基于在线评论词向量表征的产品属性提取

李良强, 袁 华, 叶 开, 钱 宇*, 唐小我
(电子科技大学经济与管理学院, 四川 成都 611731)

摘要: 在线评论中蕴含的产品信息具有很高的电子商务应用价值. 但是, 与之相关的文本挖掘工作, 常常会面临着特征抽取以及对特征属性进行归类等问题的挑战. 基于词向量模型在表达词语的情景语义方面的优势, 提出了一种结合词向量表征和 K-means 聚类相结合的半监督方法, 用于从海量在线文本中高效挖掘出用户评论的特征, 并进一步按照这些特征的语义提取出它们的归类信息. 在真实数据集上的实验结果表明, 提出的方法可有效应用于海量在线评论中的文本属性提取工作; 与经典模型相比, 本方法从特征中提取的归类属性信息能更好地呈现出评论者表达的语义.

关键词: 在线评论; 特征抽取; 属性归类; 词向量; 聚类
中图分类号: TP273 **文献标识码**: A **文章编号**: 1000−5781(2018)05−0687−11

doi: 10.13383/j.cnki.jse.2018.05.011

Extraction product features from online reviews based on word-vector-representation

Li Liangqiang, Yuan Hua, Ye Kai, Qian Yu*, Tang Xiaowo
(School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract: The implicit product information in online reviews are of great value in e-commerce. However, text mining on online reviews often faces the challenges of feature extraction and classification of feature attributes. Referring to the advantages of word vector model in semantic expression, a semi-supervised method that combines word vector representation and K-means clustering is proposed, in order to efficiently extract reviewed features from massive online text and categorize the extracted features in terms of their semantics. The experimental results on a real data set show that the proposed method can efficiently extract text attributes from massive online reviews. Moreover, compared with the classical models, the categorization attributes extracted by the proposed method can better present the reviewers' semantics.

Key words: online review; feature extraction; aspect summarization; word vector; clustering

1 引 言

随着 Web 2.0 技术的广泛采用, 消费者很容易通过互联网发表和分享他们对于产品, 服务以及公司等方面的看法. 这些评论都以文本的方式在线发布, 其中蕴含的信息具有很高的电子商务应用价值, 如形成网络

收稿日期: 2015−09−18; 修订日期: 2016−02−25.
基金项目: 国家自然科学基金资助项目(71271044; U1233118; 71490720; 71572029).
*通信作者

口碑, 提升卖家和厂商的声誉[1,2]以及帮助企业进行商务策略的设计等[3]. 这使得从在线评论中抽取用户评论的特征, 以及对这些特征进行属性归类成为文本分析领域中的一个热门研究话题[4,5].

面向网络评论内容的属性抽取, 旨在从客户评论中挖掘出备受关注的特征(属性)信息, 并且总结基于这些特征的观点[6]. 在以往的研究中, 学者们提出了各种各样的特征抽取方法. 比较典型的有人工标注方法[7]; 从名词和名词短语中抽取[4]; 从依赖(搭配)关系进行抽取[8−11]; 使用机器学习方法进行抽取[12−14]以及使用话题模型进行特征抽取[15,16]. 在文本分析研究的初期阶段, 人工标注是一个非常精确的属性抽取方法, 但是它的效率低下[7]. 另外, 在利用名词和名词短语进行特征抽取方面的研究, Hu 等[4] 通过在名词候选特征词中引入关联规则挖掘的方法, 做出了开创性的工作. 后续出现很多该方法的改进研究[17,18]: Popescu 等[19]抽取评论中频繁出现的名词和名词短语作为候选产品特征, 同时借助搜索引擎计算互信息值来对候选特征进行评估. Li 等[20]提出了基于频繁名词和名词短语的特征抽取方法. 李实等[6]改进了关联规则并应用于中文评论中的产品特征挖掘, 也取得了较好的效果.

如果抽取出的特征很多, 那么对这些特征词进行归类可为用户提供更为具体和有价值的信息. Carenini 等[21]使用 WordNet 获得的词语相似性矩阵来进行特征归类. Guo 等[22]提出了mLSA 无监督算法. Zhai 等[14]提出了一种半监督的 SC-EM 算法进行特征归类, 并通过实验证明了该算法的可行性和优异性. 杨源等[23]在 SC-EM 算法上进行了改进, 用权重标准化 SimRank 计算不同特征之间的相似度, 得到了更好的分类结果. 这些研究都是沿着先度量特征词的相似性, 然后基于相似程度实施归类或者融合[24]的方法展开. 近年来, 由于话题模型(topic model)能同时识别出文本中描述产品特征的词语, 并在一定程度上对语义相近的词语进行聚类. 因此, 众多学者开始引入话题模型用于文本中的特征挖掘[5,15,16,25], 特别是针对产品属性特征的挖掘[13,26,27]和社会媒体中的话题特征抽取[28].

可以看出, 以往的方法都利用了词的共现和相似关系. 但是, 在通常的电商评价环境中, 不同背景的评论者都可以按照自己的想法发布评价内容[29]. 因此, 在这样生成的海量数据中抽取用户评论的产品属性将面临两方面挑战. 首先, 评论者形式各异的写作习惯, 决定了评论文本多样化的句法选择以及句子长度等. 其次, 不同的评论者针对同一个特征使用的表达方式或词语内容可能会不一样[23]. 例如在某些评论情境下,“造型”和“形状”都会指向手机的“外观”属性. 甚至错字和别字, 如“苹果”和“平果”也是指向同一个评论对象. 在通常的语境中, 这些特征词应该被归为同一个属性类别. 在以往的研究中, 按照原始词义相似性进行特征归类能够部分地解决这个问题[21,24]. 但是, 在线评论中的很多情景语义相似的特征词其原始词义并不一定相同或相似, 例如手机的“ 形状”和“外观设计”. 在用户评论用词模式具有随意性和多样(稀疏)性的情况下, 简单使用传统的方法不但会降低特征抽取的准确度, 而且使得归类后的同组特征词在语义的理解上非常困难. 因此, 需要更综合全面的方法来解决这个问题. 针对上述问题, 本文提出一种基于词向量表征的产品属性抽取方法, 将着力于提高抽取结果的准确性和归类结果的可理解性.

本文首先在文本处理过程引入词向量的神经语言模型[30]. 该模型可以有效地将文本中的词单元训练成高维空间上的向量, 这些向量的取值同时考虑了词在语料库中的分布和情景语义关系. 其次提出一个由表征词向量和 K-means 聚类相结合的方法进行产品属性归类. 该方法无论在海量文本的属性抽取, 还是在归类属性的语义保持上都有较好的表现.

2 产品属性抽取方法

图 1 给出了本文研究方法的整体框架. 该框架主要包括数据处理, 词向量训练和特征归类等工作. 数据预处理先从网页中抽取出评论文本数据集 T, 接着将文本内容进行分词等处理得到语料库 D, 最后使用语言模型对语料库中的数据进行训练, 得到所有词汇的表征向量集 $V(D)$. 而特征归类的任务是先从 D 中分离出潜在特征词集合 F, 同时从表征向量集 $V(D)$ 中得到这些词的表征向量值 $V(F)$, 并利用其将特征归类成组.

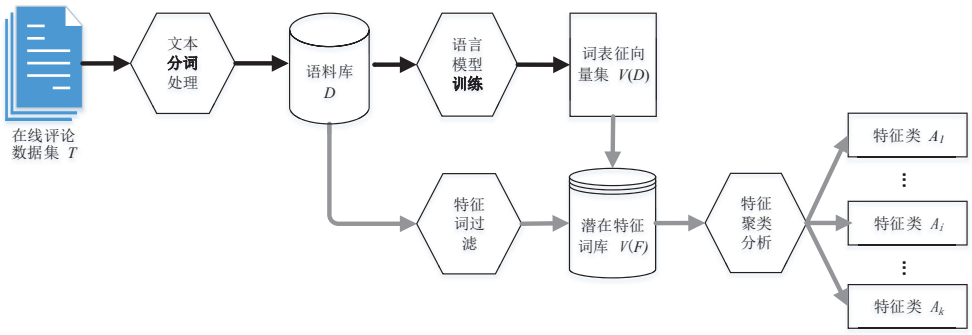


图 1　研究框架

Fig. 1　The research framework

1) 文本预处理

首先, 系统使用爬虫程序从电子商务网站中抓取出消费者的在线评论数据集 T. 然后, 再对评论文本进行必要的分句, 分词以及去除无意义符号等处理[31]. 最后, 数据集中第 i 个评论文本 $S_i, i = 1, 2, ..., |T|$ 被表示成多个有序词元素 s_{ij} 的集合

$$S_i = \{s_{i1}, s_{i2}, ..., s_{ij}, ...\}. \tag{1}$$

2) 表征词向量训练模型

分词之后, 为了实现特征词之间关系的可计算性, 将词元素用向量表示是文本分析中的重要方法. 文本分析中最常见的词向量是 one-hot representation 方法. 这种表示方法一个最大的问题是无法捕捉词与词之间的相似度; 此外, 还容易发生维数灾难[32]. Hinton[33]在 1986 年提出了一种 distributed representation 的词向量表示方法, 其基本思想是将语料库中每个词 s_j 映射到一个 K 维实数向量空间中去. 其中每个词在向量空间中的位置可以通过优化或者近似一个定义在原始文本上的目标函数来得到. 例如最大化某个词与其邻居词汇在同一句子中出现的似然概率. 这个优化过程被称为“训练”.

通过神经网络机器学习算法来获得词表征向量是文本挖掘领域的最新研究内容. 其中 Word2Vec[34]是最受关注的研究成果之一. Word2Vec 有 CBOW 模型和 Skim-gram 模型两种训练方法: 对每一个词 s_j, CBOW 模型是用其周围的词来预测 s_j 出现的概率; 而 Skim-gram 模型是用 s_j 来预测其周围词出现的概率. 一般地, 用于训练的神经网络模型有三层结构: 输入层, 隐藏层和输出层. 但是输入层和传统神经网络模型不同, 其输入的每一个节点单元是一个维度为 K 的向量, 而且该向量的每一个值均为变量, 训练过程中要对其进行更新. 训练达到稳定状态时, 这个向量就是词所对应的表征向量. 词向量训练过程如下:

(a) 将语料里的所有词随机初始化为 K 维向量;

(b) 选一个适当的窗口值作为语境(context); 输入层读入预测词附近窗口内的词, 并将它们的向量叠加;

(c) 输出层是一个巨大的二叉树, 叶节点代表所有的词. 对某个词 s_j, 给定语境, 训练的目标是使得预测词 s_j 的二进制编码概率最大.

基于此, 人们可以从大量未标注的普通文本数据中无监督地训练出词向量. 影响训练模型复杂性的因素包括输入词数, 神经网络的隐藏层节点数, 语料库词数和词向量维度等[35]. 经过神经语言模型训练后, 语料库中的词元素 s_j 可以被表征为一个 K 维的数字向量(该向量也称为 s_j 的表征向量)

$$\boldsymbol{s}_j = (s_{j_1}, s_{j_2}, ..., s_{j_K}). \tag{2}$$

词向量训练是在语料库中对聚合全局词与词共现的统计基础上进行的, 特别是在指定语料里一个词与其它词共现的频率关系[30]. 因此, 这类方法在寻找相关词和同义词工作中具有相当的优势.

3) 特征词过滤

从海量文档中挖掘出用户评论的特征词, 频繁项集挖掘是最通用且简单的方法. 但是, 如果直接应用该方法于在线评论的特征词挖掘, 则会受到数据集和方法本身的限制.

数据集的限制来自于两方面, 其一是在线评论文本中, 并不是所有的名词都是用来描述实体(特征); 其二是很多特征是所谓的隐含特征[36], 典型的评论如"(外观)非常漂亮! (系统)运行流畅."中用户评价的产品"外观"和"系统"特征就没有直接表述出来. 对于数据集的问题, 考虑到名词和名词短语在特征抽取中的重要作用[4], 以名词在评论文本中的搭配模式为基础进行分析是一个可行的思路. 用户在线评论的典型文本内容主要涉及到评价对象名词和动名词两类名词的搭配模式. 对象名词一般和观点词搭配较多, 例如"屏幕漂亮". 而动名词则和短语动词搭配较多, 例如"用来打游戏不错". 对于搭配模式的识别, 需要人工标注结果作为先验知识. 如果完全使用机器学习的方法, 则容易受到分词效果的影响.

频繁集挖掘方法本身的限制主要是阈值的设定. 如果阈值太大, 所保留的特征词信息有限; 如果阈值太小, 则计算复杂度相对较高, 而且低频词中也包含着较多的噪音. 所以对于方法的限制, 可以通过设定不同的阈值来实施探索计算, 并结合最终抽取效果评价的方法来寻找合适的阈值.

4) **特征词聚类**

聚类方法可分为: 基于划分方法, 基于层次方法以及基于密度方法等. 也有将混沌社会演化算法[37]用于文本聚类. 常用的 K-means 是划分聚类方法的代表之一. 但是它需要先给定聚类数目, 这限制了使用的灵活性. 而基于层次和基于密度的聚类算法除了计算复杂外, 在确定聚类数目时也需要额外的计算开销.

本研究选择二分 K-means 方法作为聚类算法, 因为它在大数据集中实施简单, 计算速度快, 并且可依据循环计算过程中聚类效果的变化确定合理的聚类数目. 研究表明, 二分 K-means 具有与层次方法相同的聚类质量, 且其时间复杂度优于层次聚类[38]. 为了减少人为干扰, 每次迭代计算过程中可使用词向量 $\boldsymbol{s}_j$ 和簇 C_i 的质心 $\boldsymbol{c}_i$ 之间的余弦相似性总和(sum of cosine similarity, SCS)来测量聚类效果

$$\mathrm{SCS}=\sum_{i=1}^{k}\sum_{j=1}^{|F|}\cos(\boldsymbol{c}_i,\boldsymbol{s}_j). \tag{3}$$

计算过程见算法 1. 其中 3∼7 行进行特征词过滤, 8∼16 行用二分 K-means 对向量表征的词进行聚类.

算法 1 (特征词聚类算法)

输入: 语料 D, 表征向量集合 $V(D)$和最小频繁度阈值 min_supp
输出: 聚类 C
初始化一个集合用于缓存满足条件的词 $\mathcal{S}\leftarrow\phi$
  for $s_j\in D$ do
    如果 s_j 是满足搭配模式的名词, 则 $\mathcal{S}\leftarrow s_j$
  end for
应用挖掘算法和频繁度阈值 min_supp 获得 $\mathcal{S}$ 中的 1-项频繁集 $F=FP^{(1)}(\mathcal{S})$
从 $V(D)$ 中抽取出 F 中所有词元素的表征向量集合 $V(F)\subset V(D)$
初始化簇表 C, 其中 $V(F)$ 的所有点形成一个簇
从 C 中选出一个簇, 对其进行二分试验
  for $i=1$ to 试验次数 do
    $k\leftarrow 2$, 使用基本 K-means 方法对选定的簇进行聚类
  end for
从二分试验中选择具有最小 SCS 的两个簇, 将其加入簇表 C 中, 直到 C 中包含 k 个簇
输出 C

算法 1 由模式匹配和二分 K-means 聚类构成. 其中, 模式匹配的时间复杂度为 $O(|D|\times$ 模式集大小), 二分 K-means 的时间复杂度为 $O(|V(F)|)$. 在海量文本数据库中, 模式集的大小可以控制在 1 000 以内, 而 $|D|$ 的值一般在百万级以上, 是本方法复杂性的主要影响因素.

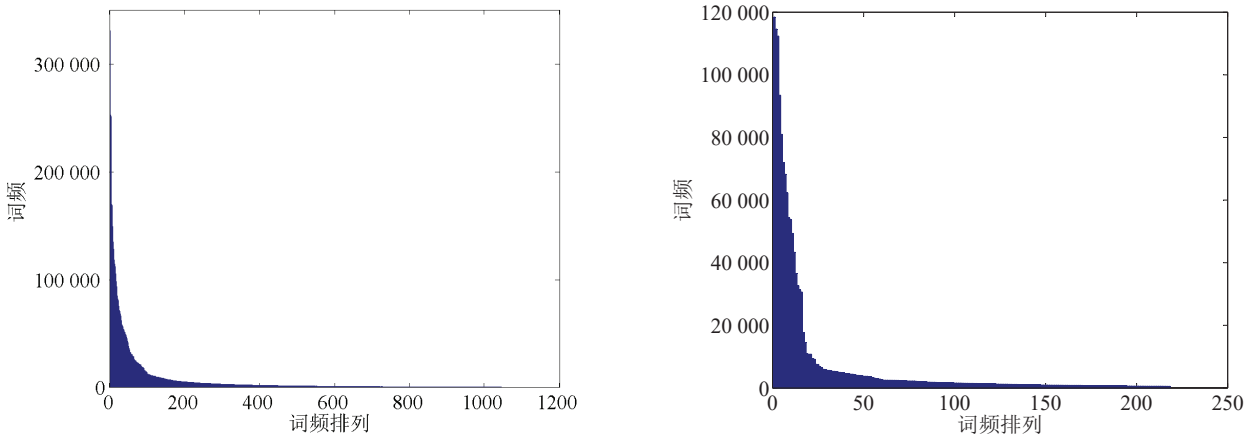
3 实验结果与分析

3.1 数据描述

本文使用的评论数据抓取自 B2C 商业购物平台京东商城(JD.com). 使用 Python 编写网络爬虫工具, 抓

取了 71 种手机产品从 2013–01～2015–01 共计 487 818 条在线评论. 其中最长的评论为 1 265 个汉字, 最短的为 2 个汉字, 平均长度 33.8 个汉字. 统计数据表明用户在 JD.com 发布了大量的评论, 但是大多是短文本评论, 长评论并不多见.

数据预处理过程中, 系统先对多句构成的评论文本按照标点位置进行整句截断; 随后引入分词程序把文本分割成基本的词元素. 抓取的全部评论中总共使用了 54 850 个词元素, 其中名词为 26 010 个. 把分词后的所有词和名词按照其词频分别排序, 其分布结果见图 2. 可以发现 JD.com 中的在线评论使用的词语分布很稀疏. 在这类数据集中实现特征抽取和属性归类, 对抽取和分类算法都是一个巨大的挑战.



(a) 全部词的词频分布

(a) The frequency distribution of all words

(b)　名词的词频分布

(b) The frequency distribution of nouns

图 2　JD.com 在线评论中的用词分布

Fig. 2　The frequency distribution of words on JD.com

3.2　特征词模式及过滤

在特征词过滤阶段, 首先处理了数据集的问题. 对于数据集的第一种限制, 系统需要从语料库 D 过滤出一些代表性的名词作为用户在线评论中的候选特征词[4]. 在候选特征词的筛选过程中, 结合用户的写作习惯, 先采用“名词+观点词”的搭配模式来过滤评论对象名词. 有两种典型模式: 一是“名词(+副词)+形容词”模式, 如“屏幕+漂亮”; 二是“(副词)+形容词+名词”模式, 如“ 很+难看+的+包装”. 进一步, 把“ 动词+动词+名词”模式中的后两者(如“用来+ 看+电影”中的“看电影”), 以及“动词+名词+观点词”的前两者(如“听+音乐+不错”中的“ 听音乐”)挑选出来, 并将那些在全部语料中出现频率较高的组合识别为动名词. 对于第二种限制, 经过对典型用户在线评论的内容分析发现, 用户之所以会隐含评价特征, 是因为其评价的对象是众所周知的, 不会引起理解混乱. 这类特征一般都是高频特征, 无需特别处理.

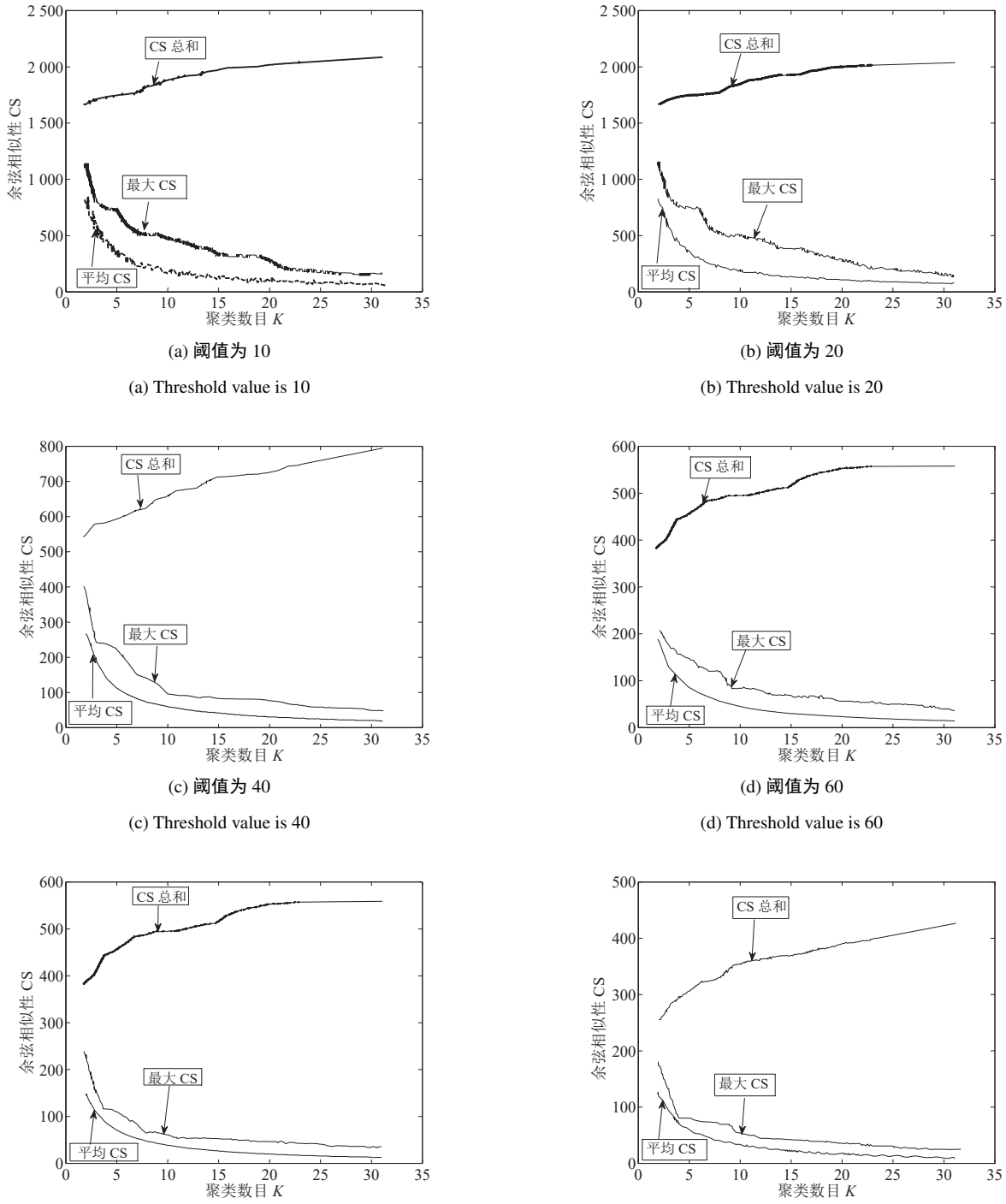
在传统的频繁集挖掘工作中, 阈值的设定依赖于专家的经验和问题的管理需求. 在缺乏先验知识的情况下, 分别计算了最小频繁度阈值(min_supp)为10 (0.002 %), 20 (0.004 %), 40 (0.008 %), 60 (0.012 %), 80 (0.016 %) 和 100 (0.020 %) 的情况下的挖掘效果. 这里主要对比了不同阈值情况下, 最大 CS(largest CS), 平均 CS(average CS)以及 CS 总和(sum of CS) 的三类聚类指标的变化, 结果见图 3. 通过比较三项指标结果, 最终选用 60 (0.012 %) 作为频繁模式挖掘过程中的支持度阈值.

3.3　表征词向量训练

已有人提出深度神经网络和递归神经网络等可用于学习词的表征向量. 但是这些方法最主要的问题是需要花很长的时间来训练模型. 在本实验中, 使用 Google 开发的开源工具软件 Word2Vector[1] 来进行词向量训练. Word2Vector 是一种用于高效学习海量文本中词的分布式表示的神经网络实现, 训练输出的结果是一个词表, 其中的每个词由一个向量来表示. 由于 Word2Vector 可以在不需要人工干预的情况下创建特征(包

[1] http://code.google.com/p/word2vec/

括词的上下文特征). 因此, 如果有足够多的数据, Word2Vector 能够基于一个词在语料中的出现情况, 高度精确地预测它的词义. 作为一种工程化方法, Word2Vector 的训练过程也需要考虑一些参数的影响.



(a) 阈值为 10

(a) Threshold value is 10

(b) 阈值为 20

(b) Threshold value is 20

(c) 阈值为 40

(c) Threshold value is 40

(d) 阈值为 60

(d) Threshold value is 60

(e) 阈值为 80

(e) Threshold value is 80

(f) 阈值为 100

(f) Threshold value is 100

图 3 不同阈值下余弦相似性(CS)变化趋势

Fig. 3 Trends of CS Under different thresholds

向量维度 K: 向量维度对训练结果有较大影响. 但是在 Word2Vec 模型分析中并没有对其给出理论上的优化结果[35]. 一般建议维度在 50 以上, 考虑到训练的时间成本问题普遍认为 100~300 较好.

语境窗口大小: 神经语言模型中要充分考虑到情景的因素, 它决定了与某个核心词相关的周围词的数量. Cui 等[39]的实验结果发现高阶 N-gram 能提高文本(情感)分类的准确率. 但是, 考察真实文本句子的表达发现, 训练效果并不是窗口越大越好, 较大的窗口也可能以更大的概率引入噪音.

本研究中 Word2Vector 工具的参数都是在经典文献推荐的范围内进行探索计算, 最终参数的设定采用在计算的时间成本和属性抽取结果之间一个较平衡的值. 其中, 维度参数 K 设置成 100. 结合实验数据的文本平均长度, 语境窗口大小设置为 5.

3.4 特征词聚类

利用 SCS 的变化趋势来确定一个合适的聚类数目 k. 当余弦距离和的测量值变化率趋弱(稳定)时, 聚类过程可以结束. 图 3 的结果显示: 当阈值在 40 以上时, SCS 的变化趋势比较相似. 因此, 本试验采用阈值为 60 时的结果, 对应的聚类数目 k 根据 SCS 的变化趋势(图3(d))以及决策需要可以取值在 10∼20 之间.

表 1 展示了 $k = 10$ 时, 属性归类结果以及与这些类相关的代表性属性特征词(同一个类中的词, 按照词频降序排序). 从表 1 中可以看到同一个簇类抽取出来的特征词与其它簇类的词相比具有更加紧密的语义相似性. 这里的语义相似性计算是基于上下文情境的. 这样, 就优于那些仅仅依靠特征词的原始语义相似性的计算结果. 换句话说, 表 1 的聚类结果具有较好的可理解性.

表 1　抽取出的属性及其特征词

Table 1　Extracted attributes and the associated feature words

外观	摄像头	声音	系统	电池
漂亮(143 468)	分辨率(49 336)	字体(80 975)	系统(78 945)	电池(71 902)
屏幕(118 447)	像素(5 330)	通话(75 524)	信号(62 326)	时间(68 213)
外观(112 388)	效果(4 971)	铃声(32 719)	软件(36 442)	待机(31 214)
手感(5 547)	摄像头(2 225)	声音(31 339)	玩游戏(5 317)	垃圾(5 631)
大气(4 266)	照片(1 121)	电话(5 861)	游戏(5 307)	电(4 375)
习惯(3 933)	续航(869)	按键(4 860)	死机(3 113)	耳机(2 934)
想象(3 060)	色彩(814)	音量(2 440)	开机(2 560)	缺点(2 426)
屏(2 425)	清晰度(638)	听筒(2 023)	视频(1 818)	情况(1 993)
样子(2 092)	镜头(154)	音乐(1 065)	2g(82)	小时(1 751)
后盖(1 634)	画质(154)	太小(960)	自带(77)	长度(1 557)
京东	品牌	价格和质量	赠品和套装	(产品)使用者
京东(27 915)	国产(59 273)	质量(93 482)	配件(2 516)	老人(43 305)
评价(9 535)	国产机(5 3691)	价格(22 154)	评(926)	朋友(14 404)
物流(9 533)	华为(6 402)	东西(21 556)	原装(899)	老爸(10 978)
货(9 394)	诺基亚(5 104)	价格便宜(15 868)	赠品(852)	妈妈(10 772)
商品(5 590)	国货(4 543)	价钱(5 173)	手机套(352)	老妈(8 226)
购物(5 056)	产品(4 541)	值(3 777)	皮套(303)	同事(6 790)
发货(4 981)	信赖(4 156)	降价(3 166)	礼品(202)	爸爸(4 912)
客服(3 383)	酷派(3 931)	宝贝(2 240)	话费(119)	老人家(3 982)
下单(1 221)	品牌(3 712)	体验(1 894)	套装(117)	老婆(2 891)
售后(535)	荣耀(444)	实体店(882)	宝(74)	老公(1 432)

3.5 结果评价

本文提出的词表征向量聚类方法(K-means +Word2Vec)与三种典型的在线评论特征挖掘方法 LDA[7], s-LDA[15] 和 HLDA[16]进行了比较. 在实验过程中, 从整个语料库中随机抽取 10 % 的评论作为测试集. 困惑度(perplexity)和宏平均准确度(macro average accuracy rate, MAAR) 被用来作为衡量特征总结的效果指标.

1) 困惑度

困惑度指标在自然语言处理中用来衡量训练出的语言模型的好坏[7]. 如果 T 是测试集, 则困惑度计算公式为

$$\text{Perplexity} = \exp\left(-\sum_{d=1}^{|T|}\ln p(W_d)\left(\sum_{d=1}^{|T|}N_d\right)^{-1}\right), \tag{4}$$

其中 W_d 表示测试集中的词语, N_d 表示文本 d 的长度, $p(W_d) = \sum_{i=1}^{k} (p(Z_i|T)p(W_d|Z_i))$ 是词语 W_d 生成文档 d 的概率(Z_i 是训练好并概率化的第 i簇).

困惑度能够在不需要人工干预的情况下对词聚类的结果进行有效的测量. 通过图 4 的结果, 可以推断, 在同样话题数目的情况下, 本文提出的方法的困惑度要小于 LDA , s-LDA, 和 HLDA 这三个模型. 因此, 本文提出的方法优于话题模型.

2) 平均准确度

如果用 a 表示正确分配的数目, b 为错误分配的数目, 则准确率为

$$p = a/(a+b). \tag{5}$$

假设共有 $|C|$ 个簇, p_j 为第 j 簇的正确率. 为了正确计算各簇中的特征词的正确数目, 宏平均准确率(MAAR)被引入到本实验中, 即

$$\mathrm{MAAR} = \frac{1}{|C|}\sum_{j=1}^{|C|} p_j. \tag{6}$$

在 Top5, Top10, Top15, Top20 和 Top25 这几个水平上, 通过使用配对 t–检验对本文提出的方法及基准方法的 MAAR 进行比较, 表2中展示了不同方法的宏平均准确率. 结果显示本文提出的方法在 MAAR 指标上显著大于其它三个用于比较的基准方法(统计显著性指标 p 值远小于 0.05).

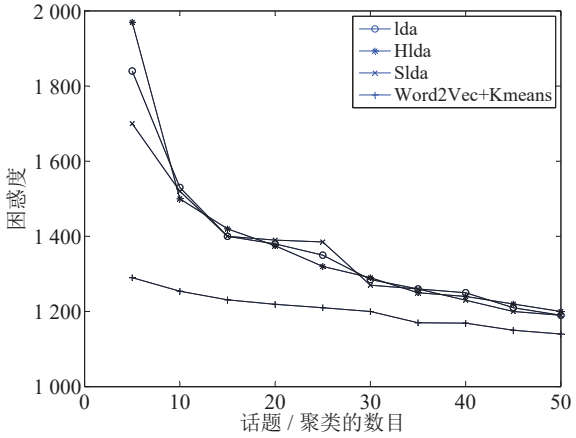


图 4 困惑度比较

Fig. 4 Comparison of perplexity

表 2 宏平均准确度结果比较

Table 2 Average MAAR

	Top5	Top10	Top15	Top20	Top25
K-means + Word2vec	0.828 5	0.735 7	0.714 3	0.672 7	0.683 3
LDA	0.723 6	0.684 2	0.647 3	0.521 7	0.563 6
HLDA	0.628 1	0.598 3	0.513 5	0.416 2	0.474 3
sLDA	0.633 4	0.595 2	0.532 3	0.427 7	0.488 2

综合关于困惑度和 MAAR 这两个指标的实验结果, 可以看出本文提出的方法无论是在评论特征抽取还是特征聚类方面都具有更好的表现.

4 结束语

Web2.0 的发展使得消费者能通过互联网渠道发表在线评论, 这些评论中蕴含着丰富的价值, 在电子商务活动中起着重要的作用. 要更好地利用在线评论中的隐藏价值, 文本特征抽取与属性归类是基础性研究

工作. 但是, 由于在线评论用词的稀疏性和表达方式的自由性, 经典的特征抽取方法在抽取时往往会丢掉词序和语义等方面的内容.

基于词向量模型在词语的序列和语义表达方面的优势, 本文提出了一种结合词向量表征和二分 K-means 聚类的特征提取和属性归类方法. 该方法首先利用评论中名词的搭配关系形成特征词的候选集合, 进而引入同时考虑了语义特性和位置分布的词向量来表征这些候选词, 并用高效的聚类方法将其迅速归类. 为了检验方法的可行性, 实验抓取了真实 B2C 电商网站上近 49 万条用户生成的评论文本作为实验数据集. 在真实数据上的实验结果表明, 本文提出的方法能有效提升海量文本中属性抽取结果的准确性和可理解性. 同时, 与 LDA 话题模型及类似方法相比较, 本文提出的方法无论是在困惑度还是在宏平均准确度上都有更好的表现.

参考文献:

[1] Vilpponen A, Winter S, Sundqvist S. Electronic word-of-mouth in online environments. Journal of Interactive Advertising, 2006, 6(2): 8–77.

[2] Cheung C, Thadani D. The impact of electronic word-of-mouth communication: A literature analysis and integrative model. Decision Support Systems, 2012, 54(1): 461–470.

[3] 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. 系统工程学报, 2014, 29(4): 562–570.
Liu Y, Liao X W, Liu Y. The impact of online review on software and platform's pricing strategies. Journal of Systems Engineering, 2014, 29(4): 562–570. (in Chinese)

[4] Hu M, Liu B. Mining and summarizing customer reviews // Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2004: 168–177.

[5] Jo Y, Oh A. Aspect and sentiment unification model for online review analysis // Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York: ACM, 2011: 815–824.

[6] 李　实, 叶　强, 李一军, 等. 中文网络客户评论的产品特征挖掘方法研究. 管理科学学报, 2009, 12(2): 142–152.
Li S, Ye Q, Li Y J, et al. Mining features of products from Chinese customer online reviews. Journal of Management Sciences in China, 2009, 12(2): 142–152. (in Chinese)

[7] Blei D, Jordan M. Modeling annotated data // Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. New York: ACM, 2003: 127–134.

[8] Shi B, Chang K. Mining chinese reviews // The 6th IEEE International Conference on Data Mining - Workshops. Hongkong: IEEE, 2006: 585–589.

[9] Qiu G, Liu B, Bu J, et al. Opinion word expansion and target extraction through double propagation. Computational linguistics, 2011, 37(1): 9–27.

[10] 王祖辉, 姜　维, 李一军, 等. 在线评论情感分析中固定搭配特征提取方法研究. 管理工程学报, 2014, 28(4): 180–186.
Wang Z H, Jiang W, Li Y J. Regular collocation features extraction method in online reviews sentiment analysis. Journal of Industrial Engineering and Engineering Management, 2014, 28(4): 180–186. (in Chinese)

[11] Maharani W, Widyantoro D H, Khodra M L. Aspect extraction in customer reviews using syntactic pattern. Procedia Computer Science, 2015, 59: 244–253.

[12] Brody S, Elhadad N. An unsupervised aspect-sentiment model for online reviews // Human Language Technologies: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2010: 804–812.

[13] Titov I, McDonald R. Modeling online reviews with multi-grain topic models // Proceedings of the 17th International Conference on World Wide Web. New York: ACM, 2008: 111–120.

[14] Zhai Z, Liu B, Xu H, et al. Grouping product features using semi-supervised learning with soft-constraints // Proceedings of the 23rd International Conference on Computational Linguistics. Stroudsburg: ACL, 2010: 1272–1280.

[15] Lin C, He Y. Joint sentiment/topic model for sentiment analysis // Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM, 2009: 375–384.

[16] Kim S, Zhang J, Chen Z, et al. A hierarchical aspect-sentiment model for online reviews // Proceedings of The 27th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2013.

[17] Lau R Y K., Li C P, et al. Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. Decision Support Systems, 2014, 65: 80-94.

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

[19] Popescu A, Etzioni O. Extracting product features and opinions from reviews // Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2005: 339–346.

[20] Li S, Zhou L, Li Y. Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures. Information Processing & Management, 2015, 51(1): 58–67.

[21] Carenini G, Ng R, Zwart E. Extracting knowledge from evaluative text // Proceedings of the 3rd international conference on Knowledge capture. New York: ACM, 2005: 11–18.

[22] Guo H, Zhu H, Guo Z, et al. Product feature categorization with multilevel latent semantic association // Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM, 2009: 1087–1096.

[23] 杨 源, 马云龙, 林鸿飞. 评论挖掘中产品属性归类问题研究. 中文信息学报, 2012, 26(3): 104–108.
Yang Y, Ma Y L, Lin H F. Clustering product features in opinion mining. Journal of Chinese Information Processing, 2012, 26(3): 104–108. (in Chinese)

[24] 王 伟, 王洪伟, 孟 园. 协同过滤推荐算法研究: 考虑在线评论情感倾向. 系统工程理论与实践, 2014, 34(12): 3238–3249.
Wang W, Wang H W, Meng Y. The collaborative filtering recommendation based on sentiment analysis of online reviews. Systems Engineering: Theory & Practice, 2014, 34(12): 3238–3249. (in Chinese)

[25] Moghaddam S, Ester M. On the design of LDA models for aspect-based opinion mining // Proceedings of the 21st ACM international conference on Information and knowledge management. New York: ACM, 2012: 803-812.

[26] Ma B Z, Zhang D S, Yan Z J, et al. An Lda and synonym lexicon based approach to product feature Eextraction from online consumer product reviews. Journal of Electronic Commerce Research, 2013, 14(4): 304-314.

[27] 彭 云, 万常选, 江腾蛟, 等. 一种词聚类LDA的商品特征提取算法. 小型微型计算机系统, 2015, 36(7): 1458-1463.
Peng Y, Wan C X, Jiang T J, et al. An algorithm based on words clustering LDA for product aspects extraction. Journal of Chinese Computer Systems, 2015, 36(7): 1458-1463. (in Chinese)

[28] 曹丽娜, 唐锡晋. 基于主题模型的BBS话题演化趋势分析. 管理科学学报, 2014, 17(11): 109–121
Cao L N, Tang X J. Trends of BBS topics based on dynamic topic model. Journal of Management Sciences in China, 2014, 17(11): 109–121 (in Chinese)

[29] 程 葳, 钟 华, 孙娇华. 网络论坛中发帖行为复杂性研究. 系统工程学报, 2009, 24(4): 385–391.
Cheng W, Zhong H, Sun J H. Research on complexity of posts in network forums. Journal of Systems Engineering, 2009, 24(4): 385–391. (in Chinese)

[30] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model. Journal of Machine Learning Research, 2003, 3: 1137–1155.

[31] Gao J, Li M, Huang C, et al. Chinese word segmentation and named entity recognition: A pragmatic approach. Computational Linguistic, 2005, 31: 531–574.

[32] Turian J, Ratinov L, Bengio Y. Word representations: A simple and general method for semi-supervised learning // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2010: 384–394.

[33] Hinton G. Learning distributed representations of concepts // Proceedings of the 8th Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum, 1986: 1–12.

[34] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 2013, 26: 3111–3119.

[35] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation // Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014). Stroudsburg: ACL, 2014: 1532–1543.

[36] McAuley J, Leskovec J. Hidden factors and hidden topics: Understanding rating dimensions with review text // Proceedings of the 7th ACM Conference on Recommender Systems. New York: ACM, 2013: 165–172.

[37] 郝占刚, 王正欧. 基于混沌社会演化算法的文本聚类新方法. 系统工程学报, 2007, 22(1): 109–112.
Hao Z G, Wang Z O. New text clustering method based on chaotic social evolutionary programming algorithm. Journal of Systems Engineering, 2007, 22(1): 109–112. (in Chinese)

[38] Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques // Proceedings of the KDD-2000 Workshop on Text Mining. New York: ACM, 2000: 109–111.

[39] Cui H, Mittal V, Datar M. Comparative experiments on sentiment classification for online product reviews // Proceedings of the 21st National Conference on Artificial Intelligence. Palo Alto. 2006: 1265–1270.

作者简介:

李良强(1981—), 男, 四川广元人, 博士生, 研究方向: 商务智能, Email: langmalee@gmail.com

袁　华(1973—), 男, 四川达州人, 博士, 副教授, 研究方向: 电子商务, 商务智能, Email: yuanhua@uestc.edu.cn

叶　开(1990—), 男, 湖北荆州人, 硕士生, 研究方向: 数据挖掘, Email: tingxueye@gmail.com

钱　宇(1978—), 女, 重庆人, 博士, 副教授, 研究方向: 运作管理, 电子商务与信息经济学, Email: qiany@uestc.edu.cn;

唐小我(1955—), 男, 四川彭州人, 博士, 教授, 研究方向: 供应链与运作管理, Email: xwtang@uestc.edu.cn.

********** **********

(上接第 636 页)

[22] 金常飞, 曹二保, 赖明勇. 双寡头零售市场绿色营销演化博弈分析. 系统工程学报, 2012, 27(3): 383–389.
Jin C F, Cao E B, Lai M Y. Analysis on green marketing strategy of duopoly retailing market based on the evolutionary game theory. Journal of Systems Engineering, 2012, 27(3): 383–389. (in Chinese)

[23] 程永宏, 熊中楷. 碳标签制度下产品碳足迹与定价决策及协调. 系统工程学报, 2016, 31(3): 386–397.
Cheng Y H, Xiong Z K. Product carbon footprint and pricing decisions and coordination under carbon labeling system. Journal of Systems Engineering, 2016, 31(3): 386–397. (in Chinese)

[24] 王一雷, 朱庆华, 夏西强. 基于消费偏好的供应链上下游联合减排协调契约博弈模型. 系统工程学报, 2017, 32(2): 188–198.
Wang Y L, Zhu Q H, Xia X Q. Supply chain upstream and downstream joint coordination contract game model based on consumer preference. Journal of Systems Engineering, 2017, 32(2): 188–198. (in Chinese)

[25] Weber T, Neuhoff K. Carbon markets and technological innovation. Journal of Environmental Economics and Management, 2010, 60(2): 115–132.

作者简介:

张　盼(1989—), 男, 湖北黄冈人, 博士, 讲师, 研究方向: 低碳经济, 物流与供应链管理, Email: zhangpanyjs@cqu.edu.cn;

熊中楷(1948—), 男, 江西南昌人, 教授, 博士生导师, 研究方向: 运营管理, 低碳经济, Email: xiongzhongkai@cqu.edu.cn.