i\}+1))\\ \text{tf\_}\{i,j\}*\log2(\text{ND}/(\text{df\_}\{i\}+1))*\varepsilon\end{cases}$$

实验过程中,文中设定 ε 的值为 3,用以扩大标题中关键词的权重。

(2)隐含语义分析。

LSA 模型假设在随机的词组下,隐藏着语义结构,利用奇异值分解[21](Singular Value Decomposition,SVD)可将原始高维向量空间模型投影到低维正交矩阵,从而获取隐藏的概念空间。与传统 VSM 假设词特征统计独立不同的是,该模型假设文本中词语之间具有紧密联系,利用统计学上相关的词的统一性获取隐藏的概念空间,连接词义相近的词和文档,来缓解传统 VSM 中同义词、多义词的问题,提高查询与文档间的语义相似度。

LSA 的一般过程可描述如下:

定义 4:假设 $\boldsymbol{B}$ 表示一个 $m\times n$ 的矩阵,则 SVD 的过程即将 $\boldsymbol{B}$ 分解为:

$$\boldsymbol{B}_{m\times n}=\boldsymbol{U}_{m\times r}\boldsymbol{\Sigma}_{r\times r}\boldsymbol{V}^{\mathrm{T}}_{r\times n}$$

其中,r 为 $\boldsymbol{B}$ 的秩,$r\leqslant\min(m,n)$;$\boldsymbol{\Sigma}=\text{diag}(\sigma_1,\sigma_2,\cdots,\sigma_r)$,$\sigma_i>0$,$\sigma_1,\sigma_2,\cdots,\sigma_r$ 按降序排列,称为奇异化因子。

假设 k 表示 $\boldsymbol{\Sigma}_{r\times r}$ 前 k 个最大的奇异化因子,则:$\boldsymbol{B}_k=\boldsymbol{U}_{m\times k}\boldsymbol{\Sigma}_{k\times k}\boldsymbol{V}^{\mathrm{T}}_{k\times n}\sim\boldsymbol{B}$,$\boldsymbol{B}_k$ 为 $\boldsymbol{B}$ 的近似矩阵。

根据定义 4,SVD 分解过程的时间复杂度为 $O(\min\{mn^2,nm^2\})$,当数据量增大到一定程度,SVD 的时间消耗对于在线环境的话题检测,显然无法满足实时性需求。因此,文中将采用 Brand[22] 等提出的增量式 SVD 算法。该算法在新的文档到来时,不再对新的数据集进行分解降维,而是将原始 SVD 的分解过程表示为 $\boldsymbol{U}_{m\times r}\boldsymbol{U}'_{r\times r}\boldsymbol{\Sigma}'_{r\times r}\boldsymbol{V}^{\mathrm{T}}_{r\times r}\boldsymbol{V}^{\mathrm{T}}_{r\times n}$,然后利用 $\boldsymbol{U}_{m\times r}\boldsymbol{U}'_{r\times r}$、$\boldsymbol{V}^{\mathrm{T}}_{r\times r}\boldsymbol{V}^{\mathrm{T}}_{r\times n}$ 来分别更新左右子空间 $\boldsymbol{U}$、$\boldsymbol{V}$ 以获取新的矩阵表示。该算法能够将传统 SVD 的时间复杂度降低到 $O(mnr)$,其中 $r\leqslant\min(m,n)^{1/2}$。

对于 k 的选择,Deerwester 最初建议取 50~350 之间。实际的应用中也发现,若 k 值取得太小,会导致多个不相关文档被误认为属于同一个主题,增加主题内部噪声;若 k 值取得太大,一方面对于一个文档数较多的主题可能会被划分为多个子主题,另一方面对于计算的开销会变大。文中经多次实验后选取 k 值为 200。

3.3 Single-pass 增量聚类

Single-pass 算法对文档输入顺序较为敏感,但考虑到在线新闻流数据的时间特性:对同一话题来自多方面的相关报道往往呈现出时间上的关联性和承接性。因此,若基于时间顺序对报道进行排序再聚类,单遍扫描的情况下即避免了顺序依赖问题。因此在进行聚类之前,将对新采集的文档数据按时间进行排序并只截取当天时间范围内的新闻报道作为研究对象,以较早的时间值为小,较新的时间值为大,升序排列。

考虑到新闻话题的不断更新,每天都会有新的话题产生,随着时间的流逝一些旧的话题也渐渐被淡忘。若是对所有的话题簇都在内存中永久保留,不仅有可能干扰聚类结果的精度,也会付出更多不必要的计算开销和存储开销。因此,需要对话题类簇进行有效的舍取。一般而言,新闻话题的热度持续时间在 7 天左右,若 7 天内某个话题的相关报道数很低,基本上可认为该话题已无法再得到广泛关注,可以忽视。文中以天为单位每天进行一次话题聚类,为避免低关注度话题的一直存在所带来的时间和空间的开销,对每个话题簇维护一个周期计数 T(初始值为 7)和周期内文档计数 P(初始值为 0),每经过一次聚类即将 T 值减 1,话题簇内每新到来一篇文档即将 P 值加 1。当某个话题簇的 T 值小于 1 时,考察该话题内新加入的相关文档总数 P 是否低于既定阈值 t_{hot},若是则将该话题簇备份到本地数据库并从当前类簇集合中去掉,否则将 T 值重新设置为 7,P 值设置为 0。

假设某个话题类簇的结构化形式为 Topic=<类簇质心 c,周期计数值 T,周期内文档计数值 P>,则改进的 Single-pass 算法具体过程如下:

输入:新的一批文档 $D=\{d_1,d_2,\cdots,d_n\}$,类簇列表 clusters,聚类阈值 t_c,类簇淘汰阈值 t_{hot};

输出:话题类簇 Topics。

过程:

```
begin:
//将所有 T 值为 0 而 P 值小于 t_hot 的话题簇过滤掉
//若 T 值为 0 而 P 值不小于 t_hot,则重新设置该话题簇,T=7,P=0
//其他情况,将 T 值减 1
if clusters is not null:
clusters=update_clusters(clusters, t_hot)
D=Sort_D_by_time(D) //将文档按时间顺序排列
for doc in D:
if clusters is null:
//以 doc 初始化为第一个类中心,并设置 T、P 的初始值
clusters.add(doc,7,0)
doc_lables.add (0)//记录文档对应的类索引
else:
//初始化记录 doc 与每个类簇相似度的列表
sims_list=list()
//计算 doc 与每个类簇的相似度
for cluster in clusters:
sim=calculate_similarity(doc,cluster.c)
sims_list.add(sim)
//获取最大相似度及其对应的类簇索引
```

```
max_sim,index=get_max_sim_index(sim_list)
if max_sim> t_c :
doc_lables. add(index)
//更新类簇中心
update_cluster_center(index,doc)
//更新周期内文档计数值 P
update_ P (index,++ P )
else:
doc_lables. add (len(clusters))
//以 doc 作为新的类簇,并初始化类中心
clusters. add(doc,7,0)
//根据 doc_lables 的分布划分 D 中文档到不同的类簇
Topics=Distribute_docs(doc_lables)
end
```

对一个有 m 篇文档的数据集,若用 Single-pass 聚类最终得到 k 个类簇,则算法的时间复杂度为 $O(mk)$,而利用层次聚类将达到 $O(m^3)$ 的时间复杂度,可见 Single-pass 算法在时间上有很大的优越性。另一方面 Single-pass 也避免了 K - means 算法输入依赖的问题。对于类簇淘汰阈值 t_{hot} 的设置,经过实验过程中的比较分析,最终发现当 t_{hot} 值至少为 200 时,可认为类簇还存在着潜在的关注度。而聚类阈值 t_c 的设置一般情况下取 0.28 较为合理。

3.4　话题热度评估与筛选

聚类结果将产生当前新闻流中的话题簇,簇中包含的话题往往较多,而公众所关注的是当前最新、最热的话题。对于热门话题往往群众参与度较高,新闻媒体也会争相报道,因此,文中以报道来源数、群众参与度、话题簇内报道数作为影响话题热度的主要因素综合量化话题热度。由于各新闻门户网站都提供了新闻评论功能,群众参与度可通过新闻评论数直接反应。

首先,基于采集得到的新闻元数据统计每个话题簇的总体参与人数及报道来源数。然后,结合簇内文档总数综合评估类簇的话题热度,并将热度值进行降序排列,取前 topN 个作为热点话题簇。

定义 5:假设 Topics 为聚类产生的话题簇,对于每个话题簇,其热度定义为:

$$\begin{aligned}\text{Hot\_Topic}(i) = {} & \alpha * \text{Doc\_Count}_i + \beta * \\ & \text{Join\_Count}_i + \gamma * \\ & \text{Source\_Count}_i (\alpha + \beta + \gamma = 1, \\ & \text{且 } \alpha,\beta,\gamma \text{ 均大于 } 0)\end{aligned}$$

其中,$\text{Join\_Count}_i = \sum j\_\text{count}$,$\text{Hot\_Topic}(i)$ 表示第 i 个类簇的话题热度,对类簇中某一新闻文档,j_count 表示评论的参与人数,Join_Count_i 即为话题总的参与度;Doc_Count_i 表示话题类中相关的文档数;Source_Count_i 为话题内新闻文档来源总数,不包括重复值;α,β,γ 分别为加权因子。

4　实　验

4.1　数据描述

为验证文中所述话题检测方法的有效性和准确性,将数据集分为两部分。一部分选取搜狗文本分类语料库中的部分数据,主要来源于 Sohu 新闻网站,共 9 个话题,17 910 篇新闻报道,用以对文中方法的性能进行评估。另一部分,选取基于文中采集规则实时采集得到的(在“2015-01-04 ~ 2015-01-15”时间段内)的 2 413 篇新闻报道,通过文中方法实现在线话题捕捉,验证其有效性。

对于话题热度的评估,考虑到群众参与度和报道来源数是话题热度最为直观的反映,而群众参与度值往往较报道数和新闻来源数要高出至少一个数量级。因此,为使得大部分话题热度值处于相近的数量级,利于客观比较,实现热点话题的加权评估,文中在多次实验后最终设置 α,β,γ 为 80,0.4,100。实际过程中,根据应用场景的不同,最佳的加权系数可能不一样。

4.2　实验结果分析

4.2.1　话题检测方法评测

(1)评测指标。

以准确率(Precision, P)、召回率(Recall, R)及 F 值(F-measure, F)作为评价指标。表 3 展现了聚类可能出现的几种情况。

表 3　聚类文档分布情况

文档数	该类相关文档数	该类不相关文档数
检测到的	A	B
未检测到的	C	D

各度量值公式为:

$$P = \frac{A}{A + B} \tag{1}$$

$$R = \frac{A}{A + C} \tag{2}$$

$$F = \frac{2P * R}{P + R} \tag{3}$$

根据上述定义,假设参照话题类 $\text{Topic\_}i$ 所含文档数为 n_i,聚类所得话题类 $\text{Test\_}j$ 所含文档数为 n_j,则定义关于话题类 $\text{Topic\_}i$ 的准确率为:

$$P(\text{Topic\_}i) = \max P(\text{Topic\_}i, \text{Test\_}j), \quad j = 1,2\cdots \tag{4}$$

$$P(\text{Topic\_}i, \text{Test\_}j) = \frac{n_{ij}}{n_j} \tag{5}$$

其中,n_{ij} 为参照类与检测类共有的文档数。

召回率采用相同的方法定义,即:

$$R(\text{Topic\_}i) = \max R(\text{Topic\_}i, \text{Test\_}j), \quad j = 1,2\cdots \tag{6}$$

$$R(\text{Topic\_}i, \text{Test\_}j) = \frac{n_{ij}}{n_i} \tag{7}$$

(2)实验结果及分析。

首先,基于搜狗的文本分类语料对传统 VSM 和文中基于改进特征权重计算的 LSA 方法进行对比分析。利用式(4)所定义的准确率为评判指标,考查不同数据规模下通过同一聚类方法进行话题检测时 VSM 和 LSA 的准确度和时间消耗,见表 4。

表 4　LSA 和 VSM 效果对比

数量	LSA	VSM
2 000	(0.612,1.884)	(0.573,6.321)
4 000	(0.727,6.816)	(0.685,37.970)
8 000	(0.733,11.432)	(0.653,103.256)
10 000	(0.758,19.044)	(0.651,198.912)

表 4 中,形如"(0.612,1.884)"的数据项表示(准确率,消耗时间)。对比表 4 可发现,基于特征加权的 LSA 文本表示方法在文本聚类的整体效果上要优于传统 VSM 的文本表示方法。主要是因为基于特征加权的 LSA 方法首先充分考虑了新闻标题特征的重要性,去除了文档中虚词、助词的干扰,再者通过 SVD 分解后获取的词-文档概念空间,充分挖掘了文档集合中词、文档之间的语义信息,从而使得文档之间的相似度计算相比于传统 VSM 受影响较小。而传统 VSM 随着数据的不断增长,数据稀疏性和同名异议问题越来越明显,影响了文本相似度的计算精度,准确率在数据规模超出 4 000 之后出现降低的趋势,而且由于维度的增长其所带来的计算时间开销也增长较快。

图 2 是经过 10 次重复实验后,基于组平均距离的层次聚类(HAC)、K - means 聚类、改进 Single-pass 聚类算法的平均性能对比。竖轴表示各性能指标的百分比。

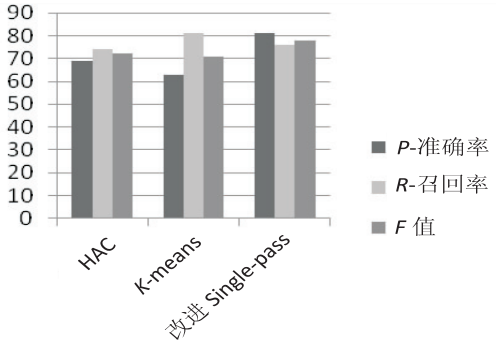

图 2　各聚类算法性能比较

从图中可以看出,文中改进的 Single-pass 算法在准确率和总体 F 值上表现突出,而此处由于是基于已分类语料,对于 K - means 算法的初始 K 值设定即设置为语料类别数 9,并从初始语料的每个类中选取一篇文档作为初始种子。但在在线环境下无法准确预计类别数 K,初始种子的选择也很难具备代表性。因此,文中方法更适用于在线环境的话题捕捉。

4.2.2　话题检测方法有效性验证

作为对文中方法有效性的验证,从新浪微博数据中心获取到"2015 年 1 月微博热门话题月报",截取其中关于"2015 年 1 月热点话题热议度、热搜度排名"(以下简称"话题参考")作为参考。对第二部分新闻网页数据进行文本预处理后,基于特征加权的 LSA 方法将原始词-文档矩阵映射到 k 维语义空间,然后使用改进的 Single-pass 算法进行话题检测,根据定义 5 对聚类结果进行话题热度评估,分别抽取的两个主题模块下前 10 个话题如表 5、表 6 所示。

表 5　news 主题模块下 top10 话题

话题热度值	话题
2 062.813	哈尔滨火灾 5 名消防员牺牲均为 90 后最小 18 岁
1 681.330	亚航失事客机寻获 5 块残骸 34 具遗体
1 265.359	讨薪致死农妇家属六问太原警方:到底犯了什么罪
1 215.887	上海踩踏事故重伤员减少至 9 人已有 20 人出院
1 201.406	南京市委书记杨卫泽被查落马揣测已传大半年
1 089.475	"布鞋院士"李小文病逝因穿布鞋讲课走红网络
1 019.715	讨薪身亡女子今日尸检家属拒绝百万元"私了"
923.124	多地社保缴费基数标准上浮 被指加重个人负担
912.409	高清图:郑州公园雕塑文字正反不一被疑贴反了
839.204	河南民警被曝借债数千万后自杀官方称非正式工

表 6　ent 主题模块下 top10 话题

话题热度值	话题
961.850	《我是歌手》收视口碑勇夺第一
856.454	《重返 20 岁》北京首映
760.168	《智取威虎山 3D》热映
716.167	奶茶妹妹与刘强东同删"爱的微博"两人疑分手
709.019	《复仇者联盟 2》发中文预告
700.248	《何以笙箫默》江苏将播 钟汉良霸道告白唐嫣
700.005	《只因单身在一起》今晚开播
650.143	英国乐队星际水手中国巡演谢幕　京沪深三地开唱
612.811	《一代宗师 3D》口碑爆棚
607.768	央视羊年春晚二审今举行 开心麻花贾玲将亮相

对比新浪微博数据中心话题月报的内容发现,文中方法能够有效捕获"何以笙箫默"、"我是歌手"、"重返 20 岁"等多个话题。同时由新浪新闻中心数据排行可发现这些话题中"上海踩踏事故"、"讨薪农妇死亡"、"亚航客机残骸打捞"都是在该段时间内引起广泛关注的热点。因此,可以证明该方法能够有效发现并筛选出当前热点话题。

5　结束语

随着移动互联网时代的到来,可穿戴设备的盛行及公众网络参与度的提高,迅速有效的在线话题发现,对于话题推送、舆情监测有着重要意义。文中在传统

VSM 的基础上,利用基于加权特征的隐含语义分析,有效克服了在线环境下传统 VSM 由于维度的不断增长使得数据稀疏和同名异议问题更加突出,从而导致文本相似度计算效果不理想的问题。同时,基于新闻报道的时间特点,提出改进的 Single-pass 算法,实现在线环境下的话题捕捉,并对聚类得到的话题簇进行加权热度评估,筛选出最终的热点话题。实验结果表明该方法是切实有效的。

在未来的工作中,将尝试通过对新闻主体特征(如人物、时间、地点等)的抽取来对新闻数据进行细化、分解,构造有关话题的知识图谱,寻求从这些特征的相互关系上来量化新闻文本之间的相似度,相信基于这些特征的组合将更有利于话题的发现。

参考文献:

[1] Allan J, Papka R, Lavrenko V. On-line new event detection and tracking[C]//Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval. [s. l.]: ACM, 1998: 37-45.

[2] Allen J. Topic detection and tracking: event-based information organization [M]. [s. l.]: Springer Science & Business Media, 2012.

[3] Connell M, Feng A, Kumaran G, et al. UMass at TDT 2004 [C]//Topic detection and tracking workshop report. [s. l.]: [s. n.], 2004.

[4] Xu R F, Peng W H, Xu J, et al. On-line new event detection using time window strategy[C]//Proc of international conference on machine learning and cybernetics. [s. l.]: IEEE, 2011: 1932-1937.

[5] Li C, Sun A, Datta A. Twevent: segment-based event detection from tweets[C]//Proceedings of the 21st ACM international conference on information and knowledge management. [s. l.]: ACM, 2012: 155-164.

[6] Abdelhaq H, Sengstock C, Gertz M. Eventweet: online localized event detection from twitter[J]. Proceedings of the VLDB Endowment, 2013, 6(12): 1326-1329.

[7] Zhang K, Zi J, Wu L G. New event detection based on indexing-tree and named entity[C]//Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval. [s. l.]: ACM, 2007: 215-222.

[8] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[9] Deerwester S C, Dumais S T, Landauer T K, et al. Indexing by latent semantic analysis[J]. JASIS, 1990, 41(6): 391-407.

[10] Reisinger J, Mooney R J. Multi-prototype vector-space models of word meaning [C]//Human language technologies: the conference of the north American chapter of the association for computational linguistics. [s. l.]: [s. n.], 2010: 109-117.

[11] Yih W, Zweig G, Platt J C. Polarity inducing latent semantic analysis[C]//Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. [s. l.]: Association for Computational Linguistics, 2012: 1212-1222.

[12] Yih W, Toutanova K, Platt J C, et al. Learning discriminative projections for text similarity measures[C]//Proceedings of the fifteen conference on computational natural language learning. Portland, Oregon, USA: [s. n.], 2011: 247-256.

[13] Valle-Lisboa J C, Mizraji E. The uncovering of hidden structures by latent semantic analysis[J]. Information Sciences, 2007, 177(19): 4122-4147.

[14] Gao N, Gao L, He Y, et al. Topic detection based on group average hierarchical clustering[C]//Proc of 2013 international conference on advanced cloud and big data. [s. l.]: IEEE Computer Society, 2013: 88-92.

[15] 李胜东,吕学强,施水才,等. 基于话题检测的自适应增量 K-means 算法[J]. 中文信息学报,2014,28(6):190-193.

[16] 柴　松. 基于 K-means 的网络话题自动检测技术研究[D]. 青岛:中国石油大学(华东),2011.

[17] 马雯雯,魏文晗,邓一贵. 基于隐含语义分析的微博话题发现方法[J]. 计算机工程与应用,2014,50(1):96-100.

[18] 周　刚,邹鸿程,熊小兵,等. MB-SinglePass:基于组合相似度的微博话题检测[J]. 计算机科学,2012,39(10):198-202.

[19] 税仪冬,瞿有利,黄厚宽. 周期分类和 Single-Pass 聚类相结合的话题识别与跟踪方法[J]. 北京交通大学学报:自然科学版,2009,33(5):85-89.

[20] 刘　群,张华平,俞鸿魁,等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展,2004,41(8):1421-1429.

[21] Menon A K, Elkan C. Fast algorithms for approximating the singular value decomposition [J]. ACM Transactions on Knowledge Discovery from Data, 2011, 5(2): 161-171.

[22] Brand M. Fast low-rank modifications of the thin singular value decomposition [J]. Linear Algebra & Its Applications, 2006, 415(1): 20-30.