Computer Science and Application 计算机科学与应用, 2017, 7(8), 738-746
Published Online August 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.78085

Hans汉斯

Research and Mine on Mobile Phone Feature Based on User Reviews

Wei Gao, Xiangling Fu

Beijing University of Posts and Telecommunications, Beijing
Email: 472434565@qq.com, fuxiangling@bupt.edu.cn

Received: Jul. 19th, 2017; accepted: Aug. 4th, 2017; published: Aug. 7th, 2017

Abstract

User online comment is becoming the important data resource for enterprise to get user's requirement in Internet environment. However, how to accurately and effectively extract the characteristics of the product and its description of the characteristics in the vast sea of comments is a difficult problem of the theory and practice. Based on the data preprocessing and feature extraction, the LinLog energy model is introduced to cluster and analyze the features of products and their descriptive information, so as to obtain the accurate evaluation of the feature. This paper applies the energy model to the evaluation of 4 mobile phones collected from Jingdong Mall, and then analyzes the clustering results, finally obtains the advantages and disadvantages of the four mobile phones. The results show that this method can extract the features of the product intuitively and effectively from the content generated by user.

Keywords

Data Mining, Feature Extraction, Phrase Match, Energy Model

基于用户评论的手机特征挖掘应用研究

高　威，傅湘玲

北京邮电大学，北京
Email: 472434565@qq.com, fuxiangling@bupt.edu.cn

收稿日期：2017年7月19日；录用日期：2017年8月4日；发布日期：2017年8月7日

摘　要

用户在线评论逐渐成为互联网环境下企业获取用户需求的重要数据资源。然而如何能够准确有效的在浩

文章引用: 高威, 傅湘玲. 基于用户评论的手机特征挖掘应用研究[J]. 计算机科学与应用, 2017, 7(8): 738-746.
DOI: 10.12677/csa.2017.78085

如烟海的评论中提炼出产品的特征及其对特征的描述，是一个理论和实践界的难题。本文对手机评论数据的预处理和特征提取的基础上，引入了LinLog能量模型，对产品特征及其描述信息进行聚类分析，从而获得该特征的准确评价。本文对采集自京东商城的4款手机评论应用能量模型，通过分析聚类结果，最终获得了这四款手机特征的优劣，结果经过验证，表明该方法能够直观有效的从用户产生内容中提炼出产品特征的优劣。

关键词

数据挖掘，特征提取，短语匹配，能量模型

1. 引言

从大量文献中可知[1]，在线评论已成为消费者的重要信息来源，以消费者的角度来看，评论可以影响消费者的选择，以制造商的角度来看，选择有效的评论可以帮助新产品的开发。根据国内外研究结果，每 100 个新产品方案中，平均只有 6.5 个能产品化，不到 15%的新产品能成功地商品化，37%进入市场的新产品在商业上是失败的[2]，究其原因，是缺少对现有商品的准确分析，从而无法把握用户需求。

目前，对用户评论的研究，主要采用的是朴素贝叶斯算法。然而，朴素贝叶斯算法需要大量的人工标注，并且在对文本进行分析时，只能判断文本的极性，无法分析文本中包含的产品特征。而 Rob[3]等人利用能量模型对音乐进行聚类，发现它比其他算法有更好的效果，能够明显区分各个类别的边界。为此本文在 LinLog 能量模型的基础上，选取对用户评论进行分析，挖掘其中潜在的信息。发现用户需求，为用户选购和厂商改进商品提供参考。

本研究的所有数据均采集自京东商城，其中 iphone 6 手机的评论 569 条，三星 892 条，华为 2068 条，小米 1069 条。主要对三星、苹果、华为和小米的手机评论进行了数据分析。

2. 技术路线

本文工作的技术路线共包括三部分，第一部分是对数据进行预处理，构造手机特征词典和情感词词典，然后根据词典，从评论中提取手机特征词和情感词，并使用基于窗口的搭配模式对特征词和情感词进行匹配，获得(名词，情感词)词对；第二部分利用 LinLog 能量模型对上一步抽取出来的词对进行聚类分析，获得图形化的聚类结果和各个词语在坐标系中的坐标；第三部分是计算名词和情感词在聚类结果中的欧几里得距离，并根据从小到大进行排序，距离的大小反映了情感词和特征词的匹配程度，距离越小，匹配度越高。最终我们根据匹配度较高的特征词和情感词，对产品进行评估，获得分析结果。整体技术路线如图 1 所示。

3. 数据预处理

3.1. 构造词典

在本研究中，为了能准确提取评论中的手机特征和用户情感，我们决定构建手机特征和用户情感词典。然后根据词典提取其中的名词集合、形容词集合。

经过前期调研，目前针对中文文本分词效果较好的分词器有“结巴分词”、“NLPIR”、“BosonNLP”、“IKAnalyzer”，其中“结巴分词”基于 python 语言，“BosonNLP”需要进行 Token 认证，“IKAnalyzer”的分词结果中不包含词性。因为我们的实验使用 java 语言，并且后期处理中，需要使用文本中词语的词性，同时从易用性方面考虑，我们最终决定采用“NLPIR”分词器，其分词精度达到了 98.45%，对网络评论分词效果较好。文本及分词结果如图 2、图 3 所示。

从图 3 可以看出，分词结果中，可以获得分词结果的词性，其中“n”属于名词，“a”属于形容词，我们下一步的工作，是利用 LPCE (LDA + PageRank + ConditionEnrtopy)模型模型系统进行词典构造。LPCE 如图 4 所示。

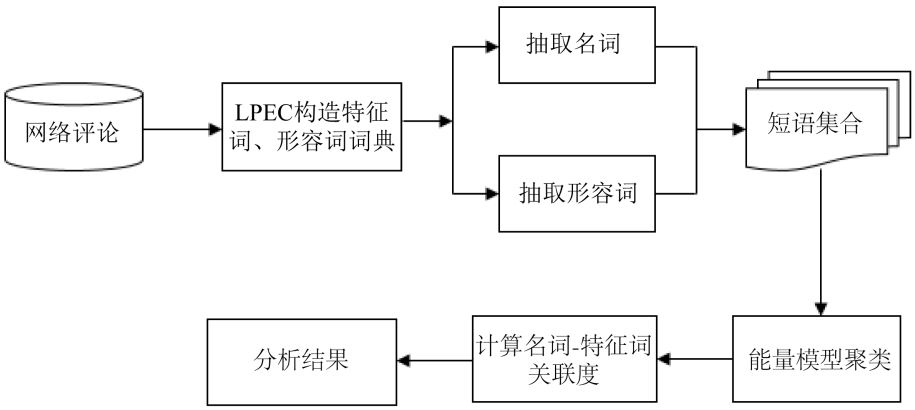

Figure 1. Technical route
图 1. 技术路线图

好好好手机好好好好玩，我是形容词，手机我买了。

Figure 2. User comments
图 2. 待分词文本

好/a 好好/z 手机/n 好好/d 好好/d 玩/v ，/wd 我/r
我/rr 买/v 了/y 。/wj 手机/n 质量/n 外观/n ，/wd
,/wd 屏幕/n 的/ude1 分辨率/n 很/d 高/a 的/ude1

Figure 3. Word segmentation
图 3. 分词结果

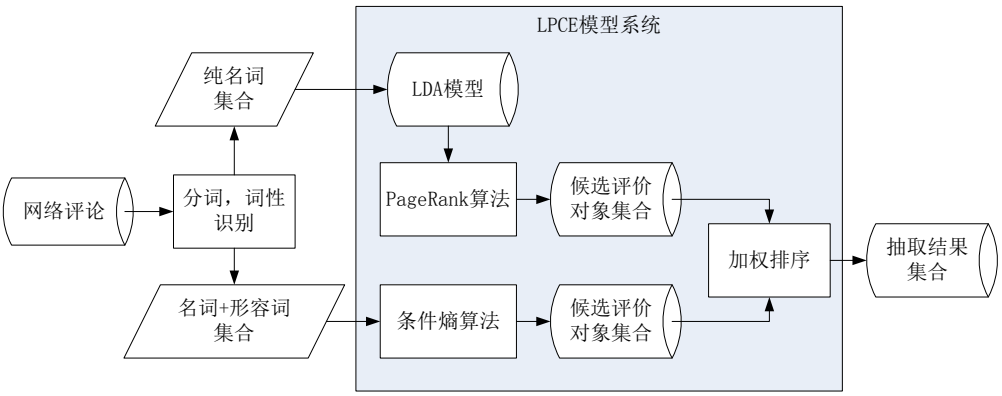

Figure 4. LPCE model
图 4. LPCE 模型系统

如图 4 所示，本系统使用基于无监督的评价对象的抽取技术，对于给定的语料，首先进行分词和词性识别。然后通过词性识别提取其中的名词及名词短语作为一个集合，无需保证集合中的名字按原文中有序。同时提取名词、名词短语与形容词作为另外一个集合，该集合要保证每个文档中提取的形容词和名词按原文中有序，以辅助后面的工作需要。对于纯名词集合通过 LDA 训练出模型[4]，再通过基于 PageRank 思想改进的算法来实现集合内所有名词的优先级排序[5]。对于名词和形容词有序的集合，则通过基于共现概率的条件熵过滤算法来实现集合内所有名词的优先级排序。在以上两个步骤产品的名词优先级排序集合基础上进行加权重排序，从而得到最终的评价对象抽取结果集合。最终得到两个词库：加权排序的主题词词库和形容词词库。

在模型验证中，采用了 1245 条测试数据，通过人工标注出了 1062 条观点句，183 条非观点句。以 LPCE 模型提取的评价对象作为词典来进行观点句的识别，并与采用 TF (Term Frequency)的词典集合进行对比实验。

实验采用 top 200 和 top 500 两组词典集合，来进行分组对比实验。上表为基于评价对象的观点句识别的评测结论，通过表 1 可以看出 LPCE 的词典集合在观点句识别的测试中无论是精度、召回还是 F 值都优于 TF 词典集合，尤其在 Top 200 词的评测中，LPCE 方法的召回和 F 值都明显高出 TF 方法 2%~3%，实验证明基于 LPCE 方法提取评价对象集合具有更优的效果。

3.2. 特征抽取

在应用 LinLog 能量模型时，我们的输入数据是<特征词，情感词>词对，所以第二步工作是根据词典提取名词和形容词，然后对其进行配对。Kim 等人[6]提出了一种在评价词的固定长度窗口内查找评价对象的方法。所以本研究决定采用固定窗口的方法，对手机特征及情感词进行匹配。

首先，我们根据手机特征词词典和情感词词典，对评论中的手机特征和情感词进行提取。因为我们要根据窗口来匹配手机特征和情感词，所以必须获得各个词语在评论中的位置。因此我们需要对评论进行分词，获得其中名词、名词短语及形容词的位置。

通过对中文文本进行语法分析，我们发现，手机特征词的构成，有三种模式。第一种是单个名词，即手机特征由一个名词描述，例如“电池”；第二种是两个名词构成名词短语来形容手机特征，例如“后盖材料”；第三种是两个名词和一个助词来描述手机特征，例如“屏幕的分辨率”。所以，我们决定根据这三种模式，对评论中的名词进行重新组织。对于情感词的提取，我们根据其词性，即形容词，来进行提取。在抽取的同时，我们记录各个词语的位置，以便使用窗口的方式对手机特征和情感进行匹配。

在进行匹配时，我们使用词典匹配得到的手机特征来对评论中的名词进行筛选，当名词属于手机特征的时候，我们记录其位置，然后以该词出发寻找大小为 3 的窗口内的形容词。这样就找出了句子中的“特征词—形容词”的搭配。如图 5 是对根据用户对 iphone 6 的评论抽取获得的匹配结果。

Table 1. Recognition of performance indicators based on evaluation object
表 1. 基于评价对象的观点句识别性能指标

评测集合性能指标	1245 条测试数据(top 200 词)		1245 条测试数据(top 500 词)	
词典集合来源	TF 方法	LPC 方法	TF 方法	LPC 方法
精度	89.2%	90.8%	88.6%	89.7%
召回	52.9%	56.4%	59.3%	60.9%
F 值	66.4%	69.6%	71%	72.5%

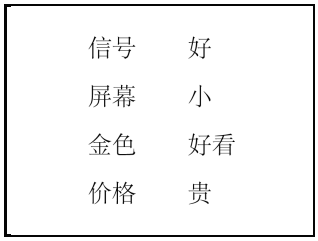

Figure 5. Noun phrases
图 5. 名词短语

本研究中，根据京东商城里各款手机的评论，从 iphone 6 中共抽取出 118 对短语，三星 GalaxyNote 4 共抽取 410 对短语，华为荣耀畅玩 4X 共抽取 1115 对，小米 note 470 对。

4. LinLog 能量模型

本研究的目的是发现用户对手机各个特征的评价，从而发现该手机所具有的优势及其不足。所以我们需要对上一步抽取的短语进行聚类，聚类的目的是将特征和情感划分至不同的类别当中，从而挖掘出特征词匹配的情感。

在 LinLog 能量模型[7]中，短语中的特征词和情感词都可以看作是图中的一个点，每个点都对周围的点具有作用力，他们吸引或排斥图中的其他点，点之间的作用力之和，构成了整幅图中的能量，LinLog 能量模型，根据点与点之间的吸引力与斥力进行聚类，最终相互之间吸引力大的点会聚为一类，与该类斥力大的点会与其他点聚为一类，从而整幅图的能量将至最低。最终图中会出现许多簇点集，每一簇都是一类，每一簇中都包含特征词和情感词，我们认为位于同一簇中的情感词与该簇中的特征词匹配度较高，即我们可以使用这些情感词来较为准确的评价该簇中的手机特征。

使用 LinLog 能量模型进行聚类，我们最终得到了两个输出，一个是表示聚类结果的展示图，从图中可以直观的观察各个类的大小及分布位置；另一个输出是图中各个点的坐标，即手机特征词和情感词的位置，根据坐标，我们可以计算特征词和情感词之间的距离，距离越近，情感词对该特征的描述越准确。

LinLog 能量模型的目的是将图中联系密切的点连接在一起，同时将联系较少的点分离。从而使得该图所具有的能量最小。能量越小，表明聚类效果越准确。在模型中，我们定义一个点的度为与该点连接的边数，一个点的度越大，该点具有的引力与斥力也越大。LinLog 能量模型计算一个图 P 所具有的能量如式 1 所示：

$$U_{EdgeLinLog(p)} = \sum_{\{u,v\}\in E} \left\| p(u) - p(v) \right\| - \sum_{\{u,v\}\in E} \deg(u)\deg(v)\ln\left\| (u) - p(v) \right\| \tag{1}$$

如式 1 所示，其中 U 表示图 P 所具有的能量，u, v 是图 P 中一条边的两个端点，$p(u)$ 和 $p(v)$ 表示这两点在图中的位置，$\left\| p(u) - p(v) \right\|$ 是这两点的欧几里得距离，deg(u)表示点 u 的度，u 的度等于该点所连接的边数。

5. 计算特征词－情感词关联度

使用 LinLog 能量模型，我们获得了各个点在图中的坐标。实验的最后一步，是通过计算特征词与情感词之间的距离，判断哪些情感词能够准确的描述手机特征。仅从上一步生成的图中，我们只能观察到一些大概的数据，并不能精确到描述特征词和情感词的关系，所以我们需要通过计算距离，来为特征词匹配较为准确的情感。在计算图中两点的距离时，我们采用欧几里得距离来度量。二维坐标系的欧几里得距离计算公式如式 2 所示：

$$dis(u,v)=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2} \tag{2}$$

在式 2 中，u, v 表示图中的两点，(x_1, y_1) 为点 u 在图中的坐标，点 (x_2, y_2) 为点 v 在图中的坐标。在计算图中点与点之间的距离时，我们首先做的是将其中的特征词与情感词分离，我们采用字典的方法，将文本中的特征词和情感词分离，然后分别计算每个特征词到各个情感词的距离，最后根据从小到大的顺序对距离进行排序，距离越小，表明该情感对该特征的描述越准确。

6. 实验结果

本研究共对 iphone 6、三星 GalaxyNote 4，华为荣耀畅玩 4X，小米 note 四款手机进行了产品分析，所有评论数据均来自京东商城，其中关于 iphone 6 手机的评论共有 569 条，三星 892 条，华为 2068 条，小米 1069 条。对四款手机应用 LinLog 能量模型得到图 6~图 9。

在计算特征词和情感词的欧几里得距离之后，我们获得的结果如图 10~图 13。

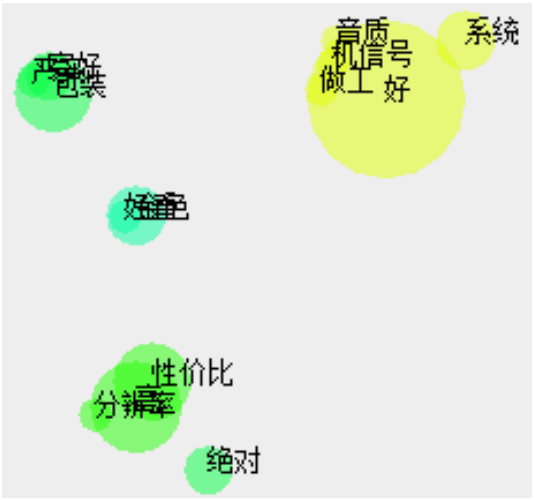

Figure 6. Iphone 6 clustering results
图 6. Iphone 6 聚类结果

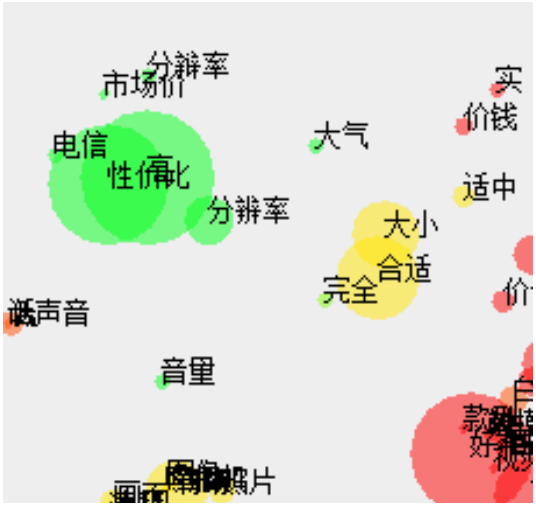

Figure 7. Huawei clustering results
图 7. 华为聚类结果

Figure 8. Samsung clustering results
图 8. 三星聚类结果

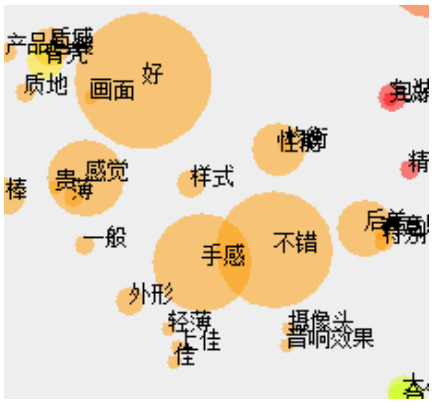

Figure 9. MI clustering results
图 9. 小米聚类结果

网络，全面，0.06346850680809675
外包装，简陋，0.06675554815449139
少，按键，0.06778238023359404
机时，长，0.06784077633210976
均匀，喷漆，0.07190985308850 36
合适，大小，0.18898938906517976
差；外壳，0.19858808514488646
外观，漂亮，0.228107961071352 76
电池，理想，0.24747562711200252
大，声音，0.251231747455775412

Figure 10. Iphone 6 Feature-Emotional distance
图 10. Iphone 6 特征－情感距离

持久，电量，0.020472795420774334
短，充电时间，0.023877795633167807
及时，信息，0.023909925285451 17
看电视，爽，0.057226711394164936
色彩，艳丽，0.059355318065093074
坏，充电器，0.06033876698506296
接收信号，弱，0.06404216780874511
手机运行，畅通，0.09112691102778182
话声音，低，0.11989941454784374
材质，精细，0.12887959285172357

Figure 11. Huawei Feature-Emotional distance
图 11. 华为特征－情感距离

按键，灵，0.043258053105001144
单手操作，方便，0.043260636427134076
机身，薄，0.043265112886434365
划算，最低价，0.04327093579187471
最新，版本，0.04954725660650378
机时，长，0.09950224123191684
程序，少，0.10524412630035747
丑，外形，0.1242887723797658
音乐，正常，0.13260758227606112
后背，烫，0.13569955426539582

Figure 12. Samsung Feature-Emotional distance
图 12. 三星特征－情感距离

省电模式，贴心，0.031681376453308985
亮度，均匀，0.03170217570035405
触感，爽利，0.03174634144108803
导航定位，慢，0.03660201006868464
背盖，别致，0.0366081132500405
粗糙，充电器，0.07308895068152514
按键，软，0.08549087301821766
数据线，坏，0.09255007977026096
难受，材质，0.09625988383246677
灵敏，触屏，0.11842824928093372

Figure 13. MI Feature-Emotional distance
图 13. 小米特征－情感距离

我们从分析结果中截取了前 10 个，来判断用户对该手机的评价，从以上分析结果可以看出，大部分用户对 iphone 6 的网络评价是全面，并且认为他的待机时间比较长，对它的喷漆也比较满意，认为其外观很漂亮，同时声音很大。但是 iphone 6 的缺点也很明显，iphone 6 的外包装比较简陋，并且用户对它的外壳不是很满意。

对于华为来说，用户对它的电量、充电时间、色彩、运行速度、材质都比较满意，但是它的充电器质量不是很好，同时通话声音偏小。

从三星手机的计算结果可以看出，它的按键灵敏，单手操作方便，机身轻薄，价格比较划算，待机时间比较长，但是外形不是很令用户满意，而且发热现象比较严重。

小米手机的统计结果显示，其省电模式比较受用户青睐，触屏也很灵敏，屏幕亮度好，但是其问题也很多，比如导航定位慢、充电器做工粗糙，数据线有损坏，手机材质不好等。

以上分析结果显示了用户对手机各个特征的评价，这些分析结果可以为用户购买手机提供参考，同时也可以为手机厂商进一步改进提供依据。

7. 总结与展望

在进行短语匹配时，我们采用了窗口匹配的方法，该方法简单，有效，但是准确度不高，所以，在未来的工作中，可以对该部分进行优化，采用准确度更高的算法，相信最后的分析结果会更加准确。目前，对 LinLog 能量模型的研究，多见于外文文献。本研究尝试性的对中文数据应用该模型，取得了不错的效果。但是，在研究过程中，还是存在可以改进的地方。

在该研究阶段，我们仅仅是利用 LinLog 能量模型对数据进行了聚类分析，未来，可以利用其它算法进行聚类，比较两者聚类结果的差别。

基金项目

国家自然科学基金项目重点项目“面向不确定性的 web 2.0 用户创作内容管理研究”(71231002)。

参考文献 (References)

[1] Berger, J., Sorensen, A.T. and Rasmussen. S.J. (2010) Positive Effects of Negative Publicity: When Negative Reviews Increase Sales. *Marketing Science*, **29**, 815-827. https://doi.org/10.1287/mksc.1090.0557

[2] 李升林，乌兰木其. 基于数据挖掘的需求分析研究[J]. 中国机械工程, 2003, 14(5): 392-395.

[3] Vignoli, F., Gulik, R.V. and Wetering, H.V.D. (2004) Mapping Music In The Palm Of Your Hand, Explore and Discover Your Collection. *International Conference on Music Information Retrieval*.

[4] 石晶，李万龙. 基于 LDA 模型的主题词抽取方法[J]. 计算机工程, 2010, 36(19): 81-83.

[5] 黄德才，戚华春. PageRank 算法研究[J]. 计算机工程, 2006, 32(4): 145-146.

[6] Kim, S.M. and Hovy, E. (2005) Automatic Detection of Opinion Bearing Words and Sentences. *Proceedings of Ijcnlp*.

[7] Noack, A. (2007) Energy Models for Graph Clustering. *Journal of Graph Algorithms & Applications*, **11**, 453-480. https://doi.org/10.7155/jgaa.00154