基于观点挖掘的笔记本电脑评论分析系统①

王卫平, 盛秋华

(中国科学技术大学 管理学院, 合肥 230026)

摘 要: 主要以商业领域的需求和应用为背景, 构建一个智能化的笔记本电脑评论分析系统. 该系统对国内大型购物网站上非结构化、自由式的笔记本电脑评论文本进行情感倾向识别和产品特征归纳, 实现了利用数据挖掘和商务智能的手段分析网络消费者对特定产品的反馈, 帮助企业管理人员了解特定产品的市场需求、制定商业决策. 实验结果证明该系统能够较准确的得出分类结果并归纳出产品特征.

关键词: 观点挖掘; 笔记本电脑评论; 情感倾向分类; 产品特征归纳

Analysis System of Notebook Computers' Comments Based on Opinion Mining

WANG Wei-Ping, SHENG Qiu-Hua

(Faculty of Management, University of Science and Technology of China, Hefei 230026, China)

Abstract: With the commercial demand and application background, this paper creates an analysis system of notebook computers' comments, which aims at those unstructured and freestyle online comments. It can classify the text sentiment orientation and conclude products features. The system uses data mining and business intelligence methods to analyze huge amounts of feedback from online customers and help corporate executives to realize the market needs, make commercial decisions. Finally, the experimental results prove that this system can obtain a relatively accurate result.

Key words: opinion mining; comments of notebook; text sentiment orientation classification; product features introduction

1 引言

随着信息化时代的来临, 以网络为渠道的商品交易量在不断增加, 以淘宝商城为例, 各大品牌笔记本制造商、代理商入驻, 新的销售渠道增加了笔记本电脑销售量的同时, 也更加便捷的提供了大量有价值的客户反馈. 企业及时了解顾客对其产品和服务的评价有助于新产品的改良与服务的改进, 增强企业的竞争力[15].

针对以上需求, 本文尝试构建一个智能化的笔记本电脑评论分析系统, 该系统面向企业管理人员, 针对国内主流电子商务网站上笔记本电脑的消费者评论文本, 按照情感倾向进行分类并完成产品特征抽取和属性归纳, 生成关于产品特征的客户评论观点摘要, 帮助企业有效的整合网络评论资源, 获得该类产品和服务的改进信息.

2 系统总体架构

充分利用网络评论中隐藏的数据资产, 提炼出有价值的信息, 对提高企业的商业决策至关重要. 这样不仅可以缩短制造商对电脑的设计周期、降低成本, 并且可以通过反馈结果及时更改解决方案, 保证新产品更加符合市场需求. 基于以上思想, 构建的笔记本电脑评论分析系统总体构架如图 1 所示.

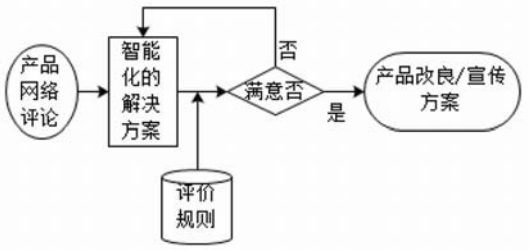

图 1 系统总体架构图

① 收稿时间:2011-12-30;收到修改稿时间:2012-03-04

2.1 系统模块

该系统内部模块间独立工作，相互承接，信息流在这些模块中依次传递，如图 2 所示.

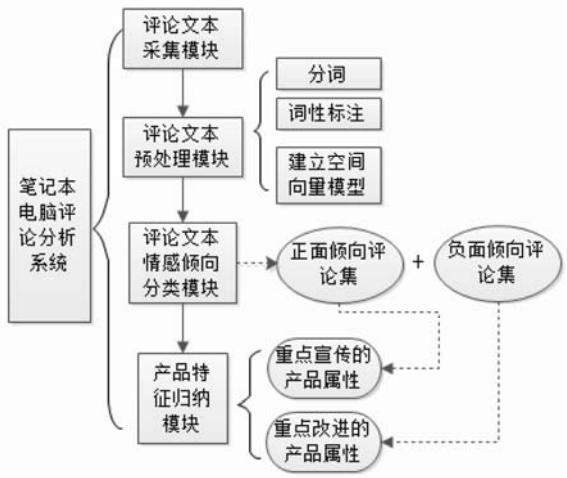

图 2 笔记本电脑评论分析系统的模块构成

(1) 数据采集模块

在固定的时间段内，选取各大 B2C 网站、论坛上关于笔记本电脑的评价详情，以及网友的产品使用心得等 web 文本.

(2) 评论文本预处理模块

该模块实现了从文本模型到数学模型的转换，预处理阶段包含三个子模块：分词，词性标注，构建空间向量模型 VSM.

(3) 文本情感倾向性分类模块

按照监督式学习算法支持向量机(SVM)构建分类引擎，对评论文本的情感倾向进行识别.

(4) 产品特征归纳模块

根据评论情感倾向分类的结果，生成意见摘要.以可视化的形式展示给企业管理人员以辅助商业决策.

3 系统核心模块的设计与实现

3.1 文本预处理模块

3.1.1 分词与词性标注

通常网络评论以自由式评论为主. 例如:“这台笔记本外观很漂亮，卖家服务态度还可以. 缺点是散热性能不好. ”这段评论前一句是对于笔记本外观的正面表述，而后一句则属于负面表述. 若合并处理，对于评论情感分类的结果造成模棱两可的影响. 为了更加精确地对情感倾向进行分类，需要进行细粒度的观点挖掘，识别观点表达的多个语义成分. 因此，采取句子为处理单元，即将消费者的评论文本以单句进行分割得到细化粒度的评论文本；再利用 ICTCLAS 进行网络评论的分词和词性标注.

3.1.2 评论文本规范化处理

系统在对电脑评论语句进行分词和词性标注之后，根据标注对象的结果，单元句中只保留形容词、副词、程度副词、否定词、感叹词；去掉数次、介词、连词、名词等无关于情感表达的词汇，将其处理得到的结果存储在特征词典中[5].

为了将非结构化的评论文本转化成计算机可以识别并处理的数学模型，本文采取了目前应用比较广泛的VSM空间向量模型. 将评论文档划分为一簇词语集合(t1, t2, ..., tn)，每个词语都赋予一定的权值 w，文档 Di 被映射为向量空间中的一个向量；表示为 D(i)=(t1, w1; ..., tj, wj; ..., tn, wn). 文档集 D=(d1, d2, ..., dn);

关键特征词集 T=(t1, t2, ..., tm);

T 集合是过滤情感倾向不相关的词汇后，最能表征文本情感倾向的关键特征词集合，tm 表示第 m 个关键特征词. 对于每个网络评论文档 di，表示为所有关键特征词的权重. di=(w1, w2, ..., wm);wi 权重的计算方法本文采取 TF-IDF 加权，计算和表达某个关键词在文本中的重要程度.

TF-IDF=TF(ti, dj)×IDFi=TF(ti, dj)×log(|N|)/|Ni|);

其中|N|表示文档总数，|Dt|表示关键词 i 在段落中出现的段落数目. 其结果用以衡量这些特征词对情感倾向分类的贡献大小.

3.2 评论文本情感倾向识别模块

3.2.1 情感倾向分类模型的构建

二分类非线性支持向量机在情感分类效果上较好[2]. 设训练样本为{Xi, Yi}, Y 为情感倾向性类别，取值为 1(正面倾向)或-1(负面倾向). 设存在两个平行的超平面将这两个类别的数据最大限度地分开:

$\omega \cdot xi+b \geq 1, yi=1$

$\omega \cdot xi+b \leq -1, yi=-1 \quad i=1,2,....,l$

由此得到的决策函数如下:

$F(x)=sgn(\omega \cdot xi+b)$

其中，ω为可变化的权值变量；x 是输入变量；b 是偏置；ω·x 表示向量 ω∈RN 与 x∈RN 的内积.

对于情感倾向性分类，系统分类模块将输入向量

映射到一个高维空间，并在该特征空间中构造最优分类面[15].

最优化问题则演变为:

$$\min\left\|\omega^2\right\|/2+C\sum_{i=1}^{l}\xi_i$$

$$s.t: y_i(\omega\cdot\phi(x_i)+b)\geq 1-\xi_i$$

$$\xi_i\geq 0,\ \ i=1,2,...l$$

$$\min\frac{1}{2}\sum_{i,\ j=1}^{n}a_j a_i y_j y_i K(x_j,x_i)-\sum_{j=1}^{n}a_j$$

$$s.t: 0\leq a_i\leq C$$

$$i=1,2,...,l$$

$$\sum_{i=1}^{l}a_i y_i=0$$

3.3 产品特征归纳模块

3.3.1 产品特征库的构建

系统构建笔记本电脑特征库，针对笔记本电脑各项特征归纳整理，形成这类产品的整体概念以及客户反馈信息参照的标准. 此特征库能容许态扩充和修改，其数据结构如下表示:

Pi <Pno(i), Pna(i), P(k)>; i=1, 2, ..., i, ..., n;　k=1, 2, ..., m;

Pi 代表产品属性库中的一成员；由属性编号 Pno(i)，属性名称 Pna(i)，所属类别 P(k)组成. 例如，中央处理器是笔记本电脑的一个属性，规定这一属性的类别为 1. 而不同消费者在评论中可以用“CPU、处理器、中央处理器指代这一属性”，即:

Pno(1)=1, Pna(i)=“CPU”, P(k)=1

O o Pno(2)=2, Pna(i)=“处理器”, P(k)=1

Pno(3)=3, Pna(i)=“中央处理器”, P(k)=1

依照上述方法对笔记本所有属性进行归纳整理，形成一个能够代表笔记本电脑整体的数据库.

3.3.2 意见摘要生成

由于是从正面情感倾向文本中获得的产品属性得到了消费者的认同，则企业考虑对这些属性加以宣传. 相反，从负面倾向评论文本选取的大多数产品属性，企业则考虑加以改进. 意见摘要流程图如图 3 所示:

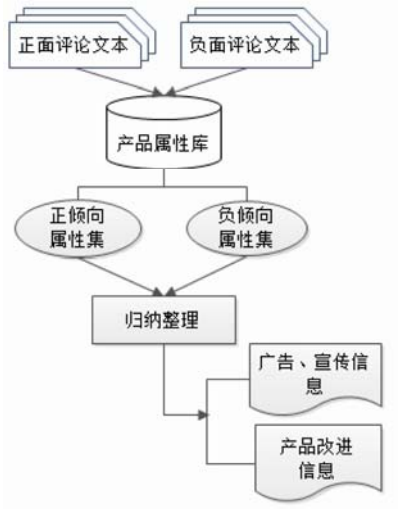

图 3　笔记本电脑特征归纳流程图

4　实验结果分析

本文选取 2011 年 6 月至 8 月，淘宝网、当当网上笔记本的评价详情，共 1200 篇评论文档. 以及一周内销售的型号为 Q-46 的三星笔记本电脑评论为预测数据集，共 45 篇. 测试数据的选择与训练数据不相关，用于减少实验结果的数据依赖性.

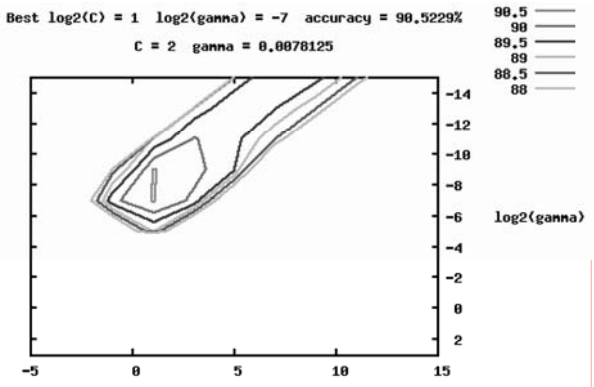

图 4　Libsvm 训练集情感分类分布图

实验结果证明，空间向量维数范围在 600-800 区间，BRF 核函数较线性核函数有更好的分类效果. 采用交叉验证选择最佳参数，支持向量机 BRF 核函数中的参数 C 值和 γ 值较稳定(C=2, gamma=0.0078)且正确率较高. 实验结果如下表 1 所示.

BRF 训练文档	400 (286)	600 (405)	800 (454)	1000 (500)
训练集正确率	87.59%	90.52%	91.03%	90.5%
预测集正确率	**68.89%**	**86.67%**	**84.44%**	**73.33%**
RBF 核 C 值	32	2	2	2
RBF 核 γ 值	3.05e-05	0.0078	0.0078	0.00195

表 1 评论文本情感倾向性分类指标

通过 libsvm 计算出的预测数据集的分类结果，本文利用查准率(Precision rate)与召回率(Recall rate)来评价评论文本的分类质量. 实验选择在不同的文档规模下，预测数据集的各项评价指标的变化如表 2 所示.

评论文本与特征词 \ 预测数据的各项评价指标	查准率 (Precision rate)		召回率 (Recall rate)		F-score	
	正类（1）	负类（-1）	正类（1）	负类（-1）	正类（1）	负类（-1）
400	94.74%	50%	58.06%	92.86%	0.7200	0.6500
600	96.30%	72.22%	83.87%	92.86%	0.8966	0.8124
800	96.15%	68.42%	80.65%	92.86%	0.8773	0.7878
1000	95.24%	54.17%	64.52%	92.86%	0.7693	0.6842

表 2 评论文本情感倾向分类结果的评估标准表

分类引擎对于负面情感倾向的识别更加精确，随着 VSM 空间向量的不断扩展，分类结果的差异主要体现在对于正倾向文本的识别上. 文档和特征词的增加，扩大了数据噪声的影响范围，使得分类引擎的性能受到限制.

特征归纳之后，对于 Q-46 型号笔记本电脑的网络销售评论中，在外观、速度、内存、性能、价格、质量等评价情况良好，客户反映趋于正面的评论；而负面评论中主要涉及到的属性包括显示屏幕、声音、鼠标等. 综上所述，对于下一代产品的研发，企业可以在这款笔记本的外观、价格、性能上面进行宣传. 对于硬件方面如散热、屏幕、音响等进行改良.

5 总结

随着市场竞争的日益加剧，构建智能型企业已经成为当今社会的一大趋势. 企业在获得大量数据时，
需要及时高效的做出反应，适应顾客需求变化的需要. 本系统是商务智能的一个实例，智能化的实现了从网络评论中挖掘出关于产品和服务改进的信息，帮助管理者减少收集处理信息的时间，及时推出符合市场需求的高质量产品，提高产品在市场上的响应能力. 但本文并没有解决指代消解问题，仅从显示的、最能够表现情感倾向的词汇类型作为特征关键词选择的标准，实际中应该有对具体的评价对象识别、对象之间情感关系的的选取，这也是进行下一步的研究的具体内容.

参考文献

1 李纲,程明结,寇广增.基于情感倾向识别的汽车评论挖掘系统构建.情报学报,2011,30(2):204–211.
2 白鸽,左万利,赵乾坤,曲仁镜.使用机器学习对汉语评论进行情感分类.吉林大学学报(理学版),2009,47(6):1260–1263.
3 王素格,李伟.面向中日关系论坛的情感分类问题研究.计算机工程与应用,2007,43(32):174–177.
4 王洪伟,尹裴,廖亚国.Web 文本情感分类研究综述.情报学报,2010,29(5):981–986.
5 胡熠,陆汝占,李学宁,段建勇,陈玉良.基于语言建模的文本情感分类研究.计算机研究与发展,2007,44(9):1469–1475.
6 吴鹏.支持向量机文本分类算法的研究及其应用[硕士学位论文].大连:大连理工大学,2009.
7 王林梅.Web 用户评价的自动情感分析[硕士学位论文].天津:天津大学,2009.

(下转第 42 页)

在图 4 中，定位误差为 1m 时，误差累积分布已达到 0.5，意味着定位均方根误差在 1m 以内的点数达到了总实验点数的一半. 而定位误差为 5m 时，误差累积分布为 1. 从以上数据可知，相对于定位点所处的区域大小，该系统定位精度较高，能较为准确的对标签进行定位，有比较理想的实用性.

5 结语

本系统在 LOS 情况下具有定位精度高、便携、系统可维护等优点，适合具有挑战性环境下的定位，若在 NLOS 传输情况下，由于信号直达路径被障碍物阻挡，延长了信号传播时间，导致定位结果偏离真实坐标点位置较大，定位精度会有所下降. 因此，后续的研究中需不断地改进定位算法，提高 NLOS 情况下的定位精度，同时可将此系统构建成一个具有精确定位和通信功能的自组织无线传感器网络，具有很大的应用背景和研究意义.

参考文献

1 Dulmage J, Cioffi R, Fitz MP, Cabric CD. Characterization of Distance Error with Received Signal Strength Ranging. IEEE Conference on Wireless Communications and Networking Conference. 2010. 1–6.

2 Wang S, Waadt A, Burnic A, Xu D, Kocks C, Bruck GH, Jung P. System Implementation Study on RSSI based Positioning in UWB Networks. IEEE Conference on 7th International Symposium on Wireless Communication Systems. 2010. 36–40.

3 余芳文,胡旭科.基于线性调频的 nanoLOC 新技术与应用研究.信息通信,2011,(2):4–6.

4 Nanotron P. nanoLOC TRX Transceiver. http://www.nanotron.com/EN/pdf/Factsheet_nanoLOC-NA5TR1.pdf.

5 Brugger M, Christ T, Kemeth F, Nagy S, Schaefer M, Pietrzyk MM. The FMCW Technology-Based Indoor Localization System. IEEE Conference on Ubiquitous Positioning Indoor Navigation and Location Based Service. 2010. 1–6.

6 王永虹,徐炜,郝立平.STM32 系列 ARM Cortex-M3 微控制器原理与实践.北京:北京航空航天大学出版社,2008.40–51.

7 Caffery J. A new approach to the geometry of TOA location. The 52nd IEEE Vehicular Technology Conference. 2000. 1943–1949.

8 朱宇,张鑫,李斌.基于嵌入式定位系统的研究.微电子学与计算机,2011,28(4):181–183.

9 钱晨,徐荣华,王钦若.基于 Linux 操作系统的设备驱动程序开发.微计算机信息,2004,20(9):131–133.

(上接第 13 页)

8 Turney P. Thumbs Up or Thumbs Down Semantic Orientation Applied To Unsupervised Classification of Reviews. Meeting of the Association for Computational Linguistics (ACL'02). 2002: 417–424.

9 Kim S, Hovy E. Determining the Sentiment of Opinions. Int'l Conf. on Computational Linguistics(CONLING'04). 2004.

10 Riloff E, Wiebe J. Learning Extraction Patterns for Subjective Expressions. Conference on Empirical Methods in Natural Language Processing(EMNLP'03). 2003.

11 Hu M, Liu B. Mining Opining Features in Customer Reviews. 19th National Conf. on Artificial Intelligence (AAAI'04). 2004: 755-760.

12 Tong R. An Operational System for Detecting and Tracking Opinions in On-line Discussion. SIGIR Workshop on Operational Text Classification. 2001

13 Liu B. Web 数据挖掘.北京:清华大学出版社,2009.296–318.

14 赵卫东.商务智能.北京:清华大学出版社,2009.

15 纪系禹.数据挖掘技术应用实例.北京:机械工业出版社,2009.