基于特征规则的在线医疗社区用户评论观点挖掘与情感分析方法

向 菲 谢耀谈

（华中科技大学同济医学院医药卫生管理学院 武汉 430030）

〔摘要〕 针对医院评价领域缺少大规模观点－情感语料库的情况，阐述国内外在线医疗社区知识共享以及基于特征的观点挖掘与分析研究现状，采用特征规则方法，基于补充情感词典，对在线医疗社区中用户关于医院的就医评价内容进行观点挖掘与情感分析并开展实证分析，结果显示该方法具有较好的挖掘效果。

〔关键词〕 在线医疗社区；用户评论；观点挖掘；情感分析

〔中图分类号〕 R－056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673－6036.2018.11.002

Method for Opinion Mining and Sentiment Analysis of User Comments of Online Medical Community Based on Characteristic Rules

XIANG Fei, XIE Yaotan, School of Medicine and Health Management, Tongji Medical College, HUST, Wuhan 430030, China

〔**Abstract**〕 In the case of the lack of large－scale opinion－sentiment corpus in the hospital comment area, the paper expounds on the current study situation of the sharing of domestic and overseas medical community knowledge and the opinion mining and sentiment analysis based on characteristics. Using the characteristic rule method, it implements opinion mining and sentiment analysis of the comments on seeking medical advice in hospitals made by users of online medical community on the basis of supplementing the sentiment lexicon and carries out empirical analysis. Results show that mining effect of the method is good.

〔**Keywords**〕 online medical community; user comments; opinion mining; sentiment analysis

1 引言

随着移动通讯技术的飞速发展，网络与实体空间的结合越来越紧密，很多虚拟与现实空间的业务逐步贯通，线上与线下活动逐步结合[1]。在线医疗社区中主要体现为网络空间与实体医疗卫生资源的对接与融合，由此衍生出在线问诊、在线预约、远程会诊、网上挂号等以实体为载体、以虚拟为途径的新兴医疗服务[2]。在这一进程中，在线医疗社区中由用户生成的数据呈现爆炸式增长，这一新兴数据源蕴含着大量有价值的信息，为知识发现提供独特的研究对象。

类似于网络电商中用户对于某一类产品的评论，在线医疗社区中也包含用户对于现实医疗服务的评价。随着通过网络途径获取实体医疗卫生服务用户数量急剧增多，越来越多的用户去在线医疗社区中撰写相关评论，往往包含用户现实体验与情感表达[3]。及时有效地从在线医疗社区中了解用户就医实际体验的观点表达与情感倾向，对于社区用户来说可以从侧面了解医疗服务质量，从而帮助自身更好地选择就医途径；对于医疗服务提供者来说有助于提升患者满意度与卫生质量。此外针对医院评

〔修回日期〕 2018－10－18

〔作者简介〕 向菲，讲师，发表论文 23 篇；通讯作者：谢耀谈，硕士。

价领域缺少大规模观点－情感语料库的情况，本研究基于特征规则与补充情感词典的观点特征挖掘与情感分析方法[4]，对于在线医疗社区中用户关于医院的就医评价进行研究，对其内容的特征进行挖掘；对各个特征的情感倾向（积极或消极）以及评论内容的情感倾向（积极或消极）进行识别；对本文方法效果进行测定。

2 国内外研究现状

2.1 在线医疗社区知识共享

在线医疗社区能够打破时间、空间上的限制，将用户与用户、用户与医护人员以及用户与医疗机构之间紧密联系起来，为其交流与合作提供统一平台[5-6]。现有研究证实用户通过其中的知识共享与交流能够有效地得到信息、情感、人际支持等社会支持[7-8]，对用户后续健康相关行为产生积极作用[9]，这为本文研究用户评论的现实意义奠定理论基础。目前关于在线医疗社区知识共享的研究多集中在知识共享影响因素[10-11]、用户参与行为[12-13]等方面，而少有研究从用户知识共享内容角度出发探索对于文本内容语义上的观点表达与情感分析。Lu[14]等对 MedHelp. org 中用户提问内容进行分析，对提问内容的利益相关者、热点话题以及情感倾向进行探究。有些学者则采用机器学习方法对癌症社区中用户帖子进行情感分类并辨别相应的情感极性[15-16]。

2.2 基于特征的观点挖掘与情感分析

观点挖掘与情感分析相关研究最早出现于 Hatzivassiloglo、Picard 等的研究中，经过多年的发展，现有研究可划分为 3 个层级：文档、句子以及方面级别[17]。根据 Liu[18]等的定义，用户观点表达包含特征（feature）、情感词（opinion word）、倾向（orientation）3 个方面。方面层级则对评论特征及情感倾向进行挖掘。从方面层级的观点提取来看，可分为显式[19-20]与隐式方面[21-22]挖掘，但是由于研究领域的限制以及现实原因，有关隐式方面观点挖掘的研究仍存在很大局限[17]。从对于观点特征提取与倾向判定的方法来看，可划分为基于规则[23]、基于统计[24]以及基于深度学习的提取[25]。从研究领域来看，多集中于电商、酒店、影视、餐饮等行业[26]。由此可知相关研究尚未涉及医疗领域，即对在线医疗社区中用户关于实际就医体验后的评论进行观点挖掘与情感分析。鉴于该领域尚未有完整标注的相关语料库，本文提出基于特征规则方法对相关评价的显式观点进行挖掘与分析。

3 研究方法

3.1 技术路线（图 1）

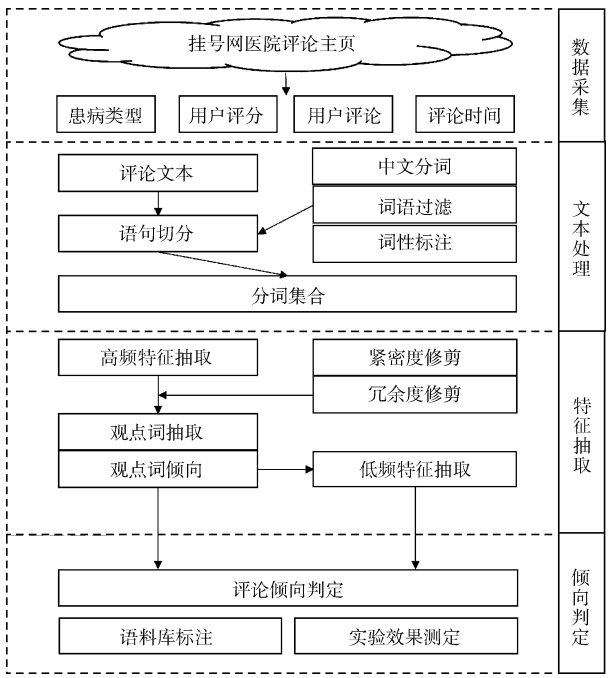

图 1 技术路线

3.2 数据收集与文本处理

利用自开发软件获取微医网（www. guahao. com）用户关于医院的评论数据，涉及的元数据包括患病类型、用户评分、用户评论文本以及评论时间。对网页数据源中的广告、无意义文本等无效数据进行去噪处理以及对获取数据中的缺省值进行填充处理。实体的特征通常以名词或名词词组的形式出现在评论文本中，鉴于中、英文的差异，文本分词与词性标注的工作就尤为重要。本文以中科院 ICTCLAS 分词系统结合停用词表对提问文本集进行词语切分处理。利用该系统词性标注功能进一步标注出各切分词语词性，根据词性标注结果去除词性为助词、虚词、连词、介词等无实际含义的词以及

各种标点符号，以此得到各评论文本带有词性标注的分词集。

3.3 特征抽取

3.3.1 高频特征抽取与修剪 本文以特征在评论中出现次数为支持特征重要程度的指标，但是评论文本中包含许多重复出现且与评论内容无关的内容对评论特征的抽取产生干扰。考虑到用户撰写相关评论的实际情况，当样本足够大时，其用词具有一定的收敛性[4]，故采取基于关联规则挖掘的方法对高频特征进行抽取。基于关联规则的挖掘方法可以描述为：给定词项集合 $I = \{i_1, i_2, \cdots, i_n\}$，D 为评论文本集合，每个评论文本都包含词项集合 I 特定的子集，对于 I 的任意两个互斥子集 $X \longrightarrow Y$，关联规则 $X \longrightarrow Y$ 满足 D 中有 $c\%$ 的评论文本同时支持 $X \longrightarrow Y$，即 $c\%$ 的评论文本同时包含 $X \longrightarrow Y$ 并集，以此挖掘出支持度大于最小支持度（minimum support）的规则 $X \longrightarrow Y$。然而由关联规则生成的高频特征并不完全能够反映评论实体的特征，存在混杂或冗余的情况。所以需要对高频特征做进一步的修剪：（1）紧密度修剪。定义构成词词间距 <3 的高频特征 f 为效应词，修剪词频 <2 的高频特征 f。（2）冗余度修剪。定义高频特征构成词单独在文本中出现的频次为净支持度，修剪净支持度 <3 的词。

3.3.2 观点词抽取 观点词可定义为人们表达主观意义上积极或消极情感的词或词组[27]。通过对评论文本的观察可以发现用户表达观点的词一般都贴近评论文本中的特征词。以往研究证实句中形容词与语句主观性之间的统计学关联[28]，由此本文以句中出现的形容词代指观点词，在此处定义特征词最临近的观点词为效应关键词。对由以上规则抽取出的观点词要进行 2 次人工过滤与筛选，以提高观点词提取的准确度。

3.3.3 低频特征抽取 高频特征是用户在评论中反复大量使用的词或词组，然而仍然可能存在一些忽略少数用户在评论中用到的特征词的情况。无论是低频特征还是高频特征，用户表达观点大多趋向一致（喜爱或厌恶）。定义未包含的高频特征但具有 1 至多个观点词的语句为低频特征语句，抽取低频特征语句中观点词附近且至少在两条评论文本中出现的名词或名词词组为低频特征。

3.4 倾向判定

在提取出所有特征的观点词后需要对其情感倾向进行判定，以表明用户积极或消极的情感。基于情感词典的判定是情感分析常用方法，能够为每个词赋予特定情感极性分数。选取基于知网的情感词典[29]，实际操作中需要考虑特定的语义或语用环境，如就医评价领域专有名词的情感倾向、否定词修饰等。鉴于以上情况，本文对现有情感词典进行领域专有词的补充，建立常用否定词词表，对评论文本中每个观点词建立对应索引，如果两个观点词之间出现否定词则将否定词与所修饰的观点词一并提取出来。根据以下公式计算观点词的情感得分，若情感分为正则判定为积极情感，否则判定为消极情感。（Score 情感得分；观点词若有否定词修饰，D 赋值为 1，否则为 0；S 为观点词在情感词典中取值）。

$$Score = (-1)^{\wedge}(D) * S$$

由此可以得到各个观点词的情感倾向。但是有些评论语句中包含 1 至多个特征与观点词，需要对这些语句的情感倾向做进一步判定。对积极关键词赋值为 1，消极关键词赋值为 -1，评论语句的情感倾向则为其中观点词情感赋值的算术加权：若 >0，为积极取向；若 <0 则为消极取向；若 =0，则与该句中的效应观点词的情感取向一致。

4 实证分析

4.1 实验数据

利用自开发 python 爬虫从微医网（www. guahao. com）医院挂号主页中采集位于武汉市的同济医院、协和医院、中南医院、普爱医院患者就医评论数据，采集时间为 2018 年 4 月 25 日。清洗、去重以及去掉无效评论共得到 1 815 条数据，对患者评论数据进行特点词与观点词人工标注并对患者打分数据进行转换处理，定义 3 ~ 5 星评价为好评，1 ~ 2 星评价为差评，构建患者评价语料库，见表 1。将表 1 中特征 - 观点数据进行可视化处理，

得到 4 家医院特征观点数据词云图，见图 2－5。

表 1　医院评论语料库

医院	特征－观点数	积极取向评论数	消极取向评论数
同济医院	816	543	57
协和医院	931	567	33
中南医院	379	249	21
普爱医院	549	408	7

图 2　同济医院特征－观点数据词云

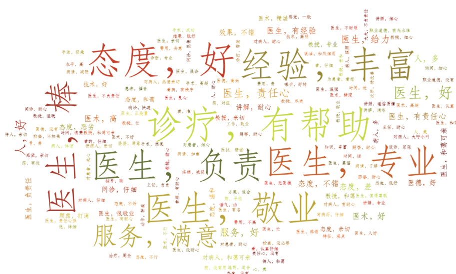

图 3　协和医院特征－观点数据词云

图 4　中南医院特征－观点数据词云

图 5　普爱医院特征－观点数据词云

可以看出 4 家医院评论出现频率较高的特征－观点为：态度，好；医生，好；诊疗，有帮助等。采用本文提出的方法对标注完成的语料库分析后提取的特征词与观点词示例，见图 6。对语料库中 4 家医院分析后得到的数据统计，见表 2。

<Positive>医生十分敬业负责
特征 观点词
<医生，敬业>
<医生，负责>
<negative>态度又差
特征 观点词
<态度，差>

图 6　观点挖掘与情感分析示例

表 2　观点挖掘与情感分析结果统计

医院	仅高频特征抽取			引入低频特征抽取		
	特征－观点数	积极取向评论数	消极取向评论数	特征－观点数	积极取向评论数	消极取向评论数
同济医院	676	394	59	766	440	78
协和医院	738	532	13	777	549	19
中南医院	309	189	19	353	210	28
普爱医院	488	275	19	568	321	27

4.2　评价指标

4.2.1　召回率（recall）　用户观点特征召回率 R_{of} 以及情感分析中评论情感倾向召回率 R_{sa} 分别为以下公式。

$$R_{of}=\frac{\text{评论中正确提取的特征观点数}}{\text{语料库中标注的特征观点总数}}$$

$$R_{sa}=\frac{\text{正确识别的情感倾向文本数}}{\text{语料库中的评论总数}}$$

4.2.2 准确率（precision） 用户观点特征召回率 P_{of} 以及情感分析中评论情感倾向召回率 P_{sa} 分别为以下公式。

$$P_{of}=\frac{\text{评论中正确提取的特征观点数}}{\text{评论中提取的特征观点总数}}$$

$$P_{sa}=\frac{\text{正确识别的情感倾向文本数}}{\text{识别情感倾向文本总数}}$$

4.2.3 F1 值 从实验结果来看通常召回率的提高往往导致准确率不同程度的降低，以此值调和准确率与召回率的平均数，权衡召回率与准确率之间的关系，计算公式如下：

$$F_1=2*\frac{recall*precision}{(recall+precision)}$$

4.3 效果评估（表 3－4）

表 3 观点特征提取效果评估（%）

医院	仅高频特征抽取			引入低频特征抽取		
	召回率	准确率	*F*1 值	召回率	准确率	*F*1 值
同济医院	69.00	83.28	71.98	75.25	80.16	81.69
协和医院	71.21	89.84	71.96	72.72	87.13	88.46
中南医院	70.71	86.73	72.89	75.20	80.74	83.63
普爱医院	75.41	84.84	78.22	81.24	78.52	81.56
平均	71.33	86.30	73.40	75.59	82.06	84.13

表 4 观点特征情感倾向效果评估（%）

医院	仅高频特征抽取			引入低频特征抽取		
	召回率	准确率	*F*1 值	召回率	准确率	*F*1 值
同济医院	67.83	89.85	71.24	75.00	86.87	88.33
协和医院	88.00	96.88	89.31	90.67	95.77	96.32
中南医院	71.48	92.79	74.67	78.15	88.66	90.67
普爱医院	65.78	92.86	70.79	76.63	91.38	92.11
平均	74.32	93.40	77.42	80.80	91.09	92.23

可以看出，观点特征提取召回率及准确率分别达到 75.59% 与 82.06%，评论情感倾向召回率及准确率分别达到 80.80% 与 91.09%，由于缺乏同领域类似实验，无法进行横向结果对比，但是从其他领域观点挖掘结果的效果来看[26]，本研究实验效果较好。此外从本研究的内部纵向对比来看，相比于仅高频特征提取，在引入低频特征提取处理后召回率都有所上升，虽然准确率下降，但是从综合评价指标 *F*1 值来看为上升，说明采用本文方法后得到的效果实现逐步提升。

5 结语

针对在线医疗社区中用户评论数据，采用基于关联规则的方法对用户的观点特征以及内容的情感倾向进行挖掘与分析，从而帮助用户更好地了解他人的就医体验与评价，为其就医选择提供参考，为了解各家医院服务质量与口碑评价提供途径。从本研究实验效果来看，在医院评论领域缺少大规模观点－情感标注预料库的情况下，本文提出的方法可行有效，可采取该方法进一步标注领域内的情感－观点语料库，为后续机器学习方法的开展奠定基础。然而本研究也存在一定的不足之处，将于后续的实验中进行改进。本研究只探索一种方法对用户评论数据进行探究与分析，缺乏不同方法间的对比分析，将在后续的实验中继续使用基于统计、机器学习等方法对在线医疗社区中患者评论数据的观点特征与情感倾向进行挖掘与分析。

参考文献

1 宁家骏.“互联网＋”行动计划的实施背景、内涵及主要内容［J］. 电子政务, 2015（6）: 32－38.

2 孟群, 尹新, 梁宸. 中国“互联网＋健康医疗”现状与发展综述［J］. 中国卫生信息管理杂志, 2017, 14

(2): 110 - 118.

3 吴江, 周露莎. 在线医疗社区中知识共享网络及知识互动行为研究 [J]. 情报科学, 2017, 35 (3): 144 - 151.

4 Hu M, Liu B. Mining and Summarizing Customer Reviews [C]. Seattle: Tenth Acm Sigkdd International Conference on Knowledge Discovery & Data Mining, 2004.

5 Eijk M V D, Faber M J, Aarts J W, et al. Using Online Health Communities to Deliver Patient - centered Care to People with Chronic Conditions [J]. Journal of Medical Internet Research, 2013, 15 (6): e115.

6 Nambisan P. Information Seeking and Social Support in Online Health Communities: impact on patients´ perceived empathy [J]. J Am Med Inform Assoc, 2011, 18 (3): 298 - 304.

7 Wang X, Zhao K. Social Support and User Engagement in Online Health Communities [C]. Beijing: International Conference for Smart Health, 2014: 97 - 110.

8 Wang X, Zuo Z Y, Zhao K. The Evolution and Diffusion of User Roles in Online Health Communities - a social support perspective [C]. Dallas: 2015 IEEE International Conference on Healthcare Informatics, 2015: 48 - 56.

9 Qiu B, Zhao K, Mitra P, et al. Get Online Support, Feel Better——sentiment analysis and dynamics in an online cancer survivor community [C]. Boston: IEEE Third International Conference on Privacy, 2012.

10 Zhao J, Wang T, Fan X. Patient Value Co - creation in Online Health Communities Social Identity Effects on Customer Knowledge Contributions and Membership Continuance Intentions in Online Health Communities [J]. Journal of Service Management, 2015, 26 (1): 72 - 96.

11 Yan Z, Wang T, Chen Y, et al. Knowledge Sharing in Online Health Communities: a social exchange theory perspective [J]. Information & Management, 2016, 53 (5): 643 - 653.

12 马骋宇. 在线医疗社区医患互动行为的实证研究——以好大夫在线为例 [J]. 中国卫生政策研究, 2016, 9 (11): 65 - 69.

13 范晓妞, 艾时钟. 在线医疗社区参与双方行为对知识交换效果影响的实证研究 [J]. 情报杂志, 2016 (7): 173 - 178.

14 Lu Y, Wu Y, Liu J, et al. Understanding Health Care Social Media Use From Different Stakeholder Perspectives: a content analysis of an online health community [J]. Journal of Medical Interent Research, 2017, 19 (e1094).

15 Biyani P, Caragea C, Mitra P, et al. Co - training over Domain - independent and Domain - dependent Features for Sentiment Analysis of an Online Cancer Support Community [C]. Niagara Falls: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis & Mining, 2013: 419 - 423.

16 Ofek N, Caragea C, Rokach L, et al. Improving Sentiment Analysis in an Online Cancer Survivor Community Using Dynamic Sentiment Lexicon [C]. Washington: International Conference on Social Intelligence & Technology, 2013: 109 - 113.

17 Liu B. Sentiment analysis: mining opinions, sentiments, and emotions [M]. Cambridge: Cambridge University Press, 2015.

18 Liu B. Mining Opinion Features in Customer Reviews [C]. San Jose: 19th National Conference on Artificial Intelligence, 2004: 755 - 760.

19 Jin W, Ho H H. A Novel Lexicalized HMM - based Learning Framework for Web Opinion Mining [C]. New York: Proceedings of International Conference on Machine Learning, 2009.

20 Chinsha T C, Joseph S. A Syntactic approach for Aspect based Opinion Mining [C]. Washington: IEEE International Conference on Semantic Computing, 2015.

21 Hai Z, Chang K, Cong G, et al. An Association - based Unified Framework for Mining Features and Opinion Words [J]. ACM Transactions on Intelligent Systems and Technology, 2015, 6 (2): 1 - 21.

22 Lixin X, Wang Z, Chen C, et al. Research on Feature - based Opinion Mining Using Topic Maps [J]. Electronic Library, 2016, 34 (3): 435 - 456.

23 Liu H, He J, Wang T, et al. Combining User Preferences and User Opinions for Accurate Recommendation [J]. Electronic Commerce Research and Applications, 2013, 12 (1): 14 - 23.

24 李晨曦, 谢罗迪. 基于 LDA 模型的文本分类与观点挖掘 [J]. 电子技术与软件工程, 2017 (4): 209 - 210.

25 Du J, Gui L, Xu R. Extracting Opinion Expression with Neural Attention [C]. Singapore: Proc. of the 5th of Chinese National Conf. on Social Media Processing, 2016.

26 韩忠明, 李梦琪, 刘雯, 等. 网络评论方面级观点挖掘方法研究综述 [J]. 软件学报, 2018 (2): 417 - 441.

27 B L. Sentiment analysis and opinion mining [M]. California: Morgan &Claypool Publisher, 2012.

28 Bruce R F, Wiebe J M. Recognizing Subjectivity: a case study in manual tagging [J]. Natural Language Engineering, 1999, 5 (2): 187 - 205.

29 Gan K W, Wang C Y, Mak B. Knowledge - based Sense Pruning Using the HowNet: an alternative to word sense disambiguation [C]. Taipei: International Symposium on Chinese Spoken Language Processing, 2002.