情报学报　2020 年 6 月　第 39 卷　第 6 期
Journal of the China Society for Scientific and Technical Information, Jun. 2020, 39(6): 630-639

DOI: 10.3772/j.issn.1000-0135.2020.06.007

基于Gaussian LDA的在线评论主题挖掘研究

国显达，那日萨，高　欢，杨心怡

（大连理工大学系统工程研究所，大连　116024）

摘　要　针对现有主题挖掘方法生成的主题分布稀疏、语义不连贯，并导致可应用性差等不足之处，提出了一种基于Gaussian LDA的在线评论主题挖掘方法。首先，通过word2vec训练得到在线评论的词向量，并基于Gaussian LDA模型获取在线评论的主题分布；然后，通过主题分布来计算评论的相似度矩阵并应用AP聚类算法实现在线评论聚类，通过分析聚类结果实现主题发现；最后，利用TextRank算法提取各主题的关键句子生成主题摘要，以完成对主题的描述。该方法可有效缓解消费者在线评论信息过载问题，通过淘宝、京东、豆瓣等平台7种不同类型产品的评论数据的实验计算证明了方法的有效性及现实应用价值。

关键词　在线评论；主题挖掘；Gaussian LDA主题模型；AP聚类；TextRank

Topic Mining of Online Reviews Based on Gaussian Latent Dirichlet Allocation

Guo Xianda, Zhao Narisa, Gao Huan and Yang Xinyi

(*Institute of Systems Engineering, Dalian University of Technology, Dalian*　116024)

Abstract: This study proposes a method based on Gaussian latent Dirichlet allocation (LDA) for online comments to overcome the limitations of the current topic mining methods, such as sparseness and semantic incoherence of generated topics, that result in a poor applicability. The word vectors of online comments are obtained by word2vec training, and the topic distribution of online comments is achieved based on the Gaussian LDA model. The topic distribution is then used to calculate the similarity matrix of comments, and the affinity propagation clustering algorithm is employed to cluster online comments. The topic discovery is realized by analyzing the clustering results. Finally, the TextRank algorithm is used to extract the key sentences of each topic to generate the topic summary so that the description of the topic can be completed. The proposed method effectively alleviates the information overload problem of consumers' online comments. The effectiveness and practical application value of the proposed method have been established through experiments and calculations performed on online product reviews from seven platforms, such as Taobao, Jingdong, and Douban.

Key words: online reviews; topic mining; Gaussian LDA; AP clustering; TextRank

1　引　言

随着电子商务的发展和大数据时代的到来，越来越多的消费者通过电子商务网站对产品进行评论，这些评论影响了消费者的购买决策和商家的销售策略，并且这些评论具有规模大和传播快的特点[1]，

收稿日期：2019-07-19

基金项目：国家自然科学基金面上项目“基于在线评论的网络消费者群体行为预测智能技术研究”（61471083）；教育部人文社会科学研究规划基金项目“基于在线评论的网络消费者群体行为机理及预测”（14YJA630044）；大连市科技创新基金项目“大连智慧城市建设中基于大数据的智能决策理论方法及支持技术研究”（2018J11CY009）。

作者简介：国显达，男，1993年生，硕士研究生，研究方向为文本挖掘与深度学习；那日萨，男，1970年生，教授，博士生导师，研究方向为文本挖掘与复杂网络，E-mail：nmgnnrs@dlut.edu.cn；高欢，女，1994年生，硕士研究生，研究方向为文本挖掘与深度学习；杨心怡，女，1997年生，本科生，研究方向为文本挖掘与复杂网络。

使得消费者在通过阅读在线评论进行决策时往往面临着信息过载的问题。面对着日益增长的海量在线评论，消费者在购买商品时需要花费大量时间和精力才能完成所有评论的阅读；特别是那些销量较高的商品，消费者阅读部分评论而做出的决策往往具有片面性，甚至会被一些排在前面的具有误导性的评论所影响而做出不正确的选择[2]。如何有效筛选出被消费者认为是有价值的目标信息，是当前在线评论处理领域研究的热点。

在线评论通常噪声影响更大、主题特征不明显，从而不能提供足够的共同信息。早期的文本处理往往使用向量空间模型（vector space model，VSM）进行文本的向量化表示，该方法没有考虑文本中深层次的语义信息和丰富的语义关联，在文本的语义联系描述方面存在不足。如何对在线评论进行数学表达，进而提取有效信息是文本聚类算法是面临的挑战之一。文本聚类技术是缓解在线评论信息负载问题的重要手段，通过将计算机语言学和聚类相结合的手段，有助于发现相似评论信息，挖掘出潜在的有效信息，进而对评论进行再组织，帮助用户找到感兴趣的评论。尽管现在聚类技术相对较成熟，将聚类应用到在线评论仍然具有挑战性。由于聚类模型的适用范围不同，各个算法在不同的数据集具有不同的特性，聚类表现出的性能也不尽相同。在线评论数据量大、评论内容参差不齐、具有未知的空间特征，所以经典的 *K*-means、DBSCAN 等聚类是否适合于文本还需进一步讨论。发现合适的聚类方法使其在理论和实践中具有较好的效果是需要解决的问题之一。

针对以上问题，本文以在线评论为核心，探索适合文本处理的表示学习方法，提出使用 Gaussian LDA（latent Dirichlet allocation）主题模型进行在线评论的特征表示；在此基础上，探索相应的聚类方法，研究各聚类算法在评论聚类的优劣势；最后在聚类模型基础上，利用 TextRank 算法计算每个句子的重要度，并生成主题摘要。

2　相关研究

2.1　主题模型

主题模型（topic modeling）是一种常见的机器学习应用，主要用于对文本进行分类。LDA 模型是比较常用的概率主题模型，可用来识别大规模文本或语料库中隐含的主题信息。在主题模型 LDA 提出以后，以其可以自动挖掘文本中的潜在主题等优势，被广泛地用于文本主题挖掘，尤其适用于大规模语料库的自动处理，并且在文本相似度处理方面具有独到的优势。Rus 等[3]讨论了基于 LDA 模型的文本相似度计算问题，成功解决了 VSM 维度高和数据稀疏的问题。刘啸剑等[4]出了一种基于图和主题模型的关键词抽取算法，首先利用 LDA 主题模型计算词与词之间的权重并构建一个带权无向词图，选择短语作为图的节点，然后再从这些短语节点中选择 Top *K* 个词作为文章的关键词。刘晓君等[5]为研究消费者在线评论的相互关系及整体演化发展，以 LDA 模型对消费者在线评论进行话题挖掘为基础，通过 Pearson 相似度确定评论间话题关系，构建了以评论为节点的复杂网络模型。Quan 等[6]则借助主题作为第三方，考察词语之间的关联性，并进一步度量两篇文本之间的相似性。de Groof 等[7]利用改进的 LDA 自动识别在线医院评论中的关键字和主题。Hu 等[8]在 LDA 模型的基础上提出了一种新的主题建模技术——Gaussian LDA，该模型将文档视为词嵌入的集合，将主题本身视为嵌入空间中的多元 Gaussian 分布，并且通过将文档表示为稠密的词向量，有效地解决了传统 LDA 模型在语义一致性问题上的欠缺。Das 等[9]在文章中验证发现 Gaussian LDA 模型的 PMI 分值较传统的 LDA 平均高出 275%，证明其能够找到更加连贯以及更有意义的主题，并且能够推断出含有许多未登录词的主题分布。

2.2　文本聚类研究

电子商务环境下的在线评论特征稀疏、描述信息能力较弱，这造成了文本的主题模糊和内容混乱，制约了在线评论的应用范围，而文本聚类可以将大量文本文档进行适当分组[10]，揭示文本内容的一致性，从而发现同一类别文本所包含的共同信息，有助于在线评论的重新组织和二次应用。中文的在线评论聚类需要针对中文文本的特点对已有的文本聚类方法进行改进和优化，以获取具有主题化并且包含更丰富信息的在线评论特征，实现更高性能的聚类算法。Frey 等[11]于 2007 年首次提出了 AP（affinity propagation）聚类算法，该算法无须在聚类初始化时定义聚类簇数，而是在迭代过程中不断搜索合适的聚类中心，自动从所有的数据点间识别聚类中心，使所有的数据点到最近的聚类中心的相似度之和最大[12]。与传统的 *K*-means 算法相比，AP 算

法不需要事先确定聚类中心的个数，且多次独立运行的聚类结果更加稳定；并且 AP 聚类中的输入数据是文本之间的相似度矩阵，避免了直接使用文本的原始特征，十分适合处理文本数据。Guan 等[13]将 AP 聚类算法应用于半监督文本聚类，并且通过定义 UFS、CFS 和 SCFS 来计算文本的相似度矩阵，得到了比传统的经典聚类算法更优的聚类结果。Rangrej 等[14]比较了几种常用文本聚类方法应用于短文本数据的性能，结果表明 AP 聚类算法优于 *K*-means、基于 SVD 和基于图的聚类方法。郭崇慧等[15]针对传统共词分析方法存在的缺陷提出了一种新的共词分析方法——GMAP 共词分析方法，即将 g 指数、互信息概念以及 AP 聚类算法融入共词分析方法中，为改进共词分析方法提供了一个新的研究思路。AP 算法聚类质量的好坏直接由文本之间的相似度决定，如果能够将在线评论转换为向量空间中的点，并通过相似度矩阵反映出在线评论间的相似关系，则 AP 算法能较好地发现评论之间的关联关系，从而实现更好的文本聚类。由此，本文提出了一种基于 Gaussian LDA 的中文在线评论 AP 聚类算法，利用 Gaussian LDA 主题模型挖掘在线评论自身的潜在主题信息，并结合内部语义信息来提高文本相似度计算的精度。

2.3 文本摘要生成研究

目前，关于缓解文本信息过载问题的研究主要集中于文本摘要生成上。评论摘要生成的研究主要基于标签，其通过一系列启发式规则或者机器学习的方法将评论按照产品属性分类，并且给予类别一个标签和数值，其中提取商品属性和商品属性对应的评价词是生成评论摘要的基础[16-17]。如 Hu 等[16]通过采用关联规则的方法用高频名词和名词词组来抽取商品属性，获得了较好的效果。莫鹏等[17]提出了一种新的基于超图的协同抽取方法，利用句子与词之间的高阶信息来生成文本摘要和关键词。在产业界，部分电子商务网站提供了基于标签的评论摘要，如淘宝。然而，该方法存在过分依赖人工规则和提取信息不全面的问题，仅依靠部分标签并不能完全展现出在线评论中蕴含的大量信息。林莉媛等[18]充分考虑评论的句子间基于情感与基于主题的联系，提出了一种基于情感的 PageRank 算法框架，用于从评论语料中抽取文本情感摘要。唐晓波等[19]提出一种以句子为粒度的微博主题挖掘方法，利用 *K*-means 聚类算法实现主题发现，然后利用改进的 LexRank 算法计算微博各主题句子的重要度值，最后组合重要度值高的句子生成微博主题摘要。何喜军等[20]利用 TextRank 模型提取技术需求关键短语，作为技术需求识别的集合。尹裴等[21]提出一种产品特征级情感分类方法，能有效识别不同领域评论中的特征观点对，并判断其情感极性。Xiao 等[22]提出一种新型方面评级预测模型，建立了从短语中的评级到情感术语的依赖关系，有效地整合评级和评论信息。尽管基于标签的评论摘要在一定程度上能够减轻消费者获取信息的负担，但这些方法是基于特征驱动方法的基础过程[23]，具有突出的局限性：这些方法仅仅提供了部分属性的评价或情感总结，其本质上还是对已有评论的高度概括，而真实的评论中蕴含着丰富的主题，其中包含的情感因素也十分复杂，这些信息对于消费者的购买决策有着重要影响[18]。

3 本文算法模型

随着电子商务的发展，如何处理在线评论信息过载，使消费者和商家可以从繁杂冗余的在线评论中获取更多有用价值的信息是目前的研究热点。Gaussian LDA 模型本身具有主题建模能力，在文本主题挖掘方面具有先天优势。基于 Gaussian LDA 模型和 AP 聚类的在线评论模型的基本思想是：首先，使用神经网络语言模型 word2vec 在大规模语料中训练得到表示在线评论语义的词向量；然后，基于 Gaussian LDA 模型对在线评论进行训练，得到在线评论的主题分布，通过主题分布来计算文本的相似度矩阵，并应用 AP 聚类算法实现在线评论聚类，将相似的评论聚成同一类别，并生成中心代表评论；最后，利用 TextRank 算法提取各产品中心代表评论的关键句子生成主题摘要，以完成对主题的描述。模型的框架如图 1 所示。

3.1 数据准备模块

数据获取：利用数据爬取工具获取天猫、京东和豆瓣产品信息，包括不同类型产品的在线评论。

数据清洗：利用爬虫工具从网上爬取的数据中会包含网址链接、可扩展标记语言（xml）、表情符号等无用信息，需将其删除，从而得到纯文本数据。

数据过滤：在线评论中有一些超短评论，不包含任何主题信息，属于噪声数据。本文设定阈值为 10[24]，将评论文本长度小于 10 的在线评论删除。

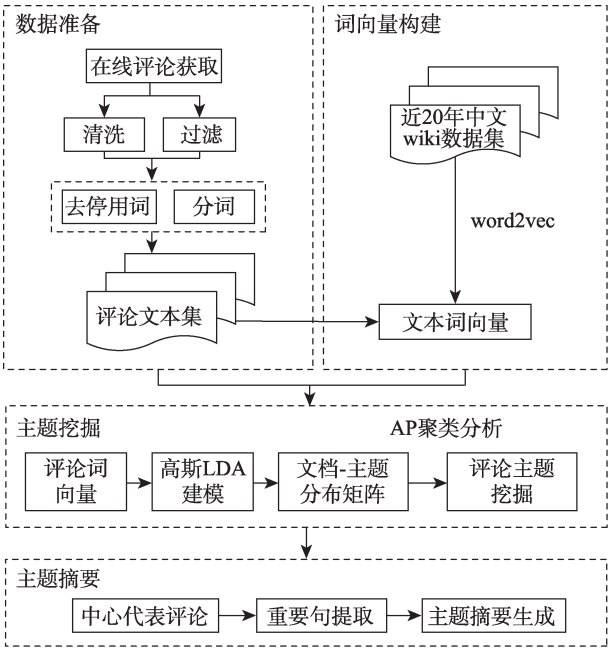

图 1　基于 Gaussian LDA 的在线评论主题挖掘算法流程

分词：利用精确模式的结巴分词工具进行分词处理。

去停用词：首先利用正则表达式匹配的方式去除标点符号，其次利用百度提供的停用词表去除停用词。

3.2　词向量构建模块

利用机器学习算法完成自然语言处理任务时，首要工作就是特征符号的数学表示，通常会用词向量来表示一个词语。word2vec 是 Google 在 2013 年开发的一款工具，其作者是 Mikolov 等[25-26]。word2vec 通过对语料库的训练，可以高效地将词表示为一个高维向量，这个高维向量在实数集上。word2vec 一共包含了两种训练模型，分别是 CBOW（continuous bag-of-words model）和 Skip-Gram[27]，其中 CBOW 模型利用词 $w(t)$ 前后各 c（这里 $c=2$）个词去预测当前词；而 Skip-Gram 模型恰好相反，它利用词 $w(t)$ 去预测它前后各 c（$c=2$）个词。模型通过利用词的上下文信息将一个词转化成一个低维实数向量，越相似的词在向量空间中越相近。本文使用的是基于 hierarchical softmax 构造的 Skip-Gram 模型，将近 30 年维基百科中文语料 data_wiki 和爬取的在线评论语料 data_set 作为 word2vec 的训练集，在尝试了不同的上下文窗口参数设置（5、10、15）和不同的词向量维度设置（50、100、150、200、250、300）后，本文最终将上下文窗口参数设置为 5[28]，词向量维度设置为 100[29]。

3.3　主题挖掘模块

3.3.1　Gaussian LDA 主题建模

经过词向量处理后的在线评论的词向量在高维空间的分布难以判断，因此采用 Gaussian LDA 模型对在线评论进行进一步的训练。

在 Gaussian LDA 中，为了获取更好的语义一致性，不再将文档看作是离散的词组成，而是由连续的词向量组成，即文档的每一个词都用一个 100 维的向量表示。每个主题 k 也不再是一个关于词项的多项分布，而是均值为 μ_k 和协方差为 $\sum_k$ 的多变量高斯分布。$\sum_k$ 服从 inverse Wishart 分布即 $\sum_k \sim W^{-1}(\psi,v)$，$\mu_k$ 服从均值为 μ、协方差为 $\sum_k / k$ 的高斯分布，即 $\mu_k \sim N(\mu, \sum_k / k)$。具体地，对于语料库中的每条在线评论，Gaussian LDA 生成在线评论的过程为

Step1　对于每个主题 $k=1$ 到 K，

Step1.1　生成主题的协方差矩阵 $\sum_k \sim W^{-1}(\psi,v)$；

Step1.2　生成主题的均值 $\mu_k \sim N(\mu, \sum_k / k)$；

Step2　对于语料库 D 中的每条评论 d，

Step2.1　从参数为 α 的 Dirichlet 分布中选择主题参数 θ_d；

Step2.2　对于评论中的每个词 $n=1,\cdots,N_d$，

Step2.2.1　从参数为 θ_d 的多项式主题分布中产生一个主题 Z_n；

Step2.2.2　生成词向量 $v_{d,n} \sim N(\mu_{z_n}, \sum_{z_n})$。

基于 Gaussian LDA 文档的生成过程，可以得到该过程的图模式（图 2），其中，$\varsigma=(\mu,\kappa,\psi,v)$，使用 Gibbs 采样推断每个词对应的主题。将 word2vec 训练得到的词 word 文本、词向量 vector 文本、train 文本输入 Gaussian LDA 模型，训练可以得到文档集的 k 个主题，同时，每条评论被表示成 k 个主题的概率向量分布生成 document_topic 文件，实现了评

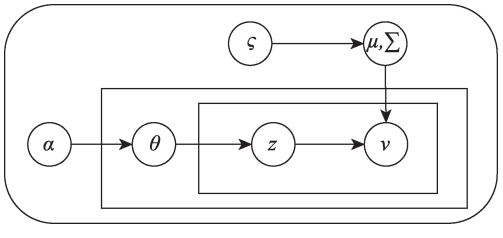

图 2　Gaussian LDA 的图模式表示

论的主题向量化表示。

3.3.2 相似度矩阵计算及AP聚类算法

本节利用Gaussian LDA模型计算得到的在线评论话题分布，确定评论间的相似度矩阵，并以此为基础进行AP聚类的训练。常用的相似度计算方法有负欧氏距离和Pearson相关系数等，本文使用负欧氏距离来计算相似度，计算公式为

$$d_{xy} = -\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{1}$$

式中，x_i、y_i表示两条不同评论在第i个主题上的概率值；d_{xy}为评论x、y之间的相似度，$-\sqrt{2} \leqslant d_{xy} \leqslant 0$，$d_{xy}$越大，则相似度越高。

再以相似度矩阵作为AP聚类的输入，将数据划分为多个不同的类别。由于AP聚类是本方法的核心内容之一，有必要对其原理进行介绍。AP聚类通过将空间中的特征分布全部转化为单一的相似度，扩大了聚类算法的应用范围，并且AP聚类并不需要指定聚类的类别个数和初始点，所有节点都当作潜在的聚类中心，有效提高了聚类结果的可靠性。

3.4 基于TextRank的中心评论摘要生成模块

基于TextRank的在线评论摘要生成模块[30]，通过选取代表评论中重要度较高的句子形成文摘，主要内容为：若评论中的某个语句与其他语句相似度较高，则该语句可被视为较重要语句。例如，给定两个句子S_i、S_j，则计算公式为

$$\text{Similarity}(S_i,S_j) = \frac{|\{t_k \vee t_k \in S_i \wedge t_k \in S_j\}|}{\log|S_i| + \log|S_j|} \tag{2}$$

式中，t_k表示候选关键词。Mihalcea等[30]曾提出关键短语的概念，即在句子中若存在相邻的关键词，则这两个关键词构成一个关键短语；若两个句子之间的相似度大于给定的阈值，就认为这两个句子语义相关并将它们连接起来，即边的权值w_{ji} = Similarity(S_i,S_j)。根据公式迭代传播权重计算每个主题下中心评论中各句子的得分，抽取重要度最高的T个句子作为候选文摘句组合重要句子，得到主题摘要，完成对主题的描述。

4 实验及结果分析

4.1 评论爬取和预处理

通过集搜客爬虫软件，对天猫购物商城、京东购物商城和豆瓣上的商品在线评论进行爬取，形成原始语料库，然后进行数据预处理：①使用结巴分词对评论进行分词处理，然后去掉多余空白、去停用词，根据预处理结果形成待用语料库；②剔除字数在10字[20]以下及无效的评论。处理后，共得到有效评论16859条，每个数据集的数据特征如表1所示。

经过数据获取和预处理之后，本文将结果作为准备好的语料库，以便于下一步的挖掘和模型训练。其中，数据集corp1主要用于算法的性能测试，而corp2和corp3则主要用于验证算法的平台和领域可移植性。

4.2 评价指标

实验使用Compactness（紧密性CP）、Separation（间隔性SP）和Davies-Bouldin Index（戴维森堡丁指数DB）[31]来评估基于Gaussian LDA的在线评论聚类效果。

1）Separation（间隔性）

以SP计算各聚类中心两两之间平均欧氏距离，

$$\text{SP} = \frac{2}{k^2 - k}\sum_{i=1}^{k}\sum_{j=i+1}^{k}\|w_i - w_j\|_2 \tag{3}$$

式中，k表示聚类结果中簇的数目；w_i、w_j表示聚类中心。SP越高，意味类间聚类距离越远。

表1 实验数据集

平台	商品领域	商品名称	数据集特征			
			评论数	最长评论字数	最短评论字数	评论平均字数
天猫（corp1）	手机类	OPPO	1803	365	15	65
		iPhone	1815	364	10	52
		小米	1840	328	10	38
	食品类	三只松鼠	1878	385	10	57
		百草味	1697	365	10	35
京东（corp2）	电脑	MacBook Air	890	340	10	37
豆瓣（corp3）	电影	战狼2	6936	188	10	40

2）Compactness（紧密性）

以 CP 测量类内每个数据点到聚类中心的平均距离，

$$\mathrm{CP}_i = \frac{1}{|\Omega_i|}\sum_{x_i \in \Omega_i} \|x_i - w_i\| \tag{4}$$

式中，Ω_i 表示集合中的聚类类别；x_i 是类中的点；w_i 表示聚类中心。作为一个紧密性的全局度量，每一个类中各点到聚类中心的平均距离为

$$\mathrm{CP} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{CP}_i \tag{5}$$

式中，k 表示聚类结果中簇的数目。理想情况下，每个聚类的成员应尽可能地接近。因此，较低的 CP 值表示聚类结果更紧凑，类内距离更小。

3）Davies-Bouldin Index（戴维森堡丁指数）

以 DB 度量每个簇类最大相似度的均值，

$$\mathrm{DB} = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{C_i + C_j}{\|w_i - w_j\|_2}\right) \tag{6}$$

式中，k 表示聚类结果中簇的数目；w_i、w_j 表示聚类中心；C_i、C_j 表示类内距离的平均距离。DB 越小，意味着类内距离越小，同时类间距离越大。

4.3　参数影响分析

利用基于 Gaussian LDA 模型的 AP 聚类算法对在线评论进行文本聚类分析，涉及 Gaussian LDA 模型主题数和 AP 聚类算法中 preference 超参数等变量的选取。为分析这些变量对于聚类结果的影响，选取数据集 corp1 中小米、iPhone、OPPO 三个手机品牌进行实验，探究以上变量对最终 AP 聚类算法簇的数目及聚类效果的影响。

4.3.1　AP 聚类算法中 preference 与簇的数目关系

使用不同的 preference，其他参数设置：word2vec 词向量维度 size=100，Gaussian LDA 迭代次数 iter=500，主题数 k=50[9]。实验结果如图 3 所示。可以看出，当 preference<−1 时，簇的数量随着 preference 的增大而缓慢增大；当 preference 的取值逐渐接近 0，簇的数量增加速度变快。为保证本文算法在不同领域数据集的稳定性，在下文实验中不妨选取 preference=−1。由此，电子商务网站可以根据实际情况灵活地调节参数来控制压缩后新的评论集评论数，通过簇中心评论将在线评论集压缩到原评论数量的 0.2%~5%，在保留原始信息的基础上有效解决信息过载问题，该方法具有重要的实际应用价值。

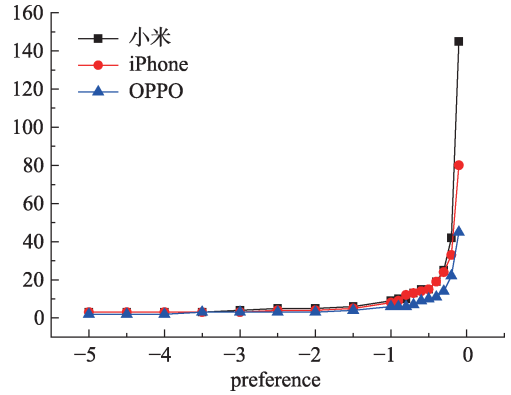

图 3　参数 preference 与簇的数目关系

4.3.2　Gaussian LDA 模型主题个数与簇的数目关系

探究主题个数对聚类结果的影响，其他参数设置：Gaussian LDA 迭代次数 iter=500，词向量维度 size=100[27]，同时设置 AP 聚类中的偏好参数 preference=−1，主题个数范围设置为 10~100，间隔为 10，实验结果如图 4 和图 5 所示。由图 4 可以看出，当主题数为 40~60 时，簇数目较为稳定；而当其小于 40 或者大于 60 时，簇数目波动较大。而图 5 显示，随着主题数的增加，衡量聚类效果的综合值 DB 值

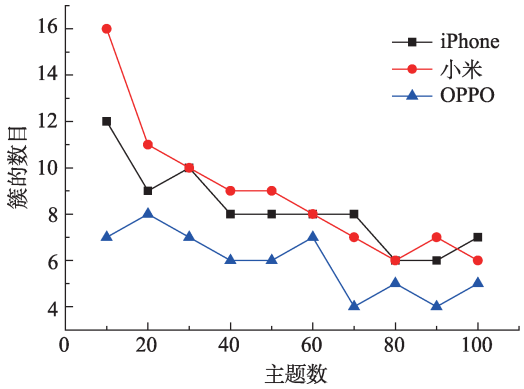

图 4　选取的主题数与簇的数目的关系

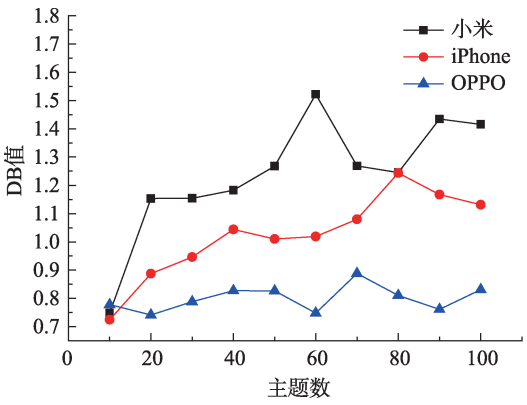

图 5　选取的主题数与评价指标的关系

波动上升，表示聚类性能逐渐降低，而在20~50范围内性能较为稳定。综合聚类稳定性及高效性，主题数的选取范围可设置为40~50。因此，不失一般性。下文实验中，设定主题数为50。

综上，该方法对于簇的数目有很好的控制，影响簇的数目的因子有偏好参数preference和主题个数k，因而只要调节好preference和k便能得到合理的簇；并且只要固定好preference和k，该方法对不同产品类型的评论都能得到比较稳定的效果。

4.4 聚类性能分析实验

首先，分析本文聚类结果的有效性和领域可移植性。为了评估本文聚类算法，利用聚类算法中最经典、聚类效果最好的*K*-means聚类算法进行对比分析。由于*K*-means算法需要事先指定聚类簇数，为了更好地进行对比分析，*K*-means算法的聚类簇数设置为AP聚类所得聚类簇数，并且*K*-means算法的输入文件同样为Gaussian LDA模型所得文档-主题向量。具体实验参数设置如下：将word2vec训练词向量的维数设置为100，Gaussian LDA模型的参数设置为：$\alpha = 1/K$，主题数K=50，迭代次数iter=500，AP聚类算法的preference=−1。选取的训练语料为corp1、corp2和corp3训练集，得到聚类结果及各项评价指标如表2所示。

表2 不同算法的聚类结果及性能比较

平台	商品领域	商品名称	数据规模	AP算法聚类结果				*K*-means算法聚类结果			
				簇数	SP	CP	DB	簇数	SP	CP	DB
天猫	手机类	OPPO	1803	6	0.162	0.175	0.826	6	0.135	0.157	1.066
		iPhone	1815	8	0.274	0.211	1.01	8	0.202	0.192	1.327
		小米	1840	9	0.223	0.232	1.268	9	0.183	0.219	1.568
	食品类	三只松鼠	1878	7	0.275	0.187	0.979	7	0.249	0.179	1.23
		百草味	1697	8	0.269	0.229	1.23	8	0.216	0.218	1.45
京东	电脑	MacBook Air	890	4	0.179	0.245	1.679	4	0.176	0.232	1.631
豆瓣	电影	战狼2	6936	25	0.255	0.22	1.253	25	0.218	0.215	1.484

由表2可以看出，AP算法的聚类性能明显高于*K*-means算法。首先，与*K*-means算法相比，AP算法的SP值高出*K*-means算法18.74%，说明AP算法的聚类结果簇与簇之间具有较高的平均间隔；其次，对于紧密性（CP），虽然*K*-means算法能产生更紧凑的聚类，但是和AP算法仅仅相差6.39%；最后，在戴维森堡丁指数DB值上，AP算法比*K*-means算法高出16.25%，说明AP算法能够产生类内距离较小，同时类间距离较大的聚类。

另外，通过对比corp1、corp2和corp3训练集的结果可以看出，本文所提方法具有较好的平台和领域可移植性。一方面体现在聚类效果综合值DB值上，在各平台和领域数据集上表现均比*K*-means算法优秀；另一方面体现在评论集压缩率上，天猫平台平均压缩率为0.42%，京东平台为0.45%，而豆瓣为0.36%。可见在不同平台上评论集压缩率相近，且随着数据规模的增大（豆瓣>天猫>京东），压缩率缓慢下降，说明本文算法能将同一数据集新增加的评论聚到现有的类别中。

其次，具体分析本文聚类算法的类别特征。以小米手机为例，分析本文聚类方法所得到的簇的特征，表3给出了评论集的类分布。由表3可以看到，本文方法既可以将评论主题谈论较频繁且语义高度相似的评论聚为一类，如簇5和簇6两簇中的评论占总评论的39.68%；同时也将主题谈论较少的评论聚为一小类，如簇2中评论数仅占总评论数的4.02%。更进一步地分析簇中心评论，限于文章篇幅，只在表4列出表3中前3条簇中心评论。由表4可以看到，对于小米手机，各条评论中簇中心评论讨论的主题各有侧重。簇中心评论1主要围绕着手机外观如“颜色”、“机身”和手机性能如“运行速度”、“指纹解锁”来评价，并且用户的主观态度是“非常喜欢”；簇2中心评论主要围绕着“屏幕”、

表3 小米手机AP聚类结果

簇的标签	簇中评论数	簇中评论占总评论的百分比/%
1	113	6.14
2	74	4.02
3	276	15.00
4	151	8.21
5	409	22.23
6	321	17.45
7	151	8.21
8	189	10.27
9	156	8.48

表4　小米手机AP聚类部分簇中心评论

簇的标签	中心评论	簇中评论数
1	双十一买的，第二天就到了，手机颜色漂亮，机身很薄，拿在手里手感很好，非常喜欢。手机用了几天才评价的，发烫的情况目前使用还没有出现过，卡顿偶尔有，运行速度很快，指纹解锁也很快，总之，这个价位买到的手机我非常满意。	113
2	手机下端有缝隙，密封手机盒内多出一张手机盒清单上名单上没有的蓝色“手机膜”（疑是拆封之后二次包装的，双十一特别送的，却在密封手机盒内，我很好奇怎么塞进去的），手机轻薄易弯曲，手机屏幕差（倾斜角度对光横条纹和波浪纹），播放视频，有时没声音，声音大时，有低噪！服务态度极差，客服不能处理时，就说：亲，稍等下。人不见了，待会又换客服，由需要重新描述下，我描述了不低于5次！	74
3	用本机现在拍不了照，小说包装吧，各方面都很精致!! 使用方面反应对于我来说够灵敏！但货到今天为止用了差不多三天，电量耗得很快，上午最多玩儿两个小时，还不是连续两个小时，到中午，差不多会掉到百分之40！剩%60用下午应该没问题！然后说个还算严重的问题，昨天晚上，也就是收到货的第二个晚上用微信看小说-：突然死机了：就是什么文字都没有呈现灰白色，3~5秒没反应，我就按关机键重新开机！然后就啥事没有，一如既往地使用！我就想哪位大神告知这是啥情况？算死机吗？然后再说拍照，照片呈现灰黑色，不怎么清晰！再说网速：大家都知道移动卡网速不咋样，但是这个手机用我的移动卡，网速还不错，这点儿我非常满意（以前的手机并没坏，就是用移动联通都时常断网，系统反应慢）才换的！然后：机身很薄，很轻，就是好滑呀，最后：我想知道是不是所有的智能机都无法录制孩子的指纹？因为我以前用的貌似也不行？我的孩子4岁，应该说从出生就有指纹，为毛录制孩子的指纹没反应？嗯，总结：裸机，一天一充电，拍照别指望有多清晰）偶尔死机，网速非常好！无线网没连接过，但以前用过米3链接无线网下载视频很快！所以应该还不错，然后无法录制孩子的指纹，看相机效果。	276

“播放声音”和“客服服务态度”来评价，用户表达了强烈的不满情绪；簇3中心评论围绕着“拍照”、“网速”和“电池耗电”等主题来评价，用户的态度为中性。本研究并未针对评论的情感极性进行分类研究，然而聚类结果却在一定程度上将情感极性不同的评论进行了划分，这是由于在Gaussian LDA模型通过word2vec来训练词向量的过程中，语义相差较大的词训练的词向量之间距离较远，进而应用AP聚类时将语义不同的评论聚为不同类别。由此可得，通过基于Gaussian LDA模型的在线评论AP聚类算法对在线评论进行数据稀疏化可以取得较为理想的效果，同一类别中评论相似，不同类别区别较大，进一步展现了同一类别中的语义联系。通过实际数据分析得到选取簇的中心评论作为新的代表评论具有合理性，为产业界的应用提供了参考建议。相比较于基于标签的评论摘要，本方法生成的新评论集保留了大量原始评论信息，若消费者对于某一评论讨论的话题感兴趣，可以进一步参考该类别中其他评论完善对该产品相关话题特性的认知。

为了能更清楚了解各主题内容，本研究利用第3.4节所介绍的TextRank算法生成主题摘要（其中阈值设为3）和关键短语，结果如表5所示。限于文章篇幅所限，这里给出部分产品的主题摘要和关键短语。至此，完成了对在线评论的主题挖掘和描

表5　不同产品在线评论主题描述

产品	评论主题摘要	关键短语
小米	给舅妈买的，很少女的粉色，拿在手里薄薄的很漂亮，舅妈非常满意，已经买过好几个小米手机送亲戚了，大家都挺喜欢的，以后还会再来；手机666小米手机很不错性价比高双11拍的1199这个价位买4+64G的手机物有所值非常喜欢小米希望未来出的产品更具创新时尚炫酷祝小米未来继续大卖；用了两天了，之前买的小米4，后来出现内存不足的问题，所以看到这个64g内存的就买了，外观很满意，够薄，玫瑰金也很漂亮，但是听歌音质不如米4，其他都不错。	运行内存；算对得起；小米手机送；小米未来；小米手机；昨天满心
OPPO	手机发货真快，物流也杠杠的，活动期间送的礼品真多，使用了几天了，指纹解锁超级速度，视频清晰，手感不错，感觉棒棒哒；特地用了一段时间才来评价的用着不错手机运行顺畅之前用的苹果6s摔坏了听朋友说这款手机挺好就拍下了总体来说好快递也很快给力；使用了一周多才来评价，拿到手机真的不错的，屏大手感好。	手机手感；手机充电速度；习惯大屏；触屏反应；物流速度；蓝牙音响
MacBook Air	开发ios必须要用mac，成功让公司给配了一个；操作系统是流畅没的说的；就是桌面不如微软的简单直观，有种电脑搬手机的套路，现在手机都可以加软件快捷方式到桌面了苹果电脑却不行，这是什么逻辑，我不懂。	要用mac；加软件快捷方式；开发ios
战狼2	我觉得有点蠢，但吴京还是认真做了这部电影的；而且凭什么说这是“要让男人看了更想做真男人的电影”；不是导演科班出身的吴京能把电影拍到这个程度，真挺不错，不过，比起这种个人英雄主义，大英雄，大勇士我是不太感冒的，相反更喜欢小人物。	会觉得；想做；大片化；去看；好莱坞大片

述。从表 4 小米手机聚类结果可大致了解小米手机评论的主题分类，表 5 则较为详细地描述了主题的具体内容。由此，证明了本文提出方法的可行性与有效性。

5 结 语

在线评论具有大数据的特点，如何让消费者快速全面地了解评论信息具有重要的实际意义。本文在对比了现有方法后，决定采用基于 Gaussian LDA 的在线评论主题挖掘方法。该方法有效地解决了传统 LDA 模型在语义一致性问题上的欠缺，能够找到更加连贯以及更有意义的主题。首先，本文以在线评论为核心，探索适合于文本处理的表示学习方法，提出使用 Gaussian LDA 主题模型进行在线评论的特征表示；其次，在此基础上，以文档主题分布矩阵为基础进行聚类分析，分析聚类结果发现待分析在线评论的潜在主题，得到每个类别的中心评论；最后，在聚类模型基础上，利用 TextRank 算法计算每个主题类别中心评论所包含句子的重要度值，组合重要度值高的句子，生成主题摘要和关键短语实现主题描述。

本文还存在一些需待深入研究的地方。一方面，本文所提聚类方法，不仅可以用于在线评论的聚类分析，还可以应用于其他问题，如主题演化分析；另一方面，本文研究方法为无监督学习，当数据含有标签或者部分有标签时，如何利用标签信息改进模型也是一个可行的研究方向。

参 考 文 献

[1] 刘洋, 廖貅武, 刘莹 . 在线评论对应用软件及平台定价策略的影响[J]. 系统工程学报, 2014, 29(4): 560-570.

[2] 王智生, 李慧颖, 孙锐 . 在线评论有用性投票的影响因素研究——基于商品类型的调节作用[J]. 管理评论, 2016, 28(7): 143-153.

[3] Rus V, Niraula N, Banjade R. Similarity measures based on latent Dirichlet allocation[C]// Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing. Heidelberg: Springer, 2013: 459-470.

[4] 刘啸剑, 谢飞, 吴信东 . 基于图和 LDA 主题模型的关键词抽取算法[J]. 情报学报, 2016, 35(6): 664-672.

[5] 刘晓君, 那日萨, 崔雪莲 . 基于隐含狄利克雷分配模型的消费者在线评论复杂网络构建及其应用[J]. 系统工程学报, 2017, 32(3): 305-312.

[6] Quan X J, Liu G, Lu Z, et al. Short text similarity based on probabilistic topics[J]. Knowledge and Information Systems, 2010, 25(3): 473-491.

[7] de Groof R, Xu H P. Automatic topic discovery of online hospital reviews using an improved LDA with variational Gibbs sampling [C]// Proceedings of the International Conference on Big Data. Boston: IEEE, 2017: 3940-3947.

[8] Hu P F, Liu W J, Jiang W, et al. Latent topic model based on Gaussian-LDA for audio retrieval[C]// Proceedings of Chinese Conference on Pattern Recognition. Heidelberg: Springer, 2012: 556-563.

[9] Das R, Zaheer M, Dyer C. Gaussian LDA for topic models with word embeddings[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 795-804.

[10] Abualigah L M, Khader A T. Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering[J]. The Journal of Supercomputing, 2017, 73(11): 4773-4795.

[11] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[12] 李一鸣, 倪丽萍, 方清华, 等 . 基于近邻传播的文本数据流聚类算法研究[J]. 计算机科学, 2016, 43(5): 223-229.

[13] Guan R C, Shi X H, Marchese M, et al. Text clustering with seeds affinity propagation[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(4): 627-637.

[14] Rangrej A, Kulkarni S, Tendulkar A V. Comparative study of clustering techniques for short text documents[C]// Proceedings of the 20th International Conference Companion on World Wide Web. New York: ACM Press, 2011: 111-112.

[15] 郭崇慧, 曹梦月 . GMAP: 一种基于 AP 聚类的共词分析方法[J]. 情报学报, 2017, 36(11): 1192-1200.

[16] Hu M Q, Liu B. Mining opinion features in customer reviews [C]// Proceedings of the 19th National Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2004: 755-760.

[17] 莫鹏, 胡珀, 黄湘冀, 等 . 基于超图的文本摘要与关键词协同抽取研究[J]. 中文信息学报, 2015, 29(6): 135-140.

[18] 林莉媛, 王中卿, 李寿山, 等 . 基于 PageRank 的中文多文档文本情感摘要[J]. 中文信息学报, 2014, 28(2): 85-90.

[19] 唐晓波, 肖璐 . 基于单句粒度的微博主题挖掘研究[J]. 情报学报, 2014, 33(6): 623-632.

[20] 何喜军, 张婷婷, 武玉英, 等 . 供需匹配视角下基于语义相似聚类的技术需求识别模型[J]. 系统工程理论与实践, 2019, 39(2): 476-485.

[21] 尹裴, 王洪伟 . 面向产品特征的中文在线评论情感分类: 以本体建模为方法[J]. 系统管理学报, 2016, 25(1): 103-114.

[22] Xiao D, Ji Y G, Li Y T, et al. Coupled matrix factorization and topic modeling for aspect mining[J]. Information Processing &

Management, 2018, 54(6): 861-873.

[23] Araque O, Corcuera-Platas I, Sánchez-Rada J F, et al. Enhancing deep learning sentiment analysis with ensemble techniques in social applications[J]. Expert Systems with Applications, 2017, 77: 236-246.

[24] 崔雪莲, 那日萨, 刘晓君 . 基于主题相似性的在线评论情感分析[J]. 系统管理学报, 2018, 27(5): 24-30.

[25] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[OL]. https://arxiv. org/pdf/1301.3781.pdf.

[26] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates, 2013: 3111-3119.

[27] 黄仁, 张卫 . 基于 word2vec 的互联网商品评论情感倾向研究[J]. 计算机科学, 2016, 43(s1): 387-389.

[28] Lauren P, Qu G Z, Zhang F, et al. Discriminant document embeddings with an extreme learning machine for classifying clinical narratives[J]. Neurocomputing, 2018, 277: 129-138.

[29] 李良强, 袁华, 叶开, 等 . 基于在线评论词向量表征的产品属性提取[J]. 系统工程学报, 2018, 33(5): 113-123.

[30] Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2004: 404-411.

[31] Fahad A, Alshatri N, Tari Z, et al. A survey of clustering algorithms for big data: Taxonomy and empirical analysis[J]. IEEE Transactions on Emerging Topics in Computing, 2014, 2(3): 267-279.

（责任编辑　王克平）