主题特征格分析：一种用户生成文本质量评估方法

钟 将[1,2],张淑芳[2,3],郭卫丽[2],李 雪[1,4]

(1. 信息物理社会可信服务计算教育部重点实验室,重庆大学,重庆 400030;2. 重庆大学计算机学院,重庆 400030;
3. 重庆电子工程职业学院,重庆 401331;4. 昆士兰大学信息技术与电子工程学院,布里斯班,澳大利亚 4072)

摘 要: 本文设计了一种用户生成文本的质量分析框架. 首先,基于主题分析构建商品类别主题特征集合. 其次,利用主题特征与商品分类的强关联关系,构建形式化概念分析的形式背景,将分类-主题概念格化简并生成主题特征格,以此构建五个质量特征并生成质量评估模型. 最后,在真实评论数据上的实验结果表明新方法具有更高预测精度.

关键词: 用户评论; 质量评估; 主题特征; 主题特征格

中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2018)09-2201-06

电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.09.022

TFLA: A Quality Analysis Framework for User Generated Contents

ZHONG Jiang[1,2], ZHANG Shu-fang[2,3], GUO Wei-li[2], LI Xue[1,4]

(1. *Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Changqing University, Chongqing* 400030, *China*;
2. *College of Computer Science, Chongqing University, Chongqing* 400030, *China*;
3. *Chongqing College of Electronic Engineering, Chongqing* 401331, *China*;
4. *School of Information Technology and Electrical Engineering, University of Queensland, Brisbane* 4072, *Australia*)

Abstract: In this paper, we design a topic-features lattices analysis (TFLA) framework based on objectivity quality dimensions. Firstly, we apply the latent Dirichlet allocation (LDA) approach to get latent topics as topic-features for each goods categories. Secondly, we construct formal background based on the strong relationship between goods categories and topic-features. So we could get generalization and instantiation relationship among the topic-features through formal concept analysis (FCA). We employ domain knowledge and relationships among topic-features to define five objective quality features. Also, we use machine learning methods to build quality evaluation models based on these quality features. Experiment results on actual comment data sets show that our new quality models' prediction results are in agreement with the artificial quality tags in most cases. The best performances could get that the mean absolute error (MAE) is 0.7 and F-measure is 0.5, which is significantly better than the conventional quality prediction model based on support vector machine (SVM) classification.

Key words: user comment; data quality; topic features; lattices of topic-features

1 引言

用户评论数据作为典型的用户生成内容,可有效降低交易双方的信息不对称,对于消费者进行决策,提供了重要的参考依据. 另一方面,从海量评论数据中归纳出消费者对商品特性的认知和评价,对于生产经营者具有十分重要的意义[1,2].

在线评论文本数据的用词随意、结构不规范,其质量取决于评价者的经验、态度和能力. Kim[3]等从评论的有用性角度来衡量评论质量,从结构、词汇、句法、语义和评论星级五个方面来建立分类器来分析评论的有用性. Otterbacher[4]从信誉度、关联性、可读性、可信度

收稿日期:2017-06-05;修回日期:2017-07-26;责任编辑:梅志强
基金项目:国家863高技术研究发展计划(No. 2015AA015308);国家重点研发计划项目(No. 2017YFB1402401);重庆市社会事业与民生保障科技创新专项(No. cstc2017shmsA20013)

和客观性五个方面共计22个特征来量化评论文本数据的质量,通过相关性分析和回归分析建立五个维度与有用性之间的映射函数. Ghose 等[5]将评论中的数据划分为客观性评论和主观性的评论数据,分别为消费者和厂商设计了有用性排名机制,该机制结合了计量经济学、文本挖掘方法. Aika 等[6]根据评论数据中出现的概念数量来度量评论数据的质量,其中,包括每句评论中包含的概念数量、总概念数量及概念的类别数量.

已有研究表明,在线评论质量主要取决于评价内容的特征,如文本可读性、关联性、有效长度等. 因此本文设计的质量评估框架主要是利用评论文本的内容来提取其客观性的质量特征.

本文主要贡献为:(1)设计了基于形式化概念分析构建评论数据中主题特征集之间的概化和例化关系;(2)设计了具有良好可解释性的质量特征维度和量化方法;(3)针对缺乏质量评估基准数据问题,通过人工方式标记了一定规模的测试基准数据.

2 问题定义

2.1 用户评论与评论短语集合

在线评论数据中的评论文本通常包含若干评论短语来表达消费者的观点和感受. 通常评论短语是针对产品某项属性特征的主观感受,常用的句法结构有:"〈属性特征词〉+〈形容词〉","〈属性特征词〉+〈副词〉+〈形容词〉","〈形容词〉+'的'+〈属性特征词〉","〈副词〉+〈形容词〉+'的'+〈属性特征词〉"等.

定义1 属性特征词集合 AW:是指用户在评价商品时所使用的属性特征词的集合 $\mathrm{AW}=\{w_1,\cdots,w_i,\cdots w_m\}$.

定义2 评论短语 vp_i:指由属性特征词和用户观点构成的二元组 $\mathrm{vp}_i=\langle \mathrm{attr}_i,\mathrm{option}_i\rangle$,其中 $\mathrm{attr}_i\in \mathrm{AW}$,而 option_i 则表示用户的主观观点. 用户评论文本可以使用评论短语集合 VP 来表示.

2.2 质量特征与质量评估函数

定义3 质量特征向量 $\boldsymbol{F}$:质量特征向量 $\boldsymbol{F}=[f_1,\cdots f_i,\cdots,f_n]$,其中每个元素对应在线评论数据的一项质量特征维度.

质量特征维度定义表示为函数 $f_i=\psi_i(\mathrm{VP},\mathrm{KD})$,其中 VP 为评论短语集合,KD 为评论数据的领域知识. ψ_i 是从评论数据中构建量化的质量特征函数.

定义4 质量评估函数 G:实现质量特征向量 $\boldsymbol{F}$ 到在线评论数据质量 R 之间的映射. 若评论短语集合为 VP_k,质量特征向量为 $\boldsymbol{F}_k$,其对应质量评估值 $g_k=G(\boldsymbol{F}_k)$.

在线评论数据质量评估的核心问题是寻找良好的数据质量特征维度构建方法,使质量特征维度的可解释性好、预测精度高.

3 用户评论数据的主题特征格构建

3.1 构建主题特征格的算法框架

本文设计的主题特征格构建框架如图1所示. 首先,将在线评论中对于每件产品评论数据合并为单个文本文档,并利用句法规则将文档转换为评论短语的集合,并以此为基础生成文档-特征词矩阵.

其次,利用 LDA 主题分析方法[7]获得文档-主题概率矩阵和主题-特征词概率矩阵. 本文称这些潜在主题为商品的主题特征,并根据其与文档和特征词的概率大小度量它们之间关联的强度.

然后,根据文档所属商品类别来构建每个商品类别与主题特征之间的强关联关系,并以此作为形式化概念分析(FCA)[8]的形式背景. 最后,将 FCA 获得的分类-特征概念格进行化简,获得文本评论数据的主题特征格.

主题特征与商品类别之间的关联概率的计算方法为:先使用式(1)获得商品类别 c_i 与主题特征 t_k 之间关联概率 $r_{i,j}$,其中 $\boldsymbol{Q}$ 为文档与主题的关联概率矩阵,D_i 为商品类别下所有商品对应的评论文档集合. 最后将所有商品类别与主题特征 t_k 间的关联概率进行归一化处理.

$$r_{i,j}=\max_{d_k\in D_i} q_{k,j},q_{k,j}\in \boldsymbol{Q} \tag{1}$$

3.2 类别-特征概念格分析

定义5 类别强关联主题特征:令 $R_i=\{r_{i,1},\cdots,r_{i,k}\}$ 为商品类别 c_i 与主题特征之间关联概率,其从大到小排列为 $\{s_1,\cdots,s_k\}$,$l=\arg\max_{j=1..K-1}(s_j-s_{j+1})$,若 $r_{i,k}\geqslant s_l$,那么主题特征 t_k 为 c_i 的强关联主题特征. 记 ST_i 为商品类别 c_i 强关联的主题特征集.

定义6[8] 类别-主题特征形式背景 U:类别-主题特征形式背景 $U=(C,T,\boldsymbol{I})$,其中 $C=\{c_1,\cdots,c_i,\cdots,c_N\}$ 为商品分类集,$T=\{t_1,\cdots,t_j,\cdots,t_K\}$ 为主题特征集,$\boldsymbol{I}\subseteq C\times T$,$(c_i,t_j)\in\boldsymbol{I}$ 当且仅当 $t_j\in\mathrm{ST}_i$.

定义7[8] $*$ 运算:设有类别-主题特征形式背景 $U=(C,T,I)$,若 $X\subseteq C,B\subseteq T$,$*$ 运算定义见式(2).

$$\begin{aligned} X^*&=\{t_i\mid t_i\in T,\forall x\in X,(X,t_i)\in\boldsymbol{I}\}\\ B^*&=\{x\mid x\in C,\forall t_i\in B,(X,t_i)\in\boldsymbol{I}\}\end{aligned} \tag{2}$$

定义8[8] 形式概念(X,B):对于类别-主题特征形式背景 $U=(C,T,\boldsymbol{I})$,二元组(X,B)满足 $X\subseteq C,B\subseteq T$ 且 $X^*=B,B^*=X$,那么二元组(X,B)就是一个类别-特征形式背景下的形式概念. 二元组中的 X 为概念中的商品分类对象集合,称为形式概念的内涵,B 为主题特征属性集合,称为形式概念的外延.

定义9 偏序关系$\prec$:若商品评论领域的两个概念(X_1,B_1),(X_2,B_2)满足 $X_1\subseteq X_2(B_1\subseteq B_2)$,那么两个概念就存在偏序关系$(X_1,B_1)\prec(X_2,B_2)$.

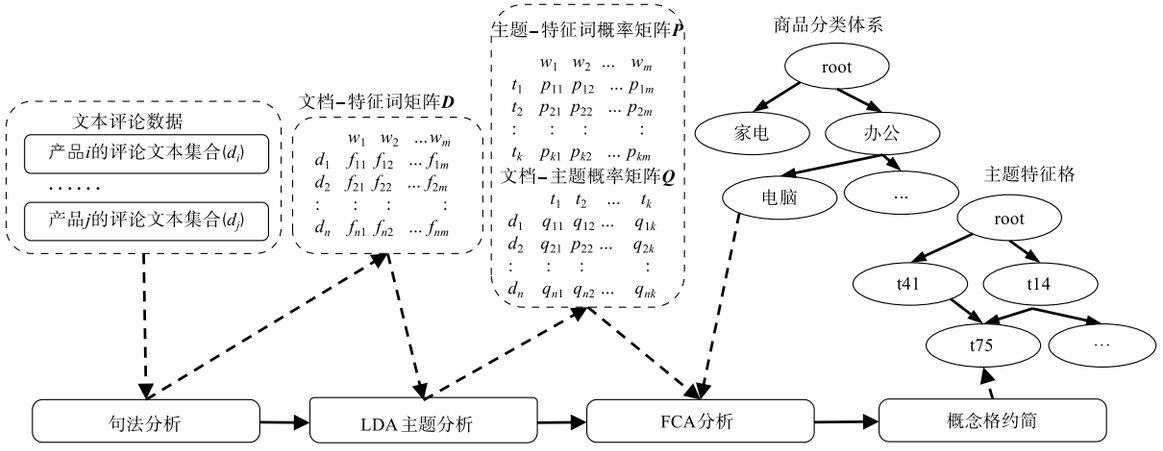

图1　主题特征概念格算法框架

定义 10　类别-特征概念格:类别-主题特征形式背景中的所有概念的偏序集记为 $L(C,T,\boldsymbol{I})$,称为类别-特征概念格.

按照上述定义,根据商品评论文档和商品分类体系构建类别-特征概念格,如算法 1 所示.

算法 1　商品类别-特征概念格生成算法

输入:商品评论文档集合 D,商品分类集合 C,评论文档与商品分类的映射函数 ϕ,潜在主题数 K

输出:类别-特征概念格 L,属性特征词集合 AW,主题特征-属性特征词概率矩阵 $\boldsymbol{P}$,商品类别-主题特征概率矩阵 $\boldsymbol{R}$

算法步骤:

Step1:for each $d_i \in D$

Step2:$d_i \rightarrow \mathrm{VP}_i$//文档转化为评论短语集合

Step3:for each vp in VP_i

Step4:　$W = W \cup \{\mathrm{vp.attr\_}w\}$

Step5:　end for

Step6:end for

Step7:for each VP_i in VP

Step8:$\mathrm{VP}_i \rightarrow \boldsymbol{t}_f$　//$\boldsymbol{t}_f$ 为属性特征词频向量

Step9:　$\boldsymbol{D}' = [\boldsymbol{D}';t_f]$

Step10:end for

Step11:$(\boldsymbol{D}',W) \xrightarrow{\mathrm{LDA}} (T,\boldsymbol{P},\boldsymbol{Q})$//主题分析

Step12:$(\boldsymbol{D}',\boldsymbol{P},\phi,C) \rightarrow \boldsymbol{R}$//分类主题关联矩阵

Step13:$\boldsymbol{R} \rightarrow \boldsymbol{I}$　//根据类别强关联主题特征生成形式分析背景 U 中的关联矩阵 $\boldsymbol{I}$

Step14:$(C,T,\boldsymbol{I}) \xrightarrow{\mathrm{FCA}} L(C,T,\boldsymbol{I})$　//生成类别-特征概念格

Step15:Return $\{L(C,T,\boldsymbol{I}),W,\boldsymbol{P},\boldsymbol{R}\}$

类别-特征概念格中具有偏序关系的两个概念 concept < child_concept 具有父子关系.

3.3　主题特征格及其性质

为获取主题特征之间的概化/例化关系,可以将类别-特征概念格 L 在主题特征集进行投影操作,构建评论特征格 L'.

定义 11　主题特征格:评论主题特征格 $L'(T)$ 是类别-特征概念格 $L(C,T,\boldsymbol{I})$ 在主题特征 T 上的投影,其节点仅包含类别-特征概念格中的主题特征集合,主题特征集之间的偏序关系与其对应形式概念之间的偏序关系保持一致.

如直接在类别-特征概念格上进行投影运算将包含大量冗余,本文先对类别-特征概念格进行化简. 化简的基本步骤是利用广度优先方式遍历每个概念节点,将其子概念节点中的内涵(主题特征集合)替换为其内涵与其所有的父概念内涵的差集,若结果为空,则将该节点删除,并将该删除节点的父节点加入其孩子节点的父节点集合. 化简之后再通过在主题特征集上的投影操作获取主题特征格.

主题特征格 L' 中若包含二元组 $T_i < T_j$,则称 T_j 为 T_i 的父节点,可将评论特征格转换为一个 Hasse 图[9]. 以下根据主题特征集之间的偏序关系来定义主题特征集的深度和距离.

定义 12　主题特征格的节点深度:若 T_i 为特征格中的主题特征集,其深度定义如式(3):

$$\mathrm{depth}(T_i,L') = \begin{cases} 0, & \text{if } T_i \text{ is root} \\ \max\limits_{T_j \in T_i.\,\mathrm{Parents}} (\mathrm{depth}(T_j,L')+1), & \text{others} \end{cases} \tag{3}$$

其中 $T_i.\,\mathrm{parents}$ 表示 T_i 父节点的集合.

定义 13　主题特征深度:设 t_i 为某个主题特征,其深度为评论特征格中所有包含该主题特征的节点深度的最大值,定义如下式:

$$\mathrm{depth}(t_i,L') = \max_{t_i \in T_j} (\mathrm{depth}(T_j,L')) \tag{4}$$

定义 14　主题特征间的距离:假设主题特征 t_i, t_j 满足条件 $t_i \in T_i, t_j \in T_j$,那么它们之间的距离定义为:

$$\mathrm{dist}(t_i,t_j,L')=\begin{cases}0,\text{if } T_i=T_j\\ \min\limits_{T_k}(\mathrm{depth}(T_i,L')+\mathrm{depth}(T_j,L')\\ \quad -2*\mathrm{depth}(T_k,L')),T_j\prec T_k,T_i\prec T_k\end{cases}\tag{5}$$

4 在线评论数据质量维度特征

本文从相关度、全面度、专业度、内聚度及可读性五个质量维度来衡量评论文本的质量. 假设某个评论对应的评论短语集合为 VP,$n=|\mathrm{VP}|$,每个质量评价维度定义如下:

相关度(relativity):是指评论数据与商品的相关程度.

若评论数据中的评论短语对象为 vp_i 中的特征词为 w_i,其关联的主题特征为 t_j,而评论涉及的商品类别为 c_k,那么评论数据相关度定义为式(6)所示.

$$\mathrm{Rela}(\mathrm{VP})=\sum_{i=1}^{n}p_{i,j}\cdot r_{k,j}/n\tag{6}$$

其中,$p_{i,j}$为特征词 w_i 与主题特征 t_j 之间的关联概率值,$r_{k,j}$为主题特征 t_j 与商品类别 c_k 关联的概率.

全面度(comprehensiveness):是指评论中有效主题特征的数量.

若评论文本评价的商品类别为 c_i,其强关联的主题集合为 ST_i,全面性计算可按照式(7)计算.

$$\mathrm{Comprehensiveness}(\mathrm{VP})=\sum_{i=1}^{n}h(\mathrm{vp}_i)/|\mathrm{ST}_j|\tag{7}$$

其中 $\mathrm{vp}_j=(w_j,\mathrm{option}_j)$ 为 VP 中的评论短语,函数 h 的定义如式(8)所示:

$$h(\mathrm{vp}_j)=\begin{cases}1, & \mathrm{tp}(w_j,P)\in\mathrm{ST}_j\\ 0, & \text{otherwise}\end{cases}\tag{8}$$

可读性(readability):指评论数据的用语是否简洁、规范和无冗余.

本文的可读性根据评论短语集合中有效的评论短语的比例来度量. 假设 VP 为评论文本数据对应的评论短语对象集合,而 VP′为去重后评论短语对象集合,可读性的计算如式(9)所示:

$$\mathrm{readability}(\mathrm{VP})=\sum_{\mathrm{vp}_i\in VP'}h(\mathrm{vp}_i)/|\mathrm{VP}|\tag{9}$$

其中函数 $h(\mathrm{vp}_i)$的定义参见式(8).

专业度(specificity):是指评论涉及主题特征的平均深度.

显然评论短语较多使用特定商品的特征词,那么其评论越专业,专业度的计算见式(10).

$$\mathrm{Spec}(\mathrm{VP})=\sum_{i=1}^{n}\mathrm{depth}(t_i,L')/n\tag{10}$$

其中 t_i 为评论短语对象 vp_i 中的属性特征词关联的主题特征.

内聚度(Cohesion):指评论所涉及主题特征的相似性. 如果评论文本中包含多段(句)评论数据,则该评论的内聚度为每段评论数据的平均值.

$$\mathrm{Cohe}(\mathrm{VP})=2*\frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathrm{sim}(\mathrm{vp}_i,\mathrm{vp}_j)}{(n*(n-1))}\tag{11}$$

其中:$\mathrm{sim}(\mathrm{vp}_i,\mathrm{vp}_j)=\mathrm{e}^{-\mathrm{dist}(t_i,t_j,L')}$,$t_i$ 和 t_j 分别是 $\mathrm{vp}_i,\mathrm{vp}_j$ 关联的主题特征.

获得文本评论数据对应的质量特征维度后,就可使用各种机器学习方法建立质量特征维度与数据质量之间的映射函数.

5 实验与结果分析

5.1 实验数据

数据来自某知名 B2C 电子商务网站上 116 个类别,109,564 件商品的 18,415,146 条在线评论数据. 实验选取了至少包含 500 条在线评论的 6,212 个商品的评论数据进行实验. 利用信息增益(information gain)选择 12000 个属性特征词.

数据质量的基准数据采用人工标注方法:由两名标注人员同时评分,当分值相同时作为有效标注,否则舍弃该标注数据. 实验中特别选择了标注人员熟悉的“笔记本电脑”、“手机”和“洗发护发”三类商品按照五分制对评论数据的质量进行标注,共获得了 3400 条测试基准数据.

5.2 实验结果

5.2.1 生成的评论特征格

主题分析 LDA 算法参数设置为:隐含主题数为 80,采用对称 Dirichlet 分布,参数 $\alpha=2,\beta=0.5$ 分别表示评论文档生成主题特征,主题特征生成特征词的 Dirichlet 分布函数的参数,吉布斯采样算法迭代次数为 5000.

表 1 中给出其中 6 个主题特征中关联概率最大的 5 个属性特征词. 如主题特征 Topic 68 关联最大的属性特征词为头发、效果、感觉、头皮和牌子等 5 个词,其中头发与该主题关联的概率为 0.282.

表 2 给出了部分商品类别的强关联主题特征,这些主题特征按照概率值大小排序. 其中与笔记本电脑关联度的主题特征有 Topic 75 和 Topic 18,通过分析 Topic 75 和 Topic 18 中关联的属性特征词发现,排序靠前的词汇包含了电脑、系统、性能、电源、接口等与笔记本电脑密切相关的词. 商品类比的强关联主题特征与我们的常识性知识具有很高的吻合度.

图 2 为实验生成的评论特征格对应的 Hasse 图的部分内容,图 3 为实验结果. 根据计算结果发现,距离根节点较近主题特征所关联的特征词通常代表一般化的概念,例如外观、样子、颜色、感觉等;反之,距离根节

点较远的主题特征所关联的特征词则更倾向于专用概念,如节点主题 Topic 75 中关联了电脑、内存等与计算机相关的专业概念. 这表明了本算法生成的评论特征格能够反映了主题特征之间的概化和例化关系.

表 1　部分属性特征词-主题特征之间的关联概率表

Topic 0		Topic 1		Topic 18		Topic 68		Topic 75		Topic 79	
颜色	0.764	包装	0.807	电源	0.157	头发	0.282	电脑	0.167	手机	0.257
红色	0.029	外包装	0.048	做工	0.119	效果	0.273	系统	0.145	功能	0.088
图片	0.027	盒子	0.028	功率	0.072	感觉	0.111	速度	0.125	电池	0.060
色彩	0.015	包装盒	0.011	接口	0.057	头皮	0.057	笔记本	0.058	屏幕	0.046
粉色	0.011	品质	0.005	用料	0.037	牌子	0.020	性能	0.042	机子	0.045

表 2　部分商品类别强关联主题特征(形式背景数据)

商品类别	与商品类别强关联的部分主题
平板电脑	topic 60、topic 70、topic 78、topic 38、topic 44、topic 79
笔记本电脑	topic 75、topic 60、topic 18、topic 48、topic 67、topic 21、topic 14
手机	topic 79、topic 59、topic 48、topic 70、topic 5、topic 21、topic 44
洗发护发	topic 68、topic 54、topic 31、topic 8、topic 17、topic 70、topic 73
厨卫五金	topic 43、topic 71、topic 66、topic 20、topic 48、topic 21

5.2.2　质量评估模型的性能分析

支持向量机(SVM)分类器在高维数据集上具有良好性质. 本文基于 SVM 分类器分别在 2000 维特征词(TFIDF)和上面提出的 5 个质量特征维度(CF5),在标记的测试基准数据上进行了十轮交叉验证. SVM分类器在 2000 维时采用线性核,在 5 个质量维度时采用了高斯核,宽度参数 σ 取值为 2.

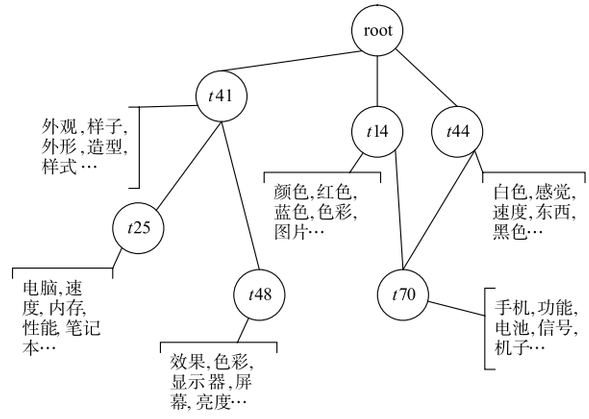

图2　评论特征格对应Hasse图(部分)示意

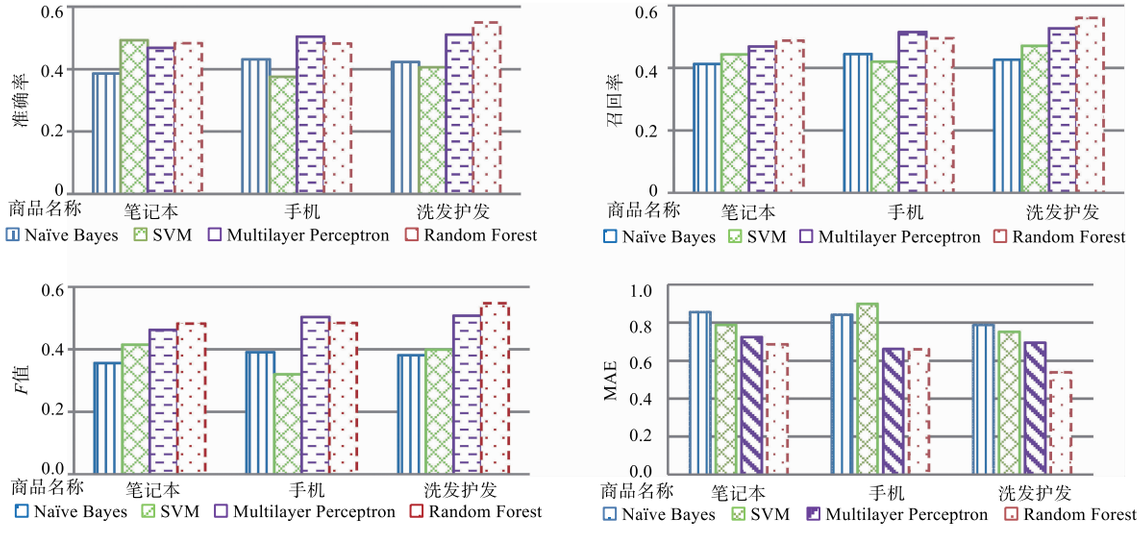

图3　采用不同分类器的质量评估性能比较

实验结果的评价指标分别为平均绝对误差(MAE)、准确率(precision)、召回率(recall)和 F 值(F-measure).

表 3 中 TFIDF 和 CF5 分别表示采用词向量特征和本文提出的 5 个质量特征维度. 结果表明:采用本文设计的 5 个质量维度预测性能均显著优于词向量特征. 特别是用户最关心的绝对误差(MAE)仅为词向量特征的 50% 左右.

为选择合适的分类器构建质量预测模型,实验中对比了朴素贝叶斯方法(Naïve Bayes)、支持向量机(SVM)、

多层感知机(multilayer perceptron)以及随机森林(random forest)四种常用的分类器. 随机森林中决策树的最大数量为 100,多层感知机隐含层为 3,学习率为 0.2.

表 3 SVM 在 CF5 和 TFIDF 上的性能比较

性能	笔记本	手机	洗发护发	加权平均
Precision(CF5)	0.493	0.375	0.406	0.425
Precision(TFIDF)	0.308	0.328	0.284	0.307
Recall(CF5)	0.444	0.420	0.471	0.445
Recall(TFIDF)	0.195	0.253	0.197	0.215
F-Measure(CF5)	0.415	0.320	0.399	0.378
F-Measure(TFIDF)	0.239	0.286	0.233	0.252
MAE(CF5)	0.789	0.899	0.752	0.813
MAE(TFIDF)	1.678	1.534	1.456	1.556

图 3 所示的实验结果可知:基于质量特征,所有数据质量预测模型的准确率、召回率以及 *F* 值均能在 0.5 左右,MAE 值均保持在 0.9 以下. 在本文定义的 5 个质量维度时,采用随机森林方法获得的综合性能最优,即在三类商品上各类指标的加权平均值均为最优. 其中平均性 MAE 值为 0.628,显著优于多层感知机模型的 0.693.

6 总结与展望

针对典型的在线用户生成数据的质量分析问题,本文设计了评论主题特征格的构建方法,以发现主题特征之间的概化/例化关系,并以此为基础定义了五种易于理解的数据质量分析特征维度. 通过测试基准数据集上的实验表明,新的质量预测模型不仅具有良好的可解释性,且预测精度显著高于基于文本分析方法的预测模型.

为进一步提高质量预测模型的精度,今后还可研究基于语用分析方法识别评论数据中的隐含商品属性特征,采用基于语义约束的主题分析方法更精确计算属性特征词与主题特征间的概率关系以及基于情感分析方法挖掘用户观点的强度和自信度等主观性度量等.

参考文献

[1] Anindya G. ,Panagiotis G. I. Estimating the helpfulness and economic impact of product reviews:mining text and reviewer characteristics[J]. IEEE Trans on Knowledge And Data Engineering,2011,23(10):1498 – 1512.

[2] 林煜明,王晓玲,朱涛,周傲英. 用户评论的质量检测与控制研究综述[J]. 软件学报,2014,25(3):506 – 527.
LIN Yu-Ming, WANG Xiao-Ling, ZHU Tao, ZHOU Ao-Ying. Survey on quality evaluation and control of online reviews[J]. Journal of Software, 2014, 25(3): 506 – 527. (in Chinese)

[3] Kim S M,Pantel P,Chklovski T,et al. Automatically assessing review helpfulness[A]. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing[C]. USA:ACM,2006. 423 – 430.

[4] Otterbacher J. 'Helpfulness' in online communities:a measure of message quality[A]. Proceedings of the SIG CHI Conference on Human Factors in Computing Systems[C]. USA:ACM,2009. 955 – 964.

[5] Ghose A, Ipeirotis P G. Designing novel review ranking systems:predicting the usefulness and impact of reviews[A]. Proceedings of the Ninth International Conference on Electronic Commerce[C]. USA:ACM,2007. 303 – 310.

[6] Aika Q,Karim B S S,et al. A concept-level approach to the analysis of online review helpfulness[J]. Computers in Human Behavior,2016,58(5):75 – 81.

[7] 王李冬,魏宝刚,袁杰. 基于概率主题模型的文档聚类[J]. 电子学报,2012,40(11):2347 – 2350.
WANG Li-dong, WEI Bao-gang, YUAN Jie. Document clustering based on probabilistic topic model[J]. Acta Electronica Sinca, 2012, 40(11): 2347 – 2350. (in Chinese)

[8] 张磊,张宏莉,韩道军,沈夏炯. 基于概念格的 RBAC 模型中角色最小化问题的理论与算法[J]. 电子学报,2014,42(12):2371 – 2379.
ZHANG Lei,ZHANG Hong-li,HAN Dao-jun,SHEN Xia-jiong. Theory and algorithm for roles mininzation problem in RBAC based on concept lattice[J]. Acta Electronica Sinca,2014,42(12):2371 – 2379. (in Chinese)

[9] Freese R. Automated lattice drawing[A]. Second International Conference on Formal Concept Analysis[C]. Berlin:Springer-Verlag,2004. 112 – 127.

作者简介

钟 将 男,1974 年出生,重庆江津人. 博士,教授,主要研究方向为数据挖掘及应用,网络信息安全.
E-mail:zhongjiang@cqu.edu.cn

张淑芳 女,1972 年出生,陕西澄城人,博士研究生,副教授,主要研究方向大数据挖掘和模拟计算.
E-mail:roseymcn2000@foxmail.com

郭卫丽 女,1990 年出生,河北行唐人,硕士,主要研究方向为数据挖掘、高性能计算.
E-mail:870188993@qq.com

李 雪 男,1955 年出生,重庆沙坪坝人,博士,教授,主要研究方向为数据挖掘,大数据.
E-mail:xueli@itee.uq.edu.au