中文网络评论中提取产品特征的研究①

祖李军, 王卫平

(中国科学技术大学 管理学院, 合肥 230026)

摘 要: 大量的网络评论已经成为挖掘用户意见、改进产品质量的重要信息来源, 而特征抽取作为后续分析的基础, 直接影响到最终意见挖掘结果的准确性. 本文提出了一种 PMI-Bootstrapping 算法, 并结合了语言规则实现中文网络评论的产品特征抽取. 首先利用语言规则产生候选特征集, 计算每个候选特征与初始给定种子集的加权平均互信息, 将满足阈值的候选特征添加到种子集中, 如此循环迭代, 直到种子集合收敛, 输出排队后的种子集合作为抽取结果. 实验证明, 该算法取得良好的准确率和召回率.

关键词: 特征抽取; 网络评论; PMI; 语言规则; 文本挖掘

Research of Extracting Product Features from Chinese Online Reviews

ZU Li-Jun, WANG Wei-Ping

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

Abstract: Now online reviews have become an important resource for mining users′ opinion and refining products. As a foundation of further analysis, features extraction influences the precision of the opinion mining results. This paper proposes a PMI-Bootstrapping algorithm which realizes extracting product features from Chinese online reviews by combining three language rules. First, utilize the language rules to get a candidate feature set. Then, calculate the weighted average PMI for each candidate feature with the seeds in the initial seed set. Add the candidate feature which satisfies the threshold to the seed set. Iterate until the seed set is convergent. Output the seed set as the extraction result. Experimental results show that the algorithm achieved very good precision and recall rate.

Key words: feature extraction; online reviews; PMI; language rules: text mining

1 引言

如今, 网络评论已经成为影响用户购买行为的重要因素之一. 随着信息技术的发展、网上支付安全性的提升以及物流的便捷, 越来越多的人选择网上购物. 与传统的购买形式不同, 网上购买时用户通常无法直接接触商品, 对商品的质量优劣、合适与否无法做出直观的判断, 因此, 网络评论可能就会左右他们最终的消费决定. 而很多用户即使选择传统的购买方式, 也经常会参考网络评论. 同样, 网络评论对于厂商也是非常重要的, 积极的评论会产生很好的广告效应, 而负面的评论不利于商品的销售, 厂商可以通过分析网络评论, 明确商品的优缺点, 有目的性地进行推广和改进.

近年来, 网络评论的研究受到广泛的关注, 意见挖掘就是这类研究的热点之一, 其目标是从海量的网络评论中分析每个用户对于商品某一属性特征的态度. 意见挖掘的核心任务可以分为三个方面: (1)抽取产品特征; (2)识别面向产品特征的观点词; (3)判断观点词的极性. 在天猫、京东等网上商城的用户评论区, 已经实现了意见挖掘的初步应用, 通过完成上述三项任务, 得到诸如“电池不错”、“系统流畅”之类的评论摘要, 使用户可以更加便捷地浏览选购. 但是这些应用也还存在一定的不足, 仅能对几个最常见的产品特征进行摘要. 而特征抽取是整个意见挖掘的基础, 只有全面准确地抽取出用户评论中所涉及的产品特征, 之后的分析才能有的放矢. 因此, 本文将尝试探讨中文网络

① 收稿时间:2013-09-30;收到修改稿时间:2013-10-24

评论的产品特征抽取.

2 相关研究

由于网络评论都是非结构化的自然语言, 句法结构和使用环境的变化都会导致抽取结果的偏差, 而机器无法直接对语意进行准确的理解, 因此需要经过分词处理和词性标注将自然语言转化成结构化或半结构化的形式, 再利用相应的方法进行抽取. 对于特征抽取的结果, 通常有两方面的诉求: 一是抽取出的特征尽可能都是正确的, 即准确率(P, Precision); 二是抽取出的特征尽可能全面, 即召回率(R, Recall). 然而这两方面又是相互矛盾的, 追求准确率的提升往往是以降低召回率为代价, 反之亦然. 实际挖掘中就需要寻求两者的一种平衡, 达到整体性能(F-Score, $2PR/(P+R)$)的最佳状态.

通过定义语言规则抽取产品特征可以获得较高的准确率. Qiu 等提出了一种双向传播方法(DP, Double Propagation), 描述了四种词汇间的关系, 定义了八个具体的抽取规则, 根据词汇依赖关系, 并考虑了代词和否定词的影响, 进行特征词和感情词的抽取[2]. Zhang 和 Liu 等对 DP 进行了改进, 结合搜索引擎中的 HITS 算法以及出现频次来对抽取产品特征进行排名, 从而进一步提高了准确率[3]. Li Zhuang 等提出一种对电影进行评论挖掘的更具体的方法, 定义了电影特征(包括电影元素和电影人员), 并利用 WordNet 抽取评论中的特征－观点对[4]. 郝博一等构建了一个产品属性挖掘系统 OPINAX, 对每个候选特征词定义了一个包含 6 个维度的特征向量, 每个维度表示一个语言规则参数, 计算特征向量的可信度并进行排队, 输出排在前列的特征词[5].

由于语言规则的定义针对性较强, 当扩展到其他领域中时, 会导致抽取结果不够理想, 而且复杂的规则也会让抽取更加困难. 在准确率可接受的情况下, 通过运用数据挖掘算法自动抽取产品特征可以获得较高的召回率. Hu 和 Liu 提出了利用关联规则、邻近剪枝、独立支持度剪枝, 抽取的频繁项作为产品特征的方法[8]. 李实等将这种方法引入到中文网络评论的特征抽取中, 结合汉语特点, 对抽取结果进行单字词的剔除, 取得了较好的效果[1]. Turney 提出一种点互信息法(PMI, Pointwise Mutual Information), 借助搜索引擎命中数目来度量候选特征词之间的联系[9]. Popescu 等利用 KnowItAll 系统定义了某些包含特定关系(如 have, part of)的模板, 计算候选特征词与这些模板的 PMI 值, 并设定阈值筛选出产品特征[10]. Zhen Hai 等利用似然比检验(LRT, Likehood Ratio Tests)和潜在语义分析(LSA, Latent Semantic Analysis)度量特征词之间、观点词之间、特征词与观点词之间的三类关系, 逐步迭代完善来实现特征抽取[11]. Kim 等采用窗口取词的方法, 利用观点词为中心, 设定某一固定半径的窗口, 抽取窗口中的名词或名词短语作为产品特征[12].

3 特征抽取

3.1 抽取流程

纵观前面提到的各种抽取方法, 都是语言学与算法的结合, 虽然各有侧重, 但绝不是严格的相互独立, 这是文本挖掘的特点所决定的. 本文的研究也是如此, 在语言规则的基础上提出了一种 PMI-Bootstrapping 算法, 在中文网络评论中实现产品特征的抽取. 抽取流程如图 1.

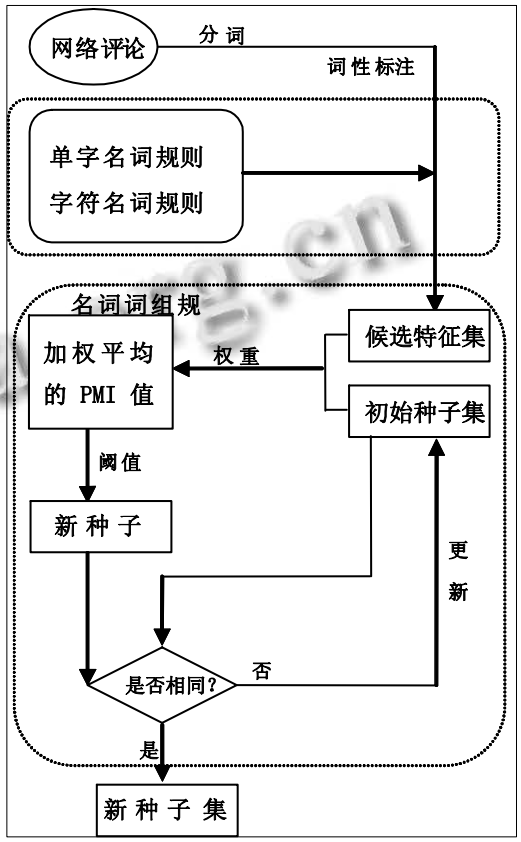

图 1 特征抽取流程

(1)首先采用中国科学院计算机所编写的中文分词

软件 ICTCLAS(http://www.ictclas.org/)对评论语料进行分词和词性标注;

(2) 利用定义语言规则进行筛选, 产生候选特征集;

(3) 给定初始的种子集, 对候选特征集利用 PMI-Bootstrapping 算法最终特征抽取的结果.

需要指出的是, 网络评论中的产品特征可以分为两种: 一种是直接用文字表达出来的, 称为显性特征, 如例 1 中的“屏幕”; 另一种是隐含在语意中的, 称为隐性特征, 如例 1 中的“买不起呀”隐含的特征是“价格”. 目前关于特征抽取的研究主要针对显性特征, 本文也是如此, 而对隐性特征的抽取尚存在难度, 暂不考虑.

例 1: 屏幕大, 看视频很清晰, 但是买不起呀!

3.2 语言规则

由于显性特征主要表现为名词, 因此, 特征抽取可以转化成对名词(ICTCLAS 的词性标注为“/n”)和名词词组的抽取. 本文提出了三条语言规则用于对分词之后的名词进行初步筛选, 形成候选特征集, 在保证规则简单且易扩展性的同时, 最大限度地将非特征名词清除, 简化后续的提取工作.

单字名词规则: 若出现仅由一个汉字组成的单字名词, 且前后无其他名词相连, 则清除该单字名词[1]. 因为分词软件会产生很多无意义的单字名词, 而汉语的音节特点决定了绝大部分的产品特征是多字名词, 并且单字名词很多情况下可以包含在多字名词的语意之中, 如“屏”包含在“屏幕”之中. 因此, 利用该规则可以有效的清除噪音, 同时防止产品特征的丢失. 如例 2, 分词之后“屏幕”、“眼”、“电池”被标注为名词, 但“眼”由于是单字名词将被消除.

例 2: 屏幕清楚, 不费眼, 电池没有担心的那么差.

字符名词规则: 若出现由英文字母或英文字母加数字构成的字符串(ICTCLAS 的词性标注为“/x”), 则保留字符串作为候选特征; 单纯的数字(ICTCLAS 的词性标注为“/m”)不算作字符串. 虽然此类字符串不会被标注为名词, 但具有名词性质, 很有可能是产品特征或其简写形式, 所以应该保留作为候选特征, 提高特征抽取的召回率, 如例 3 中的“MP3”、“CPU”.

例 3: MP3 音质有待提高, CPU 给力, 多开软件一点不卡, 安装速度快.

然而, 保留字符串不可避免地会引入噪音, 降低特征抽取的准确率. 为了降低噪音, 加入最低支持度(本文中设为 2)的限制, 只有当字符串的出现频次不低于最小支持度时, 该字符串才会被保留, 否则将被清除. 这样可以有效地过滤类似拼写错误等噪音.

名词词组规则: 若出现连续两个或连续两个以上名词(单字名词和字符名词也包含在内)直接相连的情况, 则将这些名词整体作为一个候选特征. 因为名词词组越长, 其是产品特征的可能性越大, 而将其分解成单个名词后却又无法准确表达整个词组的含义. 如例 4 中的“机身材质”由两个连续名词“机身”和“材质”组成; 例 5 中的“SIM 卡”由字符名词“SIM”和单字名词“卡”组成.

例 4: 机身材质多为塑料, 质感一般.

例 5: SIM 卡插入后, 机器过很长时间才识别.

3.3 PMI-Bootstrapping 算法

PMI 又称点互信息法, 通过搜索引擎检索时的共现条目来度量两个词语之间的联系. 词语 W_1 和 W_2 的点互信息可通过以下公式来计算:

$$PMI(W_1,\ W_2)=\log_2\left[\frac{Hits(W_1\&W_2)}{Hits(W_1)\bullet Hits(W_2)}\right] \qquad (1)$$

其中, $Hits(W_1\&W_2)$ 表示在搜索引擎中同时搜索 W_1 和 W_2 时返回结果的数目, $Hits(W_1)$ 和 $Hits(W_2)$ 表示单独搜索 W_1 或 W_2 时返回结果的数目.

不难看出, 若两词的共现次数越多, 它们的 PMI 值可能就越大, 存在某种共同性质的可能性就越高. 以此为理论基础, 若 W_1 来自候选特征集, W_2 已经确定是产品特征, 即种子特征, 当 PMI 值大于某一阈值时, 我们有理由相信 W_1 也是产品特征. 但由此也引发了两方面的问题: 首先, 确定哪些种子特征用来考查候选特征? 这些特征应该具有足够的代表性, 同时又要尽可能的简单; 其次, 阈值如何去设定? 准确的阈值才能将产品特征遴选出来, 但单一固定的阈值往往只能适应一部分的候选特征.

本文对 PMI 的理论基础加以改进: 若某候选特征是产品特征, 则它应该和已经确定的产品特征都具有较高的 PMI 值, 相反, 若某候选特征不是产品特征, 则它应该和已经确定的产品特征都具有较低的 PMI 值. 由此提出了 PMI-Bootstrapping 算法, 具体的步骤如下:

(1) 给定初始的种子集. 只需包含从候选特征集中挑选的一个代表该领域的种子特征, 如“手机”, 在

之后的每次迭代后，算法自动对种子集进行更新.

(2) 计算每个种子特征的权重. 以某个种子特征在语料中出现的频次与所有种子特征频次之和的比值作为该种子特征的权值，设第 i 个种子特征的权重为 w_i m_i，频次为，则有

$$w_i = \frac{m_i}{\sum m_i} \tag{2}$$

(3) 计算每个候选特征的加权平均 PMI 值. 设 cf_j 为第 j 个候选特征，sf_i 为第 i 个种子特征，则 cf_j 的加权平均 PMI 值为

$$\overline{PMI_j} = \sum_i w_i \bullet PMI(cf_j, sf_i) \tag{3}$$

(4) 确定阈值，加权平均 PMI 值大于阈值的候选特征将被加入到新种子集中. 设共有 N 个候选特征，则该次迭代的阈值 t 设置为

$$t = \frac{1}{N}\sum_j^N \overline{PMI_j} \tag{4}$$

(5) 比较新种子集与初始种子集. 若两者不同，将初始种子集更新为新种子集并返回第(2)步，否则，进入第(6)步.

(6) 输出新种子集.

在实际使用中，将初始种子集和新种子集完全相同作为终止条件可能过于苛刻，可以根据不同的要求，控制两者的差异在某一可接受的范围或设定迭代次数作为终止条件来提高效率. 同时，可以设置不同的系数(本文设置为 1/N)来适应挖掘需求. 以下是 PMI-Bootstrapping 算法伪代码形式:

CF: 候选特征集，N 个元素，cf_j 为第 j 个候选特征，n_j 是 cf_j 的频数.

SF: 初始种子集，M 个元素，sf_i 为第 i 个种子特征，m_i 是 sf_i 的频数，w_i 为种子特征 sf_i 的权重.

NSF: 新特征集.

t: 阈值.

1: Input: CF, SF
2: for each sf_i in SF
3: $w_i = \frac{m_i}{\sum m_i}$
4: end for
5: for each cf_j in CF
6: $\overline{PMI_j} = 0$
7: for each $\overline{PMI_j} = 0$ in SF
8: $\overline{PMI_j} += w_i \bullet PMI(cf_j, sf_i)$
9: end for
10: end for
11: $t = \frac{1}{N}\sum_j^N \overline{PMI_j}$
12: for each cf_j in CF
13: if $\overline{PMI_j} > t$ then add cf_j into NSF
14: end for
15: if NSF has the different features with SF then
16: let SF=NSF
17: goto line 2
18: else
19: output NSF
20: end if

4 实验分析

本文的实验采用一台个人电脑，具体配置为: Windows7 操作系统，4G 内存，Intel Core i5 处理器; 程序开发是基于 Microsoft Visual Basic 2008 平台; 并选择最大的中文搜索引擎——百度作为信息源计算 PMI 值.

4.1 实验语料

从太平洋电脑网(http://www.pconline.com.cn/)抓取 iphone5 手机的网络评论并选择 200 条作为实验语料. 通过人工标注的方法对评论中所涉及的产品属性进行识别，得到 82 个产品特征，为了最大限度的检验算法的适应性，对具有相同含义的特征不进行合并. 在分词软件对评论语料进行分词和词性标注之后，利用语言规则初步抽取出的候选特征 190 个，这些候选特征将经过 PMI-Bootstrapping 算法的再次抽取得到最终的产品特征.

4.2 实验结果

在表 1 中，第 1 组实验使用的是 PMI-Bootstrapping 算法，设置迭代 5 次作为终止条件. 选择两个对照实验组对比抽取结果，其种子特征集合包括手机、屏幕、电池、摄像头、操作系统、处理器和像素，共 7 个最常见的手机特征. 第 3 组实验使用的是传统的 PMI 方法，计算某个候选特征词与 7 个种子特征的 PMI 值，

以最大值作为该候选特征的 PMI 值，与阈值进行比较；第 2 组实验是在第 3 组的基础上，计算某个候选特征词与 7 个种子特征的 PMI 值，以加权平均值作为该候选特征的 PMI 值，与阈值进行比较，加权方法与本文相同．为了对三组实验进行更好的比较，阈值的设置均采用本文的方法．

表 1　特征抽取结果对比

组次	Precision	Recall	F-Score
1	67.39%	75.61%	71.26%
2	56.67%	62.20%	59.30%
3	51.72%	54.88%	53.25%

从表 1 中的数据中可以看出: PMI-Bootstrapping 算法取得了更好的准确率与召回率，比原方法都有较高的提升，与当前的其他抽取算法相比，也是具有竞争性；75.61%的召回率虽然偏低，但这是由于在人工标注时没有合并同义的产品特征引起的，若消除这一影响，采用文献[1]中的最小最大覆盖原则，得到的召回率将达到 80%以上；三组实验结果的对比也显示出，计算加权平均 PMI 值和使用迭代更新的种子集都对抽取结果产生正向的影响．

图 2 显示的是通过改变阈值使输出结果的数目逐个增加时，准确率、召回率和 F-Score 各自的变化轨迹．F-Score 作为准确率和召回率的综合指标呈现出抛物线的趋势，而本文中方法设置的阈值在最后一次迭代中，得到的 F-Score 落在抛物线的顶端区域附近，体现了本文的阈值设置方法的合理性．

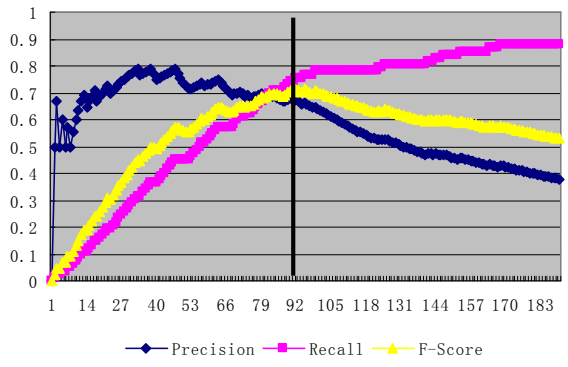

图 2　阈值设置的评价

4.3 优势与不足

本文的优势在于定义了简单通用的语言规则，可以显著的减少候选特征集的噪音；PMI-Bootstrapping 算法通过迭代获得变动的种子集和阈值，能够有效、快速地进行非监督学习，很好的解决了原 PMI 方法所存在两个问题．此外，计算候选特征与整个种子集的平均 PMI 值可以有效地修匀由于偶然情况或者搜索结果的不准确造成的误差；而利用频次加权将语料内部的信息加入到搜索引擎的搜索数据中，使得特征抽取的结果更加接近语料环境．

然而，在实验过程中也发现本文的方法还存在不足，当评论语料数量很大时，产品特征所占候选特征的比例会下降，仍以 1/N 作为阈值系数会导致抽取结果准确性降低．此时，需要将大规模的语料切分成多个较小规模(200 至 400 条评论)的语料进行分批处理，这会造成重复处理的现象．如何更加灵活地设置出适应不同规模语料的阈值将是下一步研究的重点．

5　总结

随着市场竞争的日益激烈，商家需要了解客户的需求并提供更优质的产品才能赢得市场，而从产品评论的挖掘中可以获得用户的意见以及产品的优缺之处．本文尝试了新的方法对中文网络评论进行产品特征抽取，取得了较好的效果，但对于网络评论的挖掘还有很大提升的空间，未来在这一方面的研究还将持续加深．

参考文献

1 李实,叶强,李一军,Law R.中文网络客户评论的产品特征挖掘方法研究.管理科学学报,2009,12(2):142−152.

2 Qiu G, Liu B, Bu J, Chen C. Expanding domain sentiment lexicon through double propagation. Proc. of the 21st International Joint Conference on Artificial Intelligences. 2009. 1199−1204.

3 Zhang L, Liu B, Lim SH, O'Brien-Strain E. Extracting and ranking product features in opinion documents. Proc. of the 23rd International Conference on Computational Linguistics: Posters. 2010. 1462−1470.

4 Zhuang L, Jing F, Zhu XY. Movie review mining and summarization. Proc. of CIKM. 2006.43−50.

5 郝博一,夏云庆,郑方.OPINAX:一个有效的产品属性挖掘系统.第四届全国信息检索与内容安全学术会议论文集(上). 2008.281−290.

6 张紫琼,叶强,李一军.互联网商品评论情感分析研究综述.

管理科学学报,2010,13(6):84–96.

7 易力,王丽亚.基于观点挖掘的产品可用性建模与评价.计算机工程,2008,38(16):270–274.

8 Hu M, Liu B. Mining ppining features in customer reviews. 19th National Conf. on Artificial Intelligence (AAAI'04). 2004.755–760.

9 Turney PD. 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.

10 Popescu A, Etzioni O. Extracting product features and opinions from reviews. Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP). 2005.339–346.

11 Hai Z, Chang K, Cong G. One seed to find them all: Mining opinion features via association. Proc. of CIKM. 2012.255–264.

12 Kim SM, Hovy E. Determining the sentiment of opinions. The 20th International Conference on Computational Linguistics.2004.61–66.

13 伍星,何中市,黄永文.产品评论挖掘研究综述.计算机工程与应用,2008,44(36):37–41.

14 Ding X, Liu B, Yu PS. A holistic lexicon-based approach to opinion mining. Proc. of WSDM. 2008. 231–239.

15 Yu J, Zha ZJ, Wang M, Chua TS. Aspect ranking: Identifying important product aspects from online consumer reviews. Proc. of the 49th Annual Meeting of the Association for Computational Linguistic. 2011. 1496–1505.