DOI: 10.11992/tis.201806016
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180716.1134.008.html

酒店在线评论数据的特征挖掘

秦海菲[1], 杜军平[2]
（1. 楚雄师范学院 信息科学与技术学院, 云南 楚雄 675000; 2. 北京邮电大学 计算机学院, 北京 100876）

摘　要: 论文以酒店在线评论数据为研究对象, 对酒店在线评论数据的特征挖掘进行了研究。论文首先从酒店在线评论数据的获取出发, 经过数据清洗、词性分析、特征抽取、指标确定、特征筛选、特征确定、特征校验几个环节, 实现了酒店在线评论数据特征挖掘的目的。论文以词频为基础, 融合了词性分析、聚类分析等方法, 利用词频数 (TF)、词频率 (TF_1)、词频权重 (TTW)、评论频率 (DF)、逆文档频率 (IDF) 和 TF1-IDF 等指标对候选特征词进行降维, 得出酒店在线评论数据的特征, 并对特征词进行校验, 完成了酒店在线评论数据的特征挖掘的过程。论文将为以评论为依据的客户分类、酒店分类、智能推荐奠定基础。

关键词: 酒店; 在线点评; 数据获取; 特征抽取; 特征挖掘; 聚类分析; 分类; 智能推荐
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)06−1006−09

中文引用格式: 秦海菲, 杜军平. 酒店在线评论数据的特征挖掘[J]. 智能系统学报, 2018, 13(6): 1006–1014.
英文引用格式: QIN Haifei, DU Junping. Feature mining based on online hotel review[J]. CAAI transactions on intelligent systems, 2018, 13(6): 1006–1014.

Feature mining based on online hotel review

QIN Haifei[1], DU Junping[2]
(1. School of Information Science and Technology, Chuxiong Normal University, Chuxiong 675000, China; 2. School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: In this study, the feature mining of online hotel review data is investigated. First, online hotel reviews data were obtained. To mine features from the review data, data cleaning, part-of-speech analysis, feature extraction, index determination, feature selection, feature determination, feature checking were carried out. Based on the word frequency, integrating part-of-speech analysis, and cluster analysis, the word frequency (TF), word frequency rate (TF_1), word frequency weight (TTW), comment frequency (DF), inverse document frequency (IDF), and TF1-IDF of candidate feature words were applied to reduce dimension. The online hotel review data features were obtained, and then the feature words were verified. This paper will lay a solid foundation for the classification of hotels and customers and intelligent recommendation based on online reviews.

Keywords: hotel; online review; data capture; feature extract; feature mining; cluster analysis; classification; intelligent recommendation

在网购中, 在线点评是买家购买决策的重要依据, 同时也是卖家经营信息反馈的重要环节。在线点评分为数字评分和在线评论。目前, 很多学者专注数字评分, 因为数字评分比较直观, 容易理解, 但数字评分的粒度比较粗、少, 且难于细化, 例如同时被评为 5 分的同一家酒店, 顾客对它的感受完全不一样, 有的关注环境, 有的关注设施, 有的关注服务等。不同的人关注点不同, 兴趣点也不同, 评价也亦不同。从经济学和市场理论的角度看, 产品和服务有多维属性, 由于消费者的偏好不同, 对功能和服务的期望也不同, 即用户参考评论进行决策时, 会依其偏好, 只关注或更加关注某些方面的特征。只考虑数值评分无

收稿日期: 2018−06−05.　**网络出版日期:** 2018−07−17.
基金项目: 国家自然科学基金项目 (61320106006, 61532006, 61772083).
通信作者: 杜军平. E-mail: junpingdu@126.com.

法反映用户对产品的全面和精确的评价[1]。因为某一类产品的数字评分不能为客户带来过多的信息, 但是在线评论可以表达顾客的真实感受, 能够被购买者参考和信赖。在线评论作为顾客在网络上发布的购买体验, 对其他客户的购买决策起着重要的影响, 这些体验也是企业在市场拓展和产品开发计划时要考虑的重要信息[2]。因此, 在线评论数据也变得越来越重要。

随着网络的发展, 用户生成的数据越来越多, 引起了利益双方或多方的广泛兴趣, 捕获这些数据并把它们转换为企业的核心洞察力, 可为决策、营销、分析等不同目标服务[1-4]。在线评论数据像大数据一样具有体量巨大, 增长速度快, 种类繁多, 价值密度低等特点。从在线评论数据中挖掘出顾客真正关心的酒店特征和对酒店的真实感受, 可为酒店的分类提供真实可靠的依据, 同时也为酒店的智能推荐奠定基础。

1　相关工作

1.1　在线评论数据分析

在消费者的决策过程中, 在线评论已成为非常重要的信息来源[5]。研究表明, 如果产品被他人推荐, 产品的选择次数会增加两倍, 这种影响取决于推荐来源的类型[6]。消费者在准备购买产品或服务时越来越多地寻求同行的经验, 超过60% 的消费者在购买前会咨询客户的反馈意见[6]。住宿评论决定了酒店的在线形象、销售额和未来收入[5−6]。

目前, 对在线评论的研究主要是从情感出发, 分析人们对某一产品的情感色彩和情感倾向, 从在线评论中判断出人们的喜、怒、哀、乐、批评、赞扬等, 从而判断出这一产品的受欢迎程度。在线评论挖掘属于观点挖掘, 但不同于情感挖掘, 情感挖掘只属于观点挖掘的一部分。2012 年刘冰[7]在情感分析和观点挖掘一文中对观点挖掘涉及相关技术进行了总结; 2015 年 Ravi, Guellil 等[8−9]充分阐述了观点挖掘; 2016 年 Rana[10]对观点挖掘中的方面提取技术进行了综述; 2017 年 Sun 等[11]和李建华等[12]对观点挖掘上进行进一步的总结和挖掘; 2018 年韩忠明等[13]对网络评论方面级观点挖掘方法作了综述研究。酒店是在线评论的重要内容, 且酒店在线评论数据的获取是很方便的, 可以从猫途鹰、携程、美团、大众点评、驴妈妈、微博、微信等网站上获取, 但从目前的研究看, 有影响的研究成果还比较少。

1.2　短文本分析

在线评论数据属于短文本研究。每个人每天都在应用短文本 (短信、微博、微信、评论、Tweets、facebook 等), 短文本与普通文本有很大区别。短文本是包含有限的上下文, 大多数短文本搜索查询少于 5 个单词, Tweets 是不超过 140 个字符短文本[14]。几乎所有的短文本都在 200 字以内, 在线点评数据也不例外。短文本通常不遵循语法, 自然语言处理技术 (如词性标注和句法解析等) 难于直接应用于短文本分析[15]。短文本具有稀疏性强、价值密度低, 实时性强、变化大、嘈声大、规则性弱等特点。因此, 对短文本的分析比一般的文本分析要难。目前短文本研究多数都集中在社交网络, 酒店在线评论的研究属于社交网络研究中的一部分。

2　酒店在线评论数据的特征挖掘

在线评论特征的挖掘包括数据获取、数据清洗、词性分析、特征抽取、特征词确定等环节。具体流程如图 1 所示。

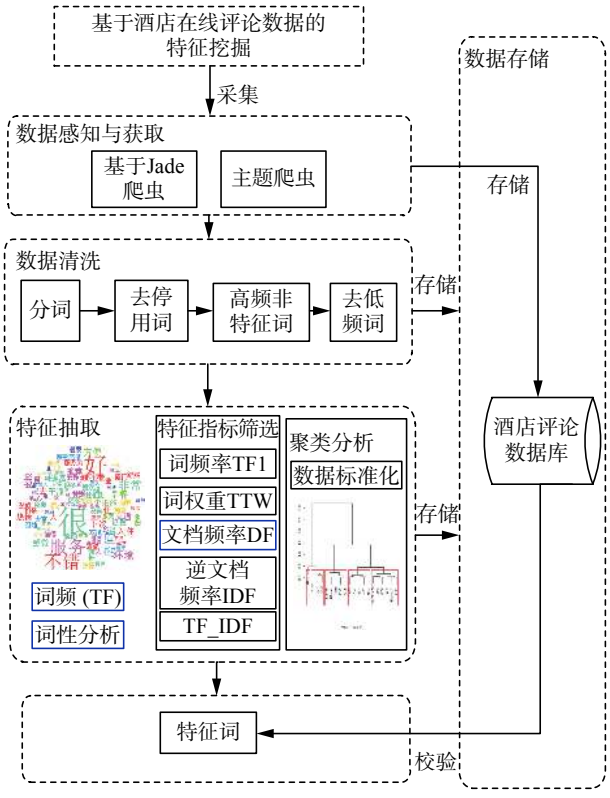


图 1　基于酒店在线评论数据的特征挖掘

Fig. 1　Feature mining based on hotel online review data

2.1　数据获取

在线点评数据包括数字、文本、图片等, 本文应用主题爬虫在猫途鹰网 (tripadvisor) 和携程网 (ctrip) 上爬取相关数据, 去除与主题无关的

各种噪音数据 (如导航条、广告信息、版权信息和其他图片、图像、声音等), 对获取到的数据进行预处理 (主要是去除无关和重复的数据) 和清洗。

2.2　数据清洗

数据清洗是保证数据质量的关键环节, 在线评论数据的清洗工作主要包括数据预处理 (去特殊标记、标点等)、分词、去停用词、去低频词、去高频非特征词, 具体步骤如图 2 所示。

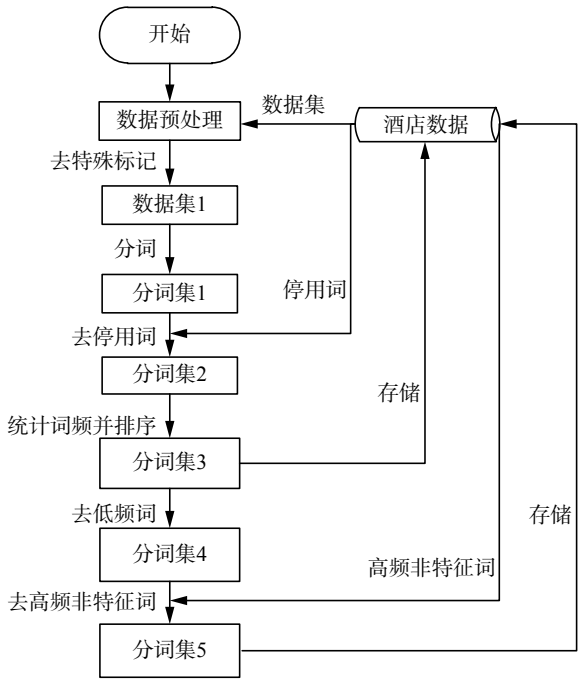


图 2　数据清洗的过程

Fig. 2　Process of data clean

文本数据预处理: 完成多余字符删除和多余数据清除。

分词: 采用中科院分词和结巴分词相结合的方式, 分词后的数据为分词集 1。

去停用词: 在分词集 1 中很多词没有实际意义, 仅仅代表一种结构, 比如介词、叹词、连词等, 把这部分词集合在一起形成停用词表。在数据清洗中需要将停用词剔除, 以降低特征向量维度, 去除停用词后的词集为分词集 2。

词频统计: 词频 (term frequency, TF) 是指词或短语在给定文档中出现的总次数, 通常认为词频越高, 其在文档中的重要度越高, 成为关键词的可能性越大[16]。在酒店评论数据中, 指在评论中某个词出现的次数。

词频排序: 对分词结果的词频进行降序排列, 排序结果为分词集 3。

去低频词: 对低频词进行剔除处理, 去除低频词后的词集是分词集 4。

去高频非特征词: 在线评论数据中, 特征不明显的高频词会削弱特征词的特性, 去除高频非特征词的词集是分词集 5。

2.3　特征词的抽取

从在线评论中提取反映评论主题的特征词 (Keyphrases, 包括单词或词组), 提取的特征词需要满足可读性相关性重要性覆盖度一致性[16]。目前比常用的特征提取方法有 TF-IDF、词频、文档频率、逆文档频率等。单独使用上述方法不能达到特征词选取的良好效果。

1) 词性分析

众多文献提出特征词通常是名词短语[16, 20],因此需要对词性进行分析。

2) 特征词指标

① 词频 (TF): 词 W 在评论中出现的次数。频数 (TF) 越高, 评论的次数越多, 关注人群越多, 关注程度也就越大。某词 W 的词频 N_W(即词 W 出现的次数) 为: $N_W = \left\{\sum_{i=0}^{N} W_i : W_i \in W\right\}$, W_i 是词 W 出现的第 i 次。

②词频率 (TF_1): 词 W 在所有词中的比重。为了与词频数区分开, 采用 TF_1 表示。

假设一条评论分词后的词集 V 是 $V = \{W_1, W_2, W_3, \cdots, W_M\}$, $W_1, W_2, W_3, \cdots, W_M$ 是评论分词 (即一条评论分成 M 个词), 有 N 条评论, 那么形成的就是一个 N 行 M 列的矩阵, N 条评论分词后构成的评论词集 V_N 是 $V_N = \sum_{i=0}^{N}\sum_{j=0}^{M} W_{ij}$ W_{ij} 是 i 行 j 列的词 (即第 i 条评论的第 j 个分词)。某一单词 W 的词频数 $\mathrm{TF}_W = \left\{\sum_{i=0}^{N}\sum_{j=0}^{M} W_{ij} : W_{ij} \in W\right\}$。词频率=某个词在评论中出现的总次数/评论的总词数, V_N(length) 是 N 条评论分词后构成的评论词集长度, TF_{1w} 是词 W 的词频率, 具体计算如式 (1) 所示:

$$\mathrm{TF}_{1W} = \frac{\mathrm{TF}_W}{V_N(\mathrm{length})} \tag{1}$$

处理后某一词 W 的词频率如式 (2) 所示:

$$\mathrm{TF}_{1W} = \frac{\left\{\sum_{i=0}^{N}\sum_{j=0}^{M} W_{ij} : W_{ij} \in W\right\}}{\sum_{i=0}^{N}\sum_{j=0}^{M} W_{ij}} \tag{2}$$

③词评权重 (TTW): 词 W 在评论中的比重。

某词 W 在一条评论中被多次提到和被多人提到, 意义是不一样的, 为了更好区分两者关系, 采用词评权重 (TTW)。假设每一条评论代表了一个点评人, 如果一个词被多个人评论, 那么代表这个词被多人关注, 这样的词可以是特征词。词评权重既考虑了词频数, 也考虑了评论人数。TF_W 是词 W 在评论中出现的次数, N_W 是含词

W 的评论条数 (假设一条评论代表一个人), $N_W=\left\{\sum_{i=0}^{N}W_i:W_i\in W\right\}$, 词 W 的词权重计算如式 (3) 所示:

$$\mathrm{TTW}_W=\frac{\mathrm{TF}_W}{\mathrm{N}_W}=\frac{\left\{\sum_{i=0}^{N}\sum_{j=0}^{M}W_{ij}:W_{ij}\in W\right\}}{\left\{\sum_{i=0}^{N}W_i:W_i\in W\right\}} \tag{3}$$

④ 评论频率 (DF): 评论频率也称文档频率, 指某条评论在总评论中的比重。

DF=包含该词的评论条数/总评论数, N 是总评论数, 评论频率计算如式如 (4) 所示:

$$\mathrm{DF}=\frac{N_W}{N}=\frac{\left\{\sum_{i=0}^{N}W_i:W_i\in W\right\}}{N} \tag{4}$$

⑤ 逆文档频率 (IDF): 衡量词或词组所在的文档在整个语料库中的频率。

逆文档频率越大表明该词越重要, 它是一个词语普遍重要性的度量[16]。IDF 的思想是: 如果包含词条 W 的评论越少, 也就是, N_W 越小, IDF 越大, 则说明词条 W 具有很好的类别区分能力。特定词语 W 的 IDF, 可以由总评论数除以包含该词语的评论, 再取对数得到。计算公式如式 (5) 所示:

$$\mathrm{IDF}_W=\mathrm{LOG}\frac{N}{N_W+1}=\mathrm{LOG}\frac{|N|}{\left|\left\{\sum_{i=0}^{N}W_i:W_i\in W\right\}+1\right|} \tag{5}$$

⑥ 特征权重值 (TF-IDF): 词频−逆文档频率 (TF-IDF) 是结合词频和逆文档频率来衡量候选关键词的重要度量。

词频−逆文档频率 (TF-IDF) 被认为是所有特征中最有效、最常用的特征之一[16]。如果某个词或短语在一篇文章中出现的频率 TF_1 高, 并且在其他文章中很少出现, 则认为该词或者短语具有很好的类别区分能力, 适合用来分类。TF-IDF 的计算如式 (6) 所示:

$$\mathrm{TF}_{\mathrm{IDF}}=\mathrm{TF}_1\times\mathrm{IDF}=\frac{\mathrm{TF}_W}{V_N(\mathrm{length})}\times\mathrm{LOG}\frac{N}{N_W}=\frac{\left\{\sum_{i=0}^{N}\sum_{j=0}^{M}W_{ij}:W_{ij}\in W\right\}}{\sum_{i=0}^{N}\sum_{j=0}^{M}W_{ij}}\times\mathrm{LOG}\frac{|N|}{\left|\left\{\sum_{i=0}^{N}W_i:W_i\in W\right\}+1\right|} \tag{6}$$

TF-IDF 值与该词的出现频率成正比, 与在整个评论中出现的次数成反比。

3) 特征词的筛选

特征词的筛选是特征词选取和降低特征词维度最有效的方法。分析各特征词指标的关系是特征词选取中重要的环节, 但各个指标之间存在有很强的相关性, 并且量纲差异较大。为了消除各指标量纲的影响和指标之间的相关性, 采用标准差标准化 (Z 标准化) 对数据进行标准化处理。计算公式如式 (7) 所示:

$$X=\frac{(x-\mu)}{\sigma} \tag{7}$$

式中: μ是所有样本数据的均值, δ是样本数据的标准差, 进一步细化后, 得到结果如式 (8) 所示。

$$x_{ij}=\frac{x_{ij}-\bar{x}_{.j}}{\sqrt{\sum_{i=1}^{N}(x-x_{.i})^2}} \tag{8}$$

聚类分析是研究样品和指标分类问题的一种多元统计方法[17−19]。在实际应用中一般有两种处理方式, 一种是根据分类问题本身的专业知识结合实际需要来选择分类方法, 并确定分类个数; 另一种是多用几种分类方法, 把结果中共性取出来, 如果用几种方法的某些结果都一样, 则说明这样的聚类确实反映事物的本质[19]。采用专业知识与多种聚类算法结合的方式对特征进行筛选, 以确定特征词。

算法 1　在线评论数据的特征挖掘聚类算法

①将候选特征词各自成一类, $\{X_1,X_2\cdots,X_N\}$;

②计算各类之间的距离 (类平均法、ward 法、最大距离法、相似分析法), 得到观测值矩阵;

③合并类间距离最小的两类为一新类。并重新计算新类与各类之间的距离, 更新矩阵表, 类的总个数依次递减, 直到为 1;

④画聚类树图;

⑤根据聚类图和专业知识决定分类的个数和成员;

4) 特征词提取方法评价

本文认为特征词能代表评价主题, N_c 为代表评价主题的特征词数, N_A 为选择的特征词数, 准确率 P 如式 (9) 所示:

$$P=\frac{N_c}{N_A} \tag{9}$$

有学者研究提出召回率不适合评论数据的评价指标, 因此本文借助别人提出的 GMM 指标, N_c 为能代表评价主题的特征词数, N_A 为所选择的特征词数, 准确率 GMM 如式 (10) 所示:

$$\mathrm{GMM}=\sqrt{\frac{N_c}{N_A}\times\frac{N_c}{N_A}} \tag{10}$$

2.4　特征词的验证

采用数据集 2 对所选特征词进行校验。

3　实验及结果分析

本文采用主题爬虫对网络数据进行抓取。获

取全国高端酒店 (四、五星级酒店) 的在线点评数据，并对在线点评数据进行处理，把全国高端酒店 (四、五星级酒店) 按数字评分进行排序，取出数字评分排在前 20 名的酒店数据为数据集 1 和数字评分排在后 20 名的酒店数据为数据集 2。应用数据集 1 进行建模分析。在建模过程中，对数据进行清洗，再对数据进行特征抽取、特征筛选和特征选取，应用数据集 2 对特征词进行校验。

根据实验需求，从网络上爬取的 50 余万条数据中筛选出前 20 名的酒店 174 449 条评论数据 (数据集 1) 和后 20 名的酒店 104 898 条数据 (数据集 2) 进行分析，采用中科院分词系统和结巴分词相结合的方法分别对数据集 1 和数据集 2 进行分词，去停用词后分别得到 4 049 078 个词条和 1 857 523 个词条，并分别对词条进行分析。由于词条数太多，再次对词条的低频词和高频非特征词进行处理，为了降低词条维度和分析的难度，采用降维、抽样的方法对词条进行分析。抽取了频数排在前 100 的词条进行词云分析，词云图如图 3 所示。

图 3　词云图

Fig. 3　Word cloud

3.1　特征词抽取与筛选

近年来，有许多学者对特征词的提取方法进行研究。特征词提取方法可以归纳为监督型和非监督型两类[20]。监督型可以用多种方法训练模型,并实时调整参数；非监督型特征词抽取主要有基于统计、基于词性规则、基于主题模型、基于词排列图的方法。本文将利用统计、词性和主题模型的方法对特征词抽取进行分析。

3.1.1　词性分析

从图 3 可以看出词条多且复杂，如果把所有的词都作为候选特征词，那么特征向量的维数将非常大，难于进行计算和分类。在词云图中很多词要与其他词组合在一起才具有实际意义。比如："服务好"、"服务不错"、"服务不好"、"服务很差"、"环境好"、"环境不错"、"环境很好"、"环境很差"、"好"、"不错"、"差"等词是对某一主题的评价。根据酒店在线评论数据的分词结果，对词性的分析如表 1 所示。

表 1　词性分析

Table 1　Part of speech analysis

词性	名词	形容词	动词	副词
词数	1 016 342	416 694	850 014	378 516
占比/%	25.10	10.29	20.99	9.35

从表 1 可以看出，在所有词语中名词、形容词、动词、副词在所有词中占到了 2/3，名词的占比是最高的，占到了 25.10%，其次是动词占 20.99% (包含动名词占总词数的 2.5%，后面分析时把这部分词归属名词处理)。对词性进行分析，分析结果如图 4 所示。

(a) 名词词云图　(b) 形容词词云图图

(c) 动词词云图　(d) 副词词云图

图 4　词性词云图

Fig. 4　Part of speech word cloud

根据图 4 名词、形容词、动词、副词的词云图可以看出名词作为评价的主题，作为候选特征词的确比较显著。除名词外，在动词词云图中最为明显的"服务"一词也可以作为评价的主题。对这部分词语的词性进行查看，"服务"属于动名词，为了特征提取的方便，把动名词归属于名词的行列。但还有一部分词语完全是动词，但也表达名词的意思，比如"装修"、"位置"、"出行"等，这一部分是动词名用，对于这一部分词需要作为特征词分析。评论数据属于短文本数据，并没有完全遵从自然语言的语法结构，动词名用或名词动用情况很常见。因此，在特征词的选取上只选取系统所分的名词是不可取的，必须根据专业、行业、常识等对词性进行进一步的筛选与确定。根据词性的词频对候选特征词的分析如表 2 所示。

表 2　候选特征词词性分析

Table 2　Part of speech analysis of candidate feature words

方法	提取特征	代表评价主题
不分词性 (TF)	很、好、不错、服务、也、是、都、房间、非常、有、环境、还、方便、入住、早餐、在、住、和、去、下次、	服务、房间、环境、早餐
名词 (TF)	房间、环境、早餐、设施、感觉、热情、交通、宾馆、服务员、性价比、前台、温泉、人、服务态度、价格、有点、水果、大堂、地理位置、味道	房间、环境、早餐、设施、交通、宾馆、服务员、性价比、前台、温泉、服务态度、价格、水果、大堂、地理位置、味道
动词 (TF)	服务、是、有、住、去、到、位置、来、推荐、满意、没有、还有、值得、会、吃、免费、选择、送、贴心、装修	服务、位置、装修
TF_ IDF	不错、服务、房间、入住、早餐、非常、环境、下次、方便、干净、设施、服务员、性价比、宾馆、前台、服务态度、感觉、热情、满意	服务、房间、早餐、环境、方便、设施、服务员、性价比、前台、服务态度
动词+名词	服务、房间、环境、早餐、设施、位置、交通、性价比、服务员、前台、服务态度、价格、卫生、水果、地理位置、大堂、温泉、味道、装修	服务、房间、环境、早餐、设施、位置、交通、性价比、服务员、前台、服务态度、价格、卫生、水果、地理位置、大堂、温泉、味道、装修

综合几种特征词提取方法, 本文先利用无监督方法 TF(词频数) 提取候选特征, 所提取的 20 个特征词能代表评价主题值有 4 个。综合 TF 和词性进行分析, 形容词、副词中没有能代表评价主题的候选特征词; 动词中代表评价主题的有 3 个, 名词中 16 个。利用 TF_IDF 提取的候选特征词代表评价主题的有 10 个。而综合无监督型的 TF、词性在无监督的情况下动词+名词提取的特征词效果与 TF_IDF 的提取效果一样, 而选择名词作为特征词, 在监督下筛选动词作为补充, 所提取的效果要比只提取名词的效果要好, 准确率和 GMM 值都达到了 87%, 而若名词+动词的筛选都在监督下完成, 所得的候选特征词与评价主题的特征词的准确率和 GMM 达到 95% 以上。具体结果如图 5 所示。

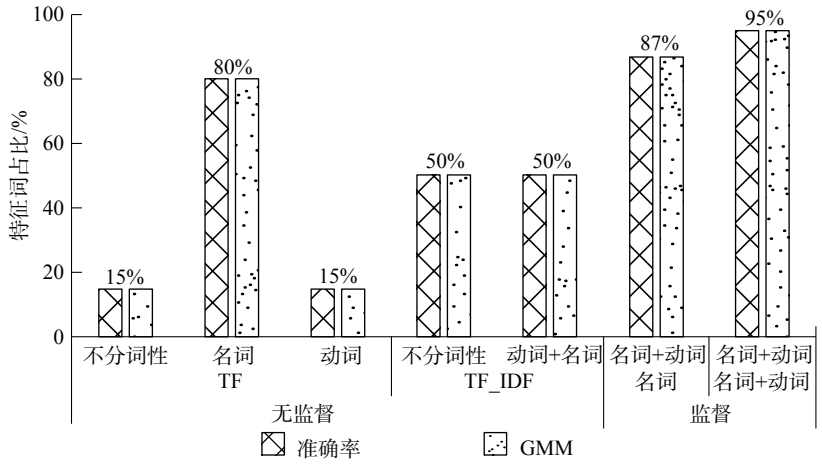


图 5　特征词提取方法对比

Fig. 5　Comparison of feature words extraction methods

3.2　特征词指标分析

虽然动词+名词结合的特征词比较适合分析, 但候选特征词的维度比较大, 各候选特征词之间的关系比较复杂, 是否具备特征词的特性还需要进一步分析, 特征词指标分析如表 3 所示。

从表 3 可以看出根据词频数 (TF)、词频率 (TF_1)、词频权重 (TTW)、评论频率 (DF)、逆文档频率 (IDF) 和 TF_1-IDF 这 6 个评价指标选取特征词时, 在各个指标上选取特征词的结果都不一样。TF 和 DF 最高的是“服务”, TF_1 最高的是“房间”, TTW 最高的是“环境”, IDF 最高的是“装修”, TF_1-IDF 最高的是“温泉”。“温泉”的 TF_1-IDF 的值是最高的, 但从专业的角度看, 温泉可能是高端型酒店的一个特征, 但不能作为最重要的评价指标。“温泉”的 TF_1-IDF 值高说明有很多高端客户在关注“温泉”, 但用“温泉”作为酒店评论数据的

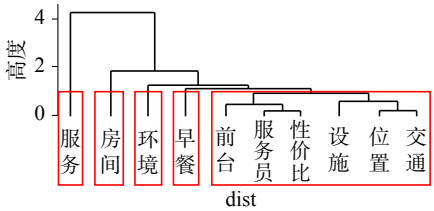
特征词是没有代表性的。从单一的指标中选取出的特征词不能完全满足特征词选择的可读性、相关性、重要性、覆盖度、一致性的要求，但各个指标对候选特征词又都有影响。因此，考虑对象酒店在线评论数据的实际情况，综合应用 TF、TF_1、TTW、DF、IDF 和 TF_1-IDF 这 6 个指标对候选特征词进行分析。从表 3 可以看出各个候选特征词在各个评价指标上的量纲是不同的，并且差距很大，TF、TF_1、TTW、DF、IDF 和 TF_1-IDF 各指标之间存在着很强的相关性。综合 19 个候选特征词的 6 个评价指标的实际情况看，降低特征词的维度是选取特征词最实用的方法。

表 3　指标分析

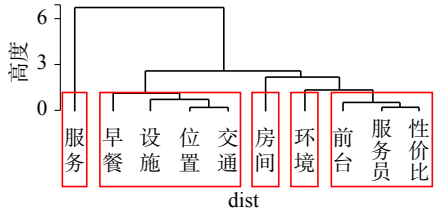
Table 3　Index analysis

候选关键词	TF	TF_1	TTW	DF	IDF	TF_1_IDF
服务	73 465	0.018	0.848	0.357	0.447	0.008
房间	46 439	0.266	0.826	0.220	0.658	0.175
环境	33 892	0.194	0.950	0.185	0.734	0.143
早餐	33 603	0.193	0.727	0.140	0.854	0.164
设施	26 501	0.152	0.597	0.091	1.043	0.158
位置	26 418	0.151	0.516	0.078	1.107	0.168
交通	26 273	0.151	0.489	0.074	1.133	0.171
服务员	25 705	0.147	0.408	0.060	1.221	0.180
性价比	25 338	0.145	0.437	0.063	1.198	0.174
前台	23 311	0.134	0.381	0.051	1.293	0.173
温泉	22 795	0.131	0.258	0.034	1.472	0.192
服务态度	21 274	0.122	0.379	0.046	1.335	0.163
价格	20 552	0.118	0.375	0.044	1.355	0.160
卫生	20 248	0.116	0.360	0.042	1.379	0.160
水果	19 386	0.111	0.351	0.039	1.409	0.157
大堂	18 792	0.108	0.318	0.034	1.465	0.158
地理位置	16 739	0.096	0.387	0.037	1.430	0.137
味道	16 729	0.096	0.348	0.033	1.477	0.142
装修	16 492	0.095	0.326	0.031	1.511	0.143

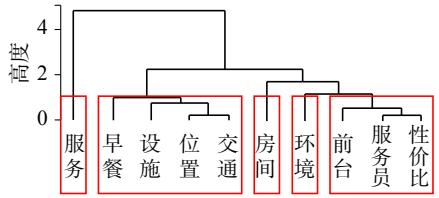
综合图 6 候选特征词的 4 个聚类树图根据聚类结果和酒店的专业知识，聚类为 5 类比较合理，把酒店在线评论候选词归并为 5 类，并对 5 类特征进行综合分析，综合 19 个候选特征词的聚类结果如表 4 所示。



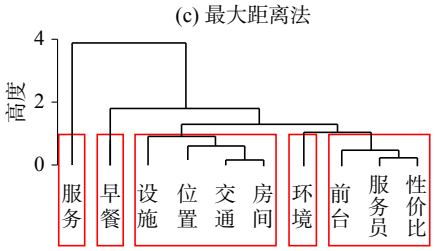
(a) 类平均法



(b) 离差平方和



(c) 最大距离法



(d) 相似分析

图 6　数据集 1 候选特征词聚类

Fig. 6　Dataset1 Candidate feature words cluster

表 4　候选特征词归类表

Table 4　Candidate feature word classification

类	候选特征词	特征词
1	服务、服务员、服务态度、前台	服务
2	房间、设施	设施
3	位置、环境、交通、地理位置	环境
4	早餐、水果、味道	餐饮
5	大堂、性价比、价格、卫生、装修、温泉	整体舒适度

通过表 4 可以看出特征词“服务”包含了“服务”、“服务员”、“服务态度”、“前台”等服务信息；特征词“设施”包含了“房间”、“设施”等硬件设施信息；特征词“环境”包含了“位置”、“环境”、“交通”、“地理位置”等信息；特征词“餐饮”包含了“早餐”、“水果”、“味道”等餐饮信息；整体舒适度包含了“大堂”、“性价比”、“价格”、“卫生”、“装修”、“温泉”等整体舒适度信息。这 5 个特征词能满足特征词选取的可读性、相关性、重要性、覆盖度、一致性的准则，因此可以作为酒店在线评论数据的特征词。

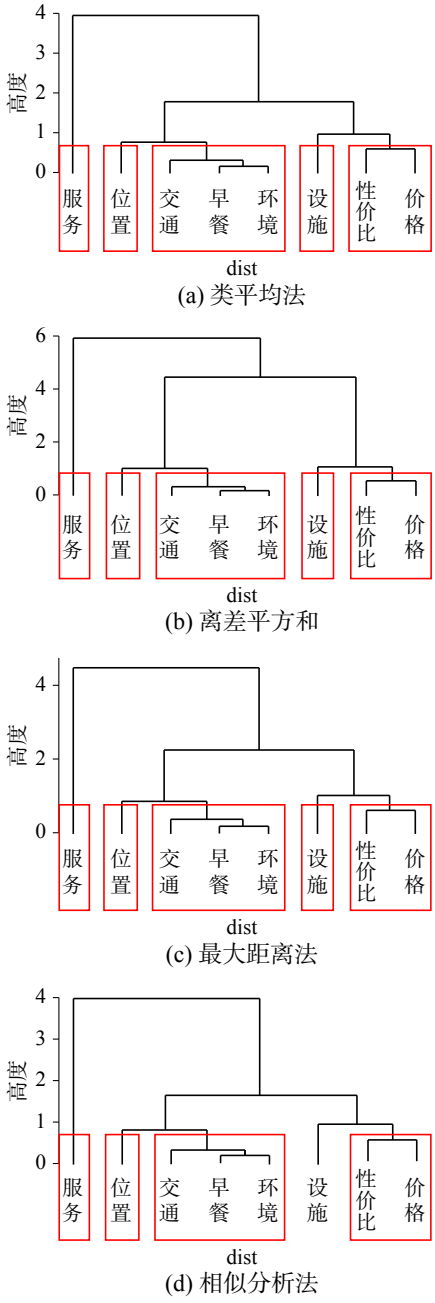
3.3　特征词的校验和选定

3.3.1　方法的验证

采用同样的方法，用数据集 2(数字评分排在后 20 家的酒店数据) 的词条进行了词性分析，处理后得到了 24 个候选特征词，计算出 24 个候选特征词的 6 个指标 (TF、TF_1、TTW、DF、IDF 和 TF_1-IDF) 的值，并对数据进行标准化后，采用 6 个指标对候选特征词进行聚类，所得的聚类结果如

图 7 所示 (为了图形清晰, 本文只选取了 TF 最高的数据进行展示)。

综合图 7 候选特征词的 4 个聚类树图, 根据聚类结果, 可以看出聚类为 5 类比较合理, 根据酒店的专业知识, 把酒店在线评论候选词归并为 5 类, 结果如表 5 所示。



(a) 类平均法

(b) 离差平方和

(c) 最大距离法

(d) 相似分析法

图 7　数据集 2 候选特征词聚类

Fig. 7　Dataset2 Candidate feature words cluster

从表 5 中可以看出, 部分被归并的候选特征词有了更细化、更相近或概括的变化, 例如设施中增加了“床”、“房”、“空调”等细化词; 环境中增加了“地段”、“出行”、“周边”、“附近”等相近词; 整体舒适度增加了“总体”、“整体”概括词。综合酒店在线评论的两个数据集和网络在线点评数据的特性, 可以看出把服务、设施、环境、餐饮和整体舒适度作为酒店在线评论数据的特征词是合理的。

表 5　后 20 名酒店特征词归类表

Table 5　The last 20 Hotel feature word

类	候选特征词	归并后特征词
1	服务、服务员、服务态度、前台等	服务
2	房间、设施、床、房、空调等	设施
3	环境、位置、交通、地理位置、地段、出行、周边、附近等	环境
4	早餐、水果、味道等	餐饮
5	大堂、性价比、价格、卫生、装修、温泉、总体、整体等	整体舒适度

综合 6 个评价指标聚类图, 对于评论数据, TF 分析的结果要比 TF_IDF 的效果好, 选取以 TF 为主, TF_1、TTW、DF、IDF、TF_1_IDF 为辅的指标聚类时, 选取 TF 排在前 10 的候选特征词聚类和选取更多的候选特征词聚类结果类似, 后面的候选特征词只是对前面结果的补充或细化。

4　结束语

本文从酒店在线点评数据出发, 对数据的感知获取、数据预处理、词性分析、特征选取、特征筛选、特征确定等进行了研究。对特征词的筛选和确定进行了分析。单个指标 (TF 或者 TF_1-IDF) 对特征词的筛选和选择效果不理想, 需要综合 TF、TF_1、TTW、DF、IDF、TF_1-IDF 多个指标进行分析。采用了无监督的聚类方法对变量进行聚类分析, 聚类时采用数据标准化消除指标相关性和量纲的影响。综合聚类分析的结果和酒店专业知识选定酒店在线评论数据的特征词, 通过将 20 家酒店作为数据集 2 对特征词进行校验, 得出酒店在线评论的特征词是服务、环境、设施、整体舒适度、餐饮。下一步将根据特征词构造更方便、快捷、可靠的分类器, 为酒店和客户进一步细分做好准备, 同时也为酒店为客户提供的个性化的智能推荐服务奠定基础。

参考文献:

[1] 吴维芳,高宝俊, 杨海霞, 等. 评论文本对酒店满意度的影响: 基于情感分析的方法[J]. 数据分析与知识发现, 2017, 1(3): 62–71.
WU Weifang, GAO Baojun, YANG Haixia, et al. The impacts of reviews on hotel satisfaction: a sentiment analysis method[J]. Data analysis and knowledge discovery, 2017, 1(3): 62–71.

[2] GAVILAN D, AVELLO M, MARTINEZ-NAVARRO G.

The influence of online ratings and reviews on hotel booking consideration[J]. Tourism management, 2018, 66: 53–61.

[3] TAN Sangsang, NA J C. Mining semantic patterns for sentiment analysis of product reviews[C]//Proceedings of the 21st International Conference on Theory and Practice of Digital Libraries Research and Advanced Technology for Digital Libraries. Thessaloniki, Greece, 2017: 382–393.

[4] PENG Honggang, ZHANG Hongyu, WANG Jianqiang. Cloud decision support model for selecting hotels on TripAdvisor.com with probabilistic linguistic information[J]. International journal of hospitality management, 2018, 68: 124–138.

[5] GAVILAN D, AVELLO M, MARTINEZ-NAVARRO G. The influence of online ratings and reviews on hotel booking consideration[J]. Tourism management, 2018, 66: 53–61.

[6] XIE K L, ZHANG Zili, ZHANG Ziqiong. The business value of online consumer reviews and management response to hotel performance[J]. International journal of hospitality management, 2014, 43: 1–12.

[7] LIU Bing. Sentiment analysis and opinion mining[J]. Synthesis lectures on human language technologies, 2012, 5(1): 1–16.

[8] RAVI K, RAVI V. A survey on opinion mining and sentiment analysis[J]. Knowledge-based systems, 2015, 89(C): 14–46.

[9] GUELLIL I, BOUKHALFA K. Social big data mining: a survey focused on opinion mining and sentiments analysis[C]//Proceedings of the 12th International Symposium on Programming and Systems. Algiers, Algeria, 2015: 1–10.

[10] RANA T A, CHEAH Y N. Aspect extraction in sentiment analysis: comparative analysis and survey[J]. Artificial intelligence review, 2016, 46(4): 459–483.

[11] SUN Shiliang, LUO Chen, CHEN Junyu. A review of natural language processing techniques for opinion mining systems[J]. Information fusion, 2017, 36: 10–25.

[12] 李建华, 刘功申, 林祥. 情感倾向性分析及应用研究综述[J]. 信息安全学报, 2017, 2(2): 48–62.
LI Jianhua, LIU Gongshen, LIN Xiang. Survey on sentiment orientation analysis and its applications[J]. Journal of cyber security, 2017, 2(2): 48–62.

[13] 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, 29(2): 417–441.
HAN Zhongming, LI Mengqi, LIU Wen, et al. Survey of studies on aspect-based opinion mining of internet[J]. Journal of software, 2018, 29(2): 417–441.

[14] YU Zheng, WANG Haixun, LIN Xuemin, et al. Understanding short texts through semantic enrichment and hashing[J]. IEEE transactions on knowledge and data engineering, 2016, 28(2): 566–579.

[15] 王仲远, 程健鹏, 王海勋, 等. 短文本理解研究[J]. 计算机研究与发展, 2016, 53(2): 262–269.
WANG Zhongyuan, CHENG Jianpeng, WANG Haixun, et al. Short text understanding: a survey[J]. Journal of computer research and development, 2016, 53(2): 262–269.

[16] 常耀成, 张宇翔, 王红, 等. 特征驱动的关键词提取算法综述[J]. 软件学报, 2018, 29(7): 2046–2070.
CHANG Yaocheng, ZHANG Yuxiang, WANG Hong, et al. Features Oriented survey of state-of-the-art keyphrase extraction algorithms[J]. Journal of software, 2018, 29(7): 2046–2070.

[17] 赵京胜, 朱巧明, 周国栋, 等. 自动关键词抽取研究综述[J]. 软件学报, 2017, 28(9): 2431–2449.
ZHAO Jingsheng, ZHU Qiaoming, ZHOU Guodong, et al. Review of research in automatic keyword extraction[J]. Journal of software, 2017, 28(9): 2431–2449.

[18] 杜政霖, 李云. 基于特征聚类集成技术的在线特征选择[J]. 计算机应用, 2017, 37(3): 866–870.
DU Zhenglin, LI Yun. Online feature selection based on feature clustering ensemble technology[J]. Journal of computer applications, 2017, 37(3): 866–870.

[19] 王斌会. 多元统计分析及 R 语言建模[M]. 4 版. 暨南大学出版社, 2016: 159–181.
WANG Binhui. Multivariate statistical analysis and modeling for R language[M]. 4th ed. Jinan University Press, 2016: 159–181.

[20] FANG Lei, LIU Biao, HUANG Minlie. Leveraging large data with weak supervision for joint feature and opinion word extraction[J]. Journal of computer science and technology, 2015, 30(4): 903–916.

作者简介：

秦海菲，女，1980 年生，副教授，主要研究方向为数据库、数据仓库、数据挖掘。

杜军平，女，1963 年生，教授，博士生导师，主要研究方向为人工智能、社交网络分析、数据挖掘、运动图像处理，主持国家“863”、“973”计划项目、国家自然科学基金重点项目、国家自然科学基金重大国际合作项目、北京市自然科学基金重点项目等多项，发表学术论文多篇。