文章编号：2095-2163(2021)01-0036-06 中图分类号：TP181 文献标志码：A

基于图卷积网络的服装评价信息分类问题的研究

姚婷婷，刘国华
（东华大学 计算机科学与技术学院，上海 201620）

摘 要： 随着互联网的快速发展以及电子设备的逐渐普及，越来越多的人选择在网上购物，买家在购买商品之后，可以通过平台提供的评价系统表达自己对服装产品的感受，因此会产生大量的服装评价信息。由于这些评价信息的标签是通过人工选择的，会受到外在因素的影响，所以具有不确定性。这些不确定性产生的误差会影响到平台以及其他用户对服装产品的评判。针对这一问题，本文研究了一种基于图卷积的分类方法，将单词、文档、主题视为节点，三者之间的关系作为边构建大型异构图网络。将该异构图作为图卷积网络模型的输入，并引入了注意力机制，根据不同邻居节点与某一特定节点的关系具有不同的重要程度，构建了关注矩阵。最后对一个公开的服装评价文本进行实验评估以及分析，实验结果表明本方法取得的分类结果优于传统神经网络。

关键词： 文本分类；文档主题生成模型；服装评价；图卷积网络；注意力机制

Research on classification of apparel comment information based on Graph Convolutional Network

YAO Tingting, LIU Guohua
(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

【Abstract】 With the rapid development of the Internet and the gradual popularization of electronic devices, more and more people choose to shop online. After buying goods, buyers can provide their own feelings about clothing products through the comment system provided by the platform, which will generate a lot of apparel comment information. Since the labels of these comment information are manually selected and will be affected by external factors, they are uncertain. The errors caused by these uncertainties will affect the judgment of the platform and other users on clothing products. To solve this problem, this paper studies a classification method based on graph convolution, which regards words, documents, and topics as nodes, and the relationship among the three as edges to build a large heterogeneous graph network. The heterogeneous graph is used as the input of the graph convolution network model, and the attention mechanism is introduced. According to the different importance of the relationship among different neighbor nodes and a specific node, the attention matrix is constructed. Finally, an experimental evaluation and analysis of a public clothing evaluation text are carried out. The experimental results show that the classification results obtained by this method are better than traditional neural networks.

【Key words】 text classification; Latent Dirichlet allocation; apparel comment; Graph Convolution Network; attention mechanism

0 引 言

随着电子商务的飞速发展以及电子设备的普及，越来越多的人选择在网上购物并且发表自己对商品的评价，服装行业亦是如此。这些服装评价信息反映了已购用户对服装的满意程度。服装评价信息中所包含的对服装特征的自然语言表述，一方面会对潜在用户的购买行为产生影响，另一方面可以为商户和电商平台对服装的市场价值的评估提供重要的评判依据，同时也为商家能不断改进服装提供了方向[1]。所以，服装评价信息对所有用户、电商平台、商户、数据研究者都具有重要意义。如果能够采用合适的算法对服装评价信息进行研究，无疑对生产生活都能提供帮助。

本文基于图卷积网络，对售卖服装的网站的评论信息进行分析。分析的意义在于，在某服装页面下的评价信息非常多的情况下，用户和商家想要查看已购用户对该服装的看法无疑会耗费大量时间和精力。虽然现在有些服装售卖网站提供了好评/差评的选项给买家进行人工选择。但是由于人工选择会有很多外在因素影响，所以具有不确定性。例如，一部分用户虽然对服装不满意，但是由于商家耐心

基金项目： 上海市工业互联网创新发展专项项目“面向纺织服装的行业级工业互联网平台项目”(2019-GYHLW-004)。

作者简介： 姚婷婷(1995-)，女，硕士研究生，主要研究方向：人工智能、自然语言处理；刘国华(1966-)，男，博士，教授，博士生导师，主要研究方向：人工智能、大数据、关系数据库理论。

通讯作者： 姚婷婷 Email：594257368@qq.com

收稿日期： 2020-10-26

的服务态度以及良好的物流体验选择了好评，同理，也会有一部分用户将基于商家不好的服务态度，物流速度慢等原因给出了差评，但是该用户群体对服装本身还是满意的，这就可能对商家和其他用户对商品的判定产生误差[2]。所以，本文采用的半监督图卷积文本分类能实现在少量标注文档的情况下实现较强的分类性能，并能可解释地学习单词和文档节点嵌入。

1　相关技术

1.1　传统文本分类

传统文本分类主要是由特征工程加分类模型两部分组成的。特征工程的主要目的是将数据转换成计算机可以理解的形式，且保留了足够用于分类的信息，能够正确表达文本的内容。词袋模型或向量空间模型是最常用的传统特征工程方法，方法中容易忽略文本的上下文关系，每个词之间彼此独立，并且无法表征语义信息[3]。而传统分类器主要作用是对特征工程处理过的信息进行分类，常见分类模型有朴素贝叶斯分类算法、KNN、SVM、最大熵和神经网络等，分别有计算量大、内存消耗大、欠拟合、分类精度低等缺点。

1.2　深度神经网络

传统的文本分类高维度高稀疏的特性，导致了其计算量大，内存占用多等缺点；特征表达能力差的特性，导致分类精度低；而且需要人工进行特征工程，耗费人力。因此图神经网络这一课题近年来受到越来越多的关注，在大量的文本数据面前将首先要研究文本表示，然后再利用 CNN 或 RNN 等神经网络模型进行文本分类，省去了人工进行特征工程的麻烦。

首先关于文本表示，学者们研究了很多有效的词嵌入方法，将文本用词向量的形式表示出来，在 2013 年 Mikolov 等人[4]发表了 2 篇关于 word2vec 的文章，同时还发布了 word2vec 工具包，跳字模型（skip-gram）和连续词袋模型（CBOW），将词嵌入模型变得更加成熟，并得到大规模应用。还有一些研究者将词嵌入聚合成文档嵌入并作为分类器的输入，至此，文本数据的表示解决了高维度高稀疏的问题。

其次，利用 CNN、RNN 等深度神经网络及其变体实现文本分类的问题。2014 年，Kim 提出的 TextCNN 主要对 CNN 的输入层做变形来进行文本分类[5]。利用训练好的词向量完成分类任务，简单的网格结构使得 TextCNN 具有计算量少，训练速度快等优势，在很多公共数据集上取得了不错的效果。但是 TextCNN 依然有局限性，其视野局限在窗口大小范围内，使其面对较长的文本序列时分类能力下降，只适合短文本分类。为了对长文本分类，且更好地表达上下文信息，Liu 等人[5]在 2016 年和 Luo 等人[6]在 2014 年使用 LSTM 来学习文本表示。CNN 和 RNN 在文本分类中都能取得显著的效果，但是可解释性不好，所以又引入了注意力机制来捕获每个词对结果的贡献程度。虽然这些方法有效地实现了利用神经网络进行文本分类，但是都忽略了全局词共现的问题，词共现中携带了不连续以及长距离的语义信息。

1.3　图神经网络

由于生活中很多数据并不具备规则的空间结构，对于这些不规则的数据，普通卷积显得难以使用固定的卷积核来适应不规则的图结构，所以研究者们又提出了一种新的图卷积模型。基于图的深度学习最早由 Gori 等人[7]在 2005 年提出，使得学习过程可直接架构于图数据之上。之后 2009 年 Scarselli 等人[8]又提出了一种监督学习的方法 GNN，基于信息传播机制，每一个节点通过相互交换信息来更新自己的节点状态，直到达到某一个稳定值。但是这种算法计算量非常大。2016 年，Defferrard 等人[9]开始探讨积分在文本分类上有好的结果的原因，从频谱上论证了方法的可行性。2016 年，Kipf 等人[10]的方法把频谱图卷积的定义进行简化，将文本文档建模为文档词图，极大提高计算效率。该模型在一系列基准数据集上取得了很好的分类结果。2019 年，Yao 等人[11]首次提出构建以单词和文档为节点的异构图网络，并没有使用注意力机制来捕获节点与节点之间的重要程度，使得在评价类文本中准确率略低于 CNN，LSTM 等神经网络模型。本文的方法基于频谱图卷积神经网络，对 Yao 等人的模型进行了进一步改进，将单词和文档作为节点构建图数据，再用 GCN 进行卷积。并引入了注意力机制，关注节点之间的重要程度。

2　服装分类模型

2.1　问题描述

已知一个用户评价信息的集合 $O=\{O_1,O_2,O_3,\cdots,O_n\}$，$O_i=\{id,class,review\}$，$(O_i\in O)$，表示每个用户的评价信息，以及一个预先定义的类别 $C=\{c_1,c_2\}$，求一个映射模型 $F(\cdot)$，使 $\forall O_i\in O\xrightarrow{F(\cdot)}C$。

本文需要对具有少量标签数的服装文本信息进行分类,提出了一种基于图神经网络的半监督文本分类的方法,从语料库中构建了一个大型异构图,图中节点为单词、文档和主题,图中的边由单词文档和主题之间的关系连接,这样可以捕捉到全局的词共现信息。再使用 Kipf 和 Welling 在 2017 年提出的图卷积网络对图中节点进行训练,还引入了 Kiran 在 2017 年提出的注意力机制,对节点之间的边添加注意力权重。使之更加适应情感分类。

2.2 异构图的构建

本次研究中,构建了一个集合了单词节点 $C=\{c_1,\cdots,c_m\}$,文档节点 $D=\{d_1,\cdots,d_n\}$,以及主题节点 $T=\{T_1,\cdots,T_i\}$ 的异构图网络 $G(V,E)$,并使用基于 Wikipedia 语料库的 word2vec 学习单词以及文档的嵌入,而潜在主题的嵌入选用单词上的概率分布来表示。如图 1 所示,将 $\{C,D,T\}$ 所有节点的集合来表示图 G。其中,G 节点数量为 m,n,i 三者之和。每个节点都被表征为特征向量。文档节点之间的边可通过 2 个节点之间的相似度评分点互信息 PMI 确定,如果 PMI 大于 0,则在文档和单词之间建立一条边,且边的权重为 $PMI(i,j)$,单词和单词之间的边采用词频-逆文档频率来判断,每个话题和文档之间,可将每个文档分配给概率最大的前 k 个主题[12]。因此,如果将文档分配给主题,则会建立文档和主题之间的边缘。

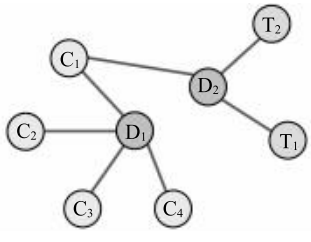

图 1 由单词、文档、主题构建的异构图

Fig. 1 Heterogeneous graph constructed from words, documents and topics

由于不同类型的节点之间特征值是不同的,所以文中对不同类型的节点分别进行卷积。研究时对不同类型的节点设为 $\emptyset=\{\varphi_d,\varphi_t\}$[13]。其中,$\varphi_d$ 表示文档和单词组成的节点类型,φ_t 表示主题节点的类型,可将同一节点类型的节点卷积后相加,对应数学运算公式可写为:

$$\boldsymbol{H}^{(l+1)}=\alpha\Big(\sum_{\varphi\in\emptyset}\widetilde{\boldsymbol{A}_{\varphi}}\cdot\boldsymbol{H}_{\varphi}^{(l)}\cdot\boldsymbol{W}_{\varphi}\Big). \tag{1}$$

2.3 图卷积分类

构建一个图 $\boldsymbol{G}(\boldsymbol{V},\boldsymbol{E})$,$\boldsymbol{V}$ 和 $\boldsymbol{E}$ 分别是节点和边的集合,设 $|V|=n$ 为节点个数,$X_i\in R^m$,其中,$i\in(1,n)$,m 是节点 i 的维度。引入节点 i 的邻接矩阵 $\boldsymbol{A}$ 以及度矩阵 $\boldsymbol{D}$,为了使每个节点卷积过程中不仅集合了邻域信息,还保留自身的信息,所以给邻接矩阵添加自环,将 $\boldsymbol{A}$ 矩阵的对角元素置 1,$\boldsymbol{A}$ 的数学表达可写为:

$$A_{ij}=\begin{cases}1, & i\in C,j\in C,PMI(i,j)>0;\\ 1, & i\in C,j\in D,C\in D;\\ 1, & i\in T,j\in D,D\ assigned\ to\ T;\\ 1, & i=j;\\ 0, & otherwise.\end{cases} \tag{2}$$

同时设置双层 GCN,其中第一层矩阵的特征值被更新为:

$$\boldsymbol{H}^0=\alpha(\widetilde{\boldsymbol{A}}\cdot\boldsymbol{X}\cdot\boldsymbol{W}_0), \tag{3}$$

其中 $\widetilde{\boldsymbol{A}}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$ 为归一化邻接矩阵;$\boldsymbol{W}_0$ 为权重矩阵;$\alpha(\cdot)$ 为激活函数。

可以获取第二层领域信息,即:

$$\boldsymbol{H}^{l+1}=\alpha(\widetilde{\boldsymbol{A}}\cdot\boldsymbol{H}^{l}\cdot\boldsymbol{W}_l). \tag{4}$$

其中,l 表示层数。

2.4 半监督分类

对于半监督分类,使用交叉熵来评估带有标签的文档,具体公式见如下:

$$L=-\sum_{l\in y_L}\sum_{f=1}^{F}Y_{lf}\ln Z_{lf}. \tag{5}$$

其中,y_L 是所有带标签节点的索引,使用梯度下降法可以更新 $\boldsymbol{W}_0$,$\boldsymbol{W}_1$ 参数矩阵;Y_{lf} 表示标注类别;Z_{df} 为预测的类别;F 是输出层的特征维数,等于类别的数量。由于本文研究是二分类问题,所以 F 等于 2。

2.5 注意力机制

在构建图的过程中,由于节点之间相关则有边,无关则没有边,但是对于某一节点,不同邻居节点对其影响是不同的,有些节点可能携带了更多有用的信息。为了区分不同邻居节点对于该节点的重要程度,本文引入了注意力机制,单词和单词之间的权重采用的是 $PMI(i,j)$ 的值表示,单词与文档之间的权重采用的是词频-逆文档频率,文档节点 i 与主题节点 j 之间的相关性将采用公式具体如下:

$$E_{ij}=\alpha([\boldsymbol{W}h_i\,\|\,\boldsymbol{W}h_j]), \tag{6}$$

其中,符号“‖”表示将节点 v_i、v_j 变换后的特征进行拼接,函数 $\alpha(\cdot)$ 作用是把拼接后的特征映射到一个实数上。假设一个特征为 F 的输入节点满足 $h_i\in R^F$,一个特征为 F' 的输出节点满足 $h'_j\in R^{F'}$,

则要对所有节点训练一个 $\boldsymbol{W} \in R^{F \times F'}$ 的权重矩阵，$\boldsymbol{W}$ 即输入与输出的关系[14]。

由于节点 v_i 只与其邻居节点有关联，所以研究中的注意力系数表达的是目标节点 v_i 与其邻居节点 $j \in N_i$ 之间的关系。为了便于计算和比较，文中采用了 softmax 函数对 v_i 与所有邻居节点的注意力系数进行正则化，最终能得到本次研究中的注意力机制为：

$$a_{ij} = \frac{exp(LeakyReLU(\alpha[\boldsymbol{W}h_i \parallel \boldsymbol{W}h_j]))}{\sum_{k \in N_i} exp(LeakyReLU(\alpha[\boldsymbol{W}h_i \parallel \boldsymbol{W}h_k]))}, \tag{7}$$

研究过程中，原先的邻接矩阵只是简单地将有关联的边置 1，加上注意力机制后文中的邻接矩阵变成了传播矩阵，定义为：

$$\widetilde{\boldsymbol{B}_\varphi} = \begin{cases} a_{ij}, & \text{节点 } v_i \text{、} v_j \text{ 之间有边,} \\ 0, & otherwise. \end{cases} \tag{8}$$

最后，加入了注意力机制的输出层为：

$$\boldsymbol{H}^0 = \alpha(\widetilde{\boldsymbol{A}} \cdot \boldsymbol{X} \cdot \boldsymbol{W}_0), \tag{9}$$

$$\boldsymbol{H}^{(l+1)} = \alpha(\sum_{\varphi \in \emptyset} \widetilde{\boldsymbol{B}_\varphi} \cdot \boldsymbol{H}_\varphi{}^{(l)} \cdot \boldsymbol{W}_\varphi). \tag{10}$$

3　实验与评估

3.1　参数设置

文中采用的数据集来自于 kaggle 网站的公开数据集 Womens Clothing E-Commerce Reviews，该数据集有 11 个字段，详见表 1。研究中选取了第五列的评价文本，以及第七列的文本标签用于实验。经统计在该数据集中，共有 23 486 条评论数据。其中有 19 314 个好评，以及 4 172 个差评，为了能够使好评和差评数量均衡，随机选取了 4 172 个好评，以及全部的差评进行了实验，共计 8 344 条文本数据。至此，单词结点数有 4 557 个，设置词嵌入维度为 200，主题数为 15。训练集输入 70%、即 5 840 条数据，测试集为 30%、即 2 504 条数据，窗口大小 20，对文本进行 200 个 *epoch* 的训练，如果损失函数超过 10 个 *epoch* 没有减少，就停止训练。学习率设置为 0.02，*dropout* 为 0.5。

3.2　准确度

模型分类的结果最终会被归为以下 4 类：

(1) *TP*：将正类预测为正类数。

(2) *TN*：将负类预测为负类数。

(3) *FP*：将负类预测为正类数。

(4) *FN*：将正类预测为负类数。

表 1　服装评价数据集字段

Tab. 1　Fields of clothing evaluation data set

编号	字段	字段含义
1	Id	编号
2	Clothing ID	产品唯一 ID
3	Age	用户年龄
4	Title	评论标题
5	Review Text	评论内容
6	Rating	评分
7	Recommended IND	是否推荐
8	Positive Feedback Count	正反馈计数
9	Division Name	产品码数(小码、均码、大码)
10	Department Name	产品大类别
11	Class Name	产品类别名称

研究时对模型的准确率做出评价，通过以下公式计算得到最终的准确率为 0.764 38，其中需用到的公式可写为：

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \tag{11}$$

3.3　实验结果

本文还使用了最常见的神经网络的分类模型 CNN 以及 RNN 对文本数据进行了分类，与本文的模型进行了对比，对比结果见表 2。可以看到本次研究结果是优于传统的文本分类的，这表明了本文提出的方法对半监督文本分类具有不错的分类效果。究其原因有以下三点：

(1) GCN 考虑了文档与文档之间、单词与单词之间、以及文档与主题之间的词共现关系。

(2) CNN 是将中心像素点与相邻像素点求均值来实现空间特征的提取，而 GCN 是利用图的拉普拉斯矩阵的特征值和特征向量来研究图的性质，通过聚合所有二阶领域的信息加权平均，通过图的边来传递节点的信息，使节点既保留了自身特征又聚合了邻居节点特征，将标签信息在图上传播。

(3) 本文引入的注意力机制使中心节点在聚合过程中关注到节点之间的关系的重要程度的影响，使得本文构建的模型更加适应情感分类。

表 2　不同算法准确度对比

Tab. 2　Comparison of accuracy of different algorithms

算法	准确度
CNN	0.751 31
RNN	0.757 42
GCN	0.764 38

实验通过改变第一层嵌入维度的大小，来观察

对模型的影响,结果如图 2 所示。该结果表明,随着嵌入维度的增加,本文模型分类准确度先增加后减少,这是由于一开始随着嵌入维度的增加,嵌入能更好地将标签信息传播到整个图中,而当到达峰值 200 维的时候,词向量的增加反而会影响分类的性能和速度。

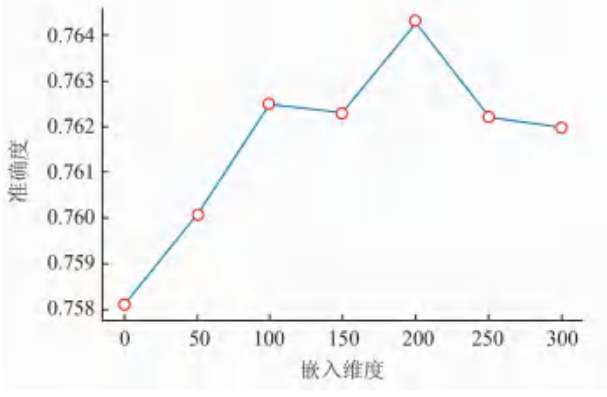

图 2 第一层嵌入维度的大小对模型的影响

Fig. 2 The impact of the size of the first – level embedding dimension on the model

实验通过改变不同比例的训练数据来观察该指标对模型的影响,如图 3 所示。由图 3 可以得出结论,准确度随着训练标签的增加而增加,但是同时还发现 GCN 在较少的训练数据的情况下,也能具有良好的性能。这是因为 GCN 是半监督分类,以及图结构可以很好地将标签信息传播到整个图中。

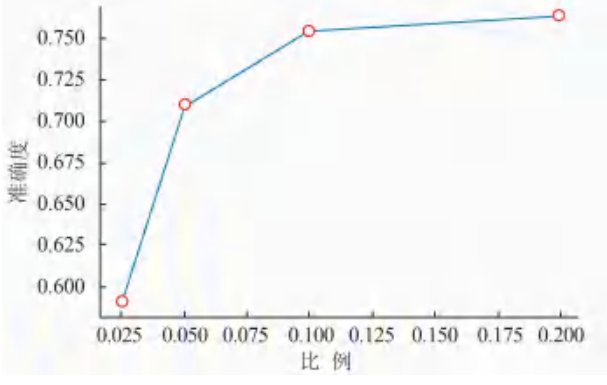

图 3 训练数据所占比例对模型的影响

Fig. 3 The influence of the proportion of training data on the model

实验通过改变滑动窗口大小对模型进行评估,如图 4 所示。图 4 表明随着窗口的增大准确度先增大,这是因为此时窗口的增大包含了更多的全局信息,但是到达峰值 15 窗口后,再增加只能为添加更多的无关节点增加新的边,所以准确度不再增长。

图 5 显示了主题数对模型准确性的影响,可以观察到,准确度现随着主题数的增加而增加,因为主题数很好地丰富了异构图表示的语义信息,当主题数到达 15 的时候,准确度随着主题数的增加而减少,证明过多的主题反而会影响图卷积分类模型的性能。

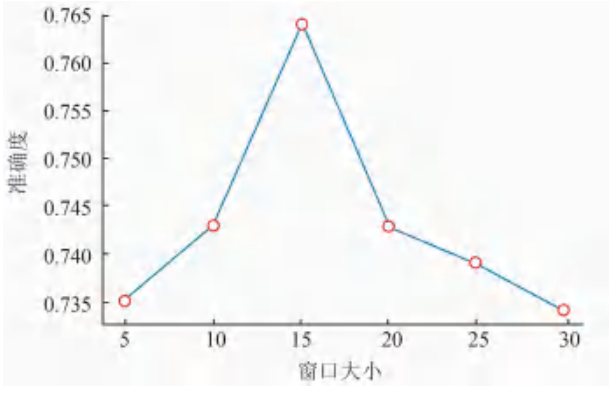

图 4 滑动窗口大小对模型的影响

Fig. 4 The effect of sliding window size on the model

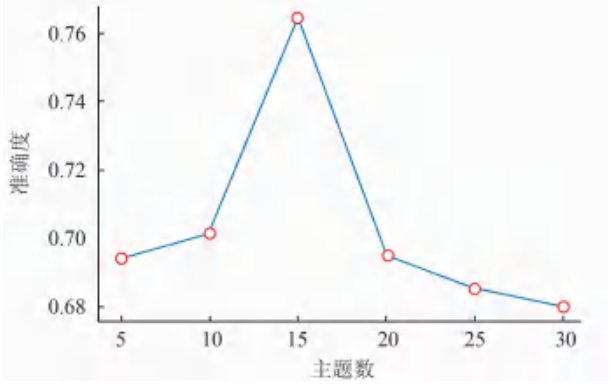

图 5 主题数对模型的影响

Fig. 5 The effect of the number of topics on the model

4 结束语

本文改进了图卷积文本分类的方法,为语料库构建基于单词、文档、主题的异构图网络,将文档分类转化成节点分类。并进行了实验,取得了不错的效果。该模型的研究在很大程度上丰富了异构图表达的语义信息,能很好地利用有限的标记文档,能有效实现语义信息在图上传播。因此,对服装评价信息进行正确分类一方面对电商平台制造更多高品质服装提供方向,另一方面对用户具有重要参考意义。所以图卷积文本分类具有较高的研究价值。

参考文献

[1] 高永兵, 王亮, 胡文江. 淘宝商品评价属性分类研究[J]. 微型机与应用, 2014, 33(11):8-11,15.

[2] 李宏媛, 陶然. 服装电商评论情感分析研究[J]. 智能计算机与应用, 2017, 7(1):27-30,34.

(下转第 45 页)