# 智能信息系统中手机产品评论的情感倾向分析

楼博文[1],许歆艺[1],蔡 根[2],张 辰[1],刘功申[1]
(1. 上海交通大学 信息安全工程学院,上海 200240;
2. 华东师范大学 计算机科学技术系,上海 200241)

**摘 要**:目前国内用户购买和使用大量不同手机产品。为帮助手机生产商识别用户评论的情感倾向、为其他潜在的手机用户提供手机产品购买建议,文中通过模块设计构建一个处理手机产品评论的智能信息系统,该系统用于挖掘和分析针对手机产品的评论信息。其中情感倾向分析是该系统的核心环节,因此文中研究并提出了一种基于条件随机场的针对手机产品的情感倾向识别方法,并通过采用多种实验手段寻找并验证该识别方法的有效性,从而完成对手机产品评论的高效、准确的自动识别。

**关键词**:手机产品;信息系统;条件随机场;情感倾向
**中图分类号**:TP31 **文献标识码**:A **文章编号**:1673-629X(2013)12-0022-04
doi:10.3969/j.issn.1673-629X.2013.12.005

# Analysis on Sentiment Orientation of Mobile Phone Product Reviews in Intelligent Information System

LOU Bo-wen[1], XU Xin-yi[1], CAI Gen[2], ZHANG Chen[1], LIU Gong-shen[1]
(1. School of Information Security Engineering, Shanghai Jiaotong University, Shanghai 200240, China;
2. Department of Computer Science and Technology, East China Normal University, Shanghai 200241, China)

**Abstract**: Nowadays, more kinds of mobile phone products are used by consumers in China. Use module design to build an intelligent information system, which can help manufacturers recognize sentiment orientation of product reviews, and provide other consumers for suggestions on buying mobile phone products. This system has the function of mining and analyzing mobile phone product reviews, and the sentiment orientation is the key module of the system. Therefore, put forward a method based on conditional random fields to analyze sentiment orientation of mobile phone product reviews. For accurate and efficient automatic recognition, many methods are adopted in the experiment and the efficiency of the method is verified.

**Key words**: mobile phone product; information system; conditional random fields; sentiment orientation

## 0 引 言

当今,随着手机产品的不断丰富以及互联网的发展,越来越多的消费者用户选择购买多款手机,并且在互联网论坛上发表对所购买产品的评论。这些评论主要包括对单一产品不同属性的评价、单一产品整体性能评价、多种不同产品在某一具体属性上的比较以及多种不同产品在整体性能上的比较等。与此同时,产品的生产者也迫切希望能够从大量产品评论中有效识别出用户的观点,并且能够提取出对产品改进的有用信息。而对于购买产品的其他潜在用户,这些用户也希望从之前产品使用的评论中获得有效信息,并以此做出更加明智的消费决策。

## 1 国内手机使用情况介绍

伴随着全球进入智能手机时代,国内手机生产商也不断涌入智能手机市场。根据艾瑞咨询研究《2011~2012 年中国智能手机市场研究报告》[1] 显示,2011 年中国智能手机出货量达到 7 210 万台,智能手机渗透率达到 13%,并且目前中国的手机用户规模已经超过 10 亿。此外,移动互联网第三方数据挖掘和整合营销机构艾媒咨询(iiMedia Research)发布了《2012Q3 中国智能手机市场季度监测报告》[2]。该报告显示在

**收稿日期**:2013-03-06 **修回日期**:2013-06-10 **网络出版时间**:2013-09-29
**基金项目**:"国家级大学生创新创业训练计划"创新训练项目(201210248039,IPP5106);国家自然科学基金资助项目(61272441)
**作者简介**:楼博文(1991-),男,研究方向为自然语言处理、信息系统管理;刘功申,博士,副教授,研究方向为网络舆情、内容安全。
**网络出版地址**:http://www.cnki.net/kcms/detail/61.1450.TP.20130929.1541.039.html

2012 年第三季度中,中国智能手机用户继续保持增长态势。截止 2012 年第三季度,中国智能手机用户数达到 3.3 亿人,环比增长 13.8%。发展到现在,智能手机市场的竞争更加激烈,目前已形成了苹果、谷歌、"微软和诺基亚组合"的三足鼎立的局面。

国内针对手机产品的论坛和网站,如手机中国(http://dp.cnmo.com/)、太平洋电脑网(http://product.pconline.com.cn/mobile/)和中关村在线(http://mobile.zol.com.cn/)等,都有大量对于不同型号产品的评论,各手机生产商也关心着用户对手机产品使用情况的反馈。

## 2　针对手机产品评论的智能信息系统设计

由于临时性、非结构化评论信息的存在,为使信息得到有效管理和处理,构建智能信息系统。该系统有四个模块,分别为评论采集模块、评论预处理模块、情感倾向识别模块和评论结果反馈模块。各个模块相互合作,数据在各模块之间传递。

### 2.1　评论采集模块

该模块自动检测、挖掘、抽取相关论坛中对手机产品的评论信息,对一些论坛异构化的数据进行搜集整理,将采集完的有效评论统一存放到系统中的数据库中,等待其他模块调用处理。

### 2.2　评论预处理模块

该模块获取系统数据库中存放的产品评论信息,选取一定数量的评论并进行人工语料标记,并去除和手机产品无关的评论信息。

### 2.3　情感倾向识别模块

该模块主要对数据库中没有标记的语料进行情感倾向预测。此外,该模块将首先进行手机产品品牌识别,再对评论中涉及到的手机产品属性,如屏幕、价格、操作系统进行识别,最后进行评论的情感倾向识别,力求给用户反馈最大化的信息量。

### 2.4　评论结果反馈模块

通过上述一系列的系统处理后,该模块生成用户对某一具体手机产品的体验报告,包含用户对特定手机品牌的接受度(即正面情感)、拒绝度(即负面情感)、哪些手机属性有待改善等信息。

## 3　利用 CRFs 对手机产品评论情感倾向识别

文中构建的智能信息系统中最核心的模块是情感倾向识别模块,而情感倾向识别是目前自然语言处理领域的热点研究方向之一。网络上针对手机产品的评论信息有一些共同点,如文本长度较短(160 字以内)、除整体评价外都会提到手机产品的某个具体属性等。

文中采用条件随机场机器学习方法进行针对手机产品评论的情感倾向识别。条件随机场(Conditional Random Fields, CRFs)最早由 John Lafferty 等人提出[3]。目前 CRFs 可应用在自然语言处理中,并且在浅层句法分析、序列标注、命名实体识别、中文分词等任务中都有很好的表现。情感倾向识别在某种程度上可以作为情感的文本分类来处理,将 CRFs 预测序列标注的特点应用到情感倾向识别中。

CRFs 在 HMMs[4](隐式马尔可夫模型)与 MEMs[5](最大熵模型)的基础上进行了改进。HMMs 的输出是独立假设的,忽略了上下文特征,这将影响特征的选取。MEMs 解决了特征选取问题,但 MEMs 在每一个节点都需要归一化,因此只能找到局部的最优值,这将导致标注偏置问题(即忽略了训练语料中未出现的标记情况)发生。CRFs 使用条件特征,可以对特征进行全局归一化。它不是在给定当前状态的条件下定义下一个状态的分布,而是在给定需要标记的观察序列的条件下,计算整个标记序列的联合概率,从而避免了 HMMs 的输出独立假设问题。而且 CRFs 很好地解决了 MEMs 的标注偏置问题。因此,在现实的序列标注任务中,CRFs 性能往往都优于 HMMs 和 MEMs。

使用 CRFs 进行手机产品评论情感倾向识别的流程如图 1 所示。

### 3.1　线性链 CRFs 模型

CRFs 是一个无向图上的概率分布模型。最常用的一类 CRFs 为一阶链式结构[6],即线性链 CRFs 模型(linear chain CRFs),它计算给定观察值条件下输出状态的条件概率,是一种判别式模型,不需要独立性假设。

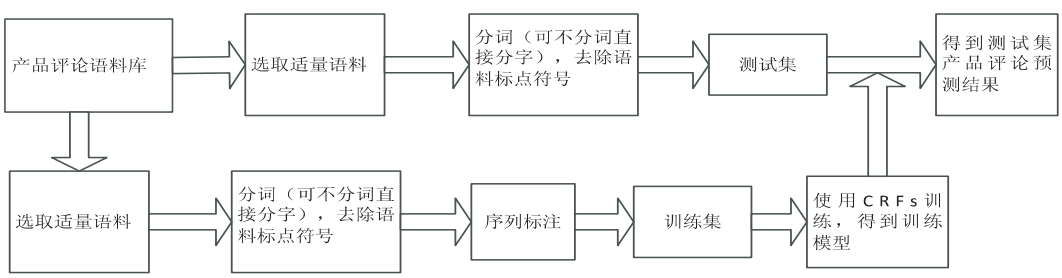


图 1　情感倾向识别流程

定义被观测的输入数据序列(如该文档中的词序列)为 $X=(x_1,x_2,\cdots,x_n)$,定义 $Y=(y_1,y_2,\cdots,y_n)$ 为待预测的标记序列,可以理解为是一个状态集合,每一个状态都与一个标号有关。其中,$x_i$ 表示 $X$ 的第 $i$ 个分量;$y_i$ 是 $x_i$ 对应的标记状态。因此,在一个输入数据序列给定的情况下,线性链 CRFs 定义标记状态序列 $Y$ 的条件概率为:

$$P(y \mid x)=\frac{1}{Z(x)}\exp\{\sum_{i=1}^{n}\sum_{k=1}^{K}\lambda_k f_k(y_i,y_{i-1},x)\}$$

其中,$Z(x)$ 是归一化参数;$f_k(y_i,y_{i-1},x)$ 是特征函数;$\lambda_k$ 为通过训练样本得到的特征函数的权重值。特征函数有两种形式,其中一种特征函数只与当前状态相关,另一种特征函数还与当前状态的前一个状态有关。

特征函数的权重 $\lambda_k$ 变化范围为 $-\infty$ 到 $+\infty$,可以使用最大似然估计法通过模型训练获得。设权重集合 $\Lambda=\{\lambda_i\}$,则最大似然估计:

$$L(\Lambda)=\sum_j[\log\frac{1}{Z(x^{(j)})}+\sum_k\lambda_k f_k(y^{(j)},x^{(j)})]$$

### 3.2 序列标注

文中智能系统使用 CRF++[7] 构建 CRFs 模型。为使 CRF++能够有效提取训练集句子特征,训练集中的句子分词后可以采用如图 2 格式。第一列是句子分词后句子中的各个单词,第二列可以是它们分别对应的词性,也可以为空,最后一列是词的分类标注。

对于图 2,将一条手机产品评论中的每个词作为第一列,将该评论的情感倾向标注作为第二列。每个词都标注为这个评论的情感倾向类别:正面,这样短文本就转化为一个标注后的序列。测试集中的每条手机产品评论句子分词后,将每个词作为第一列,句与句之间空行隔开,第二列情感倾向类别为空,使用 CRF++提取训练集特征进行预测。

| | |
|---|---|
| 应用 | 正面 |
| 丰富 | 正面 |
| 这个 | 正面 |
| 必须 | 正面 |
| 有 | 正面 |
| 而且 | 正面 |
| 做 | 正面 |
| 的 | 正面 |
| 还 | 正面 |
| 很 | 正面 |
| 不错 | 正面 |

图 2 序列标注格式

### 3.3 特征模板

使用 CRFs 进行机器学习需要特征模板,提取的不同特征对情感倾向分类的结果影响不同。CRF++拥有两类特征模板,并根据模板生成特征函数集合。其中一类是 Unigram 模板,如果用 $L$ 表示输出标注的种类数目、$N$ 表示模板生成的特征字符数目,则其特征函数总数为 $L*N$。另一类是 Bigram 模板,其特征函数总数为 $L*L*N$。采用图 3 模板对两列的训练集抽取特征,采用图 4 模板对三列的训练集抽取特征。

```
# Unigram

U1:%x[-2,0]
U2:%x[-1,0]
U3:%x[0,0]
U4:%x[1,0]
U5:%x[2,0]
U6:%x[-2,0]/%x[-1,0]
U7:%x[-1,0]/%x[0,0]
U8:%x[0,0]/%x[1,0]
U9:%x[1,0]/%x[2,0]

U10:%x[-2,0]/%x[-1,0]/%x[0,0]
U11:%x[-1,0]/%x[0,0]/%x[1,0]
U12:%x[0,0]/%x[1,0]/%x[2,0]

# Bigram
B
```

图 3 特征模板(针对两列序列)

```
# Unigram

U00:%x[-2,0]
U01:%x[-1,0]
U02:%x[0,0]
U03:%x[1,0]
U04:%x[2,0]
U05:%x[-1,0]/%x[0,0]
U06:%x[0,0]/%x[1,0]

U10:%x[-2,1]
U11:%x[-1,1]
U12:%x[0,1]
U13:%x[1,1]
U14:%x[2,1]
U15:%x[-2,1]/%x[-1,1]
U16:%x[-1,1]/%x[0,1]
U17:%x[0,1]/%x[1,1]
U18:%x[1,1]/%x[2,1]

U20:%x[-2,1]/%x[-1,1]/%x[0,1]
U21:%x[-1,1]/%x[0,1]/%x[1,1]
U22:%x[0,1]/%x[1,1]/%x[2,1]

# Bigram
B
```

图 4 特征模板(针对三列序列)

## 4　实验结果和分析

针对手机产品评论的情感倾向识别实验的语料来自中关村在线手机论坛,选择所评论的手机产品为苹果公司的 Iphone 系列手机和诺基亚公司的 Lumia 系列手机。对大量字符数目小于 160 且包含手机具体属性的评论进行了测评。

实验环境为操作系统 Windows Vista,系统配置为 CPU 2.4 GHz,内存 2 GB。智能系统识别具体产品品牌后,进行训练集和测试集的构建,实验训练集和测试集情况见表 1。

表 1　实验训练集和测试集介绍

| 手机产品 | 训练集数量 | 测试集数量 |
|---|---|---|
| Iphone | 150 | 50 |
| Lumia | 150 | 50 |

实验分词和特征词提取使用基于 CRFs 的 Stanford Word Segmenter[8],该系统在第二届国际汉语分词评测大赛 SIGHAN bakeoff 2005[9] 上对四种词类标准和语料库(台湾"中研院"、香港城市大学、北大标准、微软亚洲研究院)取得了较好的评测结果。实验对于 CRFs 模型的建立采用工具 CRF++-0.57。情感倾向分类评价使用常见分类任务中的三个指标:准确率(precision),召回率(recall)和 $F_1$ 值。其中 $F_1$ 值是综合了准确率和召回率,是主要评价标准,计算公式为:

$$F_1=\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}\times 100\%$$

采用三种方法在同一台配置的电脑上进行对比实验,以取得最佳的针对手机产品评论情感分析效果。三种方法分别为:

(1)文献[10]中对训练集不分词而标记每个字的方法,同样去除标点。

(2)使用 Stanford Word Segmenter 构建 CRFs 模型再进行情感倾向预测,训练集中仅有两列,一列分词,一列标记结果,特征模板选用图 3 所示模板。

(3)使用 Stanford Word Segmenter 构建 CRFs 模型再进行情感倾向预测,训练集中有方法(2)中的两列,还包含对所分词的词性列。使用 Stanford POS Tagger[11] 进行词性标记,特征模板选用图 4 所示模板。

三种情感倾向识别方法实验结果见表 2。

表 2　三种情感倾向识别方法实验结果

| 方法 | 准确率/% | 查全率/% | $F_1$ 值/% |
|---|---|---|---|
| 方法(1) | 85.13 | 86.21 | 85.67 |
| 方法(2) | 92.17 | 87.43 | 89.74 |
| 方法(3) | 83.81 | 83.24 | 83.52 |

文献[12]使用 SVM 监督性算法对手机产品评论进行了情感倾向识别,选择的手机产品为 MOTO1200,得到的总体准确率为 78%,查全率为 78%,$F_1$ 值为 78%。由表 2 实验结果可得,对于 Iphone 和 Lumia 等手机产品评论的情感倾向识别中,使用 CRFs 方法能对情感倾向分类中的准确率、查全率和 $F_1$ 值带来一定程度的提升。此外,使用图 3 所示的二列模板,即对于手机产品评论的情感倾向分类使用实验方法(2),三个衡量指标取得了更高的值。

## 5　结束语

文中所构建的针对手机产品评论的智能信息系统,其核心模块在于对评论信息的情感倾向识别。基于 CRFs 的机器学习模型在针对手机产品评论的情感倾向识别上取得了不错的结果,尤其是使用 Stanford Word Segmenter 对评论句子进行分词,再标记序列构建两列训练集后能够对测试评论集进行更加有效的情感倾向预测。

文中的情感倾向分类为三类,分别为:正面、中立和负面。但在产品论坛、购物网站中,很多评论都带有感情程度,尤其对于正面情感和负面情感,未来应该在这两大类情感中进一步细分情感程度,这样会对手机产品生产商和用户提供更多有意义的情感倾向信息反馈,也将使系统最大化地利用好所获取的每一条信息的价值。

**参考文献:**

[1] 艾瑞咨询. 2011-2012 年中国智能手机市场研究报告[EB/OL]. 2012. http://www.iresearch.com.cn/Report/1668.html.

[2] 艾媒咨询. 2012Q3 中国智能手机市场季度监测报告[EB/OL]. 2012/http://share.iimedia.cn/repd/85.

[3] Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th international conference on machine learning. Williamstown, MA, USA: Morgan Kaufmann Publishers Inc., 2001: 282-289.

[4] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[5] McCallum A, Freitag D, Pereira F C N. Maximum entropy Markov models for information extraction and segmentation[C]//Proceedings of the 17th international conference on machine learning. Stanford University, Stanford, CA, USA: Morgan Kaufmann Publishers Inc., 2000: 591-598.

[6] Levow G A. Automatic prosodic labeling with conditional random fields and rich acoustic features[C]//Proceedings of the third international joint conference on natural language processing. Hyderabad, India: [s.n.], 2008: 217-224.

(下转第 29 页)

源的子线程的句柄,这一句柄可以通过 CreateThread 函数的返回值得到。

有了这个链表之后,可以在每个子程序的运行之初通过 SuspendThread 函数将各自链表中的除了当前线程的其他子线程都挂起,而在每个线程结束时再通过 ResumeThread 函数将它们再一一唤醒(当然,在执行挂起和唤醒操作时需要使用上文提到的线程同步技术对它们进行保护)。

因为被挂起的线程,CPU 是不会分配时间片给它们的,所以这样就可以避免上述问题中对时间片的浪费。

这种同步技术主要适用于需要对同一硬件资源进行操作的多个线程,因为一般情况下硬件资源如通讯端口地址都是比较稳定的,所以可以对不同的端口建立不同的链表,用于存储特定的线程句柄。对于在同一程序中存在多个线程都需要频繁并且较长时间地操作同一硬件资源时,这一方法的效果就会有所体现。

## 4　结束语

文中主要分析了四种较常用的 Win32 环境下的多线程技术,并分析了各种同步技术的特点。最后,在常用的线程同步技术的基础上提出了一种改进的线程同步技术算法,这种同步技术的使用面可能没用常用的线程同步技术那么广,但是针对特定的情况,这一同步技术可以很好地提高程序的运行效率,提高了 CPU 时间片的利用率。

各个技术之间不存在哪个技术好,哪个技术不好,只有适合与不适合。在不同的程序环境下需要由程序员来选择合适的同步技术。甚至,在一个复杂的程序中程序员需要使用多种线程同步技术来进行协同的工作。

**参考文献:**


[1] 马魁涛,蔡　颖,郭宝峰. Win32 进程间信息共享的实现方法研究[J]. 计算机应用与软件,2007,24(12):119-120.

[2] 刘红海,候向华,蔡　勇. 基于 Win32 的多线程技术及其应用[J]. 计算机工程与设计,2003,24(10):113-115.

[3] 郝晓艳,杨　波,孙奕奇. Win32 下线程同步及其在 MFC 中的处理[J]. 济南大学学报(自然科学版),2002,16(4):332-335.

[4] 朱华章,孟凤珍. 基于 Win32 的多线程 PCI 设备驱动程序设计[J]. 计算机工程与应用,2004(4):107-111.

[5] Feng Jing,Zhong Hu,Ao Guoqiang,et al. Principles and application of the real-time hardware-in-the-loop simulation platform based on multi-thread and CAN [C]//Proc of ISIE. [s. l.]:IEEE,2008:2225-2230.

[6] Wang Yongwen,Zheng Qianbing,Dou Qiang,et al. Low power design for a multi-core multi-thread microprocessor [C]// Proc of GreenCom. Hangzhou:IEEE,2010:351-356.

[7] Liu Hong, Zheng Jianxiang, Chen Ying. The application of multi-thread-based embedded system in the fire monitor [C]//Proc of ISECS. Washington DC:IEEE Computer Society,2009:506-508.

[8] 孙云霞. Win32 多线程编程的控制技术[J]. 电脑编程技巧与维护,2008(17):13-15.

[9] 张红玲. Win32 环境下的多线程技术[J]. 沈阳师范学院学报(自然科学版),2000,18(4):11-16.

[10] 王日宏. 基于 VC 的 Win32 多线程同步问题[J]. 计算机系统应用,2004(7):60-62.

[11] 袁松茂. Win32 线程同步对象浅析[J]. 株洲工学院学报,2003,17(2):41-45.


(上接第 25 页)


[7] Kudo T. CRF++ 0.57: Yet another crf toolkit[EB/OL]. 2012. http://crfpp.googlecode.com/svn/trunk/doc/index.html.

[8] Tseng H,Chang Pichuan,Andrew G,et al. A conditional random field word segmenter for sighan bakeoff 2005[C]//Proceedings of the fourth SIGHAN workshop on Chinese language processing. Jeju Island,Korea:[s. n.],2005:168-171.

[9] SIGHAN bakeoff 2005. Second international Chinese word segmentation bakeoff result summary[EB/OL]. 2005. http://www.sighan.org/bakeoff2005/data/results.php.htm.

[10] 滕少华. 基于 CRFs 的中文分词和短文本分类技术[D]. 北京:清华大学,2009.

[11] Toutanova K,Manning C D. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger[C]//Proceedings of joint SIGDAT conference on empirical methods in natural language processing and very large corpora. [s. l.]: Association for Computational Linguistics,2000:63-70.

[12] 李　实,叶　强,李一军,等. 挖掘中文网络客户评论的产品特征及情感倾向[J]. 计算机应用研究,2010,27(8):3016-3019.

# 智能信息系统中手机产品评论的情感倾向分析

作者： 楼博文， 许歆艺， 蔡根， 张辰， 刘功申， LOU Bo-wen， XU Xin-yi， CAI Gen， ZHANG Chen， LIU Gong-shen

作者单位： 楼博文,许歆艺,张辰,刘功申,LOU Bo-wen,XU Xin-yi,ZHANG Chen,LIU Gong-shen(上海交通大学 信息安全工程学院,上海,200240)， 蔡根,CAI Gen(华东师范大学 计算机科学技术系,上海,200241)

刊名： 计算机技术与发展

ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2013(12)