基于关联规则和极性分析的商品评论挖掘

倪茂树，林鸿飞

（大连理工大学计算机科学与工程系，大连，116024）

摘　要： 随着电子商务的迅速发展，消费者在网络上发表的关于商品的评价变得越来越多。但对于潜在的消费者或者商家来说，完全阅读这些评论十分困难。本文针对这一问题，提出一种基于关联规则和极性分析的商品评论挖掘算法。首先确定评论中消费者经常提及的商品特征，然后将所有评价特征的句子提取出来。最后利用词与词之间的依存关系，准确定位每一个观点词的极性位置，计算该观点词在句子中的极性以及整个评论句的极性。实验表明，与人工标识的结果相比，该算法具有一定的合理性和有效性。

关键词： 商品评论；关联规则；极性分析；句法分析；语义倾向性

Mining Product Reviews Based on Association Rule and Polar Analysis

Maoshu Ni, Hongfei Lin

(Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024)

Abstract: With the rapid development of e-commerce, more and more product reviews are presented on the web by the customers. However, for potential customers or merchants, it is difficult to read those reviews completely. Focusing on this problem, this paper presents an algorithm for product reviews mining based on association rules and polar analysis. First, identify the product features which the customers often mention in the review. Second, extract all the sentences containing review features. Last, make use of the dependent relationship between words to accurately locate the polar position of each opinion word, and calculate the polarity of the opinion word in the sentence and the polarity of the whole review sentence. The experimental results indicate that the algorithm is both reasonable and effective compared with the results of manual annotation.

key words: Product Reviews; Association Rule; Polar Analysis; Syntactical Parser; Semantic Orientation

---

基金资助：国家自然科学基金资助项目（编号：60373095，60673039）和国家863高科技计划资助项目（编号：2006AA01Z151）。

作者简介：倪茂树(1983-)，男，江苏扬州，硕士生；林鸿飞(1962-)，男，博士，教授，博士生导师，研究方向为搜索引擎、文本挖掘和自然语言理解，hflin@dlut.edu.cn 。

1 引言

随着电子商务的迅速发展，网上购物已变得不再陌生，越来越多的人足不出户就能买到自己想要的商品。为了更好地服务网上购物的消费者以及增加消费者的购物经验，许多购物网站联合商家为消费者提供了发表评论的平台，这样，消费者就能及时的将对商品的评论反馈给商家以及那些潜在的消费者。但是，随着商品评论呈指数级增长，全部阅读这些评论以至于帮助消费者做决定将变得十分困难，所以急需一种有效的商品评论挖掘方法。

商品评论的挖掘旨在为消费总结所有评论的观点，而观点的总结不仅需要筛选包含评价商品特征的句子，而且还要将句子中褒贬义词汇摘取出来，通过确定词汇的语义倾向性进而判断评论句的语义倾向性。

上世纪90年代以来，语义倾向性研究在国外才得到普遍关注，并迅速发展起来。文献[1]在1997年首先开始了词汇的语义倾向性研究。他们主要是针对形容词作倾向性分析，利用词汇之间的连词（and，or，but等）训练生成词汇间的同意或翻译倾向的连接图，然后用聚类的方法将词汇聚成褒义和贬义两类。2003年，文献[2]采用计算基准词对与词汇相似度的方法识别词汇倾向性，他们选择了七对褒贬倾向比较强烈的词汇，计算待定词与每个基准词的SO-PMI（Semantic orientation-poinrwise mutual information）值来判定词汇的倾向性。2004年，文献[3]利用WordNet计算词汇倾向性，同样先选择基准词，然后判别待定词与基准词在WordNet中是否为同义词，得出词汇的倾向性。

然而，挖掘商品特征以及抽取观点直到2003年才有所涉及。NEC公司的Kusha等人于2003年所开发的ReviewSeer，通过对评论性文章的语义倾向分析，为商品的受欢迎程度进行打分评价，为商家及其消费者提供了非常重要的商业信息[4]。Minqing Hu和Bing Liu[5]在2004年首次提到利用关联规则挖掘算法挖掘商品评论中的特征；随后在文献[6]里，他们综合考虑挖掘的特征，抽取评价该特征的观点词，利用WordNet中的同义词和反义词集预测观点词的语义倾向性，最后利用统计的方式确定句子的语义倾向性。

目前，在中文词汇倾向性研究和商品评论挖掘才刚刚起步，由于中文和英文的差异，传统的基于统计的方法很难准确地表达句子的观点，因此，借助自然语言处理技术，对句子的成分和结构进行语法分析，不仅增强语义理解的可靠性，而且还能提高极性分析的准确性。

基于前人的研究工作，本文提出一种基于关联规则和极性分析的商品评论挖掘方法，首先从网络上下载一部分商品的评论，进行去噪，然后利用关联规则挖掘算法提取商品评论中的特征，随后筛选评论的句子，去掉与特征评论无关的句子。以筛选后的句子为商品的评论句，通过对句子中词与词之间的依存关系进行深层分析，研究关键成分的依存修饰关系，计算观点词的语义倾向性，最后总结评论句的语义倾向性。

2 商品的特征挖掘

2.1 关联规则挖掘

关联规则是对一个事物和其它事物的相互依存和关联关系的一种描述，作为数据挖掘中的一个重要研究领域，关联规则得到了学术界极大的关注。R.Agrawal 等人[7]在 1993 年首先提出了关联规则的概念和模型。关联规则定义如下：

设 $I=\{i_1,i_2,\cdots i_m\}$ 是项的集合。设与任务相关的数据 D 是数据库事务的集合，其中每个事务 T 是项的结合，使得 $T\subseteq I$。每个事务有一个表示符，称作 TID。设 A 是一个项集，事务 T 包含 A 当且仅当 $A\subseteq T$。关联规则就是形如 $A\Rightarrow B$ 的蕴含式，其中 $A\subset I$，$B\subset I$，并且 $A\cap B=\varnothing$。规则 $A\Rightarrow B$ 事务集 D 中成立，具有支持度 s，其中 s 是 D 中事务包含 $A\cup B$（即 A 和 B 二者）的百

分比。它是概率 $P(A \cup B)$。规则 $A \Rightarrow B$ 在事务集 D 中具有可信度 c，如果 D 中包含 A 的事务同时也包含 B 的百分比是 c。这是条件概率 $P(B \mid A)$，即是：

$$\text{support}(A \Rightarrow B)=P(A \cup B) \tag{1}$$

$$\text{confidence}(A \Rightarrow B)=P(B \mid A) \tag{2}$$

关联规则的挖掘一般可分为两个步骤：(1)找出所有支持度大于等于最小支持度阈值的频繁特征项集；(2)由频繁模式生成满足可信度阈值的关联规则。

2.2 商品特征挖掘及评论句筛选

由于网络评论的特殊性，评论中包含一些自然语言处理无法理解的字符及其标号，半自动化地对评论进行去噪是挖掘之前必须做的工作。一个有效的评论句不仅包含句子的主题，而且还要求有相应的评论，即表达观点的词（其中包含消费者表达个人情感的词，比如，某一评论句结束前出现，“郁闷死了”，“郁闷”一词很好地表达了消费者对商品的评价）。商品的特征（也就是一般评论句的主题）基本上都是名词，哈工大的词性标注系统很好地标注了这些潜在的商品特征（对一些特殊商品需要考虑添加专业词典），借助词性标注的结果，挖掘商品特征及评论句筛选主要分为以下四个步骤：

- 对所有的句子进行词性标注；
- 提取句子中的所有名词，利用关联规则挖掘算法寻找频繁 1-项集[7]；
- 频繁项集的项进一步筛剪（确定商品的特征）；
- 依据确定后的特征词从所有的句子中抽取出包含特征词的句子，称为评论句。

3 观点词的语义倾向性

3.1 知网简介

《知网》是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。

知网的基本思想是：设想所有的概念都可以分解成各种各样的义原，同时应该有一个有限的义原集合，其中的义原组合成一个无限的概念集合。如果能够把握这一有限的义原集合，并利用它来描述概念之间的关系以及属性与属性之间的关系，就有可能建立所设想的知识系统[8]。

3.2 观点词的语义倾向性

观点词在这里可以理解为极性词，即句子中带有情感色彩的词。观点词是判断评论句语义倾向性的首要前提。2003年，Hong Yu等人[9]挑选出若干极性较强的形容词构建一个种子集合，通过计算词语与种子集合中某些词同时出现的概率，判断新词的语义倾向性。

文本采用文献[10]中的词汇语义倾向性计算原理，即为每个词汇赋予一个语义倾向的度量值，其大小由这个单词与基准词的语义关系的紧密程度有关。基准词指褒贬态度非常明显、强烈，具有代表性的词语。与褒义基准词联系越紧密，则词语的褒义倾向越强烈，反之，与贬义基准词联系越紧密，词语的贬义倾向越明显。文本选择知网中已标注“良”和“莠”的词汇作为基准词的标准集，总共6952个词条，其中褒义词3361个，贬义词3591个。

知网中每一个词汇都由它的义原组成，所以义原的相似度计算是概念相似度计算的基础[11]。知网中义原间的上下位关系将同类的义原组成一棵树，所以可以通过义原在树中的语义距离计算相似度。假设两个义原在这个层次体系中的路径距离为d，则这两个义原之间的语义距离：

$$Sim(p_1,p_2)=\frac{\alpha}{d+\alpha} \tag{3}$$

其中 p_1 和 p_2 表示两个义原（primitive），d是 p_1 和 p_2 在义原层次体系中的路径长度，是一个正整数。α 是一个可调节的参数。

最后将待定词汇（观点词）与所有可能近义词的相似度求和，计算公式如下：

$$O(w)=\sum_{i=1}^{kp}Sim(wp_i,w)-\sum_{j=1}^{km}Sim(wn_j,w) \tag{4}$$

4 句子的极性分析

4.1 句法分析器

哈尔滨工业大学信息检索研究室开发的Deparser句法分析器将输入的句子进行两个处理。第一，对句子进行分词和词性标注，并在句子的每个词及词性的前面加上序号，句子的末尾增加一个句尾标志“<EOS>”，由其支配全句的核心词；第二，输出句子中所有词与词之间的依存关系，依存关系中，每个关系以一个依存对表示，依存对中的第一个词是核心词，支配第二个词。如：“[3]买_[1]我(SBV)”，这个依存对表示“我”和“买”存在依存关系SBV，其中“买”是这个关系的核心成分，“我”依存于“买”。举个完整的例子如下：

距离很远的天空都照的很清楚的，但愿我用佳能Ａ７１０用的很愉快。

分词及其词性标注结果：

[1]距离/n [2]很/d [3]远/a [4]的/ue [5]天空/n [6]都/d [7]照/vg [8]的/ue [9]很/d [10]清楚/a [11]的/ue [12]，/wp [13]但/c [14]愿/vz [15]我/rh [16]用/p [17]佳能/nz [18]Ａ７１０/ws [19]用/vg [20]的/ue [21]很/d [22]愉快/a [23]。/wp [24]<EOS>/<EOS>

依存关系对为：

[10]清楚_[9]很(ADV)　[7]照_[6]都(ADV)
[3]远_[2]很(ADV)　[22]愉快_[21]很(ADV)
[19]用_[18]Ａ７１０(SBV)　[16]用_[17]佳能(POB)
[19]用_[16]用(ADV)　[4]的_[3]远(DE)
[5]天空_[4]的(ATT)　[19]用_[15]我(SBV)
[19]用_[14]愿(ADV)　[19]用_[13]但(CNJ)
[11]的_[10]清楚(DE)　[7]照_[5]天空(SBV)
[7]照_[1]距离(SBV)　[8]的_[7]照(DE)
[20]的_[19]用(DE)　[22]愉快_[20]的(ATT)
[11]的_[22]愉快(NOT)　[8]的_[11]的(COO)
[24]<EOS>_[8]的(HED)

每一个依存对括号中的符号表示两个词之间的修饰关系（所有修饰关系见 Deparser 句法分析器文档）。对于每一个关系对，如：“[22]愉快_[21]很(ADV)”，称“[22]愉快”为关系对左侧，“[21]很”为关系对右侧。

由于网络评论的特殊性，很多评论的断句很不合理，句子冗长，导致汉语分析器的分析效果很不稳定。本文人工对评论进行了一些预处理，尤其对结构复杂的句子，进行了断句。针对一些网络新用词以及商品特征的专有名词，则统一添加到扩展词典中。

4.2 观点词的极性定位

一个完整的句子中包含很多的词语，名词、动词、形容词、副词等等，但是对句子语义倾向性起影响作用的主要集中在一些形容词和副词上。比如："机器不错，屏幕大看的舒服颜色鲜艳挺值的"这句话，词性标注后的结果是："[1]机器/n [2]不错/a [3]，/wp [4]屏幕/n [5]大/a [6]看/vg [7]的/ue [8]舒服/a [9]颜色/n [10]鲜艳/a [11]挺/d [12]值/vg [13]的/ue [14]。/wp [15]<EOS>/<EOS>"，很显然，"不错"、"舒服"、"鲜艳"、这四个词语对句子的语义倾向性起决定性作用，所以将所有的形容词当作观点词。根据分析的情况，仔细研究句子的依存关系后发现，出现观点词的依存关系对主要有三种：ADV（状中结构）、DE（的字结构）和VOB（动宾结构）。

因为某些观点词不止出现在一个关系对中，为了不重复计算观点词的极性，需要定位到最能表达该观点词极性的关系对，设计算法如下（对每一个观点词）：

1) 搜索所有的依存关系对，如果在依存关系对的左侧发现该观点词：
   - 如果是ADV结构，记ADV结构中的右侧词为副词adverb，继续寻找包含adverb的关系对，如果关系对中包含修饰副词的强调词或者否定词，一并记录下来，至此，ADV结构中观点词的修饰词寻找完毕，将副词adverb和修饰副词的词语一起作为判断该观点词极性的依据；
   - 如果是ATT结构，修饰该观点词的词语为"的"字，继续寻找含有该"的"字的关系对，记录找到关系对中的修饰词，将该修饰词结合原观点词一起作为计算原观点词的极性依据。
2) 如果在关系对的右侧发现该观点词：
   - 如果是DE结构，继续寻找该结构右侧词（观点词）的关系对，如果关系对中有包含修饰该词的强调词或否定词，按照1）中第一种情况考虑；
   - 如果是VOB结构，记录结构中的宾语，观点词的极性就是宾语的极性。

针对这一算法，举例如下：

原句：照片在电脑上的效果真的特别清晰，照相的时候也是一样的。

观点词的极性分析：真 特别 清晰

综合这三个词最终确定该观点词在句子中的极性。

4.3 程度副词和否定词对观点词极性的影响

文献[12]将程度副词考虑到文本的倾向性识别中。程度副词分为绝对程度副词和相对程度副词两类。无论是绝对还是相对程度副词都对句子的语义强度产生很大的影响。例如：

这个手机挺好的；这个手机非常好；这个手机极其好

很显然，这三个句子的语义强度是依次递增的。为了准确的表达消费者对商品的评论，考虑修饰观点词的程度副词是非常必须的。但程度副词的强度有很大的差距，本文根据程度副词的等级设定四个参数，最高到最低依次为1.5，1.2，0.8，0.6，程度副词分类表见文献[12]。

否定词同样对句子的语义强度产生很大的影响，如，"不"、"没有"等否定词在修饰观点词的时候起到了反向的作用，但是一般情况下并没有直接对原词的语义取反向。如，"不美丽"并不代表"难看"，只是对"美丽"起到了弱化语义的作用，所以本文为否定词设置一个参数-0.5。

4.4 句子的极性分析

根据3.2和3.3的分析，一个句子的极性（语义倾向性）不仅考虑观点词的原极性（根据HowNet确定的语义倾向性），而且还要考虑修饰观点词的程度副词和否定词（如果有的话）。累加经过极性定位后的观点词极性，如果值为负数，则表明句子的极性为贬义，反之为褒义。

例如：我觉得还不错，用的很方便，价格也还可以。

观点词为：不错，方便

修饰观点词的程度副词和否定词：还 不错；很 方便。

设不错的原极性值为0.8，方便为0.6，则整个句子的极性值为：0.8*0.8+1.2*0.6=1.36，显然这个句子的极性为褒义。

5 实验与分析

5.1 商品评论的获取

国外有很多的网站提供网上评论的功能，如 www.amazon.com 网站。国内也有一些网站提供了消费者发表评论的平台，如大众点评网，但是针对商品的评论还是非常少。文本所选的评论句来源于 http://www.it168.com 网站——全原创 IT 主流资讯网，从中挑选了四个商品的精华评论（网站上推荐的评论），共计 166 篇评论。四个商品分别为三个手机以及一台数码相机，都是当前比较流行的款式，分别用 A710，A1200，N72 和 V3 表示。

5.2 实验

首先利用关联规则挖掘评论中提及到的商品特征，设定支持度为2（即出现2词以上的名词），再人工删除一部分噪音词，最终每一个商品的特征个数见表2（注：由于网络评论的特殊性，消费者表达的习惯并不一样，所以有些特征从意思上讲是重复的，但表达形式却不一样）。

网络评论存在很多的噪音，有很多的句子并不是商品评论的句子，本文将包含商品特征的句子设定为评论句，这样就删除了一部分无用的句子，再通过人工断句和处理成标准的可以利用句法分析器分析的句子，最终每一个商品的评论句条数见表 2。

实验对所有商品的挑选的评论句进行极性分析，并和人工标注的结果比较，结果如表 3 所示。

表 2 商品特征数和评论句数

Tab.2 The number of features and review sentences

商品名称	A710	A1200	N72	V3
特征数	90	130	114	90
评论句条数	286	262	317	234

表 3 实验结果

Tab.1The results of experiment

商品名称		A710	A1200	N72	V3
原极性	正	168	157	233	69
	负	118	105	84	165
实验极性	正	155	150	219	65
	负	81	85	61	120
准确率（%）	正	96.13	96.67	95.89	96.92
	负	93.83	92.94	93.44	94.17
召回率（%）	正	92.26	95.54	93.99	94.20
	负	68.64	80.95	72.62	72.73

5.3 分析

实验结果表明，正极性的准确率都在95%以上，负极性的准确率平均也在93%以上，可见考虑修饰观点词的程度副词或者否定词是非常必要的。尤其相对于本文前一篇论文，简单利用统计的方法，句子语义倾向性判断的正确率才平均达到70%，对比实验结果不难证明，在判断句子极性方面，单纯利用统计学方法是不可靠的。通过阅读评论也表明，在所有的评论中，带有程度副词以及否定词的句子比例较高，这也正好符合了消费者评论的习惯（通过阅读很多评论），喜欢用程度副词强调商品特征的优点以及用否定词批评商品特征的缺点。

召回率方面，负极性的召回率明显没有正极性的高，原因是算法无法判断观点词的动态极性，另一方面也说明了中文表达的多样性。比如：对于这句话，“就是 LCD 屏内和镜头易进灰”，简短的一句话很难表达该句的极性，单纯用算法来识别也判断不出正确的极性。然而，联系实际商品不难发现，其实是在讲该商品某一特征的缺陷，极性为负。还有一种情况就是评论者的语气词，比如说“唉”、“啊”等词，有的时候表现为褒义，有的时候表现为贬义，这也影响了负极性的召回率。

6 总结

本文针对网络上出现的商品评论，综合利用关联规则挖掘算法以及观点词极性分析算法总结消费者对商品特征的评价。实验结果表明，相对于本人前一篇论文所做的工作，设计的算法取得了很大的改进，不仅提高了准确率，而且还细化了观点词和句子的极性分析思想，放弃传统的基于统计的方法，充分了利用了词与词之间的依存关系，定位每一个观点词在句子的极性位置，提取观点词的强调前缀和否定前缀，为准确理解句子的极性提供了依据。但是，由于网络评论的自由性，在本次实验中，需要大量人工干预，比如筛选特征、处理断句等，这在大规模的语料处理上必然带来很多的困难，有没有很好的算法对商品评论进行预处理还需要进一步研究。另外，消费者经常在一个很长的句子中不止提及一个商品特征，定位每一个商品特征评论的极性也有待进一步研究，只有那样，才能很好地将所有的评论和特征相对应。

参考文献：

[1] V. Hatzivassiloglou, K. R. McKeown. Predicting the semantic orientation of adjectives[A]. In: Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics[C], Madrid, ES, 1997:174-181.

[2] P. D. Turney, M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems, 2003, 21(4):315-346.

[3] J. Kamps, M. marx, R. J. Mokken, and M. D. Rijke. Using WordNet to measuresemantic orientation of adjectives[A]. In: Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation[C], Lisbon, 2004, 1115-1118.

[4] K. Dave, S. Lawrence, DM Pennock. Mining the peanut gallery: opinion exaction and semantic classification pf product reviews[A]. WWW2003, 519-528.

[5] Hu, M., Liu, B. Mining Opinion Features in Customer Reviews. In the Proceedings of AAAI (American Association for artificial intelligence)'04, San Jose, California. 2004:755-760.

[6] Hu, M., Liu, B. 2004. Mining and Summarizing Customer Reviews. In the Proceedings of KDD (Knowledge Discovery and Data Mining)'04. 2004:168-177.

[7] Agrawal R, Imielinski T, Swami. Mining association rules between sets of items in large database[A]. Proc. 1993 ACM-SIGMOD int'l Conf. Management of Data (SIGMOD'93) [C]. Washington DC: 1993. 207-216.

[8] 董振东. 语义关系的表达和知识系统的建造[J]. 语言文学应用，1998，27(3): 76-82.

[9] H.Yu, V. Hatzivassiloglou. Towards answering opinion questions separating facts from opinions and identifying the polarity of opinion sentences[A]. Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing[C]. Sappro, Japan, 2003, 129-136.

[10] 朱嫣岚, 闵锦, 周雅倩, 黄萱菁, 等. 基于 How-net 的词汇语义倾向计算[J]. 中文信息学报, 2006, 20(1): 14-20.

[11] 刘群, 李素建. 基于知网的词汇语义相似度计算. http://www.keenage.com.

[12] 徐琳宏, 林鸿飞, 杨志豪. 基于语义理解的文本倾向性识别. 中文信息学报, 2007, 21(1): 96-100.