计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300172

基于 CiteSpace 的中文评论文本研究现状与趋势分析

李建兰　潘　岳　李小聪　刘子维　王天宇
合肥工业大学管理学院　合肥 230009

摘　要　自然语言处理一直是人工智能领域中的热点话题,其中基于评论的文本分析吸引了学者的注意。通过对国内关于评论文本分析的文献进行可视化分析,进而掌握该领域的研究现状和前沿发展趋势。以中国知网为数据来源,共选取 453 篇有效的核心期刊论文,使用 CiteSpace 软件绘制知识图谱并加以分析。分析结果显示:该领域的文献数量在近 15 年内整体呈上升趋势;作者之间、研究机构之间的合作关系并不紧密,尚未形成具有凝聚力的研究群体;情感分析、在线评论、深度学习是目前研究的主要热点。从初期的理论基础发展以及应用方向上的扩展,到后期在分析手段和模型上做出改进,学者们对该领域的研究逐渐深入。未来各研究者及研究机构之间的合作关系还需加强,以深度学习为代表的各类模型未来将持续发展和改善。

关键词:文本分析;情感分析;特征提取;评论;CiteSpace
中图法分类号　TP391.1

Chinese Commentary Text Research Status and Trend Analysis Based on CiteSpace

LI Jian-lan,PAN Yue,LI Xiao-cong,LIU Zi-wei and WANG Tian-yu
School of Management,Hefei University of Technology,Hefei 230009,China

Abstract　Natural language processing (NLP) has been a hot topic in the field of artificial intelligence (AI) recently,among which commentary-based text analysis has also attracts the attention of scholars. In this study,the research status and frontier development trend can be grasped through a visual analysis of the domestic literature on comment text analysis. A total of 453 valid core journal papers on the field are selected from CNKI as the data source. CiteSpace software is used to draw the knowledge map and analyze it. Analysis results show that the number of literature in this field has been on the rise in past 15 years. The cooperation among authors and among research institutions is not close,and a cohesive research group has not been formed. Sentiment analysis,online comments and deep learning are the main research hotspots at present. From the initial development of theoretical basis and the expansion of application direction,to the improvement of analysis methods and models in the later stage,scholars have gradually deepened the research in this field. In the future,the cooperative relationship between researchers and research institutions needs to be strengthened,and various models which are represented by deep learning will continue to develop and improve in the future.

Keywords　Text analysis,Sentiment analysis,Feature extraction,Comment,CiteSpace

Web2.0 推动了用户生成内容(User Generated Content,UGC)的发展,其中更是包括蕴含着用户情感的各类评论,例如电商网站中的购物体验评论、影音平台上的观看评论、微博中为舆论发声评论等,这些评论中蕴含着大量有价值的信息。文献[1]指出商品在线评论的情感倾向会对消费者购买决策产生影响,文献[2]则研究了评论对产品销售业绩的影响。由此可以看出,对评论进行有效的分析可以挖掘出更有价值的潜在信息。评论文本分析是自然语言处理中的一个研究方向,随着深度学习模型被广泛地应用到实际问题处理当中以及对用户生成内容中潜在价值的追寻,评论文本分析也在吸引着学者们的注意。Jin 等[3]就手机的负面评论进行分析,来帮助手机产品的设计人员获取用户需求,并且对之后的新产品设计决策提供有价值的意见。Zhang 等[4]从情感分析的角度入手,提出一种划分模型来解决在微博的热点话题下,关于用户群体分类的问题。

在上述背景下,更需要对评论文本分析领域的现状利用深度分析,为之后的研究工作提出启示性建议。基于此,本文以评论文本分析领域入手,收集国内在该领域相关的文献,再利用 CiteSpace 软件进行处理,绘制与研究机构、作者、关键词有关的知识图谱,通过对处理结果进行分析,得出目前国内在评论文本分析领域的研究热点以及未来发展趋势。

1　研究现状

1.1　数据来源与研究方法

1.1.1　数据来源

本次研究选取中国知网平台为切入点,来充分了解国内在评论文本分析领域的研究现状。首先需要在平台中的高级检索选取期刊检索,检索条件为:主题是“文本分析”或“情感

基金项目:合肥工业大学省级大学生创新项目(S202010359165)
This work was supported by the Provincial College Student Innovation Project of Hefei University of Technology(S202010359165).
通信作者:李建兰(1577251726@qq.com)

分析”或“特征提取”并“评论”。由于人工检索时发现在2006年之前，国内相关的文献较少，因此将时间跨度选取为2006－2020年。期刊来源中选中核心期刊检索条件。之后进一步通过人工筛选，剔除与研究主题无关的文献，最终共得到有效文献453篇。

1.1.2 研究方法

本次研究所选用分析工具是由美国Drexel大学的陈超美教授开发的可视化分析应用软件CiteSpace，该软件适用于多元、分时、动态的复杂网络分析[5]。由于CiteSpace的使用者能够通过输入与研究主题相关的文献来探测该研究主题的现状及演进情况[6]，CiteSpace目前已经被广泛应用于探测研究现状并洞察研究的变化趋势。

1.2 年发文量

通过分析检索所得文献数据的发布时间，我们得到自2006年以来国内在评论文本分析领域的发文量分布图，如图1所示。2006－2012年为起步阶段，发文量较少，且大多数为综述，主要学习国外相关研究工作。2012－2017年为稳定发展阶段，并在2017年达到峰值66篇。2017－2020年阶段内发文量先增后减，整体呈增长趋势，并于2020年达到整体最高值105篇。

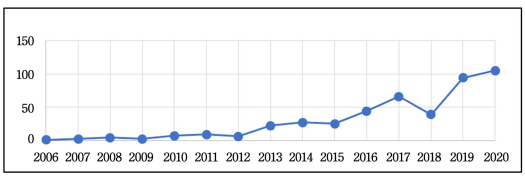

图1 2006－2020年发文量折线图

Fig. 1 Line chart of the number of papers published in 2006－2020

1.3 机构共线

图2所示为对选取文献的所属机构进行共现分析之后得出的图谱。通过分析数据以及图谱可以发现，目前在该领域的研究中，中山大学咨询管理学院、同济大学经济与管理学院发表文献数量依次为20篇和13篇。由1.4小节图3也可以发现，这两所研究机构以自己为核心形成了一个较为简单的合作网络，且研究机构大多是各成一派，并未形成具有凝聚力的研究群体。

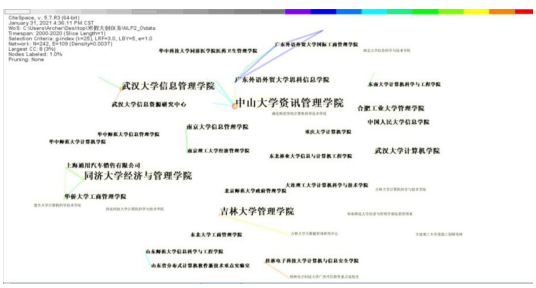

图2 研究机构共现图谱

Fig. 2 Research institution co-occurrence map

1.4 作者共线

图3显示的是高产作者合作关系知识图谱。其中发表文献数量最多的两位研究学者分别是来自中山大学的徐健和来自同济大学的王洪伟，数量分别为14篇和11篇。其中发表相关文献数量在5篇以上的仅有3人，多数学者仅发表1篇相关论文，这在一定程度上说明目前国内学者在该领域的研究并不够深入和具体。此外，正如图中所示，作者的Density为0.0045，即各个学者之间的合作关系并不紧密，目前也仅有以王洪伟为核心形成的一个较为复杂的合作网络。

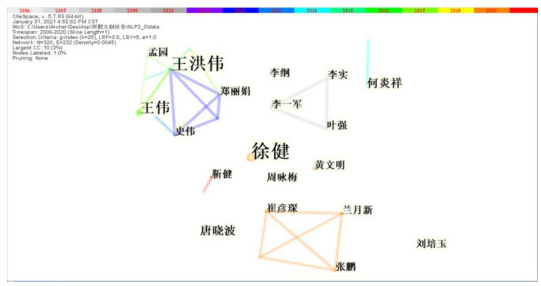

图3 作者共现图谱

Fig. 3 Authors co-occurrence map

2 研究热点

2.1 关键词共现

使用CiteSpace进行关键词共现分析后得到410个关键词，关键词网络图谱如图4所示，其中节点大小代表词频的大小，线段代表关键词之间的联系。表1列举了其中排名前10的关键词，包括关键词的词频、中心性、首次出现的时间。中心性指的是网络中经过某点并连接这两点的最短路径占这两点之间的最短路径线总数之比，中心性越高，其在网络中越重要。

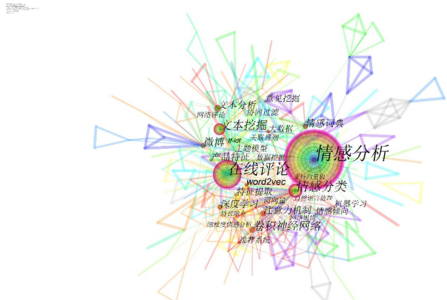

图4 关键词共现图谱

Fig. 4 Keywords co-occurrence map

表1 词频排名前10的关键词

Table 1 Top 10 keywords in terms of frequency

关键词	词频	中心性	年份
情感分析	232	0.65	2007
在线评论	90	0.35	2008
情感分类	36	0.21	2009
文本挖掘	27	0.12	2013
卷积神经网络	22	0.06	2017
微博	17	0.15	2012
深度学习	14	0.03	2017
文本分析	14	0.05	2008
产品特征	13	0.02	2009

2.1.1 研究热点比较

关键词是一篇论文重点的凝练，通过对关键词词频和中心性的统计比较分析，可以得出该领域内目前的研究热点和未来趋势。本文通过对关键词的词频进行排序，结合其中心性，可以发现情感分析、情感分类、文本挖掘是文本分析领域的研究分支；在线评论、微博则是主要分析的文本对象；深度学习等为文本分析需要使用的方法和模型。

2.1.2 研究热点脉络演进

通过对关键词首次出现的年份进行梳理，得到关于年份的关键词数量变化，如图 5 所示。计算可得 2008 年(250%)、2010 年(271%)、2013 年(174%)为 3 个变动幅度最大的年份。因此，将关键词划分为起步、摸索、稳定、发展 4 个阶段，选取其中词频排名前 4 的关键词。

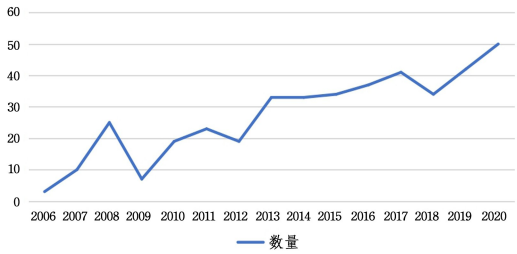

图 5 关键词数量的年份变化图

Fig. 5 Year change chart of the number of keywords

表 2 研究热点脉络演进

Table 2 Evolution of research hotspots

阶段	关键词总量	关键词	年份	词频
2006－2008	38	情感分析	2007	232
		在线评论	2008	90
		文本分析	2008	14
		中文信息处理	2007	4
2009－2010	26	情感分类	2009	36
		产品特征	2009	13
		意见挖掘	2009	10
		机器学习	2010	8
2011－2013	75	文本挖掘	2013	27
		微博	2012	17
		情感倾向	2011	9
		协同过滤	2013	8
2014－2020	271	卷积神经网络	2017	22
		深度学习	2017	14
		情感词典	2014	13
		注意力机制	2019	13

起步阶段(2006－2008 年)：由于文本分析刚刚兴起，领域相关关键词数量较少，但发展迅速，研究主题主要聚焦于情感分析、在线评论、文本分析和中文信息处理。在此阶段，文本主要以在线评论为主，通过特征标注、特征提取等将文本转化成符号化的句子，再利用基于机器学习的标准分类器对文本进行情感趋势判断[7]。

摸索阶段(2009－2010 年)：此阶段的关键词总量较低，但逐年增长，研究热点主要为情感分类、产品特征、意见挖掘、机器学习。由于机器学习的不断发展和完善，此阶段更多地基于机器学习的方法对文本中的产品特征进行提取，进而进行文本分析，且此阶段的语料库并不完善，语料标注和规范化有待提高[8]。

稳定阶段(2011－2013 年)：关键词数量呈现稳步上升趋势，该领域内热点聚焦于文本挖掘、微博、情感倾向和协同过滤。在基于评论的文本分析的背景下，出现了对微博等社交网络用户生成内容的分析，表现为对用户社交行为的特征分析和对微博的舆情分析等[9-10]。

发展阶段(2014－2020 年)：此阶段出现的关键词总量较多，且随时间变化呈逐步递增趋势，此阶段对卷积神经网络、深度学习、情感词典和注意力机制的研究较多。随着深度学习的发展，文本分析也更多地运用了神经网络等模型训练的方法。Cheng 等提出了一种结合注意力机制的多通道 CNN 和双向门控循环单元(MC-AttCNN-AttBiGRU)的神经网络模型进行文本情感倾向分析[11]。情感词典也在此阶段得到了一定的运用和发展[12]。

2.2 关键词聚类

CiteSpace 软件的关键词聚类是按照相关程度快速聚合，并按照聚类规模大小进行排序的。以本次生成聚类图为例，编号＃0－＃9 代表聚类的排序，＃0 为最大聚类，即该聚类(情感分析)所含文献量最多。前 5 位聚类词分别是情感分析、卷积神经网络、微博、在线评论以及用户评论。根据 CiteSpace 软件的统计，关键词“情感分析”出现次数最多，出现次数为 220 次；第二名是“在线评论”，出现次数为 56 次；第三名为“情感分类”，出现次数为 36 次。“卷积神经网络”与“用户评论”分别是第四名、第五名，出现次数分别是 22 次和 20 次。可以发现，关键词的次数与生成的聚类排名有较大的关联。为更直观了解关键词聚类的特点和分布，生成时间序列图，如图 6 所示。

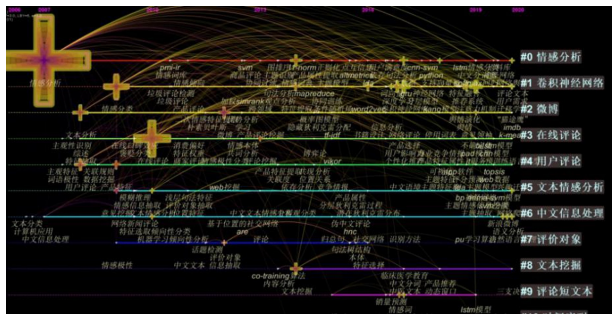

图 6 关键词聚类时间线图谱

Fig. 6 Keyword clustering timeline graph

CiteSpace 中的时间序列图是以时间为轴，展现关键词随时间变化特征的时序图，右边为每一聚类代表的关键词。图中节点的大小表示该节点关键词的重要程度，节点所在位置为该关键词出现的时间[13]。从垂直方向分析，从上到下按照聚类规模由大到小排列，不同的颜色代表不同的聚类，聚类前编号数字越大，聚类内的关键词越少；从水平方向分析，聚类内的关键词按时间分布，可以展现聚类内部随时间变化的研究情况。

聚类“情感分析”是本分析的重点方向。可以看出在发展过程中，研究热点集中在情感分析中的各种文本情感分类方法。2010－2013 年间，热点词有 PMI-IR 算法和 SVM，两者是文本情感分类发展路上的经典算法和分类器，如张清亮和徐健在 2011 年利用 PMI-IR 算法来判断情感词的情感倾向性和情感强度[14]，王刚与杨善林在 2013 年选取 SVM 作为基学习器进行网络商品评论情感分析研究[15]；2016－2019 年间，研究热点词有 CNN-SVM、LSTM 情感分类、softmax 回归等，说明文本分类的研究热点逐渐往更深层次、更精准方向发展，如 LSTM 情感分类模型是一种结合深度学习原理的准确高效分类模型。

3 研究前沿

3.1 时区图谱

图 7 是以首篇论文发表时间 2006 年为起始点，并选择时间切片为 2 绘制出来的图谱。其中圆圈代表首次出现的关键

词，圆圈大小表示与该关键词相关的文献数量，线条代表每个关键词之间的联系。时区图谱可以直观反映该评论文本分析领域内新研究热点的变化趋势。

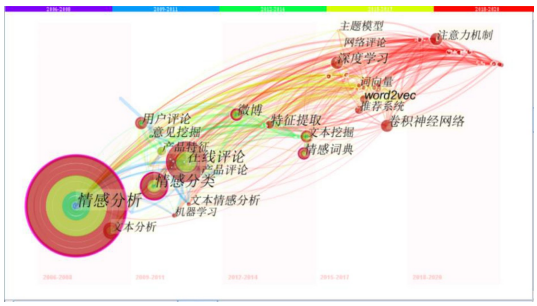

图 7 关键词时区图谱

Fig. 7 Keyword timezone graph

根据图 7 可以将时区分为 3 个阶段进行分析。

2006－2011 年：该阶段以“情感分析”为主要切入点，来分析文本所蕴含的情感方向、情感值。之后又由评论文本的分析延伸出其他方向的应用，例如是与商业销售相关的“产品评论”；通过对评论的分析还可以进一步来挖掘有价值的信息与用户意见，还可以分析出与产品相关的特征。该阶段也是国内机器学习首次出现在评论文本分析领域当中[16]。

2012－2015 年：微博在该时间阶段逐渐成为大众社交中心，其中含有大量由用户生成的短文本信息，适宜做与短文本相关的文本分析工作。因此该阶段“微博”成为学者们获得实验数据的主要平台[17]，在微博上获取了与某产品相关的微博评论之后对其进行情感分析，来挖掘用户关注的产品信息。同时，为了完善对诸如电影影评、商品评论等领域的评论文本分析工作，学者们开始建立情感词典，根据生成的情感词典来判断文本的情感倾向性。

2017－2020 年：随着技术的发展，不断涌现更多可以协助评论文本分析的算法与工具，如图 7 中所示的“卷积神经网络”算法，同时由谷歌公司研发的产生词向量的 Word2vec 模型在此阶段也在得到大量应用。此外，电商领域在该阶段也在迅猛发展，衍生出各式推荐系统和算法[18-19]，相关学者在考虑用户评论的基础上提出了新型推荐算法。目前在评论文本领域中，“注意力机制”与深度学习算法相结合的神经网络模型成为了研究热点，可以协助在大量信息中筛选出具有价值的信息。

3.2 突显词(Burst item)

突显词分析是 CiteSpace 软件提供的一项关于提取一段时间内使用频次突增或受到该领域研究学者强烈关注的专业术语，它代表这一领域的热点和研究前沿，具有较大的研究潜力，是分析发展趋势的重要依据。突显词生成过程中，因选取文献数目不多，为细化内容处理，将 γ 系数缩减为 0.5，结果发现 15 个突显词。图 8 是使用 CiteSpace 分析得出的关键词突显图谱，按照时间变化将突显词分为 3 个阶段。

理论积累期(2007－2008 年)：突显词为“中文信息处理”和“综述”。当时，国内评论文本分析处于起步阶段，中文信息处理关注度较高，引入机器学习进行情感分类取得明显成效[20]。为梳理国内外研究进度和方向，更好地开展评论文本分析工作，学者们通过总结研究成果，提出不足与展望，相关成果包括情感分析研究综述[7]以及产品评论挖掘研究综述[21]等。

快速增长期(2009－2014 年)：根据突显词“产品特征”“产品评论”等不难看出，当时的研究趋势与线上评论关系密切。资料显示，2008 年 6 月，中国网民数量达到 2.53 亿，超过 25%的普及率使得中国成为全球网民第一大国[22]。在这样的发展背景下，以淘宝商城为代表的电子商务发展最为迅猛。提供丰富商品的线上商城为人们带来便捷的同时，也带来了消费者的选择困难和商家巨大的竞争压力。为解决这两类难题，分析用户评论情感倾向、提取产品特征成为当时的研究关注重点，如挖掘中文网络客户评论的产品特征及情感倾向[23]。同时，借助互联网的发展，微博以便捷、即时信息等特点成为网民们创造内容的主要地方。通过分析微博的用户评论可以实现很多功能，比如可以进行舆情的分析与控制[9]。

研究深入期(2015－2018 年)：随着人工智能和互联网的发展，文本分析的研究也开始向更广阔和更深入的方向探索，比如可以利用用户言论情感分析识别社交网络的意见领袖[24]。另外，细粒度情感分析的研究趋势表示该领域开始注重情感分析细节，为更准确分析情感倾向打好基础。比如在基于评论数据的酒店服务质量的细粒度分析中，可以准确地对酒店评论进行极性判断，从而得到有效评价[25]。2018 年生成的“特征提取”和“深度学习”体现了评论文本分析的新的研究趋势，即利用深度学习技术进行情感分析与特征提取，说明未来的文本分析将逐步结合人工智能技术，使得计算机更准确地识别人类情感并得出结果。

从突显时间看，第二阶段的突显词持续时间最长，即此阶段的关键词受关注时间最长，是研究发展过程中的重要组成部分，为后续技术发展革新打下坚实基础。对于突显强度，第三阶段的“特征提取”突显强度最高，说明在该研究领域特征提取是重点研究方向，且开始时间为 2018 年，是大概率的研究趋势，可以预见未来学者将会在特征提取方面加大科研力度。

图 8 关键词突显图谱

Fig. 8 Keyword burst graph

结束语 未来可从以下几方面开展进一步研究。

(1)目前各个研究学者之间、研究机构之间的合作关系比较缺乏，未来还需要进一步加强研究之间的合作。

(2)文本分析中的语料库的准确、完善对实验结果十分重要，随着网络用语的流行，中文语料越发丰富，但关于此内容的研究尚少，因此构建一个标准的语料库是未来亟待解决的问题。

(3)随着人们对产品质量要求的不断提高，互联网上不仅存在大量短评论文本，也在不断产生产品测评等长评论文本供消费者参考。主要针对短评论文本的评论分析模型在面对长文本评论时会极大降低结果准确度。近些年来，深度学习与文本分析模型的结合给长文本分析带来了解决思路。可以

预见，深度学习与文本分析领域深度融合将会加快长文本分析中存在问题的攻克速度，未来对长、短评论文本的研究将会均衡全面发展。

(4)深度学习是当前在该研究领域的焦点所在，其深层结构可以学习复杂的函数模型并进行自然语言处理等任务。目前很多研究学者都在持续探索深度学习模型在处理评论文本分析时的应用。因此未来在该领域还将继续推出更具有效性的模型。

参 考 文 献

[1] YANG，ZAO L. A decision-making algorithm for online shopping using deep-learning-based opinion pairs mining and q-rung orthopair fuzzy interaction Heronian mean operators[J]. International Journal of Intelligent Systems，2020，35(5)：783-825.

[2] LI X，WU C，MAI F. The Effect of Online Reviews on Product Sales：A Joint Sentiment-Topic Analysis[J]. Information & Management，2018，56(2)：172-184.

[3] JIN J，ZHANG L X，LIU X E，et al. Use case extraction from product reviews for customer requirement analysis[J]. Information Studies：Theory & Application，2020，43(1)：104-111.

[4] ZHANG M Y，ZHU G L，ZHANG S X，et al. Grouping microblog users of trending topics based on sentiment analysis[J]. Data Analysis and Knowledge Discovery，2021，5(2)：43-49.

[5] CHEN Y，CHEN C M，LIU Z Y，et al. The methodology function of CiteSpace mapping knowledge domains[J]. Studies in Science of Science，2015，33(2)：242-253.

[6] CHEN C. CiteSpace II：Detecting and visualizing emerging trends and transient patterns in scientific literature[J]. Journal of the Association for Information Science & Technology，2006，57(3)：359-377.

[7] ZHOU L Z，HE Y K，WANG J Y. Survey on research of sentiment analysis[J]. Journalof Computer Applications，2008(11)：2725-2728.

[8] ZHANGZ Q，YE Q，LI Y J. Literature review on sentiment analysis of online product reviews[J]. Journal of Management Sciences in China，2010，13(6)：84-96.

[9] LI Y，HAN B，ZHAO J. Analysing microblogging public opinions based on short text and sentiment analysis[J]. Computer Applications and Software，2013，30(12)：240-243.

[10] LAI Q N，MA H，SONG W J，et al. Alalysis of user interactive social behavior between microblog and BBS in a university[J]，Journal on Communications，2013，34(S2)：99-106.

[11] CHEN Y，YAO L B，ZHANG G H，et al. Text sentiment orientation analysis of multi-channels CNN and BIGRU based on attention mechanism[J]. Journal of Computer Research and Development，2020，57(12)：2583-2595.

[12] YANG X，YANG Y F，JIAO W，et al. Sentiment analysis of homestay comments based on domain dictionary[J]. Science Technology and Engineering，2020，20(7)：2794-2800.

[13] CHEN C M. Science Mapping：A systematic Review of the Literature[J]. Journal of Data and Information Science，2017，2(2)：1-40.

[14] ZHANG Q L，XU J. Research on automatic extraction of web sentiment words[J]. Data Analysis and Knowledge Discovery，2011(10)：24-28.

[15] WANG G，YANG S L. Study of sentiment analysis of product reviews in internet based on RS-SVM[J]. Computer Science，2013，40(S2)：274-277.

[16] ZHOU J，LINC，LI B C. Research of sentiment classification for netnews comments by machine learning[J]. Journal of Computer Applications，2010，30(4)：1011-1014.

[17] TANG X B，WANG H Y. Research on microblogging products reviews mining model[J]. Journal of Intelligence，2013，32(2)：107-111，127.

[18] HUANG W M，WEI W C，ZHANG J，et al. Recommendation method based on attention mechanism and review text deep model[J]. Computer Engineering，2019，45(9)：176-182.

[19] CHEN J Y，WU Y Y，LIN X. Double layered recommendation algorithm based on fast density clustering with graph-based filtering& Applications[J]. Control Theory & Applications，2019，36(4)：542-552.

[20] XU J，DING Y X，WANG X L. Sentiment classification for Chinese news using machine learning methods[J]. Journal of Chinese Information Processing，2007(6)：95-100.

[21] WU X，HEZ S，HUANG Y W. Product review mining：a survey[J]. Computer Engineering and Applications，2008，44(36)：37-41.

[22] FANG X D，CHEN S. Twenty five years of internet in China[J]. Modern Communication(Journal of Communication University of China)，2019，41(4)：1-10.

[23] LIS，YE Q，LI Y J，LUO S Q. Mining product features and sentiment orientation from Chinese customer[J]. Application Research of Computers，2010，27(8)：3016-3019.

[24] ZHU M R，LIN X K，LU T，et al. Identification of opinion leaders in social networks based on sentimental analysis：evidence from an automotive forum[J]. Information Studies：Theory & Application，2017，40(6)：76-81.

[25] SUN C W，REN Z L，YANG J J，et al. Fine-grained analysis of hotel service quality based on review data[J]. Computer Applications and Software，2019，36(7)：32-38.

LI Jian-lan，born in 2000，undergraduate. Her main research interests include electronic commerce and so on.