DOI: 10.13873/J.1000-9787(2016)12-0041-03

细粒度情感分析的酒店评论研究

李　鸣[1,2]，吴　波[2]，宋　阳[3]，朱梦尧[1]，徐志广[2]，张宏俊[2]

(1.上海大学 通信与信息工程学院，上海 200444；
2.中国科学院 上海高等研究院，上海 201210；3.西安航天恒星科技实业(集团)公司，陕西 西安 710061)

摘　要：酒店在线评论细粒度挖掘具有重要研究意义。以酒店在线评论具体特征属性和情感分类为研究目标，应用 Apriori 算法和情感词典匹配算法，对重庆雾都宾馆在线评论数据深入挖掘，挖掘出用户最关注的酒店十大特征和满意度结果，进一步挖掘出商务出差等五种不同出游类型人最关注的酒店五大特征和满意度结果。这种方法不仅能对酒店领域评论进行分析，同样能够应用于其他领域。

关键词：酒店在线评论；特征挖掘；情感分析；细粒度；情感词典匹配

中图分类号：TP391　　文献标识码：A　　文章编号：1000-9787(2016)12-0041-03

Research on hotel reviews based on fine-grained sentiment analysis

LI Ming[1,2], WU Bo[2], SONG Yang[3], ZHU Meng-yao[1], XU Zhi-guang[2], ZHANG Hong-jun[2]

(1. School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China; 2. Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China; 3. Xi'an Space Star Technology Group Co Ltd, Xi'an 710061, China)

Abstract: Fine-grained mining of hotel online reviews are of great importance. Specific feature and emotional attributes of hotel online reviews can be taken as research targets, using Apriori algorithm and semantic lexicon matching algorithm, online reviews data of Chongqing Wu Du Hotel are mined, ten features that most users concerned and satisfaction results of the hotel can be inferred and five features of the hotel that five different kinds of travellers such as bussiness man most concerned together with corresponding degree of satisfaction results can also be mined in further exploration. This method can be applied in other fields.

Key words: hotel online reviews; feature mining; sentiment analysis; fine-grained; semantic lexicon matching

0　引　言

随着电子商务的快速发展，越来越多的人在网络上预订酒店并对入住体验进行在线评论。这些评论不仅有利于潜在的酒店消费者参考，也有利于商家有针对性地改善服务质量。然而，酒店评论信息量庞大冗杂，给于消费者和商家查找有用的信息带来了极大的麻烦，如何方便快捷地挖掘出评论中有价值的信息逐渐成为研究热点。情感分析能从评论中获取用户的喜怒哀乐，了解用户对酒店的喜好程度。

传统的情感分析主要采用两类方法，基于情感词典的方法和基于机器学习的方法。2002 年，Turney P D[1] 提出了基于种子词汇发现情感词的方法。Pang B 等人[2] 采用了贝叶斯、最大熵、支持向量机(SVM)等机器学习的方法来构造分类器，并对这几种方法进行了对比。Kobayashi N 等人[3] 构建了一个模式库，收录了 8 种命中率比较高且较准的模式用来提取评价主体、评价方面和评价之间的关系。Marrese-Taylor E 等人[4] 考虑到用户对不同的产品发表的评论不同，找出旅游领域的特征，构造出更准确的自然语言处理模型用于旅游领域的挖掘。

然而，前面基于篇章、句子级别的粗粒度情感分析由于没有考虑情感所针对的具体对象，无法满足用户了解酒店各个特征属性的需求。李杰等人[5] 对特征提取的研究进行了全面的概括，文献[6,7]着重对酒店细粒度的情感分析进行研究：通过关联规则方法识别出评价对象特征词、情感词以及情感修饰词，并找出他们之间的关系，计算出相应的情感值，构建相关领域的属性词表和情感词表。这些方法在英文领域取得了不错的成果，但是在中文语言下的适应性不是很理想。

收稿日期：2016-03-02

本文在前人研究的基础上，将 Apriori 关联规则算法应用于中文酒店评论领域，并结合酒店领域情感词典做分类，最终实现了更为准确的评论挖掘。通过对重庆雾都宾馆的评论数据进行属性特征挖掘，实现细粒度属性分类，挖掘出用户最关注的酒店十大特征及满意度结果，进一步挖掘出商务出差等五种不同出游类型人最关注的酒店五大特征及满意度结果。这些结果对潜在的酒店用户具有重要的参考价值，同时对于商家有针对性地改善服务质量有积极作用。

1 算法框架

图 1 为本文的算法框架图。特征挖掘模块挖掘出用户关注的酒店特征，并通过查找合并同义词进行特征过滤。观点句识别与情感分类模块根据挖掘的特征集识别出观点句，并根据用户出游类型特征将识别的观点句用情感词典匹配方法进行情感极性分类。

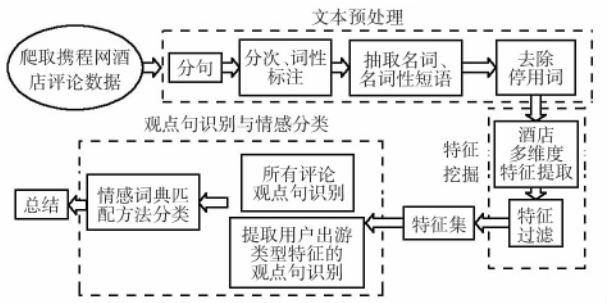

图 1 算法框架图

Fig 1 Algorithm frame

2 关键算法

2.1 Apriori 算法

本文特征挖掘模块采用了 Apriori 算法，Apriori 算法是挖掘布尔关联规则频繁项集的算法。在这个算法中，所有支持度大于最小支持度的项集称为频繁项集，简称频集。利用频繁项集性质的先验知识，通过逐层搜索的迭代方法，即将 k 项集用于探察 $k+1$ 项集，来穷尽数据集中的所有频繁项集。先找到频繁 1 项集集合 L_1，然后用 L_1 找到频繁 2 项集集合 L_2，接着用 L_2 找 L_3，直到找不到频繁 k 项集，找每个 L_k 需要一次数据库扫描。在本文中，特征挖掘模块定义最小支持度为 0.6 %，只要是在评论句子集中出现的次数大于等于 3 次，都提取出来作为候选特征集，最终经实验调优为 6 %，选出了 23 个频繁特征集。

2.2 情感词典匹配技术

情感词典匹配技术的关键是情感词典的构建和匹配算法设计。由于常用的情感词典很少包含酒店领域的专用词汇，为了提高分类准确率，需将这些专用词汇添加到情感词典中构建酒店领域情感词典。

2.2.1 构建酒店领域情感词典

本文构建的酒店领域情感词典包括：基础情感词典、酒店领域情感词典、网络情感术语词典、否定词典和程度副词词典。

基础情感词典由正面基础情感词典和负面基础情感词典组成。将 HowNet 中的正面情感词、评价词和中文情感词汇本体库中极性为“1”的词合并去重，并去掉情感倾向不显著的词条组成正面基础情感词典；将 HowNet 中的负面情感词、评价词和中文情感词汇本体库中极性为“2”的词合并去重，并去掉情感倾向不显著的词条组成负面基础情感词典。最终形成的基础情感词典含 5 821 个正面情感词，10 186 个负面情感词。

构建酒店领域专用情感词典采用了 Turney 等的点互信息法，思想是依据目标词和基准词间的点互信息，确立两词关联，预测目标词的情感分。采用 SO-PMI 算法，计算目标词与基准词的正负面点互信息之差，差值大于 0 为正面情感词，反之为负面情感词。其中，P_{set} 和 N_{set} 分别是正面和负面基准词的集合，公式如下

$$SO\text{-}PMI(Word)=\sum_{P_{word}\in P_{set}}PMI(Word,P_{word})-\sum_{N_{word}\in N_{set}}PMI(Word,N_{word}) \quad (1)$$

该实验的基准词选取方法如下：从携程网上采集了 30 万条评论数据，初始评论文本经预处理，提取形容词、副词为候选词，遍历基础情感词典库做对比，去掉和基础情感词典库重复的词，按词频由大到小排序。依据前 30 个形容词和副词的极性，选择 5 个正面基准词，5 个负面基准词。共得到 87 个正面情感词，134 个负面情感词的酒店领域专用情感词典。

网络术语情感词典：网络专用情感术语是网络中出现的风靡一时的词语，不能被传统的基础情感词典正确的识别，但是却被广泛使用。本实验以搜狗互联网词库（SogouW）的数据为基础并人工添加一些近期广泛使用的网络情感词汇来构造网络术语情感词典。否定词典由人工收集整理的 42 个否定词构成。程度副词表达了情感的强烈程度，利用 HowNet 收集的程度级别词语，并借鉴蔺璜的方法构建程度副词词典。

2.2.2 情感词典匹配算法设计

对构建好的酒店领域情感词典词语分别赋予强度值。表 1 为酒店领域情感词典词语及其相应强度值示例。

3 仿真验证

3.1 实验内容

本实验主要有两部分内容：构建情感词典和特征挖掘。

1）构建情感词典：在携程网上爬取了重庆和西安的酒店评论共 30 万条，主要提取了评论内容、用户信息、用户评分、用户出游类型和用户出游时间等信息，将这些数据用由中国科学院计算机所编写的中文分词工具 ICTCLAS 进行分词和词性标注，构建酒店领域情感词典。

表 1　酒店领域情感词典及其强度值示例

Tab 1　Sample of semantic lexicon of hotel field and corresponding intensity values

词典	强度值	示例
正面基础情感词典	1.0	好、极好、得体、满足、羡慕、吸引、爱慕
负面基础情感词典	-1.0	不良、弊病、粗暴、下流、恶心、偷偷摸摸
正面酒店领域情感词典	1.0	方便、经济、温馨、便宜、划算、便利
负面酒店领域情感词典	-1.0	简陋、发霉、坑、单薄、潮湿、异味
正面网络术语情感词典	1.0	给力、稀饭、顶、酷毙了、屌、逼格
负面网络术语情感词典	-1.0	伤不起、坑爹、倒、奇葩、抓狂、也是醉了
否定词	-1.0	不、不是、不够、否、白白、不必、非、没
	2.0	百分之百、倍加、极其、极端、最、最为
程度副词	1.5	很、太、实在、更、更加、颇为、尤其
(根据程度不同分了四类)	1.0	较、较为、进一步、大不了、多、不太
	0.5	稍、稍微、略、略微、有点、有点儿、有些

2) 特征挖掘: 选择了重庆雾都宾馆由商务出差、情侣出游、家庭亲子、朋友出游、独自出行 5 种出游类型用户评价且评价内容丰富的数据各 100 条。对 500 条评价数据进行特征挖掘,挖掘出该宾馆的 23 项频繁特征项集,并根据频繁特征项识别出观点句子并分类。最后分别对这 5 种不同出游类型的用户评论进行分析,得到每种出游类型的人关注的酒店特征和相应的评价,并统计出结果。

3.2　实验结果与分析

对重庆雾都宾馆 500 条评论数据进行挖掘得到的酒店频繁特征项集如图 2 所示。图中不仅可以看出用户对酒店地理位置、服务、房间、交通等一般特征比较关注,还可以看出用户对该酒店提供的浴缸、衣帽间等特有服务也很有兴趣。该酒店管理者可以通过这些评论继续改进自己的特色服务,用户也可能因这些特色服务而被吸引消费。

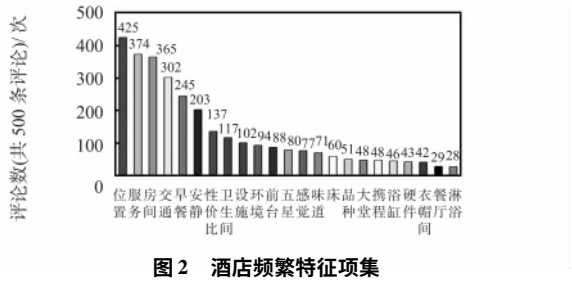

图 2　酒店频繁特征项集

Fig 2　Frequent features item sets of hotel

图 3 为挖掘重庆雾都宾馆 500 条评论数据得到的用户最关注的酒店 10 个特征和满意度。由图可知,用户最关注该酒店的房间、位置、服务、早餐等,对位置、安静和交通非常满意,对服务满意度比较低,酒店应该针对这些满意度低的方面做出相应的改善来提高酒店的核心竞争力。

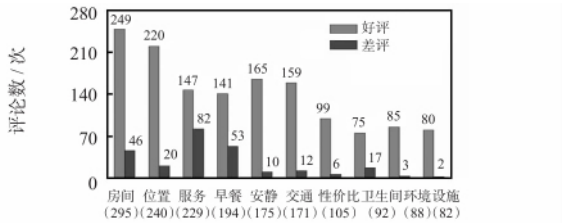

图 3　用户最关注的酒店十大特征和满意度

Fig 3　Ten features of hotel that most users concerned and satisfaction results

图 4 为商务出差、情侣出游等五种不同出游类型的人最关注的酒店五大特征和满意度。由图可知,商务出差最关注服务质量但是对服务不满意; 情侣出游对安静比较关注且非常满意等。酒店管理者可以根据不同出游类型的客户评论做出相应的改善,对客户比较满意的特色服务大力推广,客户也可以根据相应的出游类型评论来选择适合自己的酒店。

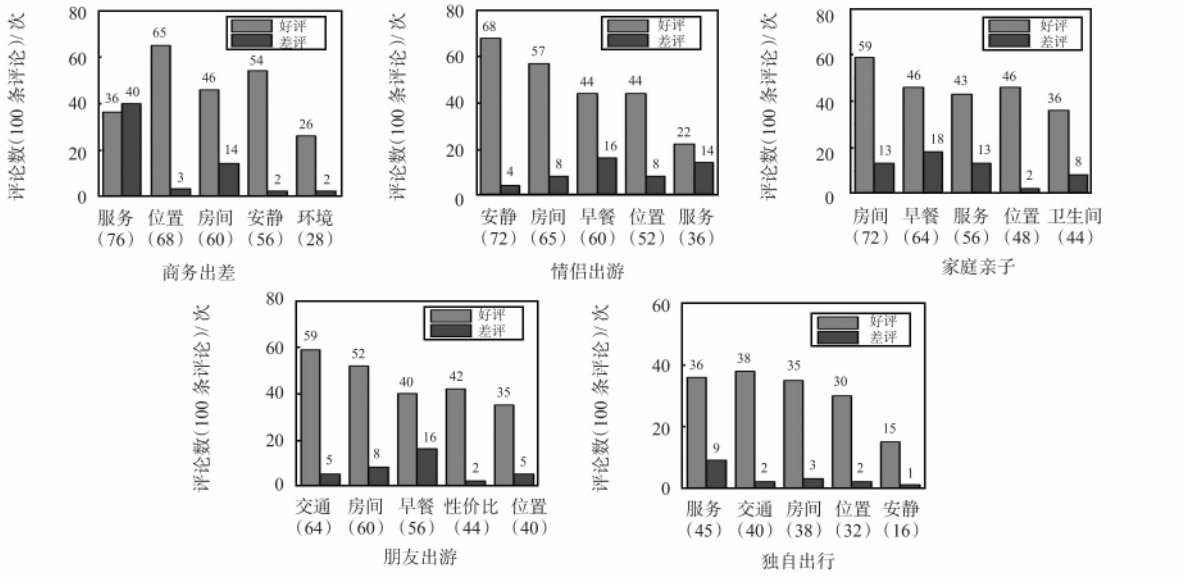

图 4　五种不同出游类型人最关注的酒店五大特征及其满意度

Fig 4　Five features of hotel that five different kinds of travellers most concerned together with degree of satisfaction

(下转第 47 页)

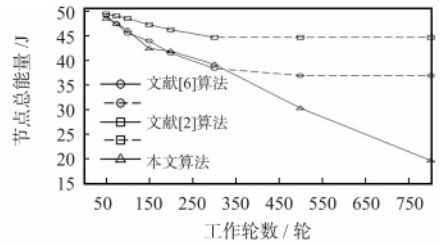

图 5　网络剩余能量随工作轮数的变化

Fig 5　Residual energy of network change with working rounds

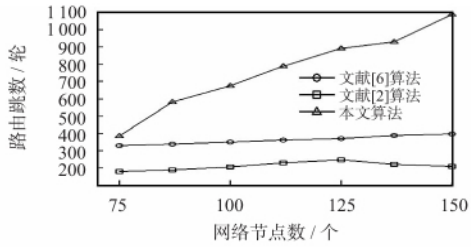

图 6　网络生存周期

Fig 6　Network lifecycle

了节点出现空洞的情况。为了减少路由跳数进一步均衡能量消耗,提出了节点能量阈值随网络总能量变化的策略。仿真结果表明: 本文提出的改进协议有效均衡了节点的能量消耗,延长了网络的生存周期。

参考文献:

[1] 刘海燕,李道全,王怀彩,等. 两类无线自组网路由协议的比较研究 [J]. 网络安全技术与应用,2009(3):15 - 17.

[2] Kapp B,Kung H T. GPSR: Greedy perimeter stateless routing for wireless networks [C] // Proc of The 6th Annual International Conference on Mobile Computing and Networking,New York: ACM,2000:243 - 254.

[3] 沈丹丹,王立华,王　宇,等. 一种基于分簇的无线传感器网络 GPSR 协议 [J]. 传感器与微系统,2015,34(12):124 - 130.

[4] 丁心体,彭新光. 一种基于 GPSR 协议的能量均衡路由 [J]. 传感器与微系统,2013,32(4):12 - 15.

[5] 王建新,赵湘宁,刘辉宇. 一种基于两跳邻居信息的贪婪地理路由算法 [J]. 电子学报,2008,36(10):193 - 199.

[6] 朱全政,杨　乐. 能量角度联合自适应路由修复新算法 [J]. 计算机应用研究,2014,31(6):1779 - 1782.

[7] 曹海英,元　元,刘志强. 基于节点剩余能量和最大角度的无线传感器网络路由算法 [J]. 传感器与微系统,2015,34(1):120 - 123.

[8] 薛　明,高德民. 无线传感器网络最大生命期聚合树路由算法 [J]. 传感器与微系统,2014,33(1):130 - 133.

[9] 吴三斌,柳　强,李成博,等. 基于能量均衡的无线传感器网络路由算法 [J]. 计算机应用研究,2012,29(4):1465 - 1469.

[10] 何杏宇,周亦敏,杨桂松,等. 无线传感器网络能量感知增强树型路由协议研究 [J]. 传感技术学报,2015,28(4):551 - 556.

[11] 陈雪娇,李向阳. WSNs 中 LEACH 协议的研究及改进 [J]. 计算机应用,2009,29(12):3241 - 3243.

作者简介:

何燕清(1990 -),男,湖北十堰人,硕士研究生,研究方向为无线传感器网络、物联网。

邓华秋,通讯作者,E-mail: hqdeng@ scut. edu. cn。

(上接第 43 页)

4　结　论

本文在对重庆和西安 30 万条酒店评论挖掘的基础上,构建了酒店领域情感词典。以重庆雾都宾馆的评论数据为例,挖掘出用户最关注的酒店十大特征及满意度结果,进一步挖掘出商务出差等五种不同出游类型人最关注的酒店五大特征及满意度结果。这些结果表明细粒度情感分析具有巨大价值: 一方面,酒店管理者不仅可以了解用户对酒店具体特征的满意度,还可以了解不同类型用户对酒店的需求,更能有针对性地改善服务; 另一方面,帮助用户了解酒店各个特征优劣,从而帮助用户更加明智的做出决策。

参考文献:

[1] Turney P D,Littman M L. Measuring praise and criticism: Inference of semantic orientation from association [J]. ACM Transactions on Information Systems(TOIS),2003,21(4):315 - 346.

[2] Pang B,Lee L,Vaithyanathan S. Thumbs up: Sentiment classification using machine learning techniques [C] // Proceedings of Association for Computational Linguistics Conference on Empirical Methods in Natural Language Processing,ACL'02,2002:79 - 86.

[3] Kobayashi N,Inui K,Matsumoto Y,et al. Collecting evaluative expressions for opinion extraction [M] // Berlin Heidelberg: Springer 2005:596 - 605.

[4] Marrese-Taylor E,Velásquez J D,Bravo-Marquez F. A novel deterministic approach for aspect-based opinion mining in tourism products reviews [J]. Expert Systems with Applications,2014,41(17):7764 - 7775.

[5] 李　杰,周　萍. 语音情感识别中特征参数的研究进展 [J]. 传感器与微系统,2012,31(2):4 - 7.

[6] Kanayama H,Nasukawa T. Fully automatic lexicon expansion for domain-oriented sentiment analysis [C] // Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,Association for Computational Linguistics,2006:355 - 363.

[7] Hu M,Liu B. Mining and summarizing customer reviews [C] // Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining,ACM,2004:168 - 177.

作者简介:

李　鸣(1990 -),女,湖北随州人,硕士,研究方向为酒店在线评论数据的情感倾向分析。