

# 面向竞争力的特征比较网络: 情感分析方法<sup>①</sup>

王伟<sup>1,2</sup>, 王洪伟<sup>2\*</sup>

(1. 华侨大学工商管理学院, 泉州 362021; 2. 同济大学经济与管理学院, 上海 200092)

摘要: 比较观点广泛存在于在线评论之中, 这些比较观点是用户表达情感的常用方式, 也是产品竞争力的重要表现. 依据在线评论的文本信息, 结合文本挖掘和情感分析技术, 提取在线评论中特征级的“比较观点对”. 利用特征比较观点, 分别构建产品特征的单边有向、双边有向以及多边有向比较网络, 并根据情感强度确定网络边的权重. 借助成熟的网络分析算法( PageRank 和 HITS) 对特征比较网络进行计算. 产品特征比较网络能够识别产品的竞争优势, 也能分析用户对每个特征的关注度, 并量化评价得分. 实验发现, 特征比较网络与销售排名显著相关, 因而有助于销量的预测.

关键词: 竞争力; 比较网络; 在线评论; 情感分析; 产品特征; 网络分析

中图分类号: TP18 文献标识码: A 文章编号: 1007-9807(2016)09-0109-18

## 0 引言

比较是人们认识事物的基本方法. 将多个事物加以对照, 可以发现它们在某些方面的相似或差异及其原因. 尤其在商业领域, 常言“不怕不识货, 就怕货比货”, 比较更是验证商品竞争力的重要手段. 消费者认识一个产品常借助于与其他产品比较来实现, 因为只有比较鉴别, 才能对产品的质量和性能有更加深入的认识. 对于商家来说, 通过比较能够体现出与其他产品的竞争优势和不足, 为市场营销策略以及产

品改进提供依据.

近年来, 社会化媒体日益普及, 为消费者发布自己对于产品的观点和体验提供了前所未有的平台, 其中相当一部分观点是以比较的形式表达的, 这为比较信息的获取提供了便利的渠道. 例如, 图 1 显示了中文评论中的特征级比较观点, 该评论比较了不同产品的价格、口碑以及防水性能等. 类似的, 图 2 展示了亚马逊网站上的一条典型的英文评论, 该评论中比较了 T4i 的图像分辨率比 20D 高; G1X 的体积比 T4i 小; T4i 的快门比 G1X 快等.

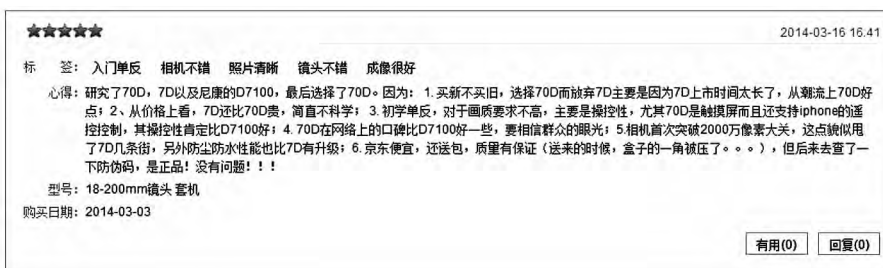


图 1 中文在线评论中的产品特征比较

Fig. 1 Feature-level comparison in Chinese online review

① 收稿日期: 2014-05-27; 修订日期: 2015-04-20.

基金项目: 国家自然科学基金资助项目(71371144); 上海市哲学社会科学规划课题一般资助项目(2013BGL004); 华侨大学高层次人才科研启动资金资助项目(16SKBS102).

通信作者: 王洪伟(1973—), 男, 辽宁人, 博士, 教授, 博士生导师. Email: hwwang@tongji.edu.cn

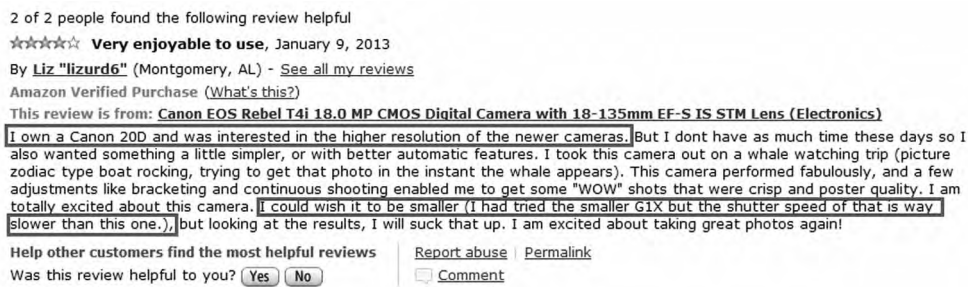


图2 英文在线评论中的产品特征比较

Fig. 2 Feature-level comparison in English online review

目前,从在线评论中提取比较信息的研究较少,尤其是细粒度的基于特征级的比较信息研究还未有系统性的解决方法.已有研究主要是以量化指标作为分析依据,例如总体评分、有用性投票等<sup>[1-4]</sup>.但是,总体评分难以度量文本中隐含的多维特征情感信息.因此,针对产品特征级的比较信息,进行细粒度情感分析,更有助于评论价值的挖掘.本文尝试识别特征级比较观点,并采用网络理论进行分析,研究成果将有助于商家识别产品竞争力,以及有助于消费者做出更为合理的购买决策.本文的贡献如下:1)针对海量在线评论,提出识别特征级比较观点的方法;2)根据特征级比较观点,构建特征比较网络,并应用网络分析方法,对特征比较网络进行计算;3)提出基于特征比较网络的实际应用:可视化产品优劣,显示产品竞争优势,便于潜在消费者对产品进行比较;同时,与同类产品对比的直观结果,有益于企业管理者在产品改进、市场营销、个性化广告等领域提供辅助决策.

## 1 文献综述

### 1.1 在线评论的特征提取和观点识别

在线评论中往往会对产品的各种特征进行评价,学者们就如何识别评论文本中的产品特征有了深入的研究.一般有几类方法:基于词性统计的方法、基于语法结构的方法、基于语言形态学的方法等<sup>[5]</sup>.最朴素的方法是统计语料中频繁出现的名词或其组合,频繁共现的词代表了对术语的规则描述并以此为模式识别特征<sup>[5]</sup>.基于规则的特征挖掘算法<sup>[6-7]</sup>根据构成句子的词汇规则,提取类似形容词修饰名词的句法结构,被修饰的名词作为候选产品特征,而形容词作为特征观点.基于规则的算法忽视了语言的使用习惯问题,因

此有研究在此基础上提出了具有语法格式的关联规则挖掘算法<sup>[8]</sup>.一般通过语料训练得到句子模型的特征长度、概率上下限和阈值等用于特征观点识别的模式<sup>[9]</sup>.机器学习算法首先进行分词和词性标注,然后人工标注部分语料作为训练集,对模型训练得到模型参数,最后在测试集上测试.比较典型的算法有:隐马尔可夫(hidden Markov model, HMM)模型<sup>[10]</sup>、条件随机场(conditional random fields, CRFs)模型<sup>[11]</sup>、以及支持向量机(support vector machine, SVM)模型<sup>[12]</sup>.在线评论的特征项会影响分类结果,研究发现,不同的语料对于语种、写作风格、文体、句法和内容区分度上存在敏感度差异<sup>[13]</sup>.

产品特征提取往往和情感分析联系在一起,识别某一具体特征的情感倾向<sup>[14]</sup>,处理后的结果是“特征-观点对”形式.对于不同类别的产品,特征提取的算法也不尽相同,例如酒店评论中会有设施、环境等特征,这显然与电子产品的特征不一样,因而需要专门提取不同领域的特征项<sup>[15-16]</sup>.产品特征提取会出现高维特征项,高维特征项本身不影响特征抽取的准确度,但是会对特征的计算产生较大偏差.例如:对于产品特征“颜色”和“色彩”,二者实际上描述的是同一特征,需要合并.一些研究提出了不同的特征项降维方法,如最大熵模型降维<sup>[17]</sup>、多阶段降维<sup>[18]</sup>、词语相似度降维<sup>[19-20]</sup>以及混合降维方法<sup>[21]</sup>.

### 1.2 细粒度情感分析研究

情感分析(sentiment analysis)也称意见挖掘(opinion mining),是利用文本挖掘技术,对在线评论进行语义分析,旨在识别用户的情感趋向是“高兴”还是“伤悲”,或判断用户的观点是“赞同”还是“反对”<sup>[22]</sup>.情感分析涉及多种技术,例如自然语言处理、机器学习、信息抽取等.因此情感分析涵盖多个研究任务,例如文本的主客观检

测<sup>[23]</sup>;不同粒度的情感分析<sup>[9, 24]</sup>;产品“特征-观点对”提取等<sup>[25-26]</sup>。

从分析粒度来看,情感分析包括粗粒度和细粒度两种分析方法。粗粒度分析是对文本的整体情感极性进行判断,包含基于模型的方法<sup>[27]</sup>、无监督<sup>[28-29]</sup>、半监督<sup>[30]</sup>以及监督机器学习方法<sup>[21, 32]</sup>。细粒度分析是对特征观点的词语级的情感极性和强度分析<sup>[33]</sup>,通常步骤是:首先计算词典中词语的原子极性,并根据该原子极性计算词语之间的相关性得出每个评价词的情感,再综合词语的情感得到句子的情感极性和强度,最后根据句子情感计算文本情感<sup>[34]</sup>。从技术方法来看,有两种方向:一种是基于语义分析<sup>[35-36]</sup>,另一种是基于机器学习<sup>[37-38]</sup>。而从应用上看,情感分析主要应用于评论效用分析<sup>[39-41]</sup>、金融市场辅助决策<sup>[42-43]</sup>以及社会舆情监督<sup>[44-45]</sup>等。

极性相同的词汇所表达的情感强度可能不同,例如“好”和“优秀”情感极性都是正面,但是后者比前者的强度大得多。已有研究专门探讨了词语的情感强度问题。值得指出的是 SentiStrength 情感强度识别系统<sup>[46]</sup>,该系统先识别非标准格式文本中可能的情感词,然后使用 SVM 进行情感极性分类和强度检测,在社会化文本中表现优异。此外,文本情感倾向也受到社会因素<sup>[47]</sup>以及句法结构的影响<sup>[48]</sup>。

### 1.3 面向在线评论的比较观点识别

绝大多数研究忽略了评论中的比较观点。事实上,人们在交流或者评论时,常常采取比较方法来凸显自己的观点,因此识别评论文本中的比较信息,能够真实地还原评论者的意图,比较信息能够体现出评论中细微的观点差别。

比较观点识别办法是构建一套词语序列规则,来作为比较观点表达方式的匹配模版。然后依次扫描评论文本中的句子,一旦出现与序列规则匹配的句子,就加入候选比较句子。通过 SVM 或朴素贝叶斯分类器对候选比较句进行识别,提取比较关系中的比较实体和特征,最后测定观点强度<sup>[49]</sup>。但是该模式面临两个问题:1) 同一句子可能既满足比较句子模式又满足非比较句子模式,该算法很难处理这类问题;2) 在线评论的文本格式相当不规范,甚至不符合语法标准,对这类自由格式文本的识别,该方法效果较差。典型的改进算法有 CRFs 算法<sup>[50]</sup>等。

有学者提出了一种简单但实用的方法:先识别领域内所有产品名称,构建产品名称列表,然后扫描评论,如果在产品  $p_1$  的评论中出现了产品  $p_2$  的名称,就认为存在  $p_1$  与  $p_2$  的比较关系。经过一轮完整的扫描,就可以完成所有产品之间的比较信息识别<sup>[51]</sup>。该方法可以快速识别产品间的比较观点。但遗憾的是,该研究只是产品级的比较,而没有深入到产品特征级的比较关系识别。

在应用方面,已有研究把评论中的比较信息应用于市场结构识别,采用文本挖掘和语义网络工具,模拟构建了轿车市场和保健品市场的市场结构。实验证明,同传统的市场销售和调研数据分析相比,通过在线评论比较信息构建的市场结构更有效<sup>[52]</sup>。

### 1.4 文献评述及问题定义

为了突显自己的观点,用户发表评论时,常采用比较的方式来表达。但是,现有情感分析大多假设被评论的对象是某一个特定的产品,侧重于该产品的特征与观点提取、情感分类等,忽略了评论中可能包含的多个产品之间的比较信息。更进一步,在线评论常常包含不同产品就某些特征上的比较观点,细粒度的识别这些关于产品特征的比较观点,有助于深度挖掘评论的商业价值。已有研究中,产品比较网络<sup>[51]</sup>与本文最为相近,该比较网络是面向产品级的,难以提取针对产品特征的比较观点。针对现有文献的不足,本文将利用情感分析技术,结合网络分析方法,构建产品特征比较信息的识别模型。

综合文献回顾,提出以下研究问题:

- 1) 比较观点对的识别;
- 2) 根据比较观点对构建比较网络并确定边的权重;
- 3) 依据节点和边的重要度指标,验证特征比较网络的有用性。

## 2 产品特征比较网络的构建方法

### 2.1 模型概述

分析框架如图3所示,分5步:1) 数据收集与预处理;2) 特征比较观点的抽取;3) 特征比较网络的构建以及权重设置;4) 依据网络理论,对特征比较网络进行权威度计算;5) 特征比较网络的验证。

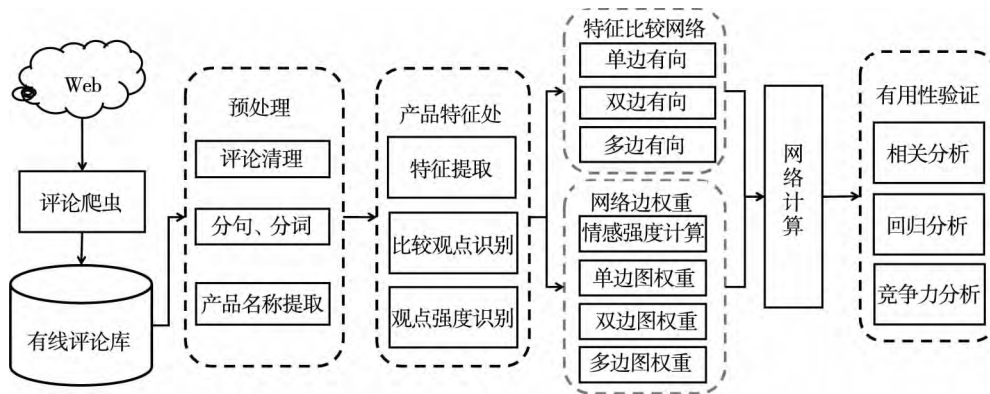


图3 产品特征比较网络的总体结构

Fig. 3 Overall structure of feature-level comparative network

### 2.2 特征比较观点的表示

在线评论包含的比较观点有2种形式:

- 1) 综合比较 例如“总体来说,X230比X220的表现出色”;
- 2) 特征比较 例如“但是,X230的键盘舒适性比X220差多了”。

第1句是综合比较,没有涉及产品特征;第2句通过与另一款产品(X220)的比较,表明特征“键盘”比“X220”差,其中又涉及比较强度,例如“差多了”就是比较强度。

将比较信息视为网络  $G = \langle V, E \rangle$ ,其中  $V$  为产品集,  $E$  为产品比较关系集。给定网络  $G = \langle V, E \rangle$ ,  $p_1, p_2$  为被比较的2个产品,  $p_1 \in V, p_2 \in V$ 。  $f_i$  为产品特征,  $f_i \in F, F$  为特征集。产品特征比较观点表示为  $\langle f_i, p_1, p_2, C \rangle, \langle f_i, p_1, p_2, C \rangle \in E$  称其为产品的比较观点对,指在  $p_1$  的评论中出现与  $p_2$  就特征  $f_i$  方面的比较,  $C$  代表特征比较的上下文。

针对每一个特征,将比较信息表示为特征比较数组  $t_i, t_i = \langle p_1, p_2, cScore \rangle$ ,其中  $i = 1, 2, \dots, m$ ,  $m$  为被比较特征数量;  $cScore$  表示该特征比较强度值,可由情感分类器计算得到,  $cScore$  正负均可,取决于  $p_1$  和  $p_2$  的情感比较方向。  $cScore > 0$  表示  $p_1$  比  $p_2$  的优势大,  $cScore < 0$  表示  $p_1$  比  $p_2$  的优势小。

如果出现3个或3个以上的产品特征比较,则分别抽取两两比较结果。

### 2.3 产品特征比较网络的构建方法

根据2.2节提取的特征比较观点,可以构建特征比较网络。假设提取出  $n$  个特征比较观点对,每个观点对都涉及2个产品  $p_i, p_j$  的比较,根据特征比较数组转换为有向边的方法<sup>[51]</sup>,考虑如下3种策略。

1) 单边有向图。基本思路是优势较强的节点输出权重到优势较弱的节点,所有关于特征  $f$  的  $n$  个数组合聚合成一条有向边。当  $\sum_{i=1}^n \{cScore_i\} > 0$  时,新建一条从节点  $p_2$  到  $p_1$  的边;当  $\sum_{i=1}^n \{cScore_i\} < 0$  时,新建一条从节点  $p_1$  到  $p_2$  的边;当  $\sum_{i=1}^n \{cScore_i\} = 0$  时,二者是并列关系,不存在连接节点  $p_1$  和  $p_2$  的边。

2) 双边有向图。基本思路是优势较强的节点输出权重到优势较弱的节点,依据特征数组情感方向,分为两类。当  $\sum_{i=1}^n \{ (cScore_i) \cdot I_{(cScore_i > 0)} \} > 0$  时,新建一条从  $p_2$  到  $p_1$  的边,称为类型I的边;当  $\sum_{i=1}^n \{ (cScore_i) \cdot II_{(cScore_i < 0)} \} < 0$  时,新建一条从  $p_1$  到  $p_2$  的边,称为类型II的边。换句话说,所有特征比较数组被分为两类,类型I是所有  $cScore_i > 0$  的特征比较数组,类型II是所有  $cScore_i < 0$  的特征比较数组。当  $cScore_i = 0$  时,不存在连接  $p_1$  和  $p_2$  的边。因此,双边有向图可能在节点  $p_1$  和  $p_2$  之间存在1条或者2条有向边(类型I,类型II,或者二者皆有)。

3) 多边有向图。基本思路是有多少个指向关系,就创建多少条边,不合并任何边。当  $\{cScore_i\} > 0$  时,新建一条从节点  $p_2$  到  $p_1$  的边;当  $\{cScore_i\} < 0$  时,新建一条从节点  $p_1$  到  $p_2$  的边;当  $\{cScore_i\} = 0$  时,不存在节点  $p_1$  与  $p_2$  之间的边。也就是,有向边的条数等于非零特征比较观点对的数量。

### 2.4 产品特征比较网络边权重计算方法

2.3节定义的特征比较有向图没有权重,或

者说,所有边的权重均相等.但是在产品特征的比较中,文本信息体现出来的情感强度是不同的,考虑情感强度的有向图更能体现文字评论中的细微情感.根据比较网络不同的创建方法,分别采用以下有向边权重计算策略.

1) 单边有向图边的权重计算.所有特征比较数组聚合成1条有向边,其权重由所有特征比较数组的情感强度均值(即  $cScore$  的均值)决定,如式(1)所示

$$w = \frac{\left| \sum_{i=1}^n \{ cScore_i \} \right|}{N_{i=1}^n (cScore_i) \neq 0} \quad (1)$$

2) 双边有向图边的权重计算.双边有向图分为类型 I 和类型 II,如式(2)和式(3)所示

$$w = \frac{\left| \sum_{i=1}^n \{ (cScore_i) I_{(cScore_i > 0)} \} \right|}{N_{i=1}^n (pScore_i - nScore_i) > 0} \quad (2)$$

$$w = \frac{\left| \sum_{i=1}^n \{ (cScore_i) II_{(cScore_i < 0)} \} \right|}{N_{i=1}^n (cScore_i) < 0} \quad (3)$$

式(2)表示所有情感极性为正的的特征比较数组聚合成1条边,其权重为该边的情感强度平均值;式(3)表示所有情感极性为负的的特征比较数组聚合成1条边,其权重为该边的情感强度绝对平均值.值得注意的是:双边有向图中,节点  $p_1$  和  $p_2$  之间可能存在1条或2条有向边(类型 I,类型 II,或者二者皆有),因此,可能只需计算式(2)和式(3)的其中1个,也可能二者都需要计算.

3) 多边有向图边的权重计算.多边有向图不存在合并边,节点  $p_1$  与  $p_2$  之间如果存在  $n$  次比

较关系,就具有  $n$  条边,且每条边的权重由比较观点的情感强度决定.所以每条边的权重可以由式(4)计算

$$w = |cScore_i| \quad (4)$$

特征比较网络的构建及其边权重的计算实例如下.给定产品  $p_1 = \text{“Lenovo Thinkpad X220”}$  和  $p_2 = \text{“lenovo Thinkpad X230”}$ .通过产品特征提取,得到3个特征比较对“键盘”、“CPU”和“电池”.并得到表1所示的特征比较数组.

表1 特征比较数组例子

Table 1 Examples of feature-level comparative arrays

序号	键盘	CPU	电池
1	{ $p_1, p_2, 3$ }	{ $p_1, p_2, 1$ }	{ $p_1, p_2, -2$ }
2	{ $p_1, p_2, 3$ }	{ $p_1, p_2, 2$ }	{ $p_1, p_2, -2$ }
3	{ $p_1, p_2, 2$ }	{ $p_1, p_2, -1$ }	{ $p_1, p_2, 3$ }
4	{ $p_1, p_2, 4$ }	{ $p_1, p_2, 0$ }	{ $p_1, p_2, 2$ }
5	{ $p_1, p_2, 1$ }	{ $p_1, p_2, -1$ }	{ $p_1, p_2, 0$ }

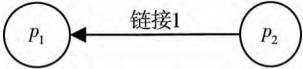
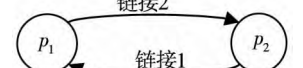
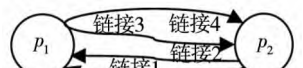
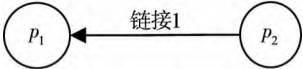
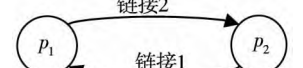
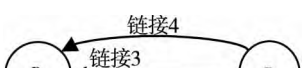
情感分类器使用5分制(-5到5),-5分表示负面情感最强,5分表示正面情感最强,0分表示中性情感.注意到,“CPU 4”和“电池 5”的  $cScore$  值为0.这表明,“CPU 4”和“电池 5”是平行关系,而不是比较关系.但是,为了处理方便,仍然把它们抽象成比较观点,而比较强度为0,连接节点的边的权重也为0.这样在处理比较观点时就不用考虑该特征比较数组对模型是否有影响,因为在网络理论中,权重为0的边对模型并无实际意义,因此在网络计算中会自动排除这类比较观点.使用表1的数据分别构建3个有向网络,结果如表2所示.

表2 产品特征比较网络的实例

Table 2 Examples of feature-level comparative networks

产品特征	网络类型	比较关系	边权重	图例
键盘	单边有向图	链接1: $3+3+2+4+1 > 0$ , 从 $p_2$ 到 $p_1$	链接1: 权重 = $(3+3+2+4+1)/5 = 2.6$	
	双边有向图	链接1: $3+3+2+4+1 > 0$ , 从 $p_2$ 到 $p_1$	链接1: 权重 = $(3+3+2+4+1)/5 = 2.6$	
	多边有向图	链接1: $3 > 0$ , 从 $p_2$ 到 $p_1$ 链接2: $3 > 0$ , 从 $p_2$ 到 $p_1$ 链接3: $2 > 0$ , 从 $p_2$ 到 $p_1$ 链接4: $4 > 0$ , 从 $p_2$ 到 $p_1$ 链接5: $1 > 0$ , 从 $p_2$ 到 $p_1$	链接1: 权重 = $ 3  = 3$ 链接2: 权重 = $ 3  = 3$ 链接3: 权重 = $ 2  = 2$ 链接4: 权重 = $ 4  = 4$ 链接5: 权重 = $ 1  = 1$	

附表2

产品特征	网络类型	比较关系	边权重	图例
处理器 (CPU)	单边有向图	链接1: $1+2+(-1)+(-1)>0$ , 从 $p_2$ 到 $p_1$	链接1: 权重 = $ 1+2+(-1)+(-1) /5=0.20$	
	双边有向图	链接1: $1+2>0$ , 从 $p_2$ 到 $p_1$ 链接2: $-1+(-1)<0$ 从 $p_1$ 到 $p_2$	链接1: 权重 = $ 1+2 /2=1.5$ 链接2: 权重 = $ -1-1 /2=1$	
	多边有向图	链接1: $1>0$ , 从 $p_2$ 到 $p_1$ 链接2: $2>0$ , 从 $p_2$ 到 $p_1$ 链接3: $-1<0$ , 从 $p_1$ 到 $p_2$ 链接4: $-1<0$ , 从 $p_1$ 到 $p_2$	链接1: 权重 = $ 1 =1$ 链接2: 权重 = $ 2 =2$ 链接3: 权重 = $ -1 =1$ 链接4: 权重 = $ -1 =1$	
电池	单边有向图	链接1: $(-2)+(-2)+3+2>0$ , 从 $p_2$ 到 $p_1$	链接1: 权重 = $ (-2)+(-2)+3+2 /5=2.0$	
	双边有向图	链接1: $-2+(-2)<0$ , 从 $p_1$ 到 $p_2$ 链接2: $3+2>0$ , 从 $p_2$ 到 $p_1$	链接1: 权重 = $ (-2)+(-2) /2=2$ 链接2: 权重 = $ 3+2 /2=2.5$	
	多边有向图	链接1: $-2<0$ , 从 $p_1$ 到 $p_2$ 链接2: $-2<0$ , 从 $p_1$ 到 $p_2$ 链接3: $3>0$ , 从 $p_2$ 到 $p_1$ 链接4: $2>0$ , 从 $p_2$ 到 $p_1$	链接1: 权重 = $ -2 =2$ 链接2: 权重 = $ -2 =2$ 链接3: 权重 = $ 3 =3$ 链接4: 权重 = $ 2 =2$	

### 3 实验设计

#### 3.1 数据采集和预处理

特征比较观点广泛存在于中英文评论中, 鉴于亚马逊网站(www.amazon.com)的在线评论比较规范和专业, 提供了丰富的打分、有用性投票等功能, 而且英语料有深入的研究基础, 因此本文实验语料来自亚马逊。本文编写了在线评论爬虫程序, 抓取数码相机这一类别的所有在线评论。抓取的信息包括产品名称、总体评分、每条评论的文字信息、评论者信息等。亚马逊网站中, 会出现不同颜色不同型号的商品, 虽然使用不同的产品ID, 但是产品为同一款, 评论列表也是相同的。这种情况下, 把这些产品进行合并。

亚马逊网站允许买家对其他评论给予评论或者打分, 把这部分评论称为子评论。之前研究都没有关注子评论, 但子评论的文字信息蕴含了丰富的产品特征比较信息, 为此本文把子评论也纳入考虑范畴。

共计抓取数码相机的9485个产品, 经过以下处理, 得到最终的实验样本: 1) 数码相机类别

中包含大量数码相机的周边产品: 三脚架、相机包、肩带、独立镜头、测距仪、测光仪等, 需要过滤这些周边产品; 2) 亚马逊网站中, 会出现同一款产品使用不同的产品ID(因为颜色、型号或者搭配不同), 但是评论是相同的, 这类产品需要合并; 3) 删除没有购买记录的产品, 同时删除没有评论的产品。经过以上处理, 得到1861个产品; 还抓取对这1861个产品的155927条一级评论, 平均每个产品获得83.79条评论; 然后抓取完整的94494条子评论信息, 平均每条评论获得0.61条子评论。时间跨度为1999-06-30至2013-06-09。所有评论来自133381个用户, 平均每个用户发表1.88条评论。

#### 3.2 产品特征比较观点的提取方法

亚马逊网站上的产品标题具有明确的层次关系, 为此, 构建了从根节点到叶节点的树结构来指示产品。根节点是品牌名称, 2级节点是系列名称, 3级节点是型号名称, 树叶是产品实体名称。任何产品只要按照从根节点到叶节点的顺序查询, 总能找到唯一的树叶与之对应, 如图4所示。

产品标题树状结构图的生成步骤如下: 1) 提取相机品牌, 根据亚马逊网站上的品牌列表, 构造

品牌层; 2) 数码相机的名称通常是“字母 + 数字”的组合形式(例如: D70, L310) 根据此规则识别产品, 构造产品层; 3) 根据品牌和产品名称, 抽取系列层和型号层; 4) 对剩余部分不能自动识别的产品进行人工识别和匹配。

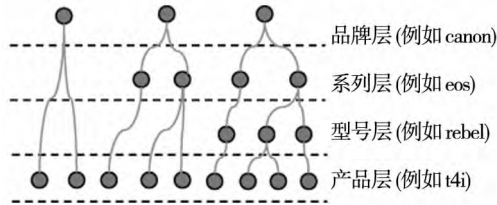


图4 产品标题树状层次结构图

Fig. 4 Tree diagram of product titles

在产品标题中提取第1个“字母 + 数字”组合作为候选产品名称. 为确保数据的准确, 还根据

数码相机品牌列表进行核对. 亚马逊网站共有126个数码相机品牌, 同时采集这126个品牌, 与生成的产品树根节点进行对比, 修正错误数据. 另外一项需要人工修正的是拼写错误, 例如把“canon”误写为“cannon”. 最后收集的数据包括29个品牌, 75个系列, 51个型号, 1861个产品.

关于数码相机的产品特征, 采用如下的方法提取. 首先对频繁产品特征进行自动提取, 结合产品使用手册, 并邀请领域专家进行修正, 构建了数码相机常见的特征列表(表3). 还分别统计了每个特征在评论中出现的次数和比例, 比例可以看作是用户对该特征的重视程度, 即客户越重视某个特征, 那么在评论中提及的次数会越多.

表3 数码相机常见的产品特征

Table 3 Most evaluated features for digital cameras

序号	特征	比较次数	比例(%)	序号	特征	比较次数	比例(%)
1	图片品质	11 286	15.981 8	15	去噪	1 473	2.085 9
2	镜头	9 291	13.156 7	16	显示屏	1 453	2.057 5
3	变焦	7 040	9.969 1	17	麦克风	1 304	1.846 6
4	电池	5 621	7.959 7	18	颜色	1 202	1.702 1
5	价格	4 351	6.161 3	19	自动对焦	866	1.226 3
6	感光度	4 271	6.048 0	20	取景器	757	1.072 0
7	视频	3 569	5.054 0	21	重量	739	1.046 5
8	传感器	2 902	4.109 4	22	对焦性能	543	0.768 9
9	操控	2 405	3.405 6	23	拍摄速度	543	0.768 9
10	大小	2 260	3.200 3	24	特效处理	319	0.451 7
11	菜单	2 052	2.905 8	25	处理器	267	0.378 1
12	快门	1 916	2.713 2	26	触摸屏	254	0.359 7
13	闪光灯	1 834	2.597 1	27	无线连接	180	0.254 9
14	有效像素	1 834	2.597 1	28	同步功能	86	0.121 8

特征比较观点的抽取, 采用以下方法进行: 1) 根据生成的产品树(见图4), 对在线评论进行搜索, 如果含有其他产品名称, 则加入候选比较观点集; 2) 根据表3中的产品特征对候选比较观点集进行匹配, 如果成功则意味着存在一个候选特征级比较观点; 3) 抽取特征比较观点, 构造  $\langle f_i, p_1, p_2 \rangle$ , 其中  $f_i$  代表特征,  $p_1$  和  $p_2$  代表2个被比较的产品; 4) 由于特征比较信息涉及上下文, 因此, 提取和保存特征比较句时, 同时提取该比较句的前一句和后一句文字信息作为评论的上下文, 构造  $\langle f_i, p_1, p_2, C \rangle$ , 其中  $C$  代表上下文信息. 根据

上面的操作, 一共提取了84 756个特征比较句, 平均每个产品有65.92个特征比较句; 平均每条评论包含0.49条比较信息.

采用 SentiStrength 计算特征比较情感值<sup>[46]</sup>. 把提取出的特征比较句输入 SentiStrength 进行查询, 得到情感强度, 即  $f_i = \{p_1, p_2, cScore\}$  中的  $cScore$ . 值得注意的是, 进行查询时, 使用关键词检测方式, 换句话说, 当一句话含有2个以上特征词时, 只需指定1个特征词, 就可以计算指定特征词的情感强度, 以避免当一句话中存在多个比较特征时计算不准确的问题.

由于用户评分习惯的差异,存在一些极端的评价,对所有  $cScore$  值均进行了标准化操作,采用下式进行标准化

$$cScore = -5 + \frac{(cScore^* - min) \times 10}{(max - min)} \quad (5)$$

其中  $cScore^*$  表示原始情感值;  $min$  表示初始情感值的下限;  $max$  表示初始情感值的上限. 标准化后的所有结果在区间  $[-5, 5]$  之间. 标准化后的数据呈现了较好的正态分布,图 5 展示了标准化后的特征比较数组分布图. 可以看到标准化后的数据较多集中在中心位置,而对边沿数据进行了有效控制.

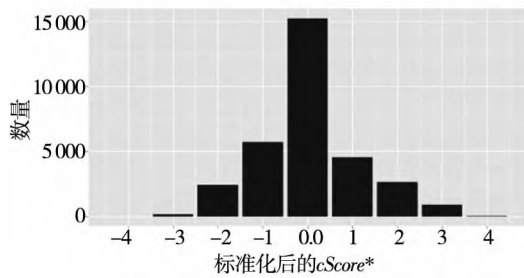


图 5 标准化后的特征比较数组统计图

Fig. 5 Statistics of normalized feature - level comparative arrays

借助 PageRank 算法和 HITS 定量计算图形中每个节点的权重信息. 采用 R 软件的 igraph 包实现 PageRank 和 HITS 的计算.

### 4 实验结果

#### 4.1 单变量相关性分析

亚马逊网站不公布销量数据,但会公布销售排名(SalesRank),因此采纳前人的研究方法<sup>[51, 53-54]</sup>,以 SalesRank 作为销售量近似的度量指标. SalesRank 服从指数分布,因此,经过指数转化,商品需求量与 SalesRank 呈现出近似的线性关系,即

$$\ln(Demand) = \alpha - \beta \ln(SalesRank) \quad (6)$$

其中  $\alpha$  是常量,表示截距;  $\beta$  是相关系数. 由于 SalesRank 的等级特征,取值越小,则销售量越大,反之亦然. 如果相关系数为负,表明自变量对销量有正向影响;反之,表明有抑制作用. 表 4 展示了不同自变量对因变量(SalesRank)的相关系数.

从表 4 看出,在特征级比较网络中,PageRank 和 HITS 的权威度都与产品销售排名 SalesRank 呈现出较强的负相关关系, spearman 系数从  $-0.085$

到  $-0.167$ (均显著相关). PageRank 比 HITS 权威度相关性更高,这可能是由于 HITS 算法需要同时兼顾内容权威度和链接权威度,而 PageRank 只是针对节点的计算. 有向比较网络图的 PageRank 比作为基准的总体评价星级相关性系数高,评价星级与 SalesRank 的系数为  $-0.131$ . 作为直接反映网络口碑效应的评论数量,其相关系数在所有自变量中是最高的( $-0.526$ ),表明评论数量与销量具有很强的相关性. 本文还考虑包含子评论的比较句数量和特征比较数量,这是前人没有尝试过的. 可以看到,包含子评论的比较句数量和特征比较数量比不包含子评论的相关系数更高, Spearman 系数分别从  $-0.469, -0.443$  分别上升到  $-0.496, -0.470$ ,这表明模型中包含子评论的效果会更好. 无向图也能很好的拟合销售数据,这表明,除了网络口碑,特征比较网络的中心度也能对产品销售形成潜在影响.

实验显示多边有向图具有更好的优势. 但是,不能认为多边有向图在所有的特征比较网络中均能表现更好,还需要更多实验验证. 表 4 中的相关系数均遵循以下排序: 无权重 PageRank > 有权重 PageRank > 无权重 HITS 权威度 > 有权重 HITS 权威度. 这可以从两方面解释: 1) PageRank 和 HITS 各自适应不同环境: HITS 算法需要同时兼顾内容权威度和链接权威度,而 PageRank 只是针对节点的计算. 显然,在本文数据下,PageRank 算法(只考虑节点)比兼顾内容和链接的算法更加有效. 这可能是由于文本中比较观点的特征造成的, HITS 算法的适应条件是那些只有链接而没有实质内容的网站(例如:网址导航),这类网站本身没有任何内容,网站所有的页面均指向外链,而在线评论文本中不存在这种没有实质内容的比较关系. 特征比较观点  $\langle f_i, p_1, p_2, c \rangle$  都具有产品特征、进行比较的 2 个产品以及比较上下文. 因此,在特征比较网络中 PageRank 比 HITS 算法更加合适; 2) 无论 HITS 算法还是 PageRank 算法,在本文的数据样本上,无权重的网络图比有权重的网络图相关性系数更高,这仅表明无权重的网络图比有权重的网络图更有实用价值. 但是,不能就此得出“在构建比较网络过程中,没有必要考虑网络边的权重”. 因为,我们只在有限的领域(数码相机),有限的实验



样本(1 861个产品)上进行实验.而在其他领域内,可能考虑网络的权重能够得到更好的结果,这需要在实验中根据产品领域和实验样本进行确定.

表4 销售排名与不同变量的相关系数

Table 4 Correlation coefficients between sales rank and different variables

链接方向	网络类型	网络规模	链接度量算法	相关系数( spearman)
有向边	单边连接图	1 295 个节点 , 34 432 条边	无权重 HITS 权威度	-0.109 ***
			有权重 HITS 权威度	-0.089 ***
			无权重 PageRank	-0.143 ***
			有权重 PageRank	-0.141 ***
	双边连接图	1 295 个节点 , 68 864 条边	无权重 HITS 权威度	-0.117 ***
			有权重 HITS 权威度	-0.095 ***
			无权重 PageRank	-0.159 ***
			有权重 PageRank	-0.137 ***
	多边连接图	1 295 个节点 , 724 447 条边	无权重 HITS 权威度	-0.108 ***
			有权重 HITS 权威度	-0.085 ***
			无权重 PageRank	-0.167 ***
			有权重 PageRank	-0.141 ***
无向边	无向网络	1 295 个节点 , 70 618 条边	节点紧度 Closeness Centrality	-0.249 ***
			节点介数 Betweenness Centrality	-0.384 ***
无链接	无图统计	1 295 个节点	总体评价星级	-0.131 ***
			评论数量	-0.526 ***
			比较句数量	-0.469 ***
			包括子评论的比较句数量	-0.496 ***
			特征比较数量	-0.443 ***
			包括子评论特征比较数量	-0.470 ***

注:\*\*\*表示  $p < 0.01$ .

#### 4.2 有用性评估

由于每个产品在不同的产品特征上可能进行多次比较,分别采用以下2种方法进行产品特征比较网络的合成计算:1)所有产品特征的权重相同(均赋值为1);2)产品特征的权重根据该特征的比较次数确定.表5展示了两种算法的特征比较网络计算结果与销售排名的相关系数对比(spearman).

表5 两种算法的特征比较网络计算结果与销售排名的相关系数对比(Spearman)

Table 5 Correlation coefficients between feature-level comparative networks and sales rank for two algorithms (Spearman)

算法	相同权重	根据评价次数确定权重
Spearman 系数	-0.5122 ***	-0.5140 ***

注:\*\*\*表示  $p < 0.01$ .

可以看到,两种算法在  $p < 0.01$  水平上显著相关,这证明了特征比较网络与产品销量的显著一致性.更进一步,采用相同权重的算法比根据评

价次数确定权重的算法相关系数稍低,这表明根据评价次数确定权重的特征比较网络比相同权重的特征比较网络能够更好拟合产品销量,可能的原因在于:产品特征在用户消费决策的影响权重是不一样的<sup>[53]</sup>.另一个可能的原因是:在评论中频繁提及的产品特征就是用户最关心的特征方面.

然而,仅仅依据相关系数并不能有效地评价特征比较网络的有用性,由于变量之间的内生性问题(endogenous),本文采用如下的计量模型来评估:联立方程模型以及两阶段最小二乘法.建立的评价模型如式(7)和式(8)所示.

$$SalesRank_{i,t} = \theta + \alpha_1 FeatureNetwork_{i,t-1} + \alpha_2 Price_{i,t-1} + \alpha_3 AvgRating_{i,t-1} + \alpha_4 SalesRank_{i,t-1} + \varphi \quad (7)$$

$$FeatureNetwork_{i,t} = \delta + \beta_1 \sum_f cScore_{f,i,t} + \beta_2 NumReview_{i,t} + \beta_3 CompFeature_{i,t} + \beta_4 FeatureNetwork_{i,t-1} + \omega \quad (8)$$

其中,  $SalesRank_{i,t}$  表示产品  $i$  在第  $t$  期的销售排名;  $FeatureNetwork_{i,t-1}$  代表特征比较网络在第  $t-1$  期的计算结果, 采用评价次数确定权重;  $Price_{i,t-1}$ ,  $AvgRating_{i,t-1}$  分别表示产品  $i$  在第  $t-1$  期的价格和平均评分;  $cScore_{f,i,t}$  表示产品  $i$  的特征  $f$  在第  $t$  期的评价得分;  $NumReview_{i,t}$  表示

评论次数;  $CompFeature_{i,t}$  表示特征比较次数;  $\varphi$  和  $\omega$  和是为了捕获每个产品的异质特征, 例如产品定位、市场营销力度、广告策略等. 采用两阶段最小二乘法估计联立方程式 (7) 和式 (8), 具体实现方法参照文献 [55]. 表 6 展示了估计结果.

表 6 两阶段最小二乘法回归结果  
Table 6 Results of two stage least squares regression

变量	最小二乘法(固定效应)	两阶段最小二乘法(联立方程)
	系数(标准差)	系数(标准差)
$FeatureNetwork_{i,t}$	-0.42 (0.13) ***	-0.33 (0.13) ***
$Price_{i,t-1}$	-0.18 (0.02) ***	-0.17 (0.02) ***
$AvgRating_{i,t-1}$	-0.09 (0.03) ***	-0.09 (0.03) ***
$SaleRank_{i,t-1}$	-0.38 (0.02) ***	-0.37 (0.03) ***
$R^2$	0.235	0.250

注: 1. 方程式 (7): 因变量为销售排名, 系数为负表示对销量有促进作用.

2. \*\*\* 表示  $p < 0.01$ , \*\* 表示  $p < 0.05$ , \* 表示  $p < 0.10$ .

表 6 分别展示了最小二乘法和两阶段最小二乘法的估计结果, 值得注意的是: 列出最小二乘法的原因在于对比, 最小二乘法估计本身是不准确的, 因为回归参数(自变量与因变量)存在内生性: 产品销量越高, 带来的用户在线评论会越多<sup>[13, 55]</sup>, 因而根据在线评论抽取的产品特征比较观点也会越丰富, 由此构建的特征比较网络就可能受到产品销量的影响. 而两阶段最小二乘法可以避免这种内生性.

两阶段最小二乘法的结果显示, 特征比较网络与产品销售排名显著正相关(显著性水平  $p < 0.01$ ), 证明在排除内生性问题后, 特征比较网络在销量预测上仍然有用.

### 4.3 回归分析及模型预测能力

根据自变量与因变量  $SalesRank$  的相关系数, 选择不同的自变量构建线性回归模型, 并比较不同模型的预测能力. 由于多边有向图与  $SalesRank$  相关系数最大, 因此选择多边有向图模型. 价格、 $SalesRank$ 、评论数量、特征比较数量等变量, 分别取其对数( $\ln$ ). 构建的 5 个线性回归模型如下.

#### 模型 1

$$\ln SalesRank_{i,t} = \alpha + \beta_1 AvgRating_{i,t-1} + \beta_2 \ln NumReview_{i,t-1} + \beta_3 \ln Price_{i,t-1}$$

#### 模型 2

$$\ln SalesRank_{i,t} = \alpha + \beta_1 AvgRating_{i,t-1} + \beta_2 \ln NumReview_{i,t-1} + \beta_3 \ln Price_{i,t-1} + \beta_4 \ln CompFeature_{i,t-1}$$

#### 模型 3

$$\ln SalesRank_{i,t} = \alpha + \beta_1 AvgRating_{i,t-1} + \beta_2 \ln NumReview_{i,t-1} + \beta_3 \ln Price_{i,t-1} + \beta_4 \ln CompFeature_{i,t-1} + \beta_5 \ln Betweenness_{i,t-1}$$

#### 模型 4

$$\ln SalesRank_{i,t} = \alpha + \beta_1 AvgRating_{i,t-1} + \beta_2 \ln NumReview_{i,t-1} + \beta_3 \ln Price_{i,t-1} + \beta_4 \ln CompFeature_{i,t-1} + \beta_5 \ln Betweenness_{i,t-1} + \beta_6 PageRank_{i,t-1}$$

#### 模型 5

$$\ln SalesRank_{i,t} = \alpha + \beta_1 AvgRating_{i,t-1} + \beta_2 \ln NumReview_{i,t-1} + \beta_3 \ln Price_{i,t-1} + \beta_4 \ln CompFeature_{i,t-1} + \beta_5 \ln Betweenness_{i,t-1} + \beta_6 PageRank_{i,t-1} + \beta_7 Authority_{i,t-1}$$

表 7 分别展示了 5 个模型的回归系数, 其中括号内的数据表示标准误差.

表 7 线性回归模型的系数  
Table 7 Coefficients of linear regression models

变量	模型 1	模型 2	模型 3	模型 4	模型 5
截距	5.117(0.104)	4.817(0.110)	4.821(0.111)	4.850(0.112)	4.868(0.118)
<i>AvgRating</i>	-0.075(0.025)	-0.047(0.025)	-0.047(0.025)	-0.052(0.025)	-0.062(0.027)
$\ln NumReview$	-0.395(0.021)	-0.230(0.030)	-0.228(0.031)	-0.231(0.031)	-0.189(0.032)
$\ln Price$	-0.186(0.023)	-0.114(0.024)	-0.114(0.024)	-0.118(0.024)	-0.095(0.025)
$\ln CompFeature$		-0.203(0.025)	-0.201(0.026)	-0.213(0.027)	-0.224(0.028)
$\ln Betweenness$			-0.003(0.012)	-0.003(0.012)	-0.010(0.013)
<i>PageRank</i>				-0.158(0.089)	-0.107(0.116)
<i>Authority</i>					0.006(0.003)
$R^2$	0.234	0.273	0.272	0.274	0.303
标准估计的误差	0.403 64	0.393 70	0.393 84	0.393 52	0.377 07

从表 7 可以看出,在线评论数量( $\ln NumReview$ )对产品销量具有较强的相关性,这表明网络口碑对于消费者的购买决策具有参照作用。相比之下,消费者总体评分(*AvgRating*)对销量的贡献小得多,这可能是因为总体评分不能体现出每个产品特征细节上的差异。出乎意料的是价格对销量的影响,5 个线性回归模型中,价格( $\ln Price$ )对销量的影响都是正向的。这可能是因为数码相机作为中高端产品,用户对于价格不敏感。另一个对 *SalesRank* 有较大影响的变量是特征比较对数量( $\ln CompFeature$ ),即用户在评论中提及的产品特征之间的比较信息越多,对销量的促进作用越大。产品的中心度( $\ln Betweenness$ )尽管也对销量有影响,但这种影响

微乎其微。从误差项可以看出,从模型 5 到模型 1 误差逐步减小,分别从 0.403 64 减小到 0.377 07,说明模型 5 比模型 1 更加可靠。

使用第  $t$  个季度的数据预测第  $t+1$  个季度的销售情况,研究模型的预测能力。在回归分析预测中,以季度为单位,从 2011 年 1 月开始,3 个月为 1 个单位,例如第 1 季度为 2011 年 1 月 1 日到 2011 年 3 月 31 日,以此类推。统计最近 10 个季度的评论,即从 2011 年 1 月到 2013 年 6 月的所有评论数据。表 8 显示了最近 10 个季度的汇总数据,从汇总数据可以看出,在线评论随时间递增。比如评论数量从第 1 季度的 9 841 条上升到第 10 季度的 12 602 条,受评论数量的影响,其他变量也呈现上升趋势。

表 8 数据概述  
Table 8 Data summary

数据概要	第 1 季度	第 2 季度	第 3 季度	第 4 季度	第 5 季度	第 6 季度	第 7 季度	第 8 季度	第 9 季度	第 10 季度
产品数	789	867	883	1 045	946	1 032	1 096	1 388	1 683	1 481
评论数	9 841	9 575	9 472	13 277	11 283	12 568	12 803	17 235	19 086	12 602
比较句子数	44 330	47 096	48 709	52 573	55 512	61 248	65 632	71 692	75 243	70 641
特征比较数	36 808	39 223	40 529	43 722	45 511	50 250	54 060	59 123	61 966	58 050
特征网络中产品数	317	301	278	309	263	290	283	332	381	327

分别使用上面 5 个线性回归模型对表 8 的数据进行回归预测,然后根据预测值和实际值计算

预测结果的均方误差。表 9 显示了 5 个线性回归模型的预测结果。

表9 模型预测能力结果均方误差

Table 9 Mean square error of the predictive models

季度 $t$	有效产品数量 $N$	模型 1	模型 2	模型 3	模型 4	模型 5
2	286	2.012 9	1.952 4	1.953 2	1.932 6	1.910 6
3	271	2.341 7	2.263 1	2.262 8	2.242 6	2.214 7
4	273	2.238 7	2.207 0	2.207 0	2.188 5	2.183 4
5	258	2.097 1	1.951 2	1.952 0	1.940 2	1.901 5
6	260	2.154 3	2.110 5	2.111 8	2.101 2	2.071 3
7	275	2.621 3	2.601 2	2.600 1	2.571 4	2.548 0
8	277	2.452 7	2.447 3	2.447 7	2.437 6	2.401 4
9	328	3.028 3	2.991 0	3.010 3	2.993 2	2.975 9
10	319	2.969 1	2.946 3	2.945 7	2.808 1	2.787 1

从表9可以看出,除了模型2和模型3的均方误差基本持平以外,模型1到模型5的均方误差是逐步减小的.换句话说,模型1到模型5的预测能力逐步增强.模型2和模型3的预测能力无显著提升,进一步印证了特征比较网络中的节点中心度对模型的贡献不大.在加入特征比较数量,以及特征比较网络的PageRank和HITS权威度后,预测性能明显提升,证明了本文提出的特征比较网络的实际效果.当然,产品销售量除了受到本文描述的变量影响外,还受相当多其他变量的影响,例如:广告、促销等.这可能是导致均方误差在第9季度增大的原因.

任何企业都非常关注产品的预期销量,这既会决定公司的生产计划,同时对物流、供应链等也有相当大影响.本文提出的销售预测模型能够从客户对产品评论的角度对未来销售量进行预测,提升了预测的准确度.

#### 4.4 产品竞争优势识别

产品竞争比较优势图是依据在线评论中体现出来的特征比较信息而构建的.图中顶点代表产品特征,极坐标系的圆心表示起点刻度,每个特征的得分都是0—5分.越接近顶点,表明该特征的竞争优势越强;反之亦然.

构建过程如下:假设产品  $p_i$  所有的特征比较数组记为

$$p_i = \bigcup_{n=1}^m (\{p_i, p_n, \epsilonScore\}_n \cup \{p_n, p_i, \epsilonScore\}_n)$$

首先进行边和节点的转换,每个节点代表1个产品特征,每条边代表2个产品之间的比较关系.2个产品之间可能会就同一特征比较多次,由于只计算汇总的比较优势,所以根据单边有向图进行计算.

限于篇幅,只选择4个商品和22个产品特征进行分析.4个商品分别是:Canon EOS Rebel T3i, Nikon D3100 Digital SLR, Canon EOS 60D和Canon EOS 7D,结果见图6所示.从图6可以看出,Nikon D3100 Digital SLR总体上有较高的满意度,但是在价格(*price*)、无线连接(*wireless*)、颜色(*color*)和视频(*video*)方面有较大的提升空间.而Canon EOS Rebel T3i有较多特征处于劣势,但是在有效像素(*pixel*)、LCD(*lcd*)、图片质量(*image quality*)和去噪(*noise*)方面有明显的优势.Canon EOS 60D在菜单(*menu*)、拍摄速度(*shooting speed*)、视频(*video*)、无线连接(*wireless*)和肩屏(*touchscreen*)等方面有显著优势.

产品特征比较网络能够为公司带来以下收益:展示了与市场同类产品相比,各个特征的竞争优势和劣势,给企业的产品改进提供了依据,同时也为市场营销策略提供了指导.图6只展示4个产品之间的比较,事实上可以扩展到任意产品.产品特征竞争优势图是从文本到可视化的分析结果,这是面向竞争力的产品特征比较网络的研究结果,为消费者准确选择产品提供支持.

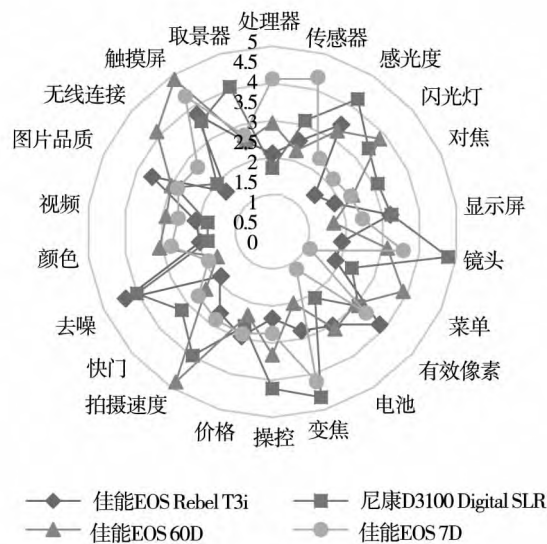


图 6 4 个产品的特征竞争优势图

Fig. 6 Visualization of competitive advantage for 4 products

4.5 产品综合竞争力模型以及启示

在 3.2 节表 3 中,统计了产品特征在评论中出现的次数以及比例,可以认为某个产品特征出现的比例越高,消费者对该特征关注度越大.利用该关注度比例,结合产品在每个特征上的竞争力,能够得到每个竞争产品的综合得分.产品综合得分是从历史评论中得到的所有消费者对产品的总体评分.比较特征得分按照多边有向图的 PageRank 计算,每个特征得分规范化为 1 分~5 分.

式(9)展示了如何由特征得分计算产品综合得分.

$$Score_p = \sum_{i=1}^n (W_i \times FS_{pi}) \tag{9}$$

其中  $Score_p$  表示产品  $p_1$  的总体得分;  $W_i$  表示第  $i$  个产品特征的权重;  $FS_{pi}$  表示产品  $p$  的第  $i$  个产品特征的比较特征得分.表 10 显示了排行榜前 15 个产品列表以及综合得分.为了不被少数特征太少的评论影响,过滤了特征比较次数小于 100 次的商品.

表 10 数码相机综合比较排行榜

Table 10 Comprehensive comparison for digital cameras

排名	产品名称	特征比较次数	SalesRank	特征比较综合得分
1	Panasonic DMC-FH25K	209	312	4.86
2	Sony NEX-6L/B	519	76	4.40
3	Panasonic LUMIX DMC-LX7K	578	60	4.12
4	Panasonic DMC-GH1K	101	3 122	3.35
5	Canon PowerShot SD780IS	251	1 357	3.17
6	Sony NEX-5RK/B	142	430	3.14
7	Canon Powershot A1200	180	826	3.10
8	Canon PowerShot A570IS	261	3 837	2.99
9	Sony Cyber-Shot DSC-RX100	885	1 037	2.90
10	Canon Powershot SX110IS	146	2 598	2.81
11	Canon PowerShot SD4500IS	147	1 810	2.79
12	Canon PowerShot SX130IS	267	409	2.77
13	Canon PowerShot SD1400IS	248	3 602	2.76
14	Sony NEX-5N	389	1 234	2.60
15	Fujifilm FinePix HS30EXR	122	484	2.48

从表 10 可以看出,并非销售排名高的产品就可以获得很高的特征比较综合得分.相反,综合排名第 1 的商品“Panasonic DMC-FH25K”销售排名仅为 312 位,但用户在评论中反映出来的满意度非常高,满分 5 分中获得了 4.86 分.为了进一步验证 SalesRank 与特征比较综合得分的关系,计算了二者之间的相关系数, Spearman 系数为  $-0.403(p < 0.001)$ ,表明总体上特征比较综合得分与 SalesRank 具有强负相关性,即特征比较综合得分与销量显著正相关.特征比较综合得分反映的是从在线评论中抽取的用户对产品的观点,可能存在的一种情况是:如果用户对产品满意的话可能并不发表评论;相反当用户对产品不太满意的时候才发表评论<sup>[56]</sup>.因此,用户评论不能严格与产品销量对等.可以通过特征比较网络计算的特征比较综合得分,是用户对该产品的综合态度,不是单纯由销量决定的,而是综合了用户观点的结果.因此,特征比较综合得分与销售排名有较强的负相关性,这与人们的经验并不冲突,只是从另一个侧面解释了该问题.

对商家而言,表 10 有助于识别竞争对手以及精准定位市场;对消费者而言,可以得出广大用户对每个产品的量化得分,辅助购买决策.

## 5 结束语

本文以比较信息为研究对象,面向在线评论,

借助情感分析技术,提出深度挖掘用户意见的方法.根据在线评论中的比较信息,提取评论中的特征比较观点对,分别采用单边有向图、双边有向图和多边有向图来构建特征比较关系.把产品抽象为网络节点,产品间的比较关系作为网络的边,构建特征比较网络.采用 PageRank 和 HITS 对特征比较网络进行分析,得到节点的权威度.产品特征比较网络可以用于识别产品的竞争优势和劣势,用于产品改进和市场营销.实验表明,基于特征比较网络的分析结果与产品销量显著相关,并且回归分析验证了本文提出的方法在销量预测上的改进,证明了特征比较网络的有用性.特征比较网络不但可以用于特征层次的比较,还能够根据用户评论内容对市场上同类产品进行总体竞争力排序.

未来的研究方向有:1) 本文提取的是显式产品特征,但是在线评论中还存在大量隐式产品特征,隐式特征比较信息的提取有待深入分析;2) 对于比较观点的识别上,本文采用关键字匹配的方法,效率有待提高,未来可以尝试采用模式识别以更加准确高效的识别比较观点;3) 本文提出的方法只在数码相机评论上进行了实验,缺乏对其他产品类别的验证,未来可以尝试把本文方法扩展到其他产品类别,进一步验证该方法的有用性;4) 除了本文提到的网络构建方法外,还有很多其他网络构建和分析算法,例如:二分网络、曲面网络等,未来可以尝试更多的网络分析算法与情感分析的结合.

## 参考文献:

- [1] Xie K L, Zhang Z, Zhang Z. The business value of online consumer reviews and management response to hotel performance [J]. *International Journal of Hospitality Management*, 2014, 43(10): 1-12.
- [2] Blal I, Sturman M C. The differential effects of the quality and quantity of online reviews on hotel room sales [J]. *Cornell Hospitality Quarterly*, 2014, 55(4): 365-375.
- [3] Chu W, Roh M, Park K. The effect of the dispersion of review ratings on evaluations of hedonic versus utilitarian products [J]. *International Journal of Electronic Commerce*, 2015, 19(2): 95-125.
- [4] Krishnamoorthy S. Linguistic features for review helpfulness prediction [J]. *Expert Systems with Applications*, 2015, 42(7): 3751-3759.
- [5] Chen L, Qi L, Wang F. Comparison of feature-level learning methods for mining online consumer reviews [J]. *Expert Systems with Applications*, 2012, 39(10): 9588-9601.

- [6] Hai Z, Chang K, Kim J J, et al. Identifying features in opinion mining via intrinsic and extrinsic domain relevance [J]. *IEEE Transactions on Knowledge Data Engineering*, 2014, 26(3): 623–634.
- [7] 王祖辉, 姜维, 李一军. 在线评论情感分析中固定搭配特征提取方法研究 [J]. *管理工程学报*, 2014, 28(4): 180–186.  
Wang Zuhui, Jiang Wei, Li Yijun. Regular collocation features extraction method in online reviews sentiment analysis [J]. *Journal of Industrial Engineering and Engineering Management*, 2014, 28(4): 180–186. (in Chinese)
- [8] Zhang L, Liu B, Lim S H, et al. Extracting and ranking product features in opinion documents [C]// *Proceedings of the 23rd International Conference on Computational Linguistics*, Stroudsburg: DIGITL LIBRARY, 2010: 1462–1470.
- [9] Wang H W, Yin P, Zheng L J, et al. Sentiment classification of online reviews: Using sentence-based language model [J]. *Journal of Experimental & Theoretical Artificial Intelligence*, 2014, 26(1): 13–31.
- [10] Borrajo L, Vieira A S, Iglesias E L. TCBR-HMM: An HMM-based text classifier with a CBR system [J]. *Applied Soft Computing*, 2015, 26(1): 463–473.
- [11] Chasin R, Woodward D, Witmer J, et al. Extracting and displaying temporal and geospatial entities from articles on historical events [J]. *Computer Journal*, 2014, 57(3): 403–426.
- [12] Liu J F, Zhang P, Lu Y J. Automatic identification of messages related to adverse drug reactions from online user reviews using feature-based classification [J]. *Iranian Journal of Public Health*, 2014, 43(11): 1519–1527.
- [13] Nassirtoussi A K, Aghabozorgi S, Wah T Y, et al. Text mining for market prediction: A systematic review [J]. *Expert Systems with Applications*, 2014, 41(16): 7653–7670.
- [14] Yu Z Z, Zheng N, Xu M. An automatic product features extracting method in Chinese customer reviews [C]// *7th International Conference on System of Systems Engineering*, Genova: IEEE, 2012: 455–459.
- [15] Peñalver-Martínez I, García-Sánchez F, Valencia-García R, et al. Feature-based opinion mining through ontologies [J]. *Expert Systems with Applications*, 2014, 41(13): 5995–6008.
- [16] Kasper W, Vela M. Sentiment analysis for hotel reviews [C]// *Computational Linguistics–Applications Conference*, Jachrenka: Free Preview, 2011: 45–52.
- [17] Wang G, Sun J, Ma J, et al. Sentiment classification: The contribution of ensemble learning [J]. *Decision Support Systems*, 2014, 57(1): 77–93.
- [18] Háva O, Skrbek M, Kordík P. Supervised two-step feature extraction for structured representation of text data [J]. *Simulation Modelling Practice and Theory*, 2013, 33(4): 132–143.
- [19] 姚长青, 杜永萍. 降维技术在专利文本聚类中的应用研究 [J]. *情报学报*, 2014, 33(5): 491–497.  
Yao Changqing, Du Yongping. Research on the dimension reduction technique in patent text clustering [J]. *Journal of the China Society for Scientific and Technical Information*, 2014, 33(5): 491–497. (in Chinese)
- [20] 王蒙, 林兰芬, 王锋. 基于伪相关反馈的短文本扩展与分类 [J]. *浙江大学学报: 工学版*, 2014, 48(10): 1835–1842.  
Wang Meng, Lin Lanfen, Wang Feng. Short text expansion and classification based on pseudo-relevance feedback [J]. *Journal of Zhejiang University(Engineering Science)*, 2014, 48(10): 1835–1842. (in Chinese)
- [21] Bharti K K, Singh P K. Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering [J]. *Expert Systems with Applications*, 2015, 42(6): 3105–3114.
- [22] 王洪伟, 郑丽娟, 尹裴. 基于句子级情感的中文网络评论的情感极性分类 [J]. *管理科学学报*, 2013, 16(9): 64–74.  
Wang Hongwei, Zheng Lijuan, Yin Pei. Classification of sentimental polarity for Chinese online reviews based on sentence level sentiment [J]. *Journal of Management Sciences in China*, 2013, 16(9): 64–74. (in Chinese)
- [23] Balahur A, Mihalcea R, Montoyo A. Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications [J]. *Computer Speech & Language*, 2014, 28(1): 1–6.

- [24]冀俊忠,张玲玲,吴晨生. 基于知识语义权重特征的朴素贝叶斯情感分类算法[J]. 北京工业大学学报,2014,40(12):1884-1890.  
Ji Junzhong,Zhang Lingling,Wu Chensheng. Semantic weight-based naive bayesian algorithm for text sentiment classification[J]. Journal of Beijing University of Technology,2014,40(12):1884-1890. (in Chinese)
- [25]Guo J L,Peng J E,Wang H C. An opinion feature extraction approach based on a multidimensional sentence analysis model[J]. Cybernetics and Systems,2013,44(5):379-401.
- [26]李 纲,刘广兴,毛 进. 一种基于句法分析的情感标签抽取方法[J]. 图书情报工作,2014,58(14):12-20.  
Li Gang,Liu Guangxing,Mao Jin. A sentiment label extraction method based on dependency parsing[J]. Library and Information Service,2014,58(14):12-20. (in Chinese)
- [27]陆 浩,牛振东,张 楠. 基于句法与主题扩展的中文微博情感倾向性分析模型[J]. 北京理工大学学报,2014,34(8):824-830.  
Lu Hao,Niu Zhendong,Zhang Nan. A model for sentiment classification of Chinese microblog based on parsing and theme extension[J]. Transactions of Beijing Institute of Technology,2014,34(8):824-830. (in Chinese)
- [28]Chen C C,Chen Z Y,Wu C Y. An unsupervised approach for person name bipolarization using principal component analysis[J]. IEEE Transactions on Knowledge and Data Engineering,2012,24(11):1963-1976.
- [29]Paltoglou G,Thelwall M. Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media[J]. ACM Transactions on Intelligent Systems and Technology,2012,3(4):1-19.
- [30]Kim K, Lee J. Sentiment visualization and classification via semi-supervised nonlinear dimensionality reduction[J]. Pattern Recognition,2014,47(2):758-768.
- [31]Moraes R, Valiati J F, Gavião Neto W P. Document-level sentiment classification: An empirical comparison between SVM and ANN[J]. Expert Systems with Applications,2013,40(2):621-633.
- [32]杜 锐,朱艳辉,鲁 琳. 基于 SVM 的中文微博观点句识别算法[J]. 湖南工业大学学报,2013,27(2):89-93.  
Du Rui,Zhu Yanhui,Lu Lin. The SVM-based algorithm for Chinese micro-blog opinion sentence identification[J]. Journal of Hunan University of Technology,2013,27(2):89-93. (in Chinese)
- [33]Kanayama H,Nasukawa T. Unsupervised lexicon induction for clause-level detection of evaluations[J]. Natural Language Engineering,2012,18(1):83-107.
- [34]史 伟,王洪伟,何绍义. 基于微博的产品评论挖掘:情感分析的方法[J]. 情报学报,2014,32(2):107-112.  
Shi Wei,Wang Hongwei,He Shaoyi. Product reviews mining from microblogging based on sentiment analysis[J]. Journal of the China Society for Scientific and Technical Information,2014,32(2):107-112. (in Chinese)
- [35]周 哲,商 琳. 一种基于动态词典和三支决策的情感分析方法[J]. 山东大学学报(工学版),2015,45(1):19-23.  
Zhou Zhe,Shang Lin. A sentiment analysis method based on dynamic lexicon and three-way decision[J]. Journal of Shandong University (Engineering Science),2015,45(1):19-23. (in Chinese)
- [36]杨佳能,阳爱民,周咏梅. 基于语义分析的中文微博情感分类方法[J]. 山东大学学报(理学版),2014,49(11):14-21.  
Yang Jianeng,Yang Aaimin,Zhou Yongmei. Sentiment classification method of Chinese micro-blog based on semantic analysis[J]. Journal of Shandong University (Natural Science),2014,49(11):14-21. (in Chinese)
- [37]Zhang Z,Ye Q,Zhang Z, et al. Sentiment classification of internet restaurant reviews written in Cantonese[J]. Expert Systems with Applications,2011,38(6):7674-7682.
- [38]曹丽娜,唐锡晋. 基于主题模型的 BBS 话题演化趋势分析[J]. 管理科学学报,2014,17(11):109-121.  
Cao Lina,Tang Xijin. Trends of BBS topics based on dynamic topic model[J]. Journal of Management Sciences in China,2014,17(11):109-121. (in Chinese)
- [39]杨 铭,祁 巍,闫相斌,等. 在线商品评论的效用分析研究[J]. 管理科学学报,2012,15(5):65-75.



- Yang Ming , Qi W , Yan X B , et al. Utility analysis for online product review [J]. *Journal of Management Sciences in China* , 2012 , 15( 5) : 65 – 75. ( in Chinese)
- [40]李常洪,高培霞,韩瑞婧. 消极情绪影响人际信任的线索效应: 基于信任博弈范式的检验 [J]. *管理科学学报* , 2014 , 17( 10) : 50 – 59.
- Li Changhong , Gao Peixia , Han Ruijing. Impacts of negative emotions on interpersonal trust: Clues effects based on trust game [J]. *Journal of Management Sciences in China* , 2014 , 17( 10) : 50 – 59. ( in Chinese)
- [41]郝媛媛,叶强,李一军. 基于影评数据的在线评论有用性影响因素研究 [J]. *管理科学学报* , 2010 , 13( 8) : 78 – 88.
- Hao Yuanyuan , Ye Qiang , Li Yijun. Research on online impact factors of customer reviews usefulness based on movie reviews data [J]. *Journal of Management Sciences in China* , 2010 , 13( 8) : 78 – 88. ( in Chinese)
- [42]蒋翠清,梁坤,丁勇,等. 基于社会媒体的股票行为预测 [J]. *中国管理科学* , 2015 , 23( 1) : 17 – 24.
- Jiang Cuiqing , Liang Kun , Ding Yong , et al. Predicting stock behavior via social media [J]. *Chinese Journal of Management Science* , 2015 , 23( 1) : 17 – 24. ( in Chinese)
- [43]刘锋,叶强,李一军. 媒体关注与投资者关注对股票收益的交互作用: 基于中国金融股的实证研究 [J]. *管理科学学报* , 2014 , 17( 1) : 72 – 85.
- Liu Feng , Ye Qiang , Li Yijun. Impacts of interactions between news attention and investor attention on stock returns: Empirical investigation on financial shares in China [J]. *Journal of Management Sciences in China* , 2014 , 17( 1) : 72 – 85. ( in Chinese)
- [44]黄卫东,陈凌云,吴美蓉. 网络舆情话题情感演化研究 [J]. *情报杂志* , 2014 , 33( 1) : 102 – 107.
- Huang Weidong , Chen Lingyun , Wu Meirong. Research on sentiment evaluation of online public opinion topic [J]. *Journal of Intelligence* , 2014 , 33( 1) : 102 – 107. ( in Chinese)
- [45]吴方照,王丙坤,黄永峰. 基于文本和社交语境的微博数据情感分类 [J]. *清华大学学报 ( 自然科学版)* , 2014 , 54( 10) : 1373 – 1376.
- Wu Fangzhao , Wang Bingkun , Huang Yongfeng. Microblog sentiment classification using both text and social context [J]. *Journal of Tsinghua University ( Science & Technology)* , 2014 , 54( 10) : 1373 – 1376. ( in Chinese)
- [46]Thelwall M , Buckley K , Paltoglou G. Sentiment strength detection for the social web [J]. *Journal of the American Society for Information Science and Technology* , 2012 , 63( 1) : 163 – 173.
- [47]殷国鹏. 消费者认为怎样的在线评论更有帮助? ——社会性因素的影响效应 [J]. *管理世界* , 2012 , ( 12) : 115 – 124.
- Yin Guopeng. What consumers think of online reviews more useful?: Effects of social factors [J]. *Management World* , 2012 , ( 12) : 115 – 124. ( in Chinese)
- [48]邱鹏,李爱萍,段利国. 基于转折句式的文本情感倾向性分析 [J]. *计算机工程与设计* , 2014 , 35( 12) : 4289 – 4295.
- Di Peng , Li Aiping , Duan Ligu. Text sentiment polarity analysis based on transition sentence [J]. *Computer Engineering and Design* , 2014 , 35( 12) : 4289 – 4295. ( in Chinese)
- [49]Jindal N , Liu B. Mining comparative sentences and relations [C]// *The Association for the Advancement of Artificial Intelligence* , Boston: AAAI Press , 2006 , 22: 1331 – 1336.
- [50]Huang G H , Yao T F , Liu Q. Mining Chinese comparative sentences and relations based on CRF algorithm [J]. *Application Research of Computers* , 2010 , 27( 6) : 2061 – 2064.
- [51]Zhang Z , Guo C , Goes P. Product comparison networks for competitive analysis of online word-of-mouth [J]. *ACM Transactions on Management Information Systems* , 2013 , 3( 4) : 1 – 22.
- [52]Netzer O , Feldman R , Goldenberg J , et al. Mine your own business: Market-structure surveillance through text mining [J]. *Marketing Science* , 2012 , 31( 3) : 521 – 543.
- [53]Archak N , Ghose A , Ipeirotis P G. Deriving the pricing power of product features by mining consumer reviews [J]. *Man-*

agement Science, 2011, 57(8): 1485 – 1509.

[54] Oestreicher-Singer G, Sundararajan A. Recommendation networks and the long tail of electronic commerce [J]. MIS Quarterly, 2012, 36(1): 65 – 83.

[55] Duan W, Gu B, Whinston A B. Do online reviews matter? — An empirical investigation of panel data [J]. Decision Support Systems, 2008, 45(4): 1007 – 1016.

[56] 李 杰, 张向前, 陈维军, 等. C2C 电子商务服装产品客户评论要素及其对满意度的影响 [J]. 管理学报, 2014, 11(2): 261 – 266.

Li Jie, Zhang Xiangqian, Chen Weijun, et al. Key content elements of online consumer review and effects on customer satisfaction for garments in C2C e-commerce [J]. Chinese Journal of Management, 2014, 11(2): 261 – 266. (in Chinese)

## Comparative network for product competition in feature-levels through sentiment analysis

WANG Wei<sup>1,2</sup>, WANG Hong-wei<sup>2</sup>

1. College of Business Administration, Huaqiao University, Quanzhou 362021, China;

2. School of Economics and Management, Tongji University, Shanghai 200092, China

**Abstract:** Comparative opinions widely exist in online reviews as a common way of expressing consumers' ideas. Mean while, such online opinions are key proxies for detecting product competitiveness. Firstly, comparative opinion pairs are extracted in feature-levels through text mining and sentiment analysis. Then, based on the comparative opinion pairs, a single-link graph, a dichotomic-link graph and a multi-link graph are built respectively, where the weights of edges are determined by the sentimental strength. Next, a feature-level comparative network is calculated by employing sophisticated network algorithms, including PageRank and Hyperlink-Induced Topic Search. The proposed comparative network can identify the strengths and weaknesses of compared products. Experimental results show that the feature-level comparative networks are correlated with SalesRank significantly, thus enabling the prediction on sales volume.

**Key words:** competitiveness; comparative network; online review; sentiment analysis; product feature; network analysis