DOI:10.11721/cqnuj20160107

基于情感分类的文本主题挖掘*

陈国良，唐万梅
(重庆师范大学 计算机与信息科学学院，重庆 401331)

摘要:互联网的电商中存在着大量的评论信息,这些带有主观情感色彩的评论信息不仅反应了客户对产品的满意程度,而且暗含了市场产品的流行趋势。针对评论信息中所蕴涵的相关主题词,提出了将文本分类和主题词挖掘相结合的方法。该方法首先使用 SVM 对情感进行分类,再通过 LDA 模型进行建模对分类后的评论信息挖掘主题词。真实数据集上的实验结果验证了本文方法的有效性,获得了良好的分类结果,能够准确地挖掘出主题词。

关键词:主题词;情感分类;LDA 模型
中图分类号:TP391 **文献标志码**:A **文章编号**:1672-6693(2016)01-0092-05

随着互联网的普及和信息技术的进步,互联网上(包括门户网站、电子商务网站、社交网站、音/视屏分享网站、论坛等)产生了海量的、由用户发表的对于诸如人物、事件、产品等目标实体的评论信息。如何有效地管理和使用这些评论信息成为当前的迫切需求;从海量的评论信息中挖掘出有效的主题信息,分析出内在的语义关联也显得尤为重要。近年来,主题模型(Topic model)成为文本挖掘领域的研究热点。

电子商务随着 Web2.0 技术的出现和发展应运而生,许多电商在互联网中获得了显著成功,带动了大量的用户在网上进行 Shopping。在用户 Shopping 的过程中,他们会对产品进行评论,用户在分享自己对某次购物经历的体验时,会涉及到两种完全相反的情感表达。分析挖掘这些评论中用户的情感信息有着潜在的商业价值,一方面,商家通过分析这些评论信息获取用户对产品质量及服务水平的相关信息,帮助他们改善产品质量、改进服务水平和改变销售策略,来满足用户的喜好,促进产品的销售;另一方面,用户可以通过参考这些带有主观情感色彩的评论内容,了解和对比自己感兴趣的产品或服务,进而作出相应的购买决策。因此,如何从评论信息中挖掘出与情感相关的主题词有着实际的应用价值,它能帮助用户做决策,又能得到对产品的反馈,还能对产品的销售情况进行预测。

本文针对如何从评论信息中获取相关主题词的问题进行了研究,考虑到正向情感和负向情感的表达对主题结果的意义完全不同,提出了一种基于情感分类的文本主题挖掘方法。首先利用 SVM(Support vector machine)对评论信息进行情感分类,再对已分类的信息利用 LDA 模型建模挖掘相关主题词。实验表明该方法能够有效地对评论信息进行主题挖掘。

1 相关工作

在情感分析方面,主要采用两种方法:一种将规则和情感词典结合使用[1],依照文档中含有的积极情感与消极情感词的数目来判定情感倾向;另一种方法是把机器学习应用到情感分析中,通过对文档中已经选择好的特征来训练模型(包含朴素贝叶斯(Naive bayes)、支持向量机(SVM)、最大熵(Max entropy))从而对情感进行分类。Pang 等人[2]对电影评论的情感极性倾向进行了研究。该文章的工作主要有 3 步:首先,进行文本预处理;其次,对一元词、二元词、词的位置等特征进行提取;最后,通过这些特征来训练模型,通过朴素贝叶斯、支持向量机等方法的比较,结果表明,支持向量机有最好的分类效果。李太白等人[3]提出了一种改进的 SVM 多类分类算

* **收稿日期**:2015-04-23 **修回日期**:2015-06-07 **网络出版时间**:2015-12-02 13:29
资助项目:重庆市自然科学基金(No. CSCT2015JCYJA00005);重庆市教委项目(No. KJ1500309;No. KJ1400519;No. KJ130602);重庆市教改项目(No. yjg123040; No. yjg152001);重庆师范大学校级项目(No. cyjg1205);重庆师范大学研究生科研创新项目(No. YKC14009)
作者简介:陈国良,研究方向为数据挖掘、机器学习,E-mail:;**通信作者**:唐万梅,教授,E-mail:1093895431@qq.com
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.n.20151202.1329.052.html

法，并应用于入侵检测中，利用 KDD CUP 1999 入侵检测数据进行实验，实验结果表明，该算法能有效提高分类准确率。谢丽星等人[4]针对中文微博消息中含有的情感信息进行了分析研究，通过将表情符号的规则、情感词典的规则以及基于 SVM 的层次结构多策略等 3 种方法进行比较研究，结果表明，基于 SVM 的层次结构多策略方法对情感分析的效果最好。

LDA 是近几年被提出的一种无监督学习技术。针对海量的文档集，LDA 可以有效地挖掘出这些信息中蕴含的主题信息。因为 LDA 模型可以把文本中的信息转化成为易于建模的数字信息，从而，它被广泛地用于主题的挖掘和情感的分析中。以下是近几年对 LDA 模型的应用研究。

王文帅等人[5]针对大规模微博数据的话题挖掘进行了研究，提出改进的 LDA 主题模型 SNLDA(Social network LDA)，实验结果表明，改进的模型能有效地从大规模微博数据中挖掘出话题信息。张晨逸等人[6]在 LDA 基础上提出了一种新的模型 MB-LDA 模型，主要工作是对微博中潜在的主题进行挖掘，实验结果表明效果良好。陈文涛等人[7]在构建用户兴趣上对 3 种不同的主题模型的性能进行了研究，其中主要思想是用户的兴趣取决于用户发表微博中的主题生成概率的大小，研究结果表明：TwitterLDA 可用于对新文档或新用户进行预测；UserLDA 和 AuthorLDA 可以对用户的社交网络关系进行更好的表达。孙艳等人[8]提出一种无监督的主题情感混合模型(UTSU)，并应用于文本情感分类中——虽然他们用实验证明了在无监督分类中效果不错，但在有监督分类中就很差。文献[9]提出了一种把词分成情感词与主题词的 TST 模型(Topic sentiment mixture)，此文认为情感词对主题发现并没有起到作用，然而主题词中包含着情感词，对主题词的表达起到很大的作用。鉴于此，本文提出首先利用 SVM 对评论信息进行情感分类，分成正向情感和负向情感两类，然后通过 LDA 模型进行建模，分别对正向情感和负向情感挖掘相关主题词，实验结果表明本文提出的方法效果良好，能够清晰地挖掘出用户对产品各方面的评论信息。

2 文本生成模型 LDA

在概率主题模型中，已经把 LDA(Latent dirichlet allocation)模型[10-12]作为一个实现标准。LDA 是一个含有单词、主题及文档三层结构的层次贝叶斯模型[13]，文档-主题服从狄利克雷分布，主题-词服从多项式分布，如图 1 所示。LDA 模型中的各种符号说明见表 1。

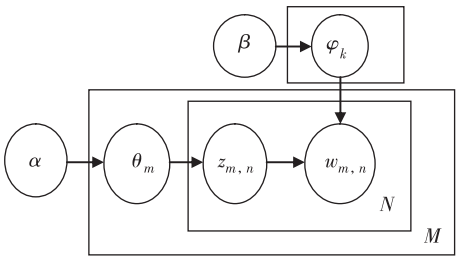

图 1　LDA 图模型的表示

表 1　符号的定义说明

参数	定义
θ_m	第 m 篇文档的主题概率分布
φ_k	主题 k 中的词项概率分布
$w_{m,n}$	第 m 篇文档中第 n 个单词
$z_{m,n}$	第 m 篇文档中第 n 个主题
z_i	第 i 个单词对应的主题的变量
$\neg i$	剔除其中的第 i 个项
$n_k^{(t)}$	第 k 个主题中出现的词项 t 次数
$n_m^{(k)}$	文档 m 中出现主题 k 的次数
$\alpha(\alpha_k)$	主题 z 的 Dirichlet 先验
$\beta(\beta_t)$	词项 t 的 Dirichlet 先验

根据 LDA 图模型很容易得到语料概率值如(1)式：

$$p(M \mid \alpha,\beta)=\prod_{m=1}^{M}\int p(\theta_m \mid \alpha)\cdot \left(\prod_{n=1}^{N}\sum_{z_{m,n}} p(z_{m,n} \mid \theta_{m,n})p(w_{m,n} \mid z_{m,n},\beta)\right)\mathrm{d}\theta_m。 \tag{1}$$

LDA 模型生成过程描述如下：

1)对主题采样 φ_k，φ_k 服从 Dirichlet(β)分布。φ_k 代表主题 k 中词项概率分布。

2)对语料库中的第 $m(m\in[1,M])$个文档进行采样：

①采样 θ_m，θ_m 服从 Dirichlet(α)分布，θ_m 代表的是主题发生的概率；

②采样文档长度 N，N 服从 Poiss(ξ)；

③对文档 m 中第 $n(n\in[1,N])$个单词进行采样，其中：

a) 选择主题 $z_{m,n}$，$z_{m,n}$服从 Multinomial(θ_m)，$z_{m,n}$代表当前选择主题；

b) 生成 $w_{m,n}$，$w_{m,n}$服从 Multinomial($\phi_{z_{m,n}}$)，$w_{m,n}$代表当前生成的单词。

2.1 模型的推导

LDA 常用的推导方法有吉布斯抽样(Gibbs sampling)、变分贝叶斯(Variational bayesian)、期望值传播(Expectation propagation)等方法。本文采用 Gibbs sampling 方法。Gibbs sampling 是一种

快速高效的MCMC(Markov chain monte carlo)抽样方法,通过迭代方式对概率分布进行推导。LDA模型推导过程如下:

$$p(z_i=k \mid z_{\neg i},w)=\frac{p(w \mid z)}{p(w_{\neg i} \mid z_{\neg i})p(w_i)} \cdot \frac{p(z)}{p(z_{\neg i})} \propto$$
$$\frac{\Delta(n_z+\beta)}{\Delta(n_{z,\neg i}+\beta)} \cdot \frac{\Delta(n_m+\alpha)}{\Delta(n_{m,\neg i}+\alpha)} \propto$$
$$\frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+\beta_t} \cdot \frac{n_{m,\neg i}^{(k)}+\alpha_k}{\left[\sum_{k=1}^{K} n_m^{(k)}+\alpha_k\right]-1}。 \tag{2}$$

对(2)式反复迭代,并对所有主题抽样,最终达到抽样稳定结果,便可得到参数计算公式:

$$\phi_{k,t}=\frac{n_k^{(t)}+\beta_t}{\sum_{t=1}^{V} n_m^{(t)}+\beta_t}, \tag{3}$$

$$\theta_{m,k}=\frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{K} n_m^{(k)}+\alpha_k}。 \tag{4}$$

至此,LDA模型通过Gibbs sampling求解出文档在主题上的概率分布,以及主题在单词上的概率分布。

3 基于文本情感分析的主题词挖掘

本文为了获得评论信息中的主题词,需要首先对评论信息进行分词、提取特征等预处理,其次通过SVM分类器对评论信息进行情感二分类(分成正向、负向两类情感),最后对分类后的两类情感进行LDA建模,挖掘出相关的主题词。流程图如图2所示。

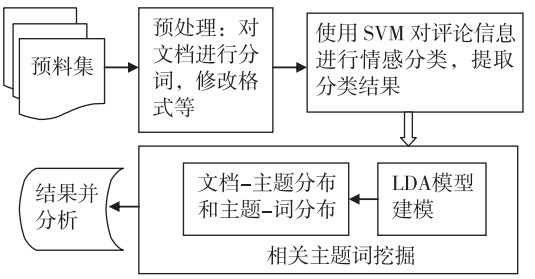

图2 评价信息相关的主题词挖掘流程

3.1 数据预处理

数据集本身包含的是原始评论信息,在使用前必须对数据进行预处理:1)去除停用词(Stopwords)。停用词是一些代词和语气助词等常用词,它们频繁出现但对于主题挖掘没有帮助,本文采用停用词词典的方法来去除评论信息中的停用词,但要把影响情感判断的词保留下来(包括“太”、“没有”、“很”);2)对评论信息进行中文分词并统计文档-词的信息。

3.2 主题挖掘

评论信息通过SVM训练的二元分类器分成正向情感和负向情感两类信息,将正向情感类作为正文档集,负向情感类作为负文档集。将正向文档集和负向文档集作为LDA的输入。通过LDA建模后,可以得到主题与词之间的分布关系,即主题-词项分布。通过主题与词的分布关系就可以对评论中蕴含的信息进行分析,做出重要的决策。

4 实验结果及分析

本文采用的数据集来源于数据堂网站[14],下载了3 000篇分别关于酒店、书、电脑等3个方面的评论数据作为实验语料集。本文选用的中文分词工具是ICTCLAS[15],分类工具是Libsvm-3.19[16],LDA模型参数设置:超参数$\alpha=50/k$,$\beta=0.1$,以上参数均为经验最优,主题数$k=10$。

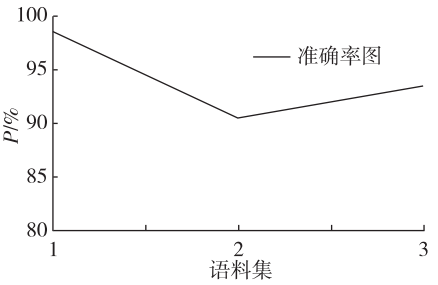

图3 情感分类效果图

4.1 情感二分类

本文只针对积极情感和消极情感进行分析(忽略中性情感)。在实验中,将语料集分为训练集1 569篇和测试集1 431篇,通过Libsvm实现SVM算法,训练二元分类器对评论信息进行分类。实验结果表明,分类的准确率均达到90%以上,如图3所示。语料集中文本分类的情况见表2。

在情感分类实验中,准确率均达到90%以上,效果良好,但是还存在

一些问题有待今后进一步研究，主要有以下两个方面：

1）中文需要分词处理，而情感分类准确率取决于分词精度。

2）中文存在“反讽”、“褒义贬用”等，目前还没有好的方法解决。

表 2　正向情感与负向情感的数量

主题	正向情感	负向情感
语料集	1 722	1 278
训练集	921	648
测试集	801	630

4.2 主题-词的挖掘

根据 LDA 模型建模可以生成主题-词项分布，就可以挖掘出评论信息中特定主题下最具有代表性的相关词项。限于空间，只列出了部分主题词表，分别是与酒店和电脑的正向情感和负向情感相关的主题词。如表 3 和表 4 所示。

表 3　酒店的正向情感类相关主题词

1	2	3
好 0.236	房间 0.093	服务 0.205 6
合适 0.081	郑州 0.047 9	坐 0.044
顶层 0.042	湖边 0.025	标准 0.044
威海 0.042	客人 0.025 1	宾馆 0.044 3
客户 0.043	点名 0.025 1	表扬 0.044 3
经理 0.043	食物 0.025	环境 0.044
价格 0.042 6	远 0.025 1	推荐 0.044
不错 0.003 8	很好 0.025	不错 0.004
酒店 0.003 8	不便宜 0.025	酒店 0.004

表 4　电脑的负向情感类相关主题词

1	2	3
太差 0.056	系统 0.106	装 0.065
清晰度 0.029 7	不能 0.037	显卡 0.065 6
问题 0.029 7	屏幕 0.037 9	驱动 0.034
开箱 0.029 7	低 0.037	问题 0.034
触摸板 0.029 7	坏 0.037 9	机子 0.034
左键 0.029 7	维修 0.03	坏 0.034
蓝屏 0.029	不好 0.037	改装 0.03
上不了 0.029	退货 0.037	偷工减料 0.03

从表 3 和表 4 上可以看到用户关注的主题词。

1）用户对酒店关注的主题词包括房间、服务、床位价格、位置、环境等。

2）用户对计算机关注的主题词包含屏幕、键盘、电池、散热、配置、无线信号等。

实验表明本文提出的方法获得了良好的分类结果，能够较为准确地挖掘出主题词。根据从评论信息中挖掘到的主题词，用户可以作出相应的购买决策；商家可以据此及时作出调整，改善产品质量和预测销售市场变化情况，调整销售策略等。

5 结语

本文对评论信息中隐含的相关主题词，提出了一种将文本情感分类和主题词挖掘相结合的方法。首先通过 SVM 对文本进行正、负情感分类，再通过 LDA 模型进行建模生成文本需要的相关主题词。该方法可以较好地挖掘出主题词。

在以后的研究工作中，将关注对情感分类的相关研究，进一步提高分类准确率，以及对本文提出的方法实现可视化。

参考文献：

[1] Wei L，Wu K H，Lee L Y，et al. Construction of an evaluation corpus for opinion extraction[C]//In NTCIR-5 Tokyo. Japan：[s. n.]，2005，12：513-520.

[2] Pang B，Lee L，Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques[C]//ACL. Philadelphia：[s. n.]，2002，02：79-86.

[3] 李太白，唐万梅. 一种改进的 SVM 多类分类算法在入侵检测中的应用[J]. 重庆师范大学学报：自然科学版，2012，9(5)：63-66.
Li T B，Tang W M. Application of an improved SVM multiclass classification to intrusion detection[J]. Journal of Chongqing Normal University：Natural Science，2012，9(5)：63-66.

[4] 谢丽星，周明，孙茂松. 基于层次结构的多策略中文微博情感分析和特征提取[J]. 中文信息学报，2012，1(1)：73-83.
Xie L X，Zhou M，Sun M S. Hierarchical structure based Hybrid approach to sentiment analysis of Chinese micro blog and its feature extraction[J]. Journal of Chinese Information，2012，1(1)：73-83.

[5] 王文帅，杜然，程耀东，等. 一种面向大规模微博数据的话题挖掘方法[J]. 计算机工程与应用，2014，50 (22)：32-37.
Wang W S，Du R，Cheng Y D，et al. Topic mining method on massive microblog data[J]. Computer Engineering and Application，2014，50(22)：32-37.

[6] 张晨逸,孙建伶,丁轶群.基于 MB-LDA 模型的微博主题挖掘[J].计算机研究与发展,2011(10):1795-1802.
Zhang C Y,Sun J L,Ding Y Q. Weibo theme mining based on MB-LDA model[J]. Journal of Computer Research and Development,2011 (10):1795-1802.

[7] 陈文涛,张小明,李舟军.构建微博用户兴趣模型的主题模型的分析[J].计算机科学,2013,4(4):127-135.
Chen W T,Zhang X M,Li Z J. Analysis of topic on modeling microblog user interestingness[J]. Computer Science,2013,4(4):127-135.

[8] 孙艳,周学广,付伟.基于主题混合模型的无监督文本情感分析[J].北京大学学报:自然科学版,2013,1(1):102-108.
Sun Y,Zhou X G,Fu W. Unsupervised topic and sentiment unification model for sentiment analysis[J]. Acta Scientiarum Naturalium Universitatis Pekinensis,2013,1(1):102-108.

[9] Mei Q Z,Ling X,Wondra M,et al. Topic sentiment mixture:Modeling facets and opinions in weblogs[C]//Proc. of the 16th Int. conference on World Wide Web. New York:ACM,2007:171-180.

[10] Blei M,Lafferty J. Text mining:theory and applications [M]. London:Chapter Topic Models,Taylor and Francis,2009.

[11] Blei D M,Ng A Y,Jordan M I. Latent dirichlet[J]. Journal of Machine Learning Research,2003,3(4/5):993-1022.

[12] Steyvers M,Griffiths T. Probabilistic topic models[M]. Latent Semantic Analysis:A Road to Meaning,Laurence Erlbaum,2005.

[13] Koller D,Friedman N. Probabilistic graphical modles:principles and techniques[M]. Cambridge:MIT Press,2009.

[14] 数据堂.数据堂页面[EB/OL].(2015-03-06)[2015-04-20]. http://datatang.com/.
Shujutang. shujutang page[EB/OL]. (2015-03-06)[2015-04-20]. http://datatang.com/.

[15] ICTCLAS.org. ICTCLAS[DB/OL].(2013-7-3)[2015-04-20]. http:www.ictclas.org/ict-clas_download.aspx.

[16] Chang C C,Lin C J. LIBSVM[EB/OL].(2014-11-05)[2015-04-20]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Text Topic Mining Based on Sentiment Classification

CHEN Guoliang, TANG Wanmei

(College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China)

Abstract: There exist a great number of comments about E-commerce on the Internet that contains personal emotions, which not only reflect product customer satisfaction, but also the market trends. With the thematic terms contained in these comments, this paper proposes an approach of combining text classification and thematic terms mining, which first classify the sentiment words by using support vector machines, then extract thematic terms from the comments dealed by LDA model. The test results with real-world datasets indicate the proposed approach of this paper is with great effectiveness that can better classify the text and dig out the thematic terms more accurately.

Key words: thematic terms; sentiment classification; LDA model

(责任编辑 游中胜)