doi:10.3969/j.issn.1673-9833.2012.05.017

基于SVM的产品评论属性特征的情感倾向分析

王文华，朱艳辉，徐叶强，杜 锐，鲁 琳，邓 程

（湖南工业大学 计算机与通信学院，湖南 株洲 412007）

摘 要：产品评论的情感倾向性分析是一个很有研究价值的领域，可以帮助客户、商家进行决策。针对产品评论中的属性词和情感词在文本中的各种关系，制定了8组特征选择规则，利用SVM算法训练模型来判断属性词和情感词的搭配识别，进而依据情感词及否定词等分析属性特征的情感倾向。实验结果表明：提出的基于SVM的搭配识别方法，在识别属性特征与情感词的搭配方面具有不错的分类效果。

关键词：支持向量机；属性搭配；情感极性分析；文本分类；中文信息处理

中图分类号：TP391 **文献标志码**：A **文章编号**：1673-9833(2012)05-0076-05

Analysis on Emotional Tendences of Attribute Characteristics in Product Reviews Based on SVM

Wang Wenhua，Zhu Yanhui，Xu Yeqiang，Du Rui，Lu Lin，Deng Cheng

(School of Computer and Communication，Hunan University of Technology，Zhuzhou Hunan 412007，China)

Abstract：Emotion analysis of product reviews is a valuable research field, it can help customers and merchants make decisions. Based on various relations of attribute words and emotion words in product review texts, develops eight sets of feature selection rules, and applies SVM algorithm training model to judge the identification of attribute words and emotions words, then analyzes the emotion tendencies of attribute characteristics based on emotion words and negative words. The experimental results show: the proposed SVM-based recognition method achieves good classification effect in the identification of attributes and emotion words.

Keywords：support vector machine；attribute collocation；emotional analysis；text categorization；chinese information processing

Web2.0时代的到来，标志着动态交互网站逐渐取代了传统的静态展现网站。随着淘宝、京东、卓越、当当等购物网站的流行，越来越多的人开始网上购物，客户和网站间的交互也越来越频繁，网上的评论信息也越来越多。据第30次中国互联网络发展状况统计报告显示[1]，截至2012年6月底，中国网民规模达5.38亿，网络购物用户截至d2012年6月底为2.10亿人，且现在依然在平缓上升。网购者往往会留下评论来对所购买的商品表达自己的情感，评论信息会成为其他潜在消费者以及商家的一个重要参考依据。然而面对浩如烟海的无结构评论信息，客户、商家无法及时准确地发现自己想要的产品参

收稿日期：2012-07-25

基金项目：国家自然科学基金资助项目（61170102），湖南省自然科学基金资助项目（10JJ3002），国家社会科学基金资助项目（12BYY045），教育部人文社会科学研究基金资助青年项目（09YJCZH019），中国包装总公司科研基金资助项目（2008-XK13）

作者简介：王文华（1987-），男，河南驻马店人，湖南工业大学硕士生，主要研究方向为文本分类与信息检索，E-mail：yueerdelei@163.com

考信息，如果单靠人工浏览的方式去浏览这些信息十分费时，而且获得的信息带有一定的片面性。因此，如何把互联网上杂乱无章的海量产品评论信息进行挖掘处理，并对评论中蕴含的情感进行情感倾向性分析成为了近些年来的一个研究热点。

1　研究背景

产品评论情感倾向性分析以Web上用户发表的非结构化产品评论文本为挖掘对象，采用自然语言处理技术，从大量文本数据中挖掘消费者对某款产品的各属性特征持肯定或否定的态度及对整个产品的情感倾向。近年来，国内外专家学者对此的研究方兴未艾。文献[2]使用语料中的特征进行打分标注，再累加分数得到整篇文章的情感倾向性。但无法解决有些特征在不同语境中的情感倾向性是不同的情况。文献[3]利用最大熵方法和朴素贝叶斯方法对新闻评论语料进行情感倾向性分析，准确率高达90%，但召回率不高。文献[4]利用最大熵模型对句法分析得到的评价对象和评价短语的路径特征进行识别，比Baseline方法的准确率高，但不规范的评论文本会影响句法分析结果。文献[5]利用标注的情感词语义角色到意见持有者的映射关系，将具相关角色格的词认为是意见持有者。文献[6]在引入条件随机场对意见持有者识别的基础上增加了上下文依存关系和位置等特征来提高准确率，其F-measure为47%左右。

目前所有分析方法的准确率、覆盖率、F-measure值都不太高。本文在已有研究的产品评论属性词抽取[7]及基础情感词词典构建[8]的基础上，结合COAE2011[9]中任务3的标准答案中的观点词，提出一种基于支持向量机的属性词与情感词搭配识别的方法，进而依据否定词、程度词和情感词的倾向来对属性特征的情感倾向性进行分析。

2　词典的构建

在产品评论文本中，评论者往往针对产品的很多属性，利用情感词来表达自己对产品的观点，因此属性词的选择和情感词的确定对评论的情感倾向性有很大的影响。例如：“诺基亚X3的屏幕大小设计非常合理，颜色也相当绚丽。”其中“屏幕”、“颜色”等属于诺基亚X3的属性，“合理”、“绚丽”为情感词。为了判断属性特征的情感倾向性，在之前研究的基础上构建了属性词词典、情感词词典、程度词词典及否定词词典。

2.1　属性词词典

针对手机领域开展研究，利用之前研究的产品评论属性词抽取[7]的方法获得了1 400个属性词，作为属性词典AttrDic。部分属性词有：屏幕、按键、电池、铃声、短信、待机时间、颜色等。

2.2　情感词词典

首先从知网（http://www.keenage.com/html/e_index.html）上下载的资源包中选取评价词和情感词作为基础情感词典，共得到8 936个词。由于产品评论中存在很多网络用语，同时人工收集整理了42个网络情感词，像“顶、NB、给力、垃圾、悲剧、不咋地”等，与基础情感词典结合在一起作为情感词集。由于文本的情感分类中，情感词所表达的情感倾向性强度不同，所以利用互信息算法对情感词赋予一定权重，并归一化到−1~1之间。得到的部分情感词及其权值见表1。

表1　带情感权重的部分情感词

Table 1　Some emotional words with emotional weight

正面情感词	权值	负面情感词	权值
好	0.939 914	板脸	−0.956 710
漂亮	0.878 707	诅咒	−0.941 896
完美	0.815 314	犹豫不决	−0.933 324
绝妙	0.815 118	假冒	−0.854 521
拔尖儿	0.790 379	贱	−0.788 520
圆润	0.780 568	卖不掉	−0.761 768
带劲	0.750 648	寒酸	−0.737 318

2.3　程度词词典

程度词通常修饰形容词或者动词，也能和名词连用表达情感倾向性，在文本分类中起着重要作用。从评论文本中及网络上收集了部分程度词，并将其分为极量、高量、中量、低量，作为程度词典DegreeDic，程度词词典见表2。

表2　程度词词典

Table 2　Degree word dictionary

程度	程度词
极量	极　极为　及其　极度　极端　至　至为　顶　过　过于　过分　分外　万分　百分之百　最　最为
高量	很　太　挺　怪　老　非常　特别　相当　十分　好　好不　甚　甚为　颇　颇为　异常　深为　满　蛮　够　多　多么　殊　特　大　大为　何等　何其　尤其　无比　尤为　不胜　更　更加　更为　更其　越　越发　备加　愈加　愈　愈发　愈为　愈益　越加　格外　益发
中量	不大　不太　不很　不甚　较　比较　较比　较为　还
低量	有点　有点儿　有些　一　稍　稍稍　稍微　稍为　稍许　略　略略　略微　略为　些微　多少

2.4　否定词词典

虽然否定词本身在语言表达中不具有情感倾向

性，但却可以使文本情感倾向性反转或影响情感倾向性程度，尤其是和程度词连用，例如：

句子1 诺基亚X3的按键不灵活。

句子2 诺基亚X3的按键很不灵活。

句子3 诺基亚X3的按键不太灵活。

3个句子中均有否定词“不”和正面情感词“灵活”，但表达的情感倾向性程度却不相同。句子1中的“不”让“灵活”的倾向性反转，不过情感强度却没有“灵活”的反义词“笨拙”那么强烈。句子2中的“不”与程度词“很”连用，表达了很强的情感倾向性。而句子3中的“不”与程度词“太”连用，却减弱了负面的情感倾向性。通过人工整理，收集了18个否定词，作为否定词词典NegDic{不，没，没有，别，甭，非，否，无，不曾，不必，不然，莫，勿，未必，未曾，未尝，无从，无需 }。

3 基于SVM的搭配识别及情感倾向性分析

支持向量机[10]（support vector machine，SVM）是在统计学习理论基础上发展而来的一种机器学习方法，基于结构风险最小化原理，通过核函数将输入空间映射到一个高维空间，并找到一个具有最大分类间隔的最优分类超平面。支持向量机在解决样本二类分类问题中有其独特的优势，并在许多实际应用中取得了较好的结果。

3.1 特征选择规则

结合属性词词典、情感词词典、程度词词典和否定词词典，针对属性词和情感词可能的搭配，以及方便进行SVM分类，将属性特征与情感词之间的关系量化，选择了如规则：

1）属性特征和情感词的距离。计算属性特征与情感词之间的汉字、英文、标点符号的长度总和，不包括空格。

2）属性特征和情感词的先后位置。如果属性特征在前面就记为1；在后面就记为0。

3）属性特征和情感词之间是否有标点符号。如果有就记为1；没有就记为0。

4）属性特征和情感词之间是否还有属性特征。如果有就记为1；没有就记为0。

5）属性特征和情感词之间是否还有情感词。如果有就记为1；没有就记为0。

6）情感词的长度。

7）属性特征和情感词之间的区域内，情感词前的3个词（如果属性特征与情感词的距离小于3，那么就按实际有多个词来计算）中是否有程度词。如果有，按照极量、高量、中量、低量分别记为1,2,3,4；如果没有，就记为0。

8）属性特征和情感词之间的区域内，情感词前的3个词（如果属性特征与情感词的距离小于3，那么就按实际有多个词来计算）中是否有否定词。如果有就记为1；没有就记为0。

例1 假设评论文本的片段为：“性价比很高，外观时尚。”其中属性词有“性价比”和“外观”，评价短语有“高”和“时尚”，它们之间的8个特征形成的原始特征模板如下：

性价比	高	01000110；
性价比	时尚	51111210；
外观	高	10100100；
外观	时尚	01000200。

其中第1列是属性特征，第2列是情感词，后面8列分别对应上述的8个规则。

3.2 搭配识别算法流程

算法的流程图如图1所示。

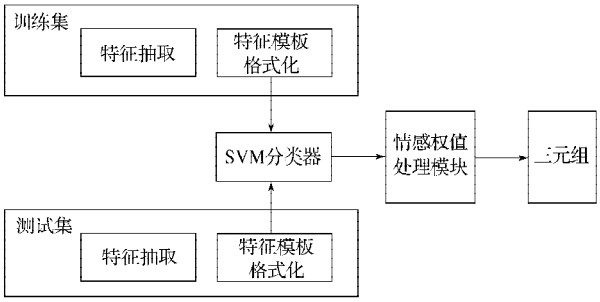

图1 搭配识别算法流程图

Fig. 1 Flowchart of collocation recognition algorithm

算法的描述如下。

1）首先利用中科院的ICTCLAS（http://www.ictclas.org/）分词系统对文本集中的评论文本进行分词。扫描分词后的每一篇评论文本中的词语跟属性特征词典与情感词词典进行比较，如果评论文本中的词语有与词典匹配的，将抽取结果分别保存为属性特征词集合和带有情感倾向性的情感词集合。然后将抽取出的两个词集做笛卡尔集，得到初始三元组<属性词，情感词，情感权重>。例如：“诺基亚X3的屏幕颜色很绚丽，但按键迟钝。”即可获得<屏幕颜色，绚丽，0.875 682>,<按键，迟钝，−0.884 202>等三元组，其中0.875 682表示正面评价，−0.884 202表示负面评价。

2）利用3.1节给定的特征规则，首先形成原始特征模版，然后用程序实现对初始三元组的扩展，再手工对所有扩展后的三元组进行标注，如果属性特

征与情感词搭配标注为 1，不搭配标注为 −1。如例 1 得到的带标注的三元组扩展格式为：

1　性价比　高　0.822 053　0 1 0 0 0 1 1 0；
−1　性价比　时尚　0.251 405　5 1 1 1 1 2 1 0；
1　外观　高　0.822 053　1 0 1 0 0 1 0 0；
1　外观　时尚　0.251 405　0 1 0 0 0 2 0 0。

去掉原始文本、三元组，保留带标注结果的扩展后的三元组，然后在每个特征前面加上序号，处理成 SVM 能识别的格式：

1　1:0　2:1　3:0　4:0　5:0　6:1　7:1　8:0；
−1　1:5　2:1　3:1　4:1　5:1　6:2　7:1　8:0；
−1　1:1　2:0　3:1　4:0　5:0　6:1　7:0　8:0；
1　1:0　2:1　3:0　4:0　5:0　6:2　7:0　8:0。

从标注集合中任意挑选一定比例的搭配和不搭配的项作为训练集，存为 train.txt。设置 SVM 训练参数，生成模型文件 train.txt.model（有关比例选择和参数设定在 4.1 节中阐述）。

3）对于识别未知文本中的搭配，用同样的方法首先处理成原始特征模版，将搭配项全部初始化为 1（设定为 −1，效果一样），同样处理成 SVM 能识别的格式。利用 train.txt.model 文件和 SVM 分类器对未知文本生成搭配结果。

4）情感权值处理模块包括两个功能：一是过滤搭配结果，删除不能搭配的属性特征和情感词的结果。不能搭配的属性特征和情感词较多，如“外观”和“大”不能搭配；二是把否定词、程度词考虑进去，并计算属性特征最终的情感权重。

3.3　情感权值处理模块

利用训练的模型文件，通过 SVM 分类器对未知文本的属性词和情感词是否搭配作出判断，而属性特征的最终情感倾向性值还需要通过情感权值处理模块来获得。

在对未知文本进行分类生成的搭配结果中，首先删除那些属性特征和情感词不能搭配的结果，在搭配的结果中选择特征选择规则 3.1 节中的 7 和 8）对应的特征项。如果 7）对应的特征项为<1,2,3,4>，则情感词相应情感权值分别乘以<1.4,1.2,0.8,0.6>；如果 8）对应特征项为 1，则情感词相应情感权值分别乘以 −0.8。对情感词情感倾向重新计算后，最终截取搭配结果模板文件的部分，得到三元组<属性词，情感词，情感最终权值>。

例如：“NokiaC2-01 反应有点慢，音质不好，像素不算很高，连拍很吃力。”其中，针对属性特征“音质”，其对应情感词“好”（权重为 0.939 914）前面出现否定词“不”，相应的情感倾向值变为 0.939 914 × (−0.8)=−0.751 931 2，情感倾向发生了变化。属性特征“反应”，其对应情感词“慢”前面出现了程度词“有点”，则把“反应慢”的情感倾向弱化了，情感词典里“慢”的权重是 −0.95，程度词“有点”属于低量程度词，其权重为 0.6，针对属性词“反应”的最终情感权重是−0.95 × 0.85=−0.807 5。对文中其他属性情感倾向采用同样的算法计算，获得最终的三元组。

4　实验及结果分析

4.1　实验数据及参数选择

从第三届中文倾向性分析评测会议[9]digital 领域的 14 799 篇评论文本中任意挑选了 400 篇作为数据源。实验中的 SVM 分类模型使用台湾大学林智仁等开发设计的 LIBSVM[11]软件包，训练集和测试集按 3:1 随机挑选。通过 3.2 节的方法人工对数据源进行标注。其中训练集中共有 5 676 条数据，其中搭配的为 1 576 条，不搭配的为 4 200 条。测试集中共有 2 750 条数据，其中搭配的总数是 750 条，不搭配的数据总数是 2 000 条。利用 3.1 节的 8 种特征规则，按照 3.2 节的算法确定训练文本格式。LIBSVM 模型参数为：

1）−s　0，SVM 类型为 C-SVC；

2）−t　2，核函数类型为 radial basis function，即 exp(-gramma*|u−v|^2)；

3）−e　0.1，epsilon 值；

4）−g　gamma 值；

5）−c　5，cost 值。

4.2　实验结果

算法是在 Eclipse 平台上，用 Java 语言实现的。实验结果采用第三届中文倾向性分析评测报告[9]中任务三的标准答案中属性特征为依据，利用准确率（precision）、召回率（recall）和 F 均值（F-measure）3 个指标评估方法的性能。介绍如下：

$$\text{准确率}=\frac{\text{提交结果中与人工标注匹配数目}}{\text{提交的所有观点句的数目}}\times 100\%,$$

$$\text{召回率}=\frac{\text{提交结果中与人工标注匹配数目}}{\text{人工标注结果中观点句的数目}}\times 100\%,$$

$$\text{F均值}=\frac{2\times\text{准确率}+\text{召回率}}{\text{准确率}+\text{召回率}}\times 100\%\ 。$$

实验结果如表 3 所示。

表 3　评价搭配实验结果

Table 3　The result of evaluate collocation experiments

准确率	召回率	F 均值
0.85	0.64	0.73

采用最近邻算法以及哈尔滨工业大学信息检索实验室开发的Deparser句法分词器依存语法的语法体系（dependency grammar）作为对比对象进行实验。实验结果如图2所示。

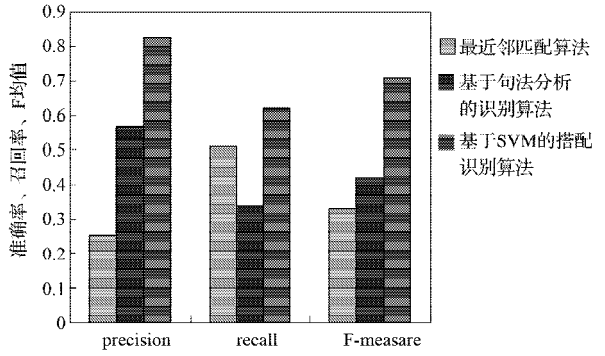

图2 实验结果对比

Fig. 2 Contrast of experiment results

实验结果表明：

1）最近邻匹配算法，准确率比较低，仅为0.25，覆盖率达0.51，而基于句法分析的搭配识别算法的准确率达0.57，但覆盖率非常低，仅为0.34，可见最近邻匹配算法能够提高覆盖率但是却降低了准确率，而基于依存句法关系的识别算法对句子的句法结构要求比较严格，由于网络上的评论文本比较口语化，依存句法关系对口语化的句子分析效果并不好，所以覆盖率比较低。

2）利用SVM机器学习方法，训练生成的model，对测试文本进行测试的效果最好，准确率达0.83，覆盖率达0.62，F值为0.71，取得了较好的性能。

5 结语

利用属性词和情感词的搭配及情感词权值可以识别属性词的情感倾向。本文首先构建了几个词典，然后利用所选特征规则，将文本中属性词与情感词在文本中各种关系进行量化，再利用支持向量机来进行属性特征和情感词的搭配识别。实验结果表明，提出的搭配识别算法取得了较好的效果。

参考文献：

[1] 中国互联网络信息中心.第30次中国互联网络发展状况统计报告[DB/OL]. [2012-06-20]. http: //www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201207/t20120723_32497.htm.

[2] Dave Kushal，Lawrence Steve，Pennock David M. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews[C]//Proceedings of the 12th International Conference on World Wide Web.[S.l.]：ACM，2003：519-528.

[3] 徐 军,丁宇新,王晓龙. 使用机器学习方法进行新闻的情感自动分类[J]. 中文信息学报，2007，21(6)：95-100.
Xu Jun，Ding Yuxin，Wang Xiaolong. Sentiment Classification for Chinese News Using Machine Learning Methods[J]. Journal of Chinese Information Processing，2007，21(6)：95-100.

[4] 樊 娜,蔡皖东,赵 煜. 基于最大熵模型的观点句主观关系提取[J]. 计算机工程，2010，36(2)：4-6.
Fan Na，Cai Wandong，Zhao Yu. Extraction of Subjective Relation in Opinion Sentences Based on Maximum Entropy Model[J]. Computer Engineering，2010，36(2)：4-6.

[5] Kim SooMin，Eduard Hovy. Extracting Opinions，Opinion Holders and Topics Expressed in Online News Media Text [C]//Proceedings of the Workshop on Sentiment and Subjectivity in Text. Pennsylvania：Association for Computational Linguistics，2006：1-8.

[6] Liu K，Zhao J. NLPR at Multilingual Opinion Analysis Task in NTCIR7[C]//Proceedings of the Seventh NTCIR Workshop. Tokyo：[s.n.], 2008：226-231.

[7] 栗春亮，朱艳辉，徐叶强. 中文产品评论中属性词抽取方法研究[J]. 计算机工程，2011，37(12)：26-28.
Li Chunliang，Zhu Yanhui，Xu Yeqiang. Research of Attribute Word Extraction Method in Chinese Product Comment[J]. Computer Engineering，2011，37(12)：26-28.

[8] 柳位平，朱艳辉，栗春亮，等. 中文基础情感词词典构建方法研究[J]. 计算机应用，2009，29(10)：2875-2877.
Liu Weiping，Zhu Yanhui，Li Chunliang，et al. Research on Building Chinese Basic Semantic Lexion[J]. Journal of Computer Applications，2009，29(10)：2875-2877.

[9] 许洪波，孙 乐，姚天昉，等. 第三届中文倾向性分析(COAE2011)总结报告[C]//第三届中文倾向性分析评测会议.济南:中国中文信息学会信息检索专业委员会，2011：1-24.
Xu Hongbo，Sun Le，Yao Tianfang，et al. Overview of the Third Chinese Opinion Analysis Evluation(COAE2011) [C]//The Third Chinese Opinion Analysis Evaluation (COAD2011). Jinan：Chinese Information Processing Society，The professional Committee of Information Retrieval，2011：1-24.

[10] Yang H W，Meng H M，Wu Z Y，et al. Modeling the Global Acoustic Correlates of Expressivity for Chinese Text-To-Speech Synthesis[C]//Workshop on Spoken Language Technology. Palm Beach：Conference Publications，2006：10-13.

[11] Chang Chihchung，Lin Chihjen. LIBSVM[DB/OL]. [2012-07-15]. http: //www.csie.ntu.edu.tw/~cjlin/libsvm/.

（责任编辑：申 剑）