·理论探索·

# 基于中文网络客户评论的消费者行为分析方法

邱云飞 王 雪 邵良杉
(辽宁工程技术大学软件学院, 辽宁 葫芦岛 125105)

〔摘 要〕网络上针对商品的评论中含有消费者的消费习惯、消费体验和消费偏好等颇有价值的信息, 这为观察和分析消费者的行为提供了很好的资料。文中设计了一个网络环境下的消费者行为分析方法。首先, 在收集的客户评论中提取产品特征、消费者信息和消费者对具体产品特征的情感倾向; 其次, 按消费者信息进行消费者群体划分, 进而探讨不同消费群体对不同产品的喜好。企业可通过该方法及时获取消费人群对产品的反馈数据并制定正确的市场营销策略。

〔关键词〕消费者行为分析; 客户特征提取; 产品特征提取; 情感倾向; 群体划分
DOI: 10.3969/j.issn.1008-0821.2012.01.002
〔中图分类号〕G252.0 〔文献标识码〕A 〔文章编号〕1008-0821 (2012) 01-0008-04

# The Method of Customers' Behavioral Analysis Based on Chinese Web Clients' Reviews

Qiu Yunfei Wang Xue Shao Liangshan
(School of Software Engineering, Liaoning Technical University, Huludao 125105, China)

〔**Abstract**〕The reviews of commodities on the web conclude lots of valuable information that are consuming habits, consuming experience and consuming preference of the customers, which can supply good materials that help analyze customers' behaves to marketing researchers. This paper designed a way to analyze customers' behaves in the web environment. Firstly, extracting features of product, customers' information in customers' reviews and customers' emotional tendency towards concrete product features which are gathered. Secondly, dividing customer groups according to customers' information. And then discussing different customers' preference to different products. Corporations can acquire customers' feedback of product in time and make excellent marketing strategies by the approach in this paper.

〔**Key words**〕customers' behavioral analysis; extract customers' features; mine features of product; emotional tendency; group division

消费者行为分析是市场营销活动的前提和基础, 只有全面深入了解消费者的消费行为规律才能制定有针对性的、具体的营销策略, 开展营销活动。在应用现有消费者行为分析方法时研究者只能通过访谈、问卷或直接观察消费者的活动来获取研究资料, 这个过程不但需要经过精心设计, 而且在样本范围、调查周期、指标选取等方面都存在着不同程度的问题, 这与这些分析方法产生的社会背景和当时的技术水平有关。

互联网技术的发展为新的消费者行为分析方法的产生创造了条件。消费者在论坛、博客和商业网站中分享其在产品或服务的取得、消费和处置过程中的体验, 这些包含了消费者偏好、消费习惯的评论文本为市场研究人员观察和分析消费者的行为提供了很好的资料。截至 2010 年 12 月底, 我国网民规模达到 4.57 亿[1], 消费者发表的中文评论越来越多, 这些海量的评论数据在规模上完全可以满足消费者行为分析对样本规模的要求。

若企业的市场研究人员对互联网上的文本、链接等信息进行提取、集成和分析, 便可以从客户发布在网上的评

收稿日期: 2011-11-23
基金项目: 国家自然科学基金 (70971059); 辽宁省创新团队项目 (2009T045)。
作者简介: 邱云飞 (1976-), 男, 副教授, 博士, 研究方向: 数据挖掘理论与应用。

论中发现消费者行为规律，不但能提高消费者行为分析的水平，缩短实施周期，而且确定指标准确、全面[2]。

文中着重研究基于中文网络客户评论的消费者行为分析方法，即向企业展现如何利用消费者的中文网络评论信息发现消费者行为规律，对其消费行为进行分析预测，助其制定正确的营销策略，提高经营管理水平。

## 1 消费者行为分析模型

依据信息的处理流程，所构建的基于中文网络客户评论的消费者行为分析流程模型（如图 1 所示）分为 3 个部分，分别是网页信息收集、数据处理与保存和消费者行为分析[3]。

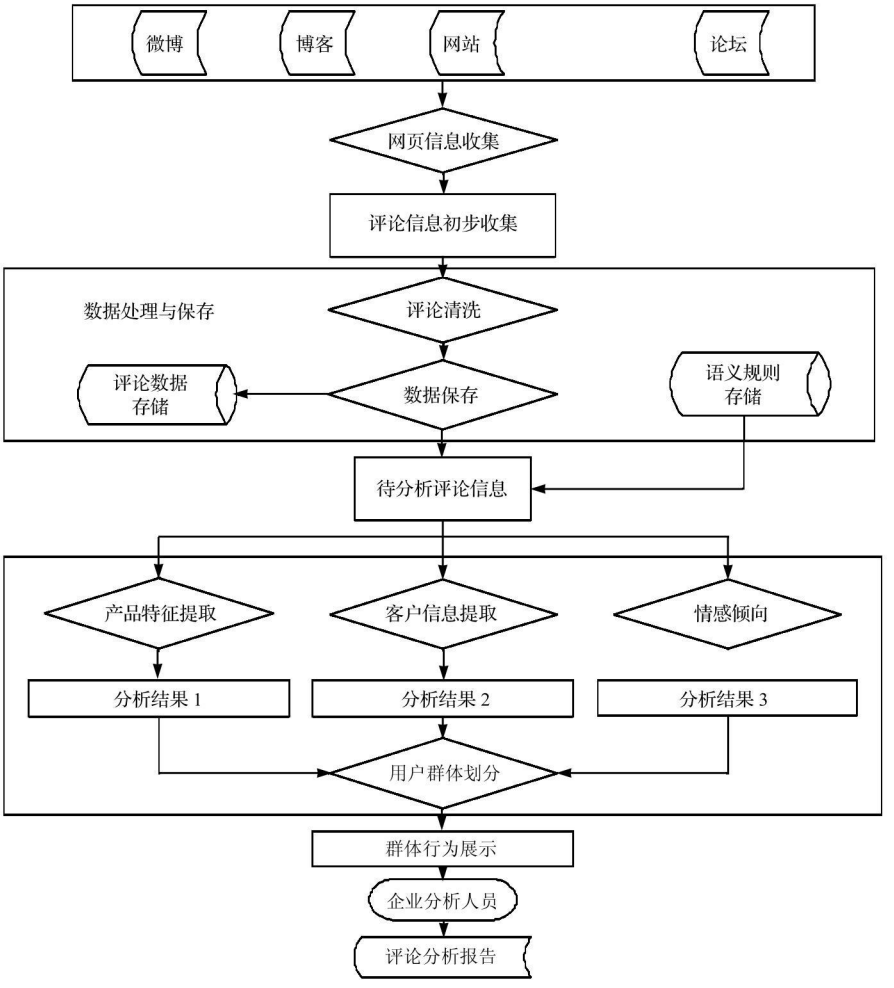


图 1 基于中文网络客户评论的消费者行为分析流程

### 1.1 客户评论信息收集

文中以 it168 网站（http://mobile.it168.com/）及天涯论坛下载的客户对两款不同型号诺基亚手机（Nokia E72、Nokia C5-03）的评论数据（以下简称帖子）为分析对象，分析总结出客户评论信息的收集过程分为以下 3 个步骤：评论初步收集、评论清洗（帖子过滤）和结果数据保存。

1.1.1 评论初步收集（以 it168 网站客户对 Nokia E72 的评论收集为例进行说明）

（1）进入 it168 网站的网址http://mobile.it168.com/，点击 Nokia E72 链接，便可进入 Nokia E72 的主页面。此后点击“点评”菜单项便可见客户对 Nokia E72 的全部评论信息。

（2）根据自身需求选定一定数量（建议大量）合适的客户评论并进行初步收集。

（3）将收集的评论数据进行保存以供后续使用。

1.1.2 评论清洗（帖子过滤）

经过解析得到的帖子，并不全是有意义的，因此在保存数据之前，需要将一些无意义的帖子，也就是所谓的“灌水帖”过滤掉。通过对帖子内容分析，发现垃圾帖主要有以下 2 类：

（1）广告帖 广告帖主要以为自己的网站或博客地址做广告而发的帖子，这种帖子多以一些无意义的回复加上广告链接存在。

（2）灌水帖 该类帖子主要以获取网站积分为目的，主要形式就是发表一些简单的无实际意义的回复，它存在的形式是多样的，如表 1 所示。

以“垃圾帖”中的广告帖为例，对于广告帖：“注册会员抢楼即可获得免费的酒店入住房券，详情请登录http://www.51gojd.cn/huodong/index.asp?refid=3166880。

表 1 灌水帖的主要形式

| 序号 | 类 别 | 形 式 |
|---|---|---|
| 1 | 空 贴 | 没有内容，只有空格占据位置。 |
| 2 | 字数少而毫无意义的帖子 | 无意义的内容，如：KK，看看，飘过，路过…等等。 |
| 3 | 纯数字帖 | 只有数字，如：3，88888，39494 等。 |
| 4 | 纯符号帖 | 只有符号，如：????，；；；，!!!，^_^等。 |
| 5 | 刷屏帖 | 以某些内容为作为基元素，通过空格、换行排版操作将帖子编辑成引人注目的符号，如脚印等。 |

通过分析可以发现该广告帖包含一个超链接，但是包含超链接的帖子又不全是广告帖，因此还需要结合“会员”、“免费”等关键词将该广告帖过滤掉。其他形式的“垃圾帖”也可以做类似的处理。

1.1.3 结果数据保存

“垃圾帖”过滤后得到的帖子满足分析需要，因此保存数据以供后续使用。

## 1.2 客户评论信息产品特征提取

评论信息中的产品特征提取是分析客户对于产品具体特征所持情感倾向的前提，是后续消费者行为分析的基础，所以产品特征的提取是关键性技术[4]。文中对具体人工提取产品特征的思路阐述如下。

(1) 将两种不同型号的手机（Nokia E72、Nokia C5-03）的评论分别集成到一起，以便后续的产品特征标注。

(2) 分别进行产品特征标注，标注对象主要以中文名词或者名词短语形式出现，但对于一些常见的口语化名词，如“机子”，“东西”，均不予以标注[5]。以 Nokia E72 为例，列出此型号手机属性的人工标注结果，如表 2 所示。

(3) 将标注结果数据保存以供后续使用。

表 2 手机（Nokia E72）属性的人工标注结果

| 商品名称 | 人工标注属性集合 | 人工标注属性数量 |
|---|---|---|
| Nokia E72 | 键盘，铃声，拍照，速度，系统，功能，摄像头，手感，售后服务，声音，机身，价格，接口，电话簿，菜单，屏幕，软件，电池，体积，游戏，外形，输入，字库，收音机，内存，语音，摄像，按键，mp3，多媒体，耳机，待机时间，版本，快捷键，兼容性，闪光灯，充电器，质量，智能，屏幕效果，桌面，运行速度，音质 | 43 |

## 1.3 客户信息提取

一般来说，客户注册所填写的性别、年龄、职业、学历和兴趣等信息均可能是不真实的，无法使用这些注册信息对客户进行群体划分，但客户发表的帖子中却可能会出现指示其身份信息的一些词语[6]。所以，有必要人工从中文网络评论中获取客户真实可靠的性别、年龄等客户相关信息[7]。下面对提取客户身份信息的思路进行说明。

(1) 识别客户名。将论坛中发表的帖子按注册客户集成到一起，即将同一个客户的所有评论合成为一个文件，然后使用该客户名标识这个文件，因为论坛中客户名惟一，所以也就将不同的客户发表的评论构成的文件区分开了。

(2) 建立评论集（以下将以提取客户的性别特征为例进行阐述）。经过深入地观察和分析客户所发表的评论，利用传统的手工标注方法对其中的评论进行性别标注。下面给出几个评论，说明如何人工识别客户性别（如表 3 所示）。性别辨识主要利用的信息包括：中文角色称谓和称谓上下文。中文角色称谓是指“老公”、“丈夫”、“老婆”、“女儿”等等的称呼性词语。但是这些称谓还要与特定的上下文结合才有助于性别的辨识。

①当表示男性的称谓词前面出现“我”，可以把“我”定制为左边界，如符合结构为〈我〉…〈老公〉或〈我〉〈老公〉…的规则，可判断其为女性。

②当表示男性称谓词之前出现“作为”之类的动词，即存在左边界，而男性称谓右边没有出现“的+ 表示女性称谓的称呼”，如符合结构〈我〉〈作为〉〈老公〉…或〈我〉〈作为〉〈爸爸〉…的规则，而不是〈我〉〈作为〉〈爸爸〉〈的〉〈女儿〉…可以判断其为男性。

③当表示男性的称谓词前面出现非第一人称称呼时，如“她”、“我的朋友”等，就无法判断客户的性别。

(3) 对所建立的评论集中的已经识别出性别的客户进行性别标注（如表 3 所示）。即为客户 1 和客户 2 标注上“女”，为客户 3 和客户 4 标注“男”。

同理，可以用类似的方法提取客户职业、学历和兴趣等等的客户身份信息，然后绘制客户评论及客户信息的二维表，这也就为后续的客户群体划分和客户行为分析作了准备。

表 3 人工识别客户性别提取表

| 客户编号 | 评 论 内 容 | 指示性别的特征 | 性别标注 |
|---|---|---|---|
| 1 | “我老公买了一个 Nokia 手机，手机像素还不错，拍出的照片效果真的很逼真!”。 | “我”、“老公” | 女 |
| 2 | “我买的这个 Nokia 手机不错，老公很喜欢”。 | “我”、“老公” | 女 |
| 3 | “我作为一名合格的老公给老婆买了一个她心爱的 Nokia 手机”。 | “我”、“作为”、“老公” | 男 |
| 4 | “我给女儿买了一个手机作为她的生日礼物，她高兴地扑到我的怀里，当时我真有一种作为爸爸的骄傲感!”。 | “我”、“作为”、“爸爸” | 男 |
| 5 | “我的朋友的老公买了一个手机，质量很好!” | “我”、“老公” | — |

### 1.4 中文客户评论情感分析

情感分析是通过分析客户对产品的每个特征所体现出来的情感态度倾向来分析消费者对产品的情感倾向的。即对每一个消费者关于产品特征的观点进行情感分级，通常划分为“非常满意”、“满意”、“一般”、“不满意”、“非常不满意”5 种级别[8-9]。下面对中文客户评论情感分析的思路进行说明。

(1) 根据《现代汉语分类词典》[10]和《汉语褒贬义词语用法词典》[11]，并结合特定的表达作者情感倾向的中文词性组合模式标注评论中能体现情感倾向的情感褒贬词。例如：在手机产品的评论中，名词“性价比”前一般是“高”，“低”等形容词，这就构成了“形容词+ 名词”的双词模式；名词“续航能力”后一般是“强”，“弱”等形容词，这就构成了“名词+ 形容词”的双词模式，如表 4 为最常用的 5 个双词模式。

(2) 根据《近代汉语：程度副词研究》[12]标注褒贬情感词的强烈程度，以便划分级别。

(3) 根据情感词的褒贬以及情感词的褒贬程度，判断所得到的情感倾向归于那个级别，以得出消费者的情感程度。例如，在“用手机看视频超清晰”这句话中，“超”属于表达情感的程度副词，而“清晰”属于褒义情感词，所以可以判断此客户的情感倾向于“非常满意”。

表 4 中文评论情感分析双词模式

| 模式类别 | 首 词 | 尾 词 |
|---|---|---|
| 模式 1 | 形容词 | 名 词 |
| 模式 2 | 副 词 | 形容词 |
| 模式 3 | 形容词 | 形容词 |
| 模式 4 | 名 词 | 形容词 |
| 模式 5 | 副 词 | 动 词 |

## 2 消费者群体识别及行为展示

上节中已经阐述了如何提取产品特征、客户信息以及情感倾向，这里可以得到一个关于“消费者”和“产品特征”的表格。这个二维表也包含消费者的“情感倾向级别”属性，这对后续的消费者群体识别及其购买行为的分析提供了必要的前提[13]。通过组合不同的特定属性，并结合一定的统计方法，企业可以对消费者群体进行识别并对其购买行为进行分析。分析方法如下：

### 2.1 统计个体消费者对产品整体的情感倾向

例如：统计女性消费者对产品 Nokia C5-03 的情感倾向，判断其对产品情感程度；统计高学历群体消费者对 Nokia C5-03 的情感倾向，判断其对产品情感程度；统计职业为学生的消费者对 Nokia C5-03 的情感倾向，判断其对产品情感程度。如此，企业可以统计具有特定属性的个体消费者对产品整体的情感倾向，这可以作为企业制定营销策略的一项依据。然后，综合各个分项结果，可以得到不同消费者群体对某产品的情感倾向。因此，按照例子可以得出，消费群体属性为 (女性，高学历群体，学生) 的人群对 Nokia C5-03 的情感倾向。

### 2.2 统计全体消费者对产品整体的情感倾向

例如：统计全体购买 Nokia E72 的消费者对产品整体的情感倾向，可以得出产品在市场上平均满意度和市场价值。

### 2.3 统计全体消费者对某个产品特征的情感倾向

例如：统计全体购买 Nokia E72 的消费者对某个产品特征的情感倾向，可以得出消费者对 Nokia E72 的某个属性 (如外观，质感，价格和性能等) 的满意度，这可以作为企业对产品进行改良的依据。

## 3 结 论

文中介绍了如何利用中文网络客户评论信息发现消费者行为规律，更准确地提取出消费者的观点及购买行为，更高效地指导生产商和服务商改进产品、改善服务、提高竞争力[14]。利用中文网络客户评论信息发现消费者的行为规律，是一个很好的营销数据搜集手段及营销策略制定的途径。若能善加利用，企业的市场研究人员一定可以从这座信息矿藏当中提炼出有价值的东西，为企业提供更大的竞争优势。

## 参 考 文 献

[1] 中国互联网络信息中心 (CNNIC). 第 27 次中国互联网络发展状况统计报告 [R]. 2011. 1.

[2] Liu B, HsuW, Ma Y. Integrating Classification and Association RuleMining [C]. KDD298, 1998: 80-86.

[3] Hu M, Liu B, Mining Opinion Features in Customer Reviews [C]. AAAI, 2004: 755-760.

(下转第 15 页)

任务创造若干合约经理代理，每个合约经理代理用配置代理定义的初始参数来承担任务。

在邀请阶段，合约经理代理向信息服务器发送请求来寻找潜在的供应商，信息服务器检索数据库，返回候选供应商名单。合约经理代理生成若干谈判代理，每一个谈判代理与一个候选供应商代理谈判。根据合约经理指示的任务目标、限制条件和谈判策略来激活谈判代理。任务发布格式如下：

```
Announce= {
    Head{AnnouncementID, AnnouncerID, Addr};
    TimeData{AnnoExpiTime, BidValiTime};
    CommData{Pric, Volu, Pena};
    TaskSpec{MateSpec, Qual, Warr, DeliTime}; }
```

在出价阶段，每个谈判代理以讨价还价方式与供应商进行交易来实现利益最大化，采用启发式学习和推导来模仿对方的偏好和兴趣。最终双方就出价达成协议，格式如下：

```
Bid= {
    Head{BidID, AnnouncerID, BidderID, Addr};
    BidSpec{BidValiTime};
    CommData{Pric, Pena};
    TaskSpec{MateSpec, Qual, Warr, DeliTime}; }
```

在签约阶段，合约经理从交涉代理处收集所有出价。配置代理在所有合约经理之间进行协调从而促使他们各自的出价达成一致。最终选择最佳的出价并提供给相应的供应商代理。

## 4 结 论

多代理技术的成功应用部分解决了供应链管理中的效率瓶颈。为进一步提高供应链管理的效率，本文提出了一种基于本体的多代理供应链管理模型。该模型包括资源层、知识管理层、代理层及应用层。并对基于本体的多代理谈判过程的工作流进行了具体描述。实验测试可利用 JADE 提供的测试平台[8]。下一步的研究将专注于如何通过知识库的完善来提高代理的谈判能力，以及如何提高供应链管理的安全性等。

## 参 考 文 献

[1] Sunil Chopra, Peter Meindl. Supply Chain Management: Strategy, Planning and Operation [M]. IIE Transactions, 2004, 34 (10): 221- 222.

[2] Weihui Dai. Consumer Oriented Supply Chain Management Based on Mobile Agent System [C]. New Trends in Information Science and Service Science (NISS), 4th International Conference on May, 2010, Gyeongju, Korea: 604- 608.

[3] Diosteanu Andreea, Cotfas L. Adrian, Smeureanu Alexandru, et al. Multi- Agents and GIS Framework for Collaborative Supply Chain Management Application [C]. 9th Roedunet International Conference (RoEduNet) on June, 2010: 157- 162.

[4] Carsten Bo¨hle, Bernd Hellingrath, Wilhelm Dangelmaier, et al. Workflow- based Agents for Supply Chain Management [C]. Industrial Informatics (INDIN), 8th IEEE International Conference, Osaka, June, 2010: 643- 648.

[5] Gong Wang, TN Wong, Xiaohuan Wang. An Adaptive Ontology - Mediated Approach to Organize Agent- based Supply Chain Negotiation [C]. Computers and Industrial Engineering (CIE), 40th International Conference on August, 2010, Awaji: 1- 6.

[6] The Open Archives Initiative Protocol for Metadata Harvesting [J/OL]. http: // www. openarchives. org/OAI/ openarchivesprotocol. html, 2010 - 07- 13.

[7] JADE [EB/OL]. http: // jade. tilab. com/, 2011- 03- 08.

[8] JADE Test Suite [J/OL]. http: // jade. tilab. com/ doc/ tutorials/ JADETestSuite. pdf/, 2011- 03- 08.

(上接第 11 页)

[4] 李实，叶强，李一军，Rob Law. 中文网络客户评论的产品特征挖掘方法研究 [J]. 管理科学学报，2009, 12 (2): 142- 152.

[5] Popescu A- M, Etizioni O, Extracting Product Features and Opinions From Reviews [C]. Proceedings of HLT - EMNLP 2005, ACL, 2005: 339- 346.

[6] Mukherjee and Liu B.. Improving Gender Classification of Blog Authors [C]. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010.

[7] 赵妍妍，秦兵，刘挺. 文本情感分析 [J]. 软件学报，2010, 21 (8): 1834- 1848.

[8] Liu B, Hu MQ, Cheng JS, Opinion Observer: Analyzing and Comparing Opinions on the Web [C]. Proceedings of the 14th international World Wide Web conference (WWW- 2005), Chiba, Japan, 2005.

[9] 姚天 ，娄德成. 汉语语句主题语义倾向分析方法的研究 [J]. 中文信息学报，2007, 21 (5): 73- 79.

[10] 董大年. 现代汉语分类词典 [M]. 上海：汉语大词典出版社，1998: 105- 110.

[11] 王国璋. 汉语褒贬义词语用法词典 [M]. 北京：华语教学出版社，2001: 123- 128.

[12] 陈群. 近代汉语：程度副词研究 [M]. 成都：巴蜀书社，2006: 34- 41.

[13] 刘鸿宇，赵妍妍，秦兵，等. 评价对象抽取及其倾向性分析 [J]. 中文信息学报，2010, 24 (1), 84- 88.

[14] 戚攻. 从社会学理论域考察网络社会群体 [J]. 探索，2001, (2): 77- 80.