

基于句法分析的评价搭配抽取及其倾向性分析*

张 铎** 朱征宇

(重庆大学计算机学院,重庆 400044)

摘 要:提出一种新的基于句法分析的方法可以提取出评价搭配。通过对大量的商品评论进行句法分析,得到了常用的评论模式集合。依靠这些模式对评价搭配进行抽取。同时,构建了一种映射集合,可以映射出隐式评价对象。对于评价对象的倾向性判别,通过 hownet 构建了情感种子词典,采用 SO-PMI 方法计算出其极性,并计算出极性强度。实验结果表明,该方法对于评价搭配的抽取和极性分析具有较好的精确率和召回率。

关键词:意见挖掘;评价搭配;倾向性判别

中图分类号:TP391 文献标识码:A doi:10.3969/j.issn.1006-6055.2013.04.010

Polarity Collocation Extraction and Sentiment Classification Based on Syntactic Analysis*

ZHANG Duo** ZHU Zhengyu

(College of Computer Science, Chongqing University, Chongqing 400044)

Abstract: A new method is proposed to extract polarity collocation. Series of templates are generalized using syntactic analysis on lots of product reviews. Then polarity collocation is extracted via the templates. In order to get implicit comment target, a mapping of comment is built targeting onto reviews. At the same time, a seed dictionary is built based on hownet, with SO-PMI used to compute the polarity and the strength. Experiment results show that the method can get a better precision and recall.

Key words: opinion mining; polarity collocation; sentiment analysis

1 引言

随着电子商务的发展,越来越多的人选择在电子商务网站上购买各种商品。电商通常鼓励用户在網上评论自己对购买商品的意见和看法,既可以为潜在的用户提供购物参考,也可以对自己的产品和服务进行相应的改进。目前很多的商品都有大量的评论,并且很多评论很长但是对于商品的具体评论却只有很少的信息,这种情况下很难让顾客获取到关于商品评论的有用信息,从而很难决定是否该购买此商品。意见挖掘技术综合运用了数据挖掘、自然语言处理等技术,可以从大量的用户评论中抽取用户对商品的评价,甚至可以抽取用户对商品各个属性的评价,经过一些技术处理,可以为顾客提供相对客观的商品评论信息,从而使得用户获得满意的购物体验。

近年来,意见挖掘受到了越来越多的关注,2004 年 Kim 和 Hovy 对意见挖掘进行了定义^[1],意见挖掘主要有四个子任务:①主题抽取 (Topic Extraction) ②意见持有者的识别 (Holder Identification);③陈述的选择 (Claim Selection);④情感分析 (Sentiment Analysis)。M. Hu 和 B. Liu^[2,3]利用词共现技术,采用数据挖掘中的关联规则方法(基于 Apriori 算法)提取产品评论中的主题词。该方法首先挖掘了频繁主题词,然后又通过这些频繁词邻近形容词发掘非频繁主题词,但是该方法仅能找出频繁的评价对象,其精确率和召回率都不太理想。宋晓雷等^[4]采用模糊匹配方法抽取候选评价对象,然后采用双向 Bootstrapping 方法识别出产品评价对象,但是这也不可避免的引入了 Bootstrapping 识别评价对象时各种噪声,使得产品属性的识别效果不太理想。Tsou 等^[5]对中国四

地(北京、香港、上海、台北)报刊上有关思维政治人物(克里、布什、小泉、陈水扁)褒贬性的新闻报道进行了分类研究,首先通过标记语料库获得文本中的极性元素,然后采用了三个衡量指标,即极性元素的散布、密度和语义强度来对每个文本进行统计,得出文本褒贬分类和强度大小的结果,其中对确定性元素之间的关系虽有所提及,但没有深入研究。Guang Qiu 等^[6]使用 Double Propagation 方法对观点词和评价对象进行抽取,但是这使用了大量基于英文处理的方法,对汉语并不适用。刘鸿宇等^[7]使用了无指导方法进行评价对象抽取和倾向性分析。评价对象抽取上,使用基于句法路径的方式抽取评价对象,同时加入三种过滤技术。在倾向性分析上,将评论分为四类,继而针对各类评论制定相应的倾向性判断规则,直接判断其极性。该方法由于采用的句法路径分析过于简单,会产生大量的噪音,虽然采用了剪枝技术进行过滤但召回率和精确率仍然有待提高,在进行极性分析时仅使用了一个小的上下文情感词表,仅能评价出评价对象的褒贬性,对于极性强度并没有一个准确的度量。

针对以上方法的不足,本文通过对大量评论进行句法分析,用得到的评论模式对评价搭配进行抽取,同时对于无显式评价对象的评论,构建了一个映射集,可以映射得到评论短语对应的评价对象。对于评价对象的倾向性判断,通过使用基于 hownet 的情感词词典和领域评价词词典,通过 SO-PMI 方法计算评价对象的褒贬程度,以便更加直观的获取评价对象的倾向性。

2 评价搭配的抽取

2.1 评价搭配的定义

评价对象是指评论针对的对象或对象的属性,由于本文目的是挖掘评价对象的属性,故此处评价对象专指评价对象的属性。评价短语是指评论中对评价对象的一个倾向性短语,包括带有褒贬性的情感词和修饰性的程度副词和否定

* 科技部国家科技支撑计划重点项目(2011BAH25B04)资助

** E-mail: zd13501@163.com

词。评价搭配就是针对某一领域,抽取出评价对象和相应的评价短语。

2.2 领域词典的建立

2.2.1 目标词典的建立

由于不同领域的商品有其独特的领域相关性,所以本文主要是对手机产品评论进行分析,建立目标词典的目的是为了能准确抽取手机这种商品的产品属性,主要采用以下方法:

- 1) 从中关村在线的手机版块上找到关于手机产品的介绍,在参数一栏里找出各个商品属性,记为属性集 S1
- 2) 在电子商务网站中,由于京东商城对商品的介绍比较详细和规范,故而可把京东商城手机版块中找出商品属性集 S2
- 3) 以上所得均为比较规范的商品介绍,但是可能出现大量的不规范的评论,通过比较大量的网络评论,可以从找出常用的商品属性网络评论词集 S3
- 4) 将 S1,S2 和 S3 合并,其中对于基本相同的商品属性,可以作为一个属性包,以规范属性作为本属性包的代表,处理时对于本属性包中的属性词均以此规范属性为代表。

2.2.2 观点词典的建立

赵妍研等^[8]总结了近年来国内外的情感词词典资源,针对简体中文的只有 hownet。知网 hownet 于为我们提供了“情感分析用词语集(beta 版)”,包括了中英文的六种情感分析词语集,分别是正面情感词语、负面情感词语、正面评价词语、负面评价词语、主张词语、和程度级别词语。本文只选取中文词语。本文构建的观点词典包括三个部分:评价词极性词典,程度副词词典和否定词词典。

1) 首先选取知网情感词典中的正面评价词语和负面评价词语集合组成原始种子词典 P1,并为每个词条添加一个重要属性:极性强度。强度包括 5 个等级:±1,±0.5,0。由 3 个人独立完成极性和强度的标注,然后取平均值。这可以保证大多数常用的极性词可以在 hownet 中直接检索到,降低了时间复杂度。

2) 然后使用 hownet 对原始观点词词典进行同义词扩展,扩展出的观点词沿袭原始观点词的极性,形成观点词词典 P2(包含正向情感词集 Pwords 和负向情感词集 Nwords)。

3) 对下载的网络评论进行分词,从中选取所有的极性词,由 3 个人独立进行极性分类,从中取分类一致的结果,忽略有歧义的词语,对得到的极性词进行强度标注,方式与标注 hownet 原始种子词典 P1 方式一致,得到观点词典 P3

4) 将进行同义词扩展后的观点词词典 P2 和获得的网络词典 P3 进行合并,得到最终的观点词种子词典 P(包含正向情感词集 Pwords 和负向情感词集 Nwords)。

2.3 评价搭配的模式规则

通过分析大量的评论语料,可以发现,构成评价搭配的基本情况大致如下:

- 1) 名词短语 + 形容词短语,例如:外形美观
- 2) 形容词短语 + 助词的 + 名词短语,例如:漂亮的造型
- 3) 动词短语 + 助词的(得) + 副词短语,例如:运行的慢
- 4) 动词短语 + 形容词短语,例如:运行流畅

5) 形容词短语 + 助词的(地) + 动词短语,例如:流畅的运行

其中,名词短语主要由如下扩展方式:

- 名词短语—>名词,例如:外形美观
- 名词短语—>名词短语 + 名词短语,例如:屏幕声音(都不错)
- 名词短语—>名词短语 + 连词 + 名词短语,例如:外观和摄像头(都不错)
- 名词短语—>名词短语 + 助词 + 名词短语,例如:屏幕的色彩(不错)

形容词短语主要的扩展方式如下:

- 形容词短语—>形容词,例如:美丽的
- 形容词短语—>副词 + 形容词短语,例如:非常漂亮
- 形容词短语—>形容词短语 + 了 + 量词短语,例如:差了一些

动词短语主要的扩展方式如下:

- 动词短语—>动词,例如:运行
- 动词短语—>动词短语 + 名词短语,例如:处理信息(非常快)

根据以上规则,用中科院计算所词性标注集标注表示的模式如下:

表 1 模式的抽取		
Table 1 Extraction patterns		
模式	评价对象	评价短语
n(或 n+n 或 n+n+n 等) + a	n(或 n+n 或 n+n+n 等)	a
n+c+n+a	c 旁的两个 n	a
n+u+n+a	u 旁的两个 n	a
n(或 n+n 或 n+n+n 等) + d+a	n(或 n+n 或 n+n+n 等)	d+a
n+c+n+d+a	c 旁的两个 n	d+a
n+u+n+d+a	u 旁的两个 n	d+a
(d+)a+u+n(+c+n..)	n(+c 旁的 n)	d+a
v(+d)+a	左方的 slot 词 n	(d)+a
(d+)a+u+v	左方的 slot 词 n	(d)+a

对上表需要做一解释,如果标注后的句子中存在此模式,则按此包括括号中词的模式抽取。对于最后的两组动词 + 形容词模式,评价对象可以取分词后本句中左方离形容词最近的名词,如“手机运行的很流畅”,但是如果名词不存在,如仅有“运行流畅”,之前的文献均不进行此项处理,而是只抽取显示评价对象,而本文尝试着通过建立一个映射集合,通过统计分析大量评论,将一些常用的无评论对象的评价词与实际的评价对象进行映射(此映射为多对多的,但最终可将评价对象规约为一个规范化用语)得出评价对象为手机速度。后面的实验结果也表明了我们的方法是令人满意的。

2.4 评价搭配的抽取过程

1) 评论的采集。使用爬虫程序从淘宝网上获取 1200 条关于手机的用户评论,其中只选取带有具体观点的评论,对于只有“不错”之类的评论,由于对本文研究无意义,故舍去,最终得到 480 条有用评论。

2) 预处理。对所得的评论进行分句、分词、词性标注。分句以短句为单位,即以“,”、“。”、“?”、“!”等作为分隔符,获取数据集 D。

3) 对于获得的数据集,对每个句子切分,按照表 1 前八个明确相符的模式进行评价搭配的抽取,得到初始评价搭配集合 $M\{\text{评价对象,评价短语}\}$,其中评价短语包括了修饰性的副词和表达极性的形容词。

4) 对于表 1 最后两种模式,如动词(+ 副词) + 形容词模式,如果在此句子切分中其左方有名词,则选取最近的名词为评价对象,对于(副词) + 形容词 + 助词 + 动词模式,也可以采用同样的方式。如果两种模式的句子切分均没有名词,则可以尝试构建一个映射集合,自动找出评价对象。此两种模式出现的方式有一定的共性,例如“不错,运行很流畅”和“赞一个,可以非常流畅的运行各种应用程序,一点都不卡”等,故若出现此种模式,且有“运行”等词语,则可将其评价对象映射为“运行速度”,以此为例,可以构建一个合适的映射集合从而找出评价对象。

3 倾向性分析

3.1 原始极性的判断

对于评价搭配集合中的形容词,首先搜索 2.2.2 所建立的观点词种子词词典,如果可以是搜索到此词,则该词极性即为种子词典中已经标注好的极性强度。对于种子词典中搜索不到词语,可以借助 SO - PMI 算法^[9]计算其极性。

SO-PMI 法利用逐点互信息计算词间的语义关联强度。两个词语 word1, word2 的相互关联强度通过以下方式计算:

$$PMI(word_1, word_2) = \log_2 \left(\frac{P(word1 \& word2)}{P(word1) \times P(word2)} \right) \quad (1)$$

$P(word1 \& word2)$ 指 word1 和 word2 同时出现的概率。如果两个词同时出现的概率较大,则 PMI 为正数,否则为负数。

式中的概率是指在同一个较大的语料库下,各个词语出现的概率,那么 $Hits(word)$ 表示 word 在这个大型语料库下出现的次数,公式可以变成下面式(2)的形式:

$$PMI(word_1, word_2) = \log_2 \left(\frac{Hits(word1 \& word2)}{Hits(word1) \times Hits(word2)} \right) \quad (2)$$

依此为基础,再依靠 2.2.2 构建的观点种子词词典中的正向情感词集 Pwords 和负向情感词集 Nwords,计算一个新词与种子词的互信息,可以得到该词的极性:

$$\begin{aligned} SO-PMI(word) &= \sum_{pword \in Pwords} PMI(word, pword) - \sum_{nword \in Nwords} PMI(word, nword) \\ &= \log \left\{ \left[\prod_{pword \in Pwords} Hits(word, pword) \times \prod_{nword \in Nwords} Hits(nword) \right] / \left[\prod_{nword \in Nwords} Hits(word, nword) \times \prod_{pword \in Pwords} Hits(pword) \right] \right\} \quad (3) \end{aligned}$$

由于种子词集包含种子词的极性,所以当一个新词 word 趋向于褒义词时,它与正向种子词出现的概率要大于与负向种子词同时出现的概率,反之亦然。就是说,当一个词语倾

向于褒义词时,那么它的 SO-PMI 值为正数,同理,贬义情感倾向的词语的 SO - PMI 值为负数。至此,我们就可以计算出观点词的原始极性。

3.2 计算词的上下文极性,即评价对象的最终倾向性

一个词的极性分为两种情况,一种是原始极性,这种极性判断可以通过利用 2.2.2 创建的观点词词集和 SO - PMI 算法得到;另外一种是上下文极性。上下文极性可以出现在两种语境中;一种是句子中含有否定前缀,如“不好看”,另外一种 是句子含有强调前缀,例如“非常好”。如果不考虑否定前缀就会导致倾向性判断错误,如果不考虑程度副词前缀,则虽然极性倾向判断正确,但是强度不够准确。

为了解决这两个问题,本文定义了两个词典:否定词词典和程度副词词典,后者中为每个程度副词标注一个表示程度强弱的值。对于抽取出的评价搭配,如果含有否定词,则将 3.1 中计算出的词语极性值符号反置。如果含有程度副词,可根据程度副词词典中的强弱值对其最终极性进行加倍或减小。

4 实验结果及分析

根据以上介绍的方法,本文使用了 C#程序,分词工具采用中科院提供的 ICTCLAS2011,并把构建的词典加入到用户的自定义词典中,以便提高分词精度。在极性计算时采用了百度语料库,通过其返回的网页数作为 Hits。本文的实验数据集从淘宝网上获取,最终整理出的待处理评论共 480 条。使用手工标注的形式对评论进行标注,包括搭配的抽取和极性标注。采取的评价标准为正确率 P,召回率 R,和调和平均值 F1。

表 2 评价对象的抽取结果实验
Table 2 Experiment results of comment target extraction

方法	Precision	Recall	F-measure
文献[7]采用方法	0.497 6	0.546	0.520 6
本文采用方法	0.653 8	0.687 6	0.670 2

这里由于文献[7]中其实验结果在 Lenient 评价结果中效果最好,故此处采纳其在 Lenient 评价中的数据作为对比,由以上数据表明本方法的效果明显好于文献[7]的方法,这是由于是文献[7]主要聚焦于抽取显式评价对象,而本文通过构建的映射集可以映射出一部分的隐式评价对象,从而大大提高了召回率。

表 3 评价短语的抽取结果实验
Table 3 Experiment results of opinion words

方法	Precision	Recall	F - measure
文献[7]采用方法	0.587 4	0.645 2	0.614 9
本文采用方法	0.724 5	0.698 4	0.711 2

由于文献[7]并没有给出其对于评价短语抽取结果的实验数据,所以我们验证了其方法的有效性,结果表明,文献[7]采用的方法提取评价短语的效果要比其抽取评价对象的效果要好,但是本文的句法分析进一步概括了评价短语中包含动词时的情况,所以进一步的提高了效果。

表 4 倾向性判断的结果实验

Table 4 Experiment results of sentiment classification

方法	Precision	Recall	F-measure
文献[7]采用方法	0.654 2	0.685 6	0.669 5
本文采用方法	0.768 2	0.702 5	0.733 9

这一块由于文献[7]仅仅是归纳了语法规则便进行极性分析,而本文通过构建情感种子词典,否定词词典能够更加准确的计算评价对象的倾向性,同时,结合构建的程度副词词典,可以更进一步的计算出情感词的极性强度,与标注结果相比,极性强度的正确率可以达到超过六成,效果还是令人满意的。

5 结束语

本文通过对大量的评论进行句法分析,使用基于模板的方法对评价搭配进行抽取,并且增添了关于动词的模板,可以丰富评价搭配的抽取。对于目前的研究只针对隐式对象抽取的情况,本文构建了一种映射集合,通过对评论词的映射可以映射出隐式评价对象,结果表明这种方法是有效的。在极性分析上,通过构建一系列的词典,同时使用基于 SO-PMI 的方法,可以更加精确的对评价对象进行极性分析,并且可以对极性的强度进行判断,这就为我们以后对评论进行进一步的处理(如通过极性强度对其进行聚类,可以判断出用户对商品的评价大致处于何种范围之中)做出了有用的贡献。

参考文献

[1] KIM S, HOVY E. Determining the Sentiment of Opinions[C]. Proceeding of the intl. Conf. on Computational Linguistics. 2004; 321-328.

[2] HU Mingqing, LIU Bing. Mining opinion features in customer reviews [C]. Proceedings of AAAI – 2004. Seattle USA, 2004; 755-760.

[3] ZHANG Lei, LIU Bing. Extracting Resource Terms for Sentiment Analysis[C]. the 5th International Joint Conference on Natural Language Processing (ICNLP – 2011). Chiang Mai, Thailand, 2011; 1 214-1 221.

[4] 宋晓雷,王素格,李红霞. 面向特定领域的产品评价对象自动识别研究[J]. 中文信息 2010, 24(1) :89-93.

[5] TSOU B. Polarity Classification of Celebrity Coverage in the Chinese Press [C]. Proceedings of the International Conference on Intelligence Analysis. McLean, USA, 2005; 243-249.

[6] QIU Guang, LIU Bing, CHEN Chun. Opinion Word Expansion and Target Extraction through Double Propagation[J]. Computational Linguistics, 2011, 37 (1) :9-27.

[7] 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.

[8] 赵妍妍,秦兵,刘挺. 文本情感分析[J]. 软件学报, 2010, 21(8) : 1 843-1 848.

[9] TURNEY P. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews[C]. Proceeding of the Meeting of the Association for Computational Linguistics (ACL'02). 2002; 417-424.

作者简介

张铎(1987-),男,硕士在读,主要研究方向:意见挖掘,自然语言处理;

朱征宇(1959-),男,博士,教授,系主任,主要研究方向:web 挖掘。

(上接第 464 页)

[2] BREUNIG M M, KRIEGEL H, NG R T, et al. LOF: identifying density-based local outliers; proceedings of 2000 ACM SIGMOD International Conference on Management of Data, New York, 2000[C]. New York: ACM Press, 2000; 93-104.

[3] 黄洪宇,林甲祥,陈崇成,等. 离群数据挖掘综述[J]. 计算机应用研究, 2006, 8; 8-11.

[4] HAN Jiawei, KAMBER M. Data mining: concepts and techniques [M]. 2nd edition. San Francisco: Morgan Kaufmann Publishers, 2006.

[5] 倪巍伟,陆介平,陈耿,等. 基于 k 均值分区的数据流离群点检测算法[J]. 计算机研究与发展, 2006, 43(9) :1 639-1 643.

[6] 杨显飞,张健沛,杨静,等. 基于距离的数据流离群点挖掘算法[J]. 计算机应用, 2010, 30(11) :2 949-2 951.

[7] 张虑平,梁永欣. 基于反 k 近邻的流数据离群点挖掘算法[J]. 计算机工程, 2009, 35(12) :11-13.

[8] 常建龙,曹锋,周傲英. 基于滑动窗口的进化数据流聚类[J]. 软件学报, 2007, 18(4) :905-918.

[9] AGARWAL C, HAN Jiawei, WANG Jianyong. A framework for clustering evolving data streams[C]. VLDB 2003; Proceedings of the 29th International Conference on Very Large Data Bases. Berlin: VLDB Endowment, 2003, 29; 81-92.

[10] HAWKINS D. Identification of Outliers[M]. London: Chapman and Hall, 1980.

作者简介

赵学良(1988-),男,硕士在读,主要研究方向:数据挖掘;

朱庆生(1956-),男,博士,教授,博导,主要研究方向:数据挖掘,图像处理,数字农业。