第43卷 第6A期
2016年6月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

基于 word2vec 和 SVMperf 的中文评论情感分类研究

张冬雯 杨鹏飞 许云峰

(河北科技大学信息科学与工程学院 石家庄 050018)

摘 要 利用有监督的机器学习的方法来对中文产品评论文本进行情感分类,该方法结合了 word2vec 和 SVMperf 两种工具。先由 word2vec 训练出语料中每个词语的词向量,通过计算相互之间的余弦距离来达到相似概念词语聚类的目的,通过相似特征聚类将高相似度领域词汇扩充到情感词典;再使用 word2vec 训练出词向量的高维度表示;然后采用主成分分析方法(PCA)对高维度向量进行降低维度处理,形成特征向量;最后使用两种方法抽取有效的情感特征,由 SVMperf 进行训练和预测,从而完成文本的情感分类。实验结果表明,采用相似概念聚类方法对词典进行扩充任务或情感分类任务都可以获得很好的效果。

关键词 情感分类,word2vec,SVMperf,语义特征,PCA

中图法分类号 TP391 **文献标识码** A

Research of Chinese Comments Sentiment Classification Based on Word2vec and SVMperf

ZHANG Dong-wen YANG Peng-fei XU Yun-feng

(School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China)

Abstract In this paper, we used the machine learning method to classify the sentiment classification of Chinese product reviews. The method combines SVMperf and word2vec. Word2vec trains out each word of the corpus of word vectors. By computing the cosine distance between each other, a similar concept word clustering is achieved, and with similar feature clustering, the vocabulary of the high similarity in the field is expanded to sentiment lexicon. The high dimensional representation of the word vector is trained out using word2vec. PCA principal component analysis method is used to reduce the dimension of the high dimensional vector, and the feature vector is formed. We used two different method to extract the effective affective feature, which is trained and predicted by SVMperf, so as to complete the sentiment classification of the text. The experimental results show that the method can obtain good results, regardless using the similar concept clustering method to expand the task or complete the emotional classification task.

Keywords Sentiment classification, Word2vec, SVMperf, Semantic features, PCA

1 研究背景

情感分类技术的任务是识别出用户在评论文本中流露出的情感信息,然后将其分为两类,即正类和负类。正类代表正面的赞赏和肯定,负类代表负面的批评和否定。常见的情感分类技术主要有两种方法:基于情感词典的无监督分类方法和基于机器学习的有监督分类方法。而基于机器学习方法的情感分类由于其更加出色的分类效果,得到了越来越多研究者的青睐,成为了主流方法。2002 年,Pang 等首次将机器学习的方法应用于情感分类领域。他们尝试使用 N-grams 模型提取出特征,并且分别利用机器学习领域的 3 种分类模型(SVM、NB 和 ME)进行了测试。他们发现选取 unigrams 作为特征集合,使用 SVM 进行分类,可以获得较好的分类效果。然而,Cui 等在 2006 年通过实验表明,当训练语料较少的时候,unigrams 确实可以获得很好的效果,

但随着语料库的增加,n-grams($n>3$)却表现出了更好的分类性能[1]。

如何抽取出复杂的特征而非简单特征以及识别出哪种类型的特征是有价值的,是机器学习方法的两个关键问题。近些年来出现了许多特征选择和提取的方法,包括 Single-Words 模型,Single-Character N-grams 模型,Multi-Word N-grams 模型和词汇_句法模型等。2011 年,Ahmed 等提出了基于规则的多元文本特征选择方法 FRN(Feature Relation Network)[2]。同年,Yao 等使用基于统计的机器学习的方法来选择特征,降低了特征向量的维度。Wang 等采用 DF(Document Frequency)、IG(Information Gain)、CHI(Chi-Squared Statistic)和 MI(Mutual Information)来选择特征[3,4]。然而以上这些方法都仅仅限于挖掘出句子中词与词之间的词汇特征和句法特征,语义特征却很少被研究。事实上,语义特征由于蕴含着词语之间更加隐含的信息,会对情感信息的识别起到更大的作用。

因此,为了提取出词与词之间的语义特征,本文使用了 word2vec[5-7] 这一工具,结合 SVMperf[8-10] 分类模型来对评论文本进行情感分类,以达到提升分类效果的目的。

张冬雯(1964—),女,博士,教授,主要研究方向为数据挖掘;杨鹏飞(1990—),男,硕士生,主要研究方向为数据挖掘,E-mail:yang2014spring@126.com;许云峰(1980—),男,副教授,硕士生导师,主要研究方向为云计算、大数据等。

2 基于 word2vec 扩充情感词典

在产品评论中，不同的消费者往往会使用不同的词语来描述同一个产品特征，更存在许多领域专属词汇，我们需要将这些相似概念聚合为同一产品特征簇[11,12]。当前比较前沿的相似概念聚类方法的准确率仍未达到实际应用的程度，由于本文主要采用情感词典提取分类特征，为了在不同领域搜索到更为全面的领域关键词，因此使用相似特征聚类的方法来寻找领域情感特征。

word2vec 是 Google 公司在 2013 年下半年发布的一款面向大众的开源 Deep Learning 学习工具。word2vec(word to vector)，顾名思义，主要目的是将单词转换成词向量，将单词映射到一个新的空间。它的核心模型就是典型的神经网络。它首先在训练文本语料库中构建一个词汇表。根据 CBOW 模型或 Skip-Gram 模型训练出每个词的词向量，然后通过计算向量之间的余弦相似度来获得文本语义上的相似度，因为余弦值不光包含着位置上的信息，也同样包含着一定的语义信息。word2vec 将对文本内容的处理简化为向量空间中的向量运算，本文利用这一功能进行了相似概念聚类的研究。

首先使用中科院计算技术研究所的 ICTCLAS 分词程序对所收集到的服装评论语料进行分词和词性标注。在完成去重和去除停用词等一系列预处理后，利用 word2vec 对语料进行训练，而后得到模型文件。word2vec 在 Linux 平台下的训练命令如下：

./word2vec -train train.txt -output vectors.bin -cbow 0 -size 200 -window 5 -negative 0 -hs 1 -sample 1e-3 -threads 12 -binary 1

以上命令表示的是输入文件是 train.txt，输出文件是 vectors.bin，不使用 CBOW 模型，默认为 Skip-Gram 模型。每个单词的向量维度是 200，训练的窗口大小为 5，不使用 NEG 方法，使用 HS 方法。-sample 指的是采样的阈值，如果一个词语在训练样本中出现的频率越高，就越会被采样。-binary 为 1 指的是训练结果用二进制存储，为 0 是普通存储。

训练完成后，就会得到 vectors.bin 这个模型文件。如果是用普通存储，打开文件后就会看到文档中词语和其对应的向量，向量的维度就是训练时设定的参数大小。word2vec 提供了一个 distance 命令来实现词语之间相似度的比较，继而达到聚类的目的。Distance 命令如下所示：

./distance vectors.bin

此命令的功能是读取模型文件中每一个词和其所对应的向量，计算所输入的词与其他所有词的余弦相似度，经过排序返回最终结果。值越大，则两个词在语义上越接近。

针对前期所收集的服装评论数据集，我们首先选取了价格、面料、尺码和款式 4 种在服装评论中最常出现也是用户评论最多的特征作为代表特征，来寻找与其语义类似的其他表述词语，组成同一产品特征簇。在使用 distance 命令得到每个代表特征词的相似词语列表后，只保留双音节名词，经过筛选和过滤，选取排名前 5 的词语作为最终的聚类结果。不同向量维度训练下的聚类结果如表 1 所列。

表 1 不同向量维度下的相似特征聚类结果

向量维度	代表特征(雏)	相似特征				
价格	200	价位	价钱	价码	价值	天价
	500	价位	价钱	价码	价值	天价
	1000	价位	价钱	价码	价值	天价
	5000	价位	价钱	价码	天价	价值
	10000	价位	价钱	价码	价值	天价
面料	200	料子	布料	质地	材质	材料
	500	料子	布料	材料	质地	材质
	1000	料子	布料	质地	材质	衣料
	5000	料子	布料	质地	材质	衣料
	10000	料子	布料	质地	材质	材料
尺码	200	尺寸	号码	型号	码号	码子
	500	尺寸	号码	码子	码号	型号
	1000	尺寸	号码	型号	码号	码子
	5000	尺寸	号码	型号	码号	码子
	10000	尺寸	号码	码子	码号	型号
款式	200	样式	款型	外观	样子	外形
	500	样式	款型	样子	外形	外观
	1000	样式	款型	外观	样子	外形
	5000	样式	款型	样子	外观	外形
	10000	样式	款型	样子	外观	式样

可以看到，在不同的向量维度下，利用 word2vec 进行相似特征聚类都可以得到很好的实验结果，只是若干特征词之间的排名次序互换而已，这并不影响聚类的准确度。根据该实验结果，对情感特征词库中原有词汇进行了扩充，使得领域专属词汇很全面地被扩充到了我们的情感词典中，之后再经过人工筛选，得到了全面的特征词库。

3 中文评论情感分类

本文采用基于机器学习的方法进行情感分类。将由 word2vec 训练出的词向量作为候选特征向量，采用基于情感词典的选择方法从候选集中筛选出有效特征。对于 SVM 分类模型，本文选择了 SVMperf 作为分类工具。

SVMperf 是 SVMlight 的作者 Joachims 在 SVMlight 的基础上采用更高级的内核算法得到的新型分类模型。SVMperf 相较于 SVMlight 有 3 点优势：分类速度更快、分类精度更高、适合大数据集。

3.1 有效特征的抽取

3.1.1 基于情感词典的方法

在得到同一产品特征簇后，继续使用 word2vec 对情感词典进行扩充。这里选取了中国知网情感词集和清华大学 IAR 情感词典作为两个原始情感词典，从中选择权重最大的若干个词作为输入词，以用 word2vec 中的 distance 命令获得与输入词语义相似的情感词，以达到相似特征聚类的目的；再将新得到的词扩充到原情感词典中，得到新的情感词典(这里的情感词典指的是广义上的情感词典，也包含否定词和程度副词等)。

接下来用 word2vec 训练语料库，因为需要生成 SVMperf 分类模型所支持的训练数据格式，也就是：

〈line〉.=.〈target〉〈feature〉:〈value〉〈feature〉:〈value〉…〈feature〉:〈value〉

〈target〉.=.{+1,-1}

〈feature〉.=.〈integer〉

〈value〉.=.〈float〉

逐行读取语料库中的每一条评论语句，判断其是否包含扩充过的情感词典中的情感词、否定词和程度副词；如果包

含，则依特征编号次序按 SVMperf 所支持的格式逐行写入数据，最终得到所需要的训练集。

3.1.2 基于词性选择的方法

在用 word2vec 对语料库进行训练时，默认对在语料库中出现次数小于 5 的词进行了剪枝。不同的词性选择方法会得到不同的分类结果[13]。例如，如果仅仅选择形容词作为训练特征，根据其得到的结果会比同时选择副词、动词以及形容词时得到的分类结果差，因为很多不同词性的词语都会带有情感特征。

与基于情感词典的特征选择类似，首先将语料库进行带有词性标注的分词，由 word2vec 训练得到模型文件，经过词性筛选，得到形容词、副词、动词、名词 4 种词性的词语和对应词向量，然后通过不同组合获得特征，按格式写入训练文件，得到训练集。

3.2 有效特征的表示

对于文本的处理，我们要将自然语言处理为机器可以识别的数字化模式。对于特征词的数字化处理，最常用的一种方法是：one-hot representation，就是把每个词表示为一个很长的向量值，其中的每一位表示一个词，总长度就是词表的大小，其中用 0 和 1 表示，1 代表存在，0 代表不在。

另一种常用的处理方式就是在深度学习中使用的 Distributed Representation。它最早是由 HInton 在 1986 年提出来的，这种方法弥补了 one-hot representation 的不足，可以使语义相近的词在距离上更加接近。可以尝试用余弦值来衡量向量之间的距离，也可以尝试用欧氏距离来衡量向量之间的距离。

word2vec 就是基于深度学习的一种词向量转换工具，它属于一种浅层次的神经网络。这种词向量的训练也是建立在语言模型上的。使用 K 维度的向量运算来对文本的内容进行处理，使用一个三层的神经网络来对语言模型进行建模，同时获得单词在向量空间的表示，使用这个表示来作为特征数学化数字表示。这样的表示方法同时可以揭示特征值之间的隐性语义。我们将所有的特征词通过匹配转换为向量值。

3.3 有效特征的维度降低

因为现有的分类模型不支持高纬度特征值作为输入，所以使用 PCA 方法对提取特征的结果进行降低维度的处理[18]。将高纬度词向量表示形式降低到一维度向量表示形式，并且尽可能地保留高纬度向量中所蕴涵的语义信息。为了尽可能降低语义信息的损失，选择生成多种高维度词向量的表示来测试语义损失程度。

利用 PCA 方法尽可能地将高维度特征中的噪音和冗余信息去除，使得保留的低维度向量地信息量尽量地大。PCA 方法虽然多用于图像处理中的特征降维，但是将它用于文本高维度降维也同样有着很好的效果。具体的 PCA 算法如下。

主要的降维步骤：

设 $X_1, X_2, \cdots, X_n$ 为原始变量，$F_1, F_2, \cdots, F_k$ 为 k 个主成分因子。

$$F_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{1n}X_n$$

方差 $Var(F_1)$ 和 F_1 包含的信息是成正比的。

$$\begin{cases} F_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{1n}X_n \\ F_2 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{2n}X_n \\ \quad \cdots \\ F_k = a_{k1}X_1 + a_{21}X_2 + \cdots + a_{kn}X_n \end{cases}$$

F_1 为第一主成分，F_i 和 F_j 不相关，$Cov(F_i, F_j) = 0$。

(1)原变量协方差矩阵的计算：

$$S_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{kj} - x_i)(x_{kj} - x_j), i, j = 1, 2, \cdots, p$$

(2)求特征值，正交化单位特征向量：

$$F_i = a_i^{\mathrm{T}} X_i$$

信息量大小：$\partial = \lambda_i / \sum_{i=1}^{m}\lambda_i$。

(3)选择主成分：

$$G(m) = \sum_{i=1}^{m}\lambda_i / \sum_{i=1}^{n}\lambda_i$$

主成分分析方法将原有的高维度特征向量投影到一个新的空间中，将原有坐标转换到了新的坐标系下面，这个新的坐标系就是原有特征向量最大线性无关向量组的特征值对应的空间的坐标系。将高维度特征向量维度降低后，原有高维度特征中所包含的语义会有一定的损失，但是在选择了最优特征维度后，这种降维方法在高维度特征向量降维中并没有太大的损失，显示出了主成分分析方法的优势。

3.4 模型的训练和分类

使用经过降低维度处理后的单维度特征向量来训练分类模型。大量的实验表明，与其它分类模型相比，支持向量机在分类中有着很好的性能[14,15]。

首先根据特征情感词典从训练文本中提取出特征词语，将其转换为词向量并经过降维处理后，输入到 SVMperf 分类器中，分类器根据训练集中的正负类标签和相应的特征值，生成分类模型[10]。

当分类模型生成后，我们就可以利用这个分类模型对数据进行正负情感分类。同样将需要分类的文本形成与训练文本相同格式的输入，然后使用之前训练好的分类模型进行正负极性的预测。

4 实验分析与讨论

4.1 实验数据

由于当前中文评论情感分类研究中的公开实验数据集还比较匮乏，而且 word2vec 这一工具需要大量的文本语料做支撑，训练语料库越大，它的训练效果越好。因此进行情感分类实验之前，需要先从互联网上爬取大量的用户评论。本文爬取的是中文亚马逊网站上的服装产品的用户评论，按照用户评论的星级进行了分类，4 星和 5 星评论作为正类，1 星和 2 星评论作为负类。实验中一共抓取了 100000 条评论文本，但是由于爬取的评论文本中正负类比例为 7:1，属于严重不平衡语料，同时也存在大量的超短评论文本，这会导致最终的分类结果出现偏差。因此为了保证分类的有效性，将实验数据集压缩为了 20000 条，10000 条正类和 10000 条负类作为本次实验的情感分类实验数据集。其中训练集和测试集的比例为 7:3。

4.2 实验结果

表 2 列出了基于情感词典选择特征的分类结果。在全局正确率这一指标上，HowNet 词典和 IAR 词典结合使用所表现出的性能略优于单独使用 IAR 词典的性能，而 HowNet 词典在所有指标上都表现出了最差的性能。这是因为 HowNet 词典中并没有包含可能会改变情感极性的否定词词典，因此会导致错误的分类。将两个词典结合作为特征，不仅包括了

必要的正面情感词和负面情感词，还包含了程度副词词典和否定词词典，既考虑了情感词本身的极性，又考虑了情感极性的反转，因此它可以获得最好的分类效果。

表 2　不同类型的特征选择方法的分类结果

特征	正/负类	准确率(%)	召回率(%)	F 值(%)	正确率(%)
HowNet 词典	正类	86.27	87.65	86.95	86.85
	负类	87.45	86.05	86.74	
IAR 词典	正类	91.36	86.70	88.97	89.25
	负类	87.35	91.80	89.52	
〈HowNet 词典，IAR 词典〉	正类	91.14	88.50	89.80	89.95
	负类	88.82	91.40	90.09	
〈形容词，副词〉	正类	87.21	82.85	84.97	85.35
	负类	83.67	87.85	85.71	
〈形容词，副词，动词〉	正类	91.38	89.00	90.17	90.30
	负类	89.28	91.60	90.42	
〈形容词，副词，名词〉	正类	91.91	85.15	88.40	88.83
	负类	86.17	92.50	89.22	
〈形容词，副词，动词，名词〉	正类	91.89	84.95	88.28	88.72
	负类	86.01	92.05	89.14	
所有频繁实词	正类	92.66	84.60	88.45	88.95
	负类	85.83	93.30	89.41	

表 2 给出了基于词性选择特征的分类结果。在 F 值和正确率这两项主要指标上，选择形容词、副词和动词作为特征的效果要比其它的词性选择策略效果更好。相反，选择形容词和副词在 4 项指标上的结果都是最差的。将所有频繁实词都看作特征获得了最高的正类准确率和负类召回率，但是其较低的正类召回率和负类准确率拉低了 F 值和整体的正确率。从实验结果可知，并不一定只有形容词才是情感词，其它词性的词也可能会带有情感信息。为了获得情感分类的最佳正确率，通过调整 SVMperf 的常规参数 C，分别对两组实验进行了性能上的优化。我们选取了两类方法中效果最好的两种特征选择方法来进行这组优化实验。实验结果如图 1 和图 2 所示。

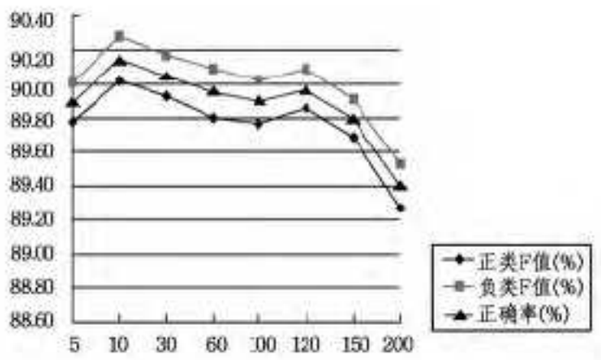

图 1　基于情感词典的特征选择方法在不同 C 值下的分类结果

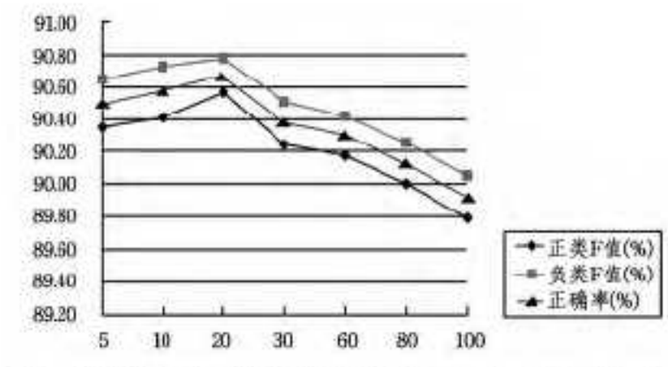

图 2　基于词性的特征选择方法在不同 C 值下的分类结果

可以看出，参数 C 的不同取值也会影响分类的效果。两种方法分别在 $C=10$ 和 $C=20$ 时，可以获得最佳的分类效果。

4.3　分析和讨论

从上述的实验数据中可以很明显地看出，借助于 word2vec 和 SVMperf 两项工具，无论是基于情感词典的特征选择方法还是基于词性选择的特征选择方法都可以获得很好的分类效果，分类正确率都在 90% 以上。在进行完一系列实验之后，本节将讨论本文的方法对于情感分类有效性的原因。

首先，由 word2vec 训练出的词向量可以抽取出语料中词语之间深层的语义关系，而并非简单的词汇特征和句法特征，这更有利于获取到句子中蕴含的情感信息；同时，输出的词向量表示是高维度的，使用主成分分析方法再对高维度向量进行降维度处理，满足了分类器的输入要求。其次，作为一个新型的线性分类 SVM 训练算法，SVMperf 在大数据集上拥有比包括 SVMlight 在内的其它 SVM 分类模型更快的分类速度和更准的分类效果。因此，基于这两项工具，本文提出的情感分类方法表现出了足够优越的性能。

结束语　不同于大多数传统的情感分类方法，本研究专注于如何抽取出词语间深层的语义特征而并非简单的词汇特征和句法特征。结合了 word2vec 和 SVMperf 两种工具来对中文评论文本进行情感分类。为了进行实验，从亚马逊上爬取了 100000 条服装产品的中文评论，经过数据筛选和清洗工作，选定了其中的 20000 条高质量评论文本作为本实验数据集。首先利用 word2vec 对相似概念进行了聚类，实验结果证明了 word2vec 工具在情感分析领域的可用性。而基于 word2vec 和 SVMperf 的情感分类方法的最好实验结果可以达到超过 90% 的正确率，这也证明了本方法在情感分类任务上可以获得很好的性能。

虽然本方法已经可以获得很好的情感分类效果，但是仍然还有进一步完善和提高的地方，未来依然有许多工作等待着去完成。为了符合 SVMperf 训练文件的格式要求，用 word2vec 训练词向量时，对词向量表示的维度进行了降低，这多少会使词向量空间中包含的信息产生损失。如何将高维度向量应用到 SVMperf 模型中是今后有待解决的一个难题。另外，我们所使用的两种特征选择方式还不足以抽取出句子中蕴含的所有情感信息，因此在今后的工作中，还需抽取出句子中有用的结构化信息或者组合评价单元作为特征来提高分类的准确度。

参 考 文 献

[1] 杨经，林世平. 基于 SVM 的文本词句情感分析[J]. 计算机应用与软件，2011，28(9)：225-228

[2] Raaijmakers S, Kraaij W. A shallow approach to subjectivity classification[C]// Proceedings of the Second International Conference on Weblogs and Social Media(CWSM). 2008：216-217

[3] Xia R, Zong C. Exploring the use of word relation features for sentiment classification[C]// Proceedings of the 23rd International Conference on Computational Linguistics(COLING). Beijing：ACL，2010，1336-1344

[4] Abbasi A, France S, Zhang Z, et al. Selecting attributes for sentiment classification using feature relation networks[J]. IEEE Transactions on Knowledge and Data Engineering(TKDE)，2011，23(3)：447-462

[5] Yao J, Wang H, Yin P. Sentiment feature identification from Chinese Online reviews[C]// Proceedings of 2011 International Conference on Computer Science and Education：Communications in Computer and Information Science. 2011：315-322

[6] Wang H, Yin P, Zheng L, et al. Sentiment classification of Online reviews：using sentence-based language model[J]. Journal of Experimental & Theoretical Artificial Intelligence，2014，26(1)：13-31

(下转第 447 页)

[3] Manzato M, Goularte R. A multimedia recommender system based on enriched user profiles[C]// Proc of the 27th Annual ACM Symp on Applied Comuputing. New York: ACM, 2012: 975-980

[4] Zhang Shao-zhong, Chen De-ren. Hybrid Graph Model with Two Layes for Personnalized Recommendation[J]. Journal of Software, 2009(20): 123-130

[5] Soo-Cheo K, Jung-Wan K, Jung-Sik C. Collaborative filtering recommender system based on social network[J]. IT Convergence and Services, 2011, 7(2): 503-510

[6] Zhao Z L, Wang C D, Wan Y Y, et al. Pipeline Item-Based Collaborative Filtering Based on MapReduce[C]// IEEE Fifth International Conference on Big Data & Cloud Computing. Dalian: 2015: 9-14

[7] Pavlov D, Pennock D. A maximum entropy approach to collaborative filtering in dynamic, sparse, high-dimensional domains[C]// Proc of the 16th Annual Conf on Neural Information Proces-sing Systems. 2002

[8] Getoor L, Sahami M. Using probabilistic relational models for collaborative filtering[C]// Proc of the Workshop Web Usage Analysis and User Profiling under KDD'99. San Diego. 1999

[9] Unger L H. Foster D P. Clustering methods for collaborative filtering[C]// Proc of the Workshop on Recommendation Systems. Menlo Park: AAAI Press, 1998: 112-125

[10] Chien Y H, George E I. A Bayesian model for collaborative filtering[C]// Proc of the 7th Int'l Workshop on Artificial Intelligence and Statistics. San Francisco: Morgan Kanfmarm, 1999

[11] Massa P, Avesani P. Trust-aware collaborative filtering for recommender systems[C]// Proc of Fourth International Conference on Trust Management. Pisa, Italy, 2006

[12] Mark Newman. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 69, 066133, Jun. 2004

[13] Du Jing-fei, Lai Jiang-yang, Shi Chuan. Multi-objective Optimization for Overlapping Community Detection[J]. Advanced Data Mining and Applications, 2013: 489-500

[14] Newman M, Leicht E. Mixture models and exploratory analysis in networks. [J]. Proc. Natl. Acad. Sci, 2007, 104(23): 9564-9569

[15] Yu Le, Wu Bin, Wang Bai. LBLP: Link-Clustering-Based Approach for Overlapping Community Detection[J]. Tsinghua Science and Technology, 2013, 18(4): 387-397

[16] Martin Rosvall and Carl Bergstrom. Maps of random walks on complex networks reveal community structure[J]. Proc. Natl. Acad. Sci., 2008, 105(4): 1118-1123

[17] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435: 814-818

[18] Shen Hua-wei, Cheng Xue-qi, Cai Kai, et al. Detect overlapping and hierarchical community structure[J]. Physica A, 2008, 388 (8): 1706-1717

[19] Zhu Xiao-jin, Ghahramani Zou-bin. Learning from Labeled and Unlabeled Data with Label Propagation[R]. Techinicalreport, CMU CALD tech report CMU-CALD-02, 2002

[20] Gregory S. Finding overlapping communities in networks by label propagation[J]. New J. Phys., 2010, 12: 103018

[21] Gower J C. A General Coefficient of Similarity and Some of Its Properties[J]. International Biometric Society, 1971, 27(4): 857-871

[22] Blondel V, Guillaue J L. Renaud Lambiotte and Etienne Lefebvre. Fast unfolding communities in large networks[J]. Journal of Statistical: Mechanics Theory & Experiment, 2008, 30(2): 155-168

[23] Sarwar B, Karypis G, Konstan J, Riedl J. Analysis of recommendation algorithm for e-commerce[C]// Proc of the 2nd ACM Conference on Electronic commerce. New York: 2000: 158-167

[24] Ziegler C, McNee S, Konstan J. Improving recommendation lists through topic diversification[C]// Proc of the 14th International Conference on World Wide Web. New York: 2005: 22-32

[25] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithm[C]// Procedings of the 10th International Conference on World Wide Web. New York: ACM Press, 2001: 285-295

(上接第 421 页)

[7] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[C]// Proceedings of Workshop at ICLR. 2013

[8] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]// Proceedings of NIPS. 2013

[9] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations[C]// Proceedings of NAACL HLT. 2013

[10] Joachims T. Training linear SVMs in linear time [C]// Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD). 2006

[11] Joachims T. A support vector method for multivariate performance measures[C]// Proceedings of the International Conference on Machine Learning(ICML). 2005

[12] Liu B, Zhang L. A survey of opinion mining and sentiment analysis[J]. Synthesis Lectures on Human Language Technologies, 2010, 2: 459-526

[13] Tang H, Tan S, Cheng X. A survey on sentiment detection of reviews[J]. Expert Systems with Applications, 2009: 10760-10773

[14] Joachims T, Yu C. Sparse kernel SVMs via Cutting-Plane training[M]// Machine Learning and Knowledge Discovery in Dtabases. Springer. Berlin Heidelbergy. 2009

[15] Tang H, Tan S, Cheng X. A survey on sentiment detection of reviews[J]. Expert Systems with Applications, 2009, 36: 10760-10773

[16] Zhai Z, Liu B, Xu H, et al. Grouping product features using semi-supervised learning with soft-constraints[C]// Proceedings of the 23rd International Conference on Computational Linguistics (COLING). Beijing: ACL, 2010: 1272-1280

[17] Zhai Z, Liu B, Xu H, et al. Clustering product features for opinion mining[C]// Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM). Hong Kong: ACM, 2011: 347-354

[18] Jose C. A Fast On-line Algorithm for PCA and Its Convergence Characteristics [J]. IEEE Transactions on Neural Network, 2000, 4(2): 299-307