文章编号：2096-1472(2019)-01-01-06
DOI:10.19644/j.cnki.issn2096-1472.2019.01.001

一种面向商品评价对象挖掘的领域词典构建法

石玉鑫，杨泽青，赵志滨，姚 兰
(东北大学计算机科学与工程学院，辽宁 沈阳 110819)

摘 要：通过挖掘商品评论中的评价对象，可以得知用户更关心商品哪些方面的属性，从而帮助企业改进商品，帮助用户选择商品。因此，商品评价对象的挖掘具有重要的意义。本文提出了一种用于商品评价对象挖掘的领域词典构建方法：首先基于LDA模型，提出了一种领域基础词典的构建方法；然后，分别提出了基于词汇之间的PMI值和基于依存句法分析的领域词典扩充方法。本文基于京东商城的洗衣液产品真实评论数据集，使用构建的词典分别进行了一级标签评价对象挖掘和二级标签评价对象挖掘的实验。实验结果表明，本文提出的方法在进行评价对象挖掘时具有良好的性能；相比一级标签评价对象，扩充后的词典对二级标签评价对象挖掘的效果有更好的提升。

关键词：领域词典；对象挖掘；商品评论；LDA；PMI
中图分类号：TP391 **文献标识码**：A

A Method on Domain Dictionary Construction for Object Mining on Commodity Comments

SHI Yuxin,YANG Zeqing,ZHAO Zhibin,YAO Lan
(*School of Computer Science and Engineering*,*Northeastern University*,*Shenyang* 110819,*China*)

Abstract: Enterprises hope to be aided by object mining on comments of their products,which reveals the clients' concerns,to improve their manufacturing.This object mining also makes sense to subsequent consumers while they are making their choice.Therefore,it is significant to mine objects of a comment.This paper proposes a method on domain dictionary construction for object mining on comments of commodity:Firstly,a method based on the LDA model,a basic domain dictionary is proposed;then,the domain dictionary expansion methods based on the PMI value of words and dependency parsing are proposed respectively.Data applied for experiments in this paper is from detergent sale data of JD.COM.The dictionaries are applied on this data set for the first-level and second-level label object mining.The experimental results prove the proposed method’s great potential in object mining.Compared with the first-level label object mining,the extensive dictionary has improved the second-level label object mining.

Keywords:domain dictionary;object mining;commodity comment;LDA;PMI

1 引言(Introduction)

在互联网中，有海量的商品评论文本。这些评论可能来自于不同的电商平台和不同的商品品类，是一种重要的资源，具有很高的研究价值。通过分析电商平台的商品评论，市场调查工作人员可以得知用户更关心商品哪些方面的属性，以及用户对这些属性持有消极的观点还是积极的观点，从而帮助公司更好地改进产品；消费者也可以通过查看这些商品评论来了解其他人的真实购物体验，有助于快速找到口碑良好的商品，做出更好的购物选择。

电商平台的商品评论是中文短文本，面向商品评论的口碑分析的基础工作是挖掘出评论所描述的商品属性，即短文本的评价对象挖掘。正因为海量的评论数据中蕴藏着非常有价值的商业信息，因此面向商品评论的评价对象挖掘备受关注。目前为止，基于领域词典的规则匹配方法是评价对象挖掘的最有效手段之一，业界普遍采用，构建领域词典是其中的关键工作内容。但是，人工构建词典的方法工作量巨大，并且难以保证词典的覆盖性，因此亟需一种有效的方法来自动构建领域词典。

针对这一问题，本文提出了一种基于隐狄利克雷分布(Latent Dirichlet Allocation，简称LDA)模型、点互信息

基金项目：本文受the National Key R&D Program of China under grant(2018YFB1004700)资助.

(Pointwise Mutual Information，简称PMI)和依存句法分析的面向商品评价对象挖掘的领域词典构建方法，目标是针对某个品类的商品评论，构建领域词典，并利用领域词典实现对该品类文本的评价对象挖掘。本文构建的领域词典包括两部分，一部分是领域基础词典，由单个的词汇构成；另一部分是领域词典的扩充，由词汇的搭配组合构成。本文的主要贡献包括：

(1)提出了构建领域基础词典的方法。将已标注的训练集按标签分为若干个文档，使用LDA模型得到每个文档中主题的概率分布，以及每个主题中词汇的概率分布，提取出主题词，从而得到该标签下的词典。对每个标签对应的文档重复上述过程，就得到了领域基础词典。

(2)基于PMI扩充领域词典。通过计算点互信息(PMI)来衡量每个文档中词汇之间的相关性，将相关性高的词汇作为词组加入每个标签对应的词组集合，得到所有标签对应的词组集合。用词组集合对领域基础词典中每个标签下的词典进行扩充，构建扩充后的领域词典。

(3)基于依存句法分析扩充领域词典。本文定义了一种新形式的词典：句法词典。通过对已标注的语料进行句法分析，可以得到一个由词组构成的句法词典；利用该词典可以对领域词典进行进一步的扩充。

本文按照如下方式组织全文。第二部分总结了近些年的评价对象挖掘、词典构建的研究成果和相关技术；第三部分明确了本文要解决的问题，并且定义了相关符号；第四部分介绍了基于LDA模型构建领域基础词典和基于PMI、依存句法分析扩充词典的具体过程；第五部分通过评价对象挖掘实验，对本文所提出方法的性能进行了评估。第六部分总结了本文的工作，并提出未来可继续改进的地方。

2 相关工作(Related work)

本文工作的核心是构建面向商品评价对象挖掘的领域词典，需要用到文本挖掘的相关技术来构建词典。现在就文本挖掘技术的最新应用，以及有关词典构建工作的最新研究成果进行总结。

文本挖掘是一个从大规模的文本数据集合中挖掘出潜在且有价值的信息的过程[1]。随着互联网的发展，网络文本数据大量涌现，这使得文本信息挖掘成为多个领域的重点研究课题。文本挖掘的主要方法有基于主题模型的方法、基于机器学习的方法、基于句法分析的方法和基于词典的方法等。Pavlinek[2]等人提出了一种基于半监督学习和LDA主题模型的文本分类方法，对文本进行分类。He[3]等人提出了一种基于依存句法分析的评论观点挖掘方法，可以有效地从评论中挖掘观点。Tomas[4]等人在Spark中实现了朴素贝叶斯、随机森林、决策树、支持向量机和Logistic回归分类器等五种分类器，并对每种分类器的分类准确度进行了评估。Mandal[5]提出了一种基于词典进行意见挖掘并计算情感极性水平的算法。在这几种文本挖掘方法中，基于词典的规则匹配方法是最有效的手段之一，并且可维护性较好，在工程上普遍采用。因此，本文要构建面向商品评价对象挖掘的领域词典。

关于领域词典的构建，有很多可行的方法，相关研究也有很多。尹文科[6]等人基于维基百科链接结构图，结合LSI算法和CPMw算法，提出了一种构建领域词典的方法，实现了领域词典的自动构建。基于大量的商品评论文本，李伟卿[7]等人提出了一种构建产品特征词典的方法。该方法在大量已标注文本数据的基础上，基于同义词词林扩展版和Word2Vec工具进行词向量训练，计算词汇的语义相似程度，对特征词汇进行总结，从而构建产品的特征词典。与其他方法相比，该方法有良好的召回率。Chen[8]等人提出了一种新颖的词典构建方法，这种方法能够使词典包含更多的长尾关键词，从而提高词典的质量。文献[9]介绍了4种构建领域情感词典的方法，并评估了每种方法所构建词典的性能。Wu[10]等人基于已标注的文本数据，利用TF−IDF算法和Word2Vec工具，构建了足球领域的情感词典。Alqasemi[11]等人基于KNN查询算法构建了观点词库，并取得了较好的实验结果。Ju[12]等人提出了一种基于条件随机场的迭代机器学习算法，目标是自动构建中文临床语料库中的症状词典。文献[13]研究了国内外几种词典系统的功能，建立了一个领域词典构建系统，并设计了总体框架和组件模块。Zhang[14]等人通过提取和构建程度副词词典、网络词典、负面词典和其他相关词典来扩展情感词典。Song[15]等人提出了一个命名实体词典半自动构建系统，该系统基于维基百科，使用主动学习技术和BM25算法，在命名实体识别实验中表现出良好的性能。文献[16]中设计了一种关系词词典的新结构，采用弱监督方法找到词典项，并填充到关系词词典中。该词典用于提取生物医学文献中有关蛋白质的词汇。文献[17]提出了一种自动构建情感词典的方法，构建的词典用于处理特定领域的情感分析任务。文章中还比较了来自不同领域的情感词典的效率。Wu[18]等人提出了一种基于数据驱动的方法，来为微博情绪分析系统构建高质量的情感词典。针对现有中文情感词汇覆盖率较低的问题，Liu[19]等人通过整合当前情感词汇，构建了一个微博情感词典。

3 问题描述(Problem description)

商品评论的评价对象挖掘是一个多标签分类问题。表1是京东商城洗衣液产品评论中的两条评论，以及它们的评价对象。评论t_1的评价对象是这款洗衣液的气味和物流/送货速度，评论t_2的评价对象是洗衣液的清洁效果，浓度和物流/送货速度。从这两条评论可以看出，“气味”“清洁效果”“浓度”和“物流/送货速度”等属性都有可能成为洗衣液产品评论中所包含的评价对象，而类似于“口感”等属性不大可能成为正常的洗衣液评论中所提及的评价对象。因此，单个领域是具有封闭性的，评论中可能涉及的评价对象数量是有限的，这些评价对象可以穷举出来。因此，基于词典的多标签分类方法能够在商品评论的评价对象挖掘工作中取得较好的效果。本文要解决的问题是，生成一个用于挖掘

商品评价对象的领域词典。

表1 产品的评论和评价对象

Tab.1 Commodity comments and evaluation objects

编号	评论	评价对象
t_1	自然清香，味道挺好闻的，物流给力！！！	气味、物流/送货速度
t_2	洗衣效果很差劲，浓度一般！但是洗衣液到的很快。	清洁效果、浓度、物流/送货速度

本文使用集合$T=\{t_1,t_2,...,t_{|T|}\}$来表示商品品类$B$的一组中文短文本集合，用集合$A=\{a_1,a_2,...,a_m\}$来表示集合$T$中可能涉及的$m$种评价对象。若商品品类$B$是洗衣液产品，则集合$A$就是洗衣液产品本身，以及外延性质的总集。

通过对关键词或词组的匹配，可以确定评论中包含了哪些评价对象。例如，关键词“清香”对应的评价对象是“气味”，关键词“洗衣效果”对应的评价对象是“清洁效果”。因此，挖掘商品评价对象的领域词典D中需要包含每个评价对象所对应的关键词集合。领域词典D可形式化表示为式(1)。

$$D=\{(a_i,WS_i)\,|\,1\le i\le m\} \tag{1}$$

其中，WS_i是评价对象a_i所对应的关键词集合，其中的元素有可能是单个词汇，也有可能是多个词汇组成的词组。

因此，本文的目标是，找到领域词典构建函数F，基于商品品类B的文本集合T，构建领域词典D。可以形式化描述为：$F:T\to D$。

4 算法描述(Algorithm description)

4.1 构建领域基础词典

首先需要对商品评论集合$T=\{t_1,t_2,...,t_{|T|}\}$进行人工标注。每条评论需要标注出其包含的评价对象，以及描述这些评价对象的文本；标注出的评价对象可能是一个，也可能是多个。标注后的任一文本 都对应一个标签集合$A_i=\{(a_{ij},c_{ij})\,|\,a_{ij}\in A,c_{ij}\in t_i\}$。标注的示例如表2所示，该文本标注了四个标签，分别是“品牌忠诚度”“洗涤效果”“价格”“物流/送货速度”等四个评价对象，以及描述它们的文本。

表2 标注示例

Tab.2 An example of labeling

示例文本	标签
一直都用立白洗衣液，衣服洗得很干净，但是这次双11的促销力度不大，快递也慢，下次再有更大的优惠，多买点儿屯着。	(品牌忠诚度，一直都用立白) (洗涤效果，衣服洗得很干净) (价格，促销力度不大) (物流/送货速度，快递也慢)

标注完成之后，需要对标注的文本进行分词，去除停用词，并将文本分为$T_1,T_2,...,T_m$等m个集合，分别是包含评价对象$a_1,a_2,...,a_m$的文本集合，任意两个集合之间都可能有交集。

本文基于LDA模型来构建领域基础词典。LDA模型是一种文档主题生成模型。在LDA模型中，一个文档以一定概率选择了一个主题，一个主题又以一定的概率选择了一个词汇，形式化表示为式(2)：

$$p(\text{word}|\text{document})=\sum_{\text{topic}} p(\text{word}|\text{topic})\times p(\text{topic}|\text{document}) \tag{2}$$

首先，要给出LDA模型的主题数。之后，将描述评价对象a_i的文本集合T_i作为一个文档，通过LDA模型对该文档的学习，可以得到该文档的文档−主题分布和主题−词汇分布，从而可以得到评价对象a_i的主题词语，这些主题词语的集合记作D_i。通过对所有文档$T_1,T_2,...,T_m$重复上述过程，就可以得到集合$D_1,D_2,...,D_m$。这些集合就构成了领域B的领域基础词典$D_{LDA}=\{(a_i,D_i)\,|\,1\le i\le m\}$。

4.2 基于PMI扩充词典

基于LDA模型构建的领域基础词典D_{LDA}只包含单个的词汇，且词汇之间都是相互独立的，不存在搭配关系。然而，如果要挖掘细粒度的评价对象，有时两个单独的词汇并不能挖掘出某个评价对象，但是它们作为词组时却可以挖掘出这个评价对象。例如，在洗衣液评论中，我们可以将“气味”这一评价对象拆分为“打开时的气味”“洗衣时的气味”“晾衣时的气味”等若干个更细粒度的评价对象。评论“打开盖子时很香，很好闻”显然包含了“打开时的气味”这一评价对象，而无论是词汇“打开”，还是词汇“香”，单独拿出来都无法挖掘出“打开时的气味”这一评价对象，而它们搭配起来却可以挖掘出这个评价对象。因此，我们需要对上一小节中得到的领域基础词典进行扩充，在词典中加入词组做关键词。

本文通过计算点互信息(PMI)来衡量两个词语之间的关联程度，从而抽取出关联程度较高的词汇组合，用这些词组对领域基础词典进行扩充。PMI从统计学的角度来衡量词语之间的语义关联程度。针对某文本集合T_i中的词汇w_j和w_k，若这两个词汇出现在同一条商品评论中，则称词汇w_j和w_k共现。w_j和w_k在T_i中的共现概率可表示为式(3)。

$$p_i(w_j,w_k)=\frac{n}{|T_i|} \tag{3}$$

其中，n是w_j和w_k共现的评论数量。w_j和w_k在T_i中的PMI值可由式(4)计算出来，其中$p(w_j)$和$p(w_k)$分别是w_j和w_k在T_i中的频率。

$$PMI_i(w_j,w_k)=\frac{p(w_j,w_k)}{p(w_j)\cdot p(w_k)} \tag{4}$$

当$PMI_i(w_j,w_k)$大于一定阈值时，可以认为集合T_i中的词汇w_j和w_k具有搭配关系，并且该搭配关系$r=(w_j,w_k)$可以描述评价对象a_i。将符合上述条件的词组构成集合$R_i=\{r_1,r_2,...,r_{|R_i|}\}$，其中任一元素都是由一对词汇构成的具有搭配关系的词组。集合R_i就是描述评价对象a_i的词组集合。对所有评价对象$a_1,a_2,...,a_m$重复上述过程，最终得到集合$D_{PMI}=\{(a_i,R_i)\,|\,1\le i\le m\}$。集合$D_{PMI}$可以对领域基础词典$D_{LDA}$进行扩充，从而得到新的领域词典。

4.3 基于依存句法分析扩充词典

除了基于PMI对领域词典进行扩充之外，还可以基于依

存句法分析对领域词典进行扩充。依存句法分析是通过分析某个句子来构建该句子的依存句法树，从而描述句子之间的依存关系。利用哈工大“语言技术平台(LTP)”得到的依存句法分析实例如图1所示。

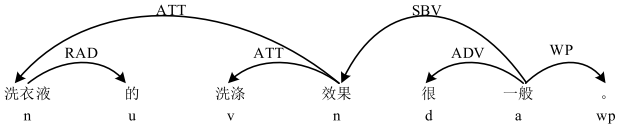

图1 依存句法分析实例

Fig.1 An example of dependency parsing

图1中的有向弧被称为依存弧，表示两个词之间存在从属关系。每个依存弧上都有一个标注，表示两个词之间的依存关系类型，每个词汇下方标注了它的词性。例如，“很”与“一般”之间存在依存关系ADV(状中结构)，“很”是程度副词，修饰形容词“一般”。“一般”是这对关系中的核心词，也叫支配词；“很”是用来修饰支配词的词语，也叫从属词。类似于“很”和“好”这样的词对，本文将其称为“依存词对”，其形式化定义如下：

定义1(依存词对)：存在依存关系的两个词语称为依存词对，形式化表示为式(5)：

$$WordPair(w_i,w_j)=(id_i,w_i,pos_i,id_j,w_j,pos_j,relation) \tag{5}$$

其中，id_i是从属词w_i的词号，即该从属词在句子中的位置；pos_i是w_i的词性；而id_j和pos_j分别是支配词w_j的词号和词性；relation是词汇w_i和w_j的依存关系类型。例如，图2中的“很”和“一般”就可以称为一个依存词对，可以形式化表示为：WordPair(很,一般)=(5,很,d,6,一般,a,ADV)。

在文本集合T_i中，某种词性组合的依存词对可能较为频繁的出现。以洗衣液产品的评论为例，评论中出现了“洁净衣领”“祛除异味”等关于产品功效的描述，均为“动词+名词”形式的依存词对。同时，多个依存词对的组合可能也会频繁出现，例如，短语“祛除顽固污渍”为“动词+形容词+名词”的形式，其中也包含“动词+名词”形式的依存词对和“形容词+名词”形式的依存词对。对于某个文本集合中类似于“动词+名词”“动词+形容词+名词”等包含一个或多个依存词对的频繁出现的词汇集合，本文称为“句法模板”，形式化定义如下：

定义2(句法模板)：在文本集合T_i中，存在文本t_l，包含词性为$\{pos_1,pos_2,...,pos_n\}$的词汇集合$W_l=\{w_1,w_2,...,w_n\}(n\geq 1)$，且对于集合$W_l$中的任意词汇$w_j$，至少存在一个词汇$w_k\in W_l$，与其存在依存关系，构成依存词对$WordPair(w_j,w_k)$或$WordPair(w_k,w_j)$。假设与$t_l$具有上述相同性质的文本集合为$T_e$，$T_e$中文本数量占$T_i$中文本数量的比例大于一个给定的阈值$\theta$，则称元组$e=(pos_1,pos_2,...,pos_n)$为文本集合$T_i$的一个句法模板，每个符合该句法模板的词组都是句法模板e的一个实例。

根据句法模板的定义，本文又给出了一种新形式词典的定义——句法词典，其形式化定义如下。

定义3(句法词典)：在文本集合T_i中，有句法模板集合$e_1,e_2,...,e_n$，其中任意一个句法模板e_j均存在描述评价对象a_i的词组集合$W(e_j,a_i)$，则这些集合可以构成一个新的集合$D_{T_i}=\{W(e_j,a_i)|1\leq j\leq n\}$。集合$D_{T_i}$就是文本集合$T_i$的一个句法词典

如果对每个文本集合$T_1,T_2,...,T_m$都构建句法词典，就可以得到文本集合T的一个句法词典$D_{DP}=\{(a_i,D_{T_i})|1\leq i\leq m\}$。为了提高词典的质量，在构建句法词典$D_{DP}$之前，需要计算文本集合$T$中每个词汇的TF−IDF值。TF−IDF是用来评估一个词汇对于一个文档重要程度的指标，TF指的是某一个给定的词语在该文档中出现的频率；IDF是逆向文档频率，是一个词语普遍重要性的度量。

将T看作一个文档，从微博上抓取一定数量的文本$WE_1,WE_2,...,WE_n$，将每条微博看作一个文档，与T组成文本集合$WE=\{T,WE_1,WE_2,...,WE_n\}$。对于词汇$w_i\in T$，它对于$T$的TF值和IDF值计算方式分别如式(6)和式(7)所示。

$$tf_i=\frac{s_i}{\sum_k s_k} \tag{6}$$

$$idf_i=\frac{|WE|}{|\{WE_j\,|\,w_i\in WE_j\}|+1} \tag{7}$$

其中，s_i是词汇w_i在本文集合T中出现的次数，$\{WE_j\,|\,w_i\in WE_j\}$是包含词汇w_i的微博文本集合。词汇w_i对于文本T的TF−IDF值计算方法如式(8)所示。

$$tfidf_i=tf_i\cdot idf_i \tag{8}$$

根据词汇的TF−IDF值，可以构建一个重要词汇词典$D_{imp}=\{w_i\,|\,tfidf_i>\theta'\}$，其中$\theta'$是一个阈值，TF−IDF值大于$\theta'$的词汇均可看作商品品类$B$的重要词汇。

根据上述定义，构造T_i的句法词典。从T_i中抽取出句法模板集合$e=\{e_1,e_2,...,e_n\}$。针对任一句子t_j中符合句法模板e_j的词组W_k=$\{w_1,w_2,...,w_n\}$，若词组满足以下两个条件之一的，即可加入词组集合$W(e_j,a_i)$：

(1)存在词汇$w_p\in W_k$，有$w_p\in D_{imp}$，且对于t_j中标注出的描述评价对象a_i的文本c_i，有$w_p\in c_i$。

(2)t_j中包含描述评价对象a_i的文本c_i，对于W_k中的任一词汇w_p，均有$w_p\in c_i$。

对T_i中所有句法模板的所有实例重复上述步骤，即可得到集合$D_{T_i}=\{W(e_j,a_i)|1\leq j\leq n\}$。用同样的方法也可以得到集合$D_{T_1},D_{T_2},...,D_{T_m}$，从而得到最终的句法词典$D_{DP}=\{(a_i,\mathrm{D}_{T_i})|1\leq i\leq m\}$。句法词典可以对领域词典进行扩充，从而得到新的领域词典。

5 实验(Experiment)

5.1 实验数据集

本文的实验数据集是京东商城洗衣液评论数据集。根据从领域专家处得到的洗衣液产品的特征码表，本文首先列出了“方便性”“品牌”“包装”“产品”“价格”“香味”“快递”“购物渠道”“产品功效”等9种评价对象，本

文称这9种评价对象为一级标签评价对象；并将每个一级标签评价对象再细分为更加细粒度的评价对象，例如“快递”可以细分为“快递(笼统)”“快递速度”“快递人员服务态度”“快递包装”等，细分完成后共有69种细粒度的评价对象，本文称这69个评价对象为二级标签评价对象。

由于实际获取到的商品评论随意性较大，会出现少量无效的评论，例如只出现标点符号的评论，或类似于“呵呵哈哈哈”这样无意义的评论，所以在进行数据预处理前需要剔除这些无效评论。剔除无效评论后，剩余的用户评论共计32400条。之后对所有有效的数据进行标注，标注内容包括每个评论所包含的一级标签评价对象、二级标签评价对象，以及每个评价对象所对应的文本。评价对象的标注是多标签标注，即一条短文本可以包含多个评价对象。由于人工标注难免有疏漏，所以对标注结果进行了细致的检查，并对百分之一的数据进行了重复标注。标注完成后，将每条评论进行分词，并剔除相应的停用词。

本文工作均采用Python 3.5语言实现，使用PyCharm开发工具，操作系统为Windows 7。洗衣液评论数据采用MongoDB数据库存储。

5.2 实验结果

本文提出了一种面向商品评价对象挖掘的领域词典构建方法，该方法可分为三部分：基于LDA模型构建领域基础词典的方法；基于PMI扩充领域词典的方法；基于依存句法分析扩充领域词典的方法。首先，使用29160条已标注的数据构建领域词典；之后，用剩余的3240条数据进行商品评价对象挖掘实验，来验证所构建领域词典的性能。

由于评价对象挖掘是一个多标签分类的过程，所以本文使用Macro-averaging评价指标来对评价对象挖掘实验的结果进行评估。Macro-averaging指标首先对各类的分类结果进行评估，然后再取所有类评估结果的均值作为整体的评估结果。Macro-averaging由三个具体指标构成：*Macro_P*,*Macro_R*和*Macro_F*，计算方法如式(9)、式(10)和式(11)所示，TP_i是实际包含评价对象a_i，预测结果也包含a_i的评论数；FP_i是实际不包含评价对象a_i，但预测结果却包含a_i的评论数；FN_i是实际包含评价对象a_i，预测结果却不包含a_i的评论数。

$$Macro\_P=\frac{1}{m}\sum_{i=1}^{m}\frac{TP_i}{TP_i+FP_i} \tag{9}$$

$$Macro\_R=\frac{1}{m}\sum_{i=1}^{m}\frac{TP_i}{TP_i+FN_i} \tag{10}$$

$$Macro\_F=\frac{2\cdot Macro\_P\cdot Macro\_R}{Macro\_P+Macro\_R} \tag{11}$$

本文将分别使用领域基础词典D_{LDA}、仅基于PMI扩充后的领域词典(LDA+PMI)、仅基于依存句法分析扩充后的领域词典(LDA+DP)、基于PMI和依存句法分析方法扩充后的领域词典(LDA+PMI+DP)等四种进行评价对象挖掘实验，并对比实验结果。本文的实验数据可挖掘到的评价对象可以分为两种，一种是一级标签评价对象，一种是二级标签评价对象，因此本文将分别对这两种评价对象进行挖掘实验。

一级标签评价对象有九种，分别是“方便性”“品牌”“包装”“产品”“价格”“香味”“快递”“购物渠道”“产品功效”。 基于PMI和依存句法分析等两种方法扩充后(LDA+PMI+DP)的一级标签领域词典的一部分如表3所示，仅列出了“香味”和“快递”等两种评价对象的部分词汇和词组。

表3 扩充后的一级标签领域词典的一部分

Tab.3 A part of the extensive first-level label domain dictionary

评价对象	词汇和词组
香味	芳香，薰衣草，味道，浸泡 味，浸泡 香，晾衣 味，晾衣 香
快递	送货，物流，快递员，小哥，配送员，包邮，邮费，快递 破了

一级标签评价对象挖掘实验结果如表4所示。从表4中的结果可以看出，与领域基础词典D_{LDA}相比，基于PMI方法和依存句法分析方法扩充后的词典的*Macro_P*指标有所降低，*Macro_R*指标有所提升，衡量整体性能的*Macro_F*指标有所提升，这说明本文提出的词典扩充方法对一级标签领域词典的整体性能是有所提升的，但是由于词典规模的扩大，随之也会出现更多的误判，导致精确率降低。同时可以看出，在只使用一种词典扩充方法的情况下，基于依存句法分析的词典扩充方法要优于基于PMI的词典扩充方法；两种扩充方法都使用时*Macro_F*指标可以达到最高，相较于只使用领域基础词典时提升了1.9个百分点。虽然扩充后的词典可以提升一级标签评价对象挖掘的性能，但是提升十分有限。

表4 一级标签评价对象挖掘实验结果

Tab.4 Result of the first-level label evaluation object mining experiment

词典构建方式	*Macro_P*	*Macro_R*	*Macro_F*
LDA	0.793	0.674	0.729
LDA+PMI	0.778	0.694	0.734
LDA+DP	0.782	0.705	0.742
LDA+PMI+DP	0.774	0.723	0.748

二级标签评价对象有69种，由一级标签评价对象细分而得。其中“香味”被分为了“香味(笼统)”“打开包装时的香味”“浸泡时的香味”“洗衣时的香味”“晾衣时的香味”“快递”被分为了“快递(笼统)”“物流/送货速度”“快递包装”“快递费用”“快递人员”。使用两种方法扩充后的二级标签领域词典的一部分如表5所示，仅列出了“香味”和“快递”等两种评价对象细分后的11个评价对象的部分词汇和词组。

表5 扩充后的二级标签领域词典的一部分

Tab.5 A part of the extensive second-level label domain dictionary

评价对象	词汇和词组
香味(笼统)	芳香，薰衣草，味道
打开包装时的香味	拧开 香，打开 香
浸泡时的香味	浸泡 味，浸泡 香
洗衣时的香味	洗衣 香，洗时 香
晾衣时的香味	晾衣 味，晾衣 香，晒衣 香
快递(笼统)	送货，物流
物流/送货速度	当天 到，快递 很快
快递包装	快递 袋子，快递 破了
快递费用	包邮，邮费
快递人员	小哥，快递员

将表5和表3对比可以看出，表3中很多对应同一评价对象的词汇在表5中被对应到不同的评价对象。同时，很多二级标签评价对象的关键词集合中词组较多，单个词汇较少。

二级标签评价对象挖掘实验结果如表6所示。表6中的各项指标变化趋势与表4中各项指标变化趋势相似。与一级标签评价对象挖掘的实验结果相比，二级标签评价对象挖掘的实验结果各项指标均有所下降。将表4和表6的实验结果进行对比可以看出，相较于一级标签评价对象挖掘实验，扩充后的词典对二级标签评价对象挖掘实验的Macro_F指标有更大的提升，相较于只使用领域基础词典时提升了4.2%，这意味着本文提出的词典扩充方法对二级标签评价对象的挖掘有更重要的意义。由于很多二级标签评价对象的关键词集合中词组较多，单个词汇较少，因此用词组扩充领域词典对于这些标签的挖掘是非常有效的。

表6 二级标签评价对象挖掘实验结果

Tab.6 Result of the second-level label evaluation object mining experiment

词典构建方式	*Macro_P*	*Macro_R*	*Macro_F*
LDA	0.684	0.577	0.626
LDA+PMI	0.656	0.649	0.652
LDA+DP	0.669	0.663	0.666
LDA+PMI+DP	0.645	0.692	0.668

6 结论(Conclusion)

本文提出了一种面向商品评价对象挖掘的词典构建方法，并使用京东商城洗衣液评论数据集进行了评价对象挖掘实验，以评估词典的性能。本文的词典分为两部分，一部分是领域基础词典，由单个的词汇构成；另一部分是领域词典的扩充，由词组构成。本文基于LDA模型从文本中提取主题词，提出了构建基础词典的方法；通过计算词汇之间的PMI值，提出了一种扩充领域词典的方法；基于依存句法分析和TF-IDF，提出了另一种扩充领域词典的方法。实验证明，扩充后的领域词典的挖掘效果好于领域基础词典单独使用的效果；用词组扩充领域词典对二级标签评价对象的挖掘意义更大。

本文的方法在针对洗衣液产品评论的评价对象挖掘实验中取得了良好的表现，将来可以使用本文方法对其他领域的短文本进行实验；同时，由于本文的方法需要大量的标注，属于有监督学习，需要耗费大量的人力物力，因此接下来将会考虑是否可以基于无监督学习的方法构建词典；本文所提出的方法只能针对特定的领域来构建词典，无法构建一个开放领域的词典，下一步将尝试是否可以得到一个跨领域的词典构建框架，来构建跨领域的词典。

参考文献(References)

[1] Mashechkin I V,Petrovskiy M I,Popov D S,et al.Applying text mining methods for data loss prevention[J].Programming & Computing Software,2015,41(1):23-30.

[2] Pavlinek M,Podgorelec V.Text classification method based on self-training and LDA topic models[J].Expert Systems with Applications,2017,80:83-93.

[3] He T,Hao R,Qi H,et al.Mining Feature-Opinion from Reviews Based on Dependency Parsing[J].International Journal of Software Engineering & Knowledge Engineering,2017,26(9n10):1581-1591.

[4] Tomas P,Virginijus M.Comparison of Na ve Bayes,Random Forest,Decision Tree,Support Vector Machines,and Logistic Regression Classifiers for Text Reviews Classification[J].Baltic Journal of Modern Computing,2013.

[5] Mandal S,Gupta S.A novel dictionary-based classification algorithm for opinion mining[C].Second International Conference on Research in Computational Intelligence and Communication Networks.IEEE,2017:175-180.

[6] 尹文科,朱明,陈天昊.基于Wiki链接结构图聚类的领域词典构建方法[J].小型微型计算机系统,2014,35(6):1286-1292.

[7] 李伟卿,王伟军.基于大规模评论数据的产品特征词典构建方法研究[J].数据分析与知识发现,2018,2(1):41-50.

[8] Chen Z,Cafarella M,Jagadish H V.Long-tail Vocabulary Dictionary Extraction from the Web[C].Proceedings of the Ninth ACM International Conference on Web Search and Data Mining,2016:625-634.

[9] Kim M,Kim J,Cui J.Performance Evaluation of Domain-Specific Sentiment Dictionary Construction Methods for Opinion Mining[J].International Journal of Database Theory and Application,2016,9:257-268.

[10] Wu J,Li Y.Research on construction of semantic dictionary

in the football field[C].IEEE,International Conference on Software Engineering Research,Management and Applications.IEEE,2017:303-306.

[11] Alqasemi F,Abdelwahab A,Abdelkader H,et al.Opinion Lexicon Automatic Construction on Arabic language[C].International Conference on Advanced Technology and Applied Sciences,2017.

[12] Ju M,Duan H,Li H.A CRF-based Method for Automatic Construction of Chinese Symptom Lexicon[C].International Conference on Information Technology in Medicine and Education.IEEE,2016:5-8.

[13] Cheng Y,Huang Y.Research and Development of Domain Dictionary Construction System[C].IEEE/WIC/ACM International Conference on Web Intelligence,2017:1162-1165.

[14] Zhang S,Wei Z,Wang Y,et al.Sentiment analysis of Chinese micro-blog text based on extended sentiment dictionary[J].Future Generation Computer Systems-The International Journal of eScience,2018(81):395-403.

[15] Song Y,Jeong S,Kim H.A Semi-automatic Construction method of a Named Entity Dictionary Based on Wikipedia[J].Journal of KIISE,2015,42(11):1397-1403.

[16] Guo X,He T,Xing Y.Construction of relational word dictionary and learning of relational rules in PPI extraction from biomedical literatures[J].International Journal of Data Mining and Bioinformatics,2016,15(2):125-144.

[17] Hangya V.Automatic Construction of Domain Specific Sentiment Lexicons for Hungarian[C].18th International Conference on Text,Speech and Dialogue,2015:183-190.

[18] Wu F,Huang Y,Song Y,et al.Towards building a high-quality microblog-specific Chinese sentiment lexicon[J].Decision Support Systems,2016,87:39-49.

[19] Liu J,Yan M,Luo J.Research on the Construction of Sentiment Lexicon Based on Chinese Microblog[C].8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC),2016:56-59.

作者简介：

石玉鑫(1994-)，男，硕士生.研究领域：WEB数据挖掘.

杨泽青(1992-)，女，硕士生.研究领域：WEB数据挖掘.

赵志滨(1975-)，男，博士，副教授.研究领域：分布式计算，WEB数据挖掘，大数据管理.

姚 兰(1977-)，女，博士，副教授.研究领域：W EB数据集成，云计算，无线传感器网络.

(上接第33页)

模拟漫反射效果，使用平行光模拟日照效果，室内物体的反射信息和间接照明可以使用Reflection Probe反射探头和Light Probe光照探头模拟。

最后在Unity中安装SteamVR插件，添加对HTC Vive虚拟现实设备的支持，编写脚本完成对虚拟场景的交互操作。

5 结论(Conclusion)

本文实验的主要硬件有：HTC Vive、Intel Xeon E5-1620v3、8G内存、显卡Nvidia GeForce GTX 1070等。主要软件有：Windows 10 64位、3ds Max2017、Substance Designer 6.0、Substance Painter 2017、Unity3d 5.6.3等。

基于上述软硬件环境，本文研究了虚拟现实技术的基本特征，以及虚拟现实技术在室内设计方面的相关工作，使虚拟现实技术与室内家居互动设计相结合，为室内家居设计提供了更真实更丰富的虚拟互动体验。

参考文献(References)

[1] 张风军,戴国忠,彭晓兰.虚拟现实的人机交互综述[J].中国科学,2016,46(12):1711-1736.

[2] 李琳,王泊谦,曾睿,等.面向虚拟环境的VR设备比较研究[J].合肥工业大学学报(自然科学版),2018,41(2):169-175.

[3] 郭云鹏,张弓,韩彰秀,等.虚拟现实技术的应用研究及发展趋势[J].电视技术,2017,41(9/10):129-134.

[4] 王丹婷,蒋友燏.古建筑三维虚拟建模与虚实交互软件实现[J].计算机应用,2017,37(S2):186-189.

[5] 刘氢.基于Unity3D和htcvive的虚拟现实游戏设计与实现[J].通信设计与应用,2017,2:43-44.

[6] 王嘉宁,王策.虚拟现实(VR)在园林景观设计中的应用[J].天津农业科学,2017,23(3):103-105.

[7] 吕屏,杨鹏飞,李旭.基于VR技术的虚拟博物馆交互设计[J].包装工程,2017,38(24):137-141.

[8] 赵鸿凯,张凯云,沈小华.基于三维虚拟现实的历史街景重现技术[J].科学技术与工程,2018,18(25):200-205.

[9] 林一,陈靖,刘越,等.基于心智模型的虚拟现实与增强现实混合式移动导览系统的用户体验设计[J].计算机学报,2015,38(2):408-422.

[10] 李微微,沈冰.基于虚拟现实在室内软装饰设计中的合理运用[J].现代电子技术,2018,41(2):148-151.

[11] 王美.虚拟现实技术在室内装饰设计中的应用[J].信息技术与信息化,2015(10):203-204.

[12] 覃 斌.虚拟现实技术在住宅室内设计互动分析中的应用[J].山西建筑,2016,42(25):7-8.

[13] 祁长兴,刘峻杭.虚拟现实在室内设计中的实际应用[J].软件工程,2017,20(4):1-3.

作者简介：

柯 健(1974-)，男，硕士，讲师.研究领域：虚拟现实技术，深度学习，数字媒体技术.

刘 畅(1982-)，女，硕士，讲师.研究领域：数字媒体技术.

周德富(1968-)，男，本科，教授.研究领域：教育技术.

王 敏(1984-)，女，硕士，讲师.研究领域：计算机应用.

夏振新(1997-)，男，专科生.研究领域：动漫制作技术.

黄冬林(1997-)，男，专科生.研究领域：软件技术.