北京大学学报(自然科学版) 第 56 卷 第 1 期 2020 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 56, No. 1 (Jan. 2020)
doi: 10.13209/j.0479-8023.2019.102

面向微博用户的消费意图识别算法

贾云龙[1] 韩东红[1,†] 林海原[1] 王国仁[2] 夏利[1]

1. 东北大学计算机科学与工程学院, 沈阳 110819; 2. 北京理工大学计算机学院, 北京 100081;
† 通信作者, E-mail: handonghong@mail.neu.edu.cn

摘要 利用迁移学习的方法, 融合京东问答平台数据与少量已标注的微博数据构建训练集, 提出一种基于注意力机制的双向长短期记忆神经网络(Attentional-Bi-LSTM)模型, 用于识别用户的隐性消费意图。针对显性意图识别问题, 提出一种结合 TF-IDF (term frequency-inverse document frequency)与句法分析中动宾关系(VOB)的消费意图对象提取算法。实验结果表明, 通过将迁移京东问答平台的数据与微博数据相融合, 可以有效地扩充训练集, 在此基础上训练的神经网络分类模型具有较高的准确率和召回率; 融合 VOB 和 TF-IDF 的显性消费意图对象提取方法的准确率达到 78.8%。

关键词 消费意图识别; 意图对象提取; 迁移学习; 注意力机制

Consumption Intent Recognition Algorithms for Weibo Users

JIA Yunlong[1], HAN Donghong[1,†], LIN Haiyuan[1], WANG Guoren[2], XIA Li[1]

1. School of Computer Science and Engineering, Northeastern University, Shenyang 110819; 2. College of Computer, Beijing Institute of Technology, Bejing 100081; † Corresponding author, E-mail: handonghong@mail.neu.edu.cn

Abstract The data set is constructed by the data of Jingdong Question Answer Platform and Weibo based on transfer learning method and a bi-directional long-term and short-term memory neural network model based on attention mechanism is proposed to identify users' implicit consumption intention. For the problem of explicit intention recognition, a new algorithm for extracting consumer intention objects is proposed, which combines TF-IDF (term frequency-inverse document frequency) with the verb-object relationship (VOB) in parsing. The experimental results show that the training set can be effectively expanded by merging the data of Jingdong Question Answer Platform and Weibo. The classification model has high accuracy and recall rate, and the method of extracting explicit consumer intent objects by fusing VOB and TF-IDF achieves 78.8% accuracy.

Key words consumption intention detection; intention object extraction; transfer learning; attention mechanism

微博作为一种新的社交网络平台, 能够使人们方便快捷地参与热点事件讨论、交流和表达情感。由此产生海量社交网络数据, 部分数据蕴含用户对某种商品的购买愿望, 即消费意图。消费意图分析指挖掘用户通过文本或行为表达的对某一产品或服务的购买意愿, 并针对社交网络用户的消费行为进行识别的研究[1], 可以应用于诸如广告推荐、市场营销等领域, 具有重大的商业应用潜力。

用户消费意图分析主要包括识别显性消费意图和识别隐性消费意图, 其中隐性消费意图的识别以及显性消费意图中的消费对象抽取等问题更具挑战性, 具体情况见表1。文本是承载情感及意愿的最好载体, 目前关于用户隐性消费意图的研究大多基于用户行为特征展开, 未考虑用户发布的文本信息内容, 因此影响分析结果的准确率。

从计算语言学的角度, 针对社交网络用户消费意图挖掘的研究刚刚起步, 特别是对隐性消费意图的研究成果鲜有问世。目前, 面向微博用户的隐性

国家重点研发计划项目(2016YFC1401900)、国家自然科学基金(61173029, 61672144, 61872072)和计算机软件新技术国家重点实验室开放课题(KFKT2018)资助

收稿日期: 2019–05–22; 修回日期: 2019–09–24

表 1　隐性消费意图与显性消费意图中的消费对象

Table 1　Consumption objects in the implicit and explicit consumption intentions

项目	隐性消费意图	显性消费意图中的消费对象
问题定义	文中蕴含或可以从中推断出的潜在购买意图	以文本形式明确表达出的购买意图对象
例子	我不喜欢用有绳吸尘器	想求购一款触屏性能好的手机，求推荐
特点	无明确触发词	有明确触发词，如“求购”
面临挑战	大多基于用户行为特征展开研究，未考虑文本内容，且缺少足够的标注语料库	需要基于搜索引擎且需要联网使用，无法离线提取消费意图对象

消费意图识别研究缺少足够的标注语料库。本文利用迁移学习方法，将源域京东商城的问答数据迁移到目标域微博训练集中，提出一种基于Attention机制的Bi-LSTM模型，预测微博用户的隐性消费意图；还提出一种基于TF-IDF与句法分析的TF-IDF-VOB显性消费意图对象提取方法。

1 相关工作

有关意图的概念，虽然在哲学和心理学层面有较多研究，但通常不关注用来表达意图的语言，或者如何从书面语言中通过计算推断意图。Goldberg等[2]从计算机语言学的角度定义“购买意图”(显性消费意图)，提出基于二元图方法的自动抽取消费意图模板，同时将文本作为消费意图识别的特征，提高了分类的准确率。Kröll等[3]给出意图分析的定义，认为意图分析与情感分析在一定程度上是正交的，例如推文“我很喜欢这款笔记本电脑”，同时表达出用户的正面情感和购买笔记本电脑的意图。他们还定义，如果一条推文具有商业意图，则该推文需要至少包含一个动词(即触发词)。Wang等[4]提出可以通过触发词识别显性消费意图的推文，并定义了意图客体即购买对象。陈浩辰[5]构建基于深度学习的用户消费意图预测模型，并提出消费意图向消费行为的转化。Ding等[6]提出基于领域自适应卷积神经网络的微博文本消费意图识别方法，并将其应用到电影票房预测的任务中。Liu等[7]利用基于Attention机制的循环神经网络结构(RNN)进行意图识别，由于RNN是基于序列的，句子中的每个词在RNN中都有隐藏状态，因此可以利用这些隐藏状态生成最后的意图类别。Ding等[8]认为消费意图识别的任务具有领域相关的特性，为此构建一个跨领域的基于深度学习的消费意图识别模型，充分利用深度神经网络和双样本检验的优势，使用基于树核的最大平均差异(TK-MMD)来提升模型的学习能力。Yang等[9]针对用户消费意图分类问题，通过一系列实验来比较支持向量机(SVM)、朴素贝叶斯(NB)和长短期记忆网络(LSTM)等模型在意图分类上的效果，并采用Wang等[4]提出的意图分类方法，将消费意图分成6类：Food&Drink，Travel，Career&education，Goods&Service，Event&Activities和Trifle。钱岳等[10]和余慧等[11]分别利用基于 Convolutional-LSTM模型和双向门控循环单元(BI-GRU)模型，进行显性消费意图预测。为解决源域标注数据不足的问题，Chen等[12]在挖掘“Intention Posts”的研究中提出在不同领域表达购买意图的方式是相似的假设，利用迁移学习方法进行意图分类，即使用多个源域的标注数据辅助分类目标域内未标注的数据。Song等[13]提出一种基于多源域和多实例的迁移学习方法。

与显性消费意图识别问题相比，关于隐性消费意图识别的研究成果很少。付博等[1]提出隐性意图识别的方法，该方法将隐性消费意图识别视为多标记分类问题，并综合使用基于用户关注行为、意图关注行为、意图转发行为以及个人信息的多种特征，利用京东商城中的评论信息作为意图行为转化的依据。Park等[14]构建平行语料库，即隐性意图的表达方式与相应的显性意图表达方式对，例如，“我好饿”对应“我想吃东西”，前者是隐性表达方式，后者是前者对应的显性表达方式。Ding等[6]提出的基于领域自适应卷积神经网络的微博文本消费意图识别方法也可以适用于隐性意图识别，利用Word Embedding挖掘词的表示信息，通过卷积和池化操作提取局部词汇信息表示，得到句子级的表示，并认为这种句子表示在不同领域有相似性。

目前关于消费意图对象抽取的研究比较少。付博等[15]比较了几种基于句法依存关系的消费意图对象提取方法，并提出一种意图对象提取算法，即首先分析具有消费意图微博中的关键词，将其视为查询，提交给搜索引擎进行搜索，最后从搜索结果中提取消费意图对象。Duan等[16]先根据依赖性解

析器自动提取候选产品，然后利用搭配抽取模型，从候选集中识别出真实的意向相关产品。Park 等[14]构建包含隐含意图文本和相应的显性意图文本的并行语料库，完成移动应用检索任务。Wang 等[17]提出从微博中挖掘趋势相关产品的方法，将“趋势”定义为用户热烈讨论的话题。

2 面向微博用户的消费意图识别

本文面向微博用户消费意图识别的框架如图 1 所示。用户的消费意图识别分为显性消费意图识别和隐性消费识别，识别其消费意图后，根据其归属继续提取其消费意图对象。

本文将消费意图识别视为二分类问题，并做如下定义：源域 $D_s=\{D_s^1, D_s^2, \ldots, D_s^n\}$ 代表 n 条京东商品问答数据，其中 D_s^i 代表第 i 条京东问答数据；目标域 $D_t=\{(D_t^1, y^1), (D_t^2, y^2), \ldots, (D_t^m, y^m)\}$ 代表少量具有标签的微博数据，其中 (D_t^i, y^i) 代表第 i 条数据，对应的标签为 y^i, $y^i \in Y$, $Y=(-1,+1)$。

2.1 基于迁移学习的目标域标注数据集构成方法

微博中具有购买意图的博文数量占 3%左右[5]。由于京东问答系统的文本数据经过预处理后均具有隐性购买意图，同时不同应用领域中对购买意图的表达方式非常接近，因此本文将京东问答数据作为源域，利用迁移学习方法，将置信度较高的京东问答数据迁移到具有小量标注数据的目标域微博训练集中，具体伪代码如算法 1 所示。

算法1 基于迁移学习的目标域标注数据集构成算法

输入 京东商品问答数据集 D_s(源域)，少量标注的微博数据集 D_t，集合分类器中基分类器的数量 N，迭代次数 M，阈值 K

输出 迁移源域数据的微博标注数据集 D_t'(目标域)

```
D_t' = D_t
对 D_t'和 D_s 进行文本预处理
For i = 1, …, M Do
    基于 word2vec 词嵌入方法和 SVM 模型，训练集成分类器中的 N 个基分类器
    For j = 1, …, Len (D_s) Do
        利用集成分类器中的 N 个基分类器分别预测 D_s^i 的类标签
        If Num (对 D_s^i 极性预测一致的基分类器)>K
            D_t'= D_t' ∪ D_s^j
        EndIf
    EndFor
EndFor
Return D_t'
```

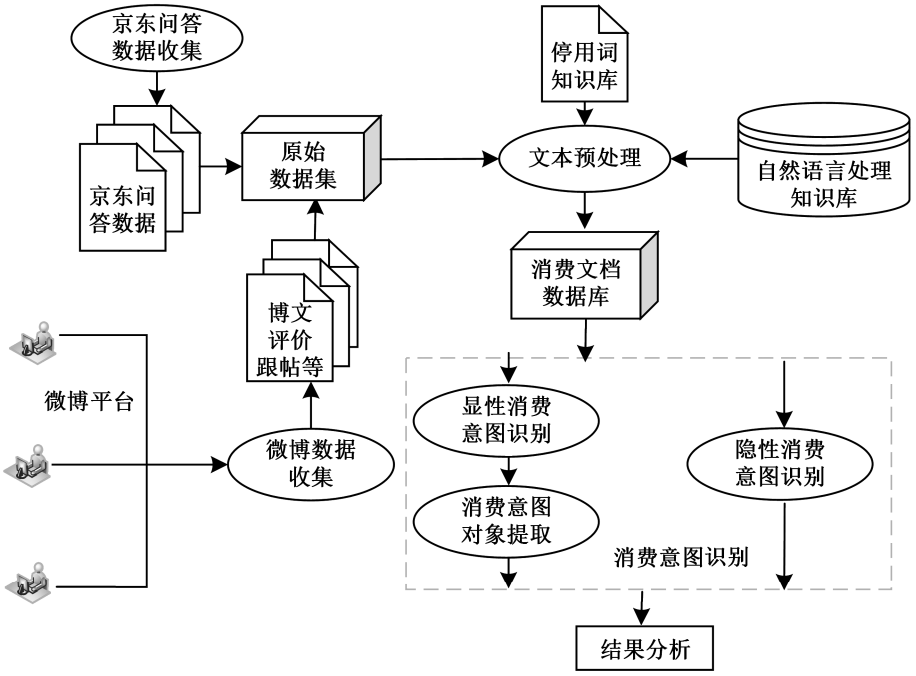

图 1 面向微博用户的消费意图识别处理框架

Fig. 1 Processing framework of consumer intention recognition for Weibo users

首先，将集合 D_t 赋予 D_t'，然后对源域 D_s 和具有少量标注的微博数据集 D_t 进行预处理。然后，将 D_t' 中的博文转换成词嵌入的形式，利用 SVM 模型训练 N 个基分类器。接着，使用 N 个基分类器对源域 D_s 中的每个博文进行极性预测，如果对第 j 个博文 D_s^j 极性的预测值一致的基分类器个数超过阈值 K，则认为该数据为置信度高的数据，将其并入微博标注数据集 D_t' 中，再使用 D_t' 重新训练分类器，重复 3~9 行，直至达到迭代次数 M。最后，输出迁移源域数据的微博标注数据集 D_t'。

2.2 基于Attentional-Bi-LSTM模型的隐性消费意图识别算法

本文使用双向 LSTM 模型(即同一个输入序列向前和向后各自训练一个 LSTM 模型)，将输出进行线性化表示，使得每一个 LSTM 单元都可以得到上下文信息，然后将 Bi-LSTM 的输出向量集合构成注意力的权重矩阵。通过学习，模型可以增加具有隐性消费意图的词语的权重，降低无关词语的权重，使分类更加准确，如图 2 所示。

2.3 基于TF-IDF-VOB的消费意图对象提取算法

目前有两种基于经验性的方式可用来分析消费意图对象在文本中的句法关系：修饰副词/名词+消费对象(ATT)和意图触发词+消费对象(VOB)。这两种方式存在一定的不足：例如当博文较短，不存在 ATT 或 VOB 关系时，无法通过这两种方式提取到意图对象；如果只是提取博文中的名词，则当名词很多时，会无法识别哪个名词是真正的意图对象。

通过分析具有显性消费意图对象的微博文本可以发现：1) 消费意图对象通常是以名词为代表；2) 消费意图对象通常与意图触发词构成动宾关系；3) 微博文本中如果包含某个商品的链接，那么这个商品则可能是用户的消费意图对象。因此，本文通过增加第 2 和 3 条中意图对象的权重，降低第 1 条中其他名词权重来提取意图对象。

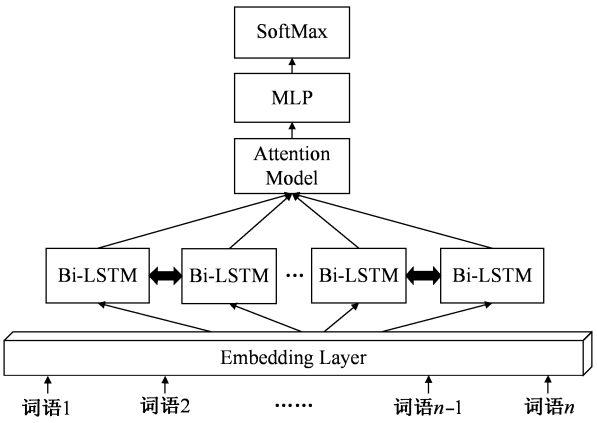

图 2　Attentional-Bi-LSTM 模型
Fig. 2　Attentional-Bi-LSTM model

消费意图对象通常与意图触发词构成 VOB 关系，且为 VOB 关系中的名词部分，故本文通过 LTP 依存句法分析器提取与意图触发词搭配的动宾关系中的名词，如图 3 所示。

本文提出一种基于 TF-IDF-VOB 的消费意图对象提取方法，其中用于衡量微博文本中每个名词重要程度的指标计算公式如下：

$$\mathrm{Score}_{ij} = \mathrm{TF}_{ij} \times \mathrm{IDF}_{ij}, \tag{1}$$

$$\mathrm{TF}_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}, \tag{2}$$

$$n_{ij} = \alpha \times n\_\mathrm{vob}_{ij} + \beta \times n\_\mathrm{others}_{ij} + \gamma \times n\_\mathrm{url}_{ij}, \tag{3}$$

$$\mathrm{IDF}_{ij} = \log \frac{N}{n_j + 1}。 \tag{4}$$

式(2)中，TF_{ij} 表示第 i 篇微博中第 j 个名词的词频，k 表示微博的数量，n_{ij} 表示第 j 个名词在该微博文本 i 相关的各部分文字中出现次数的加权和。式(3)中，n_vob_{ij} 表示第 i 篇微博中第 j 个名词，并且是与意图指示词呈现 VOB 关系的名词数量；n_others_{ij} 表示第 i 篇微博中第 j 个，且与意图指示词无 VOB 关系

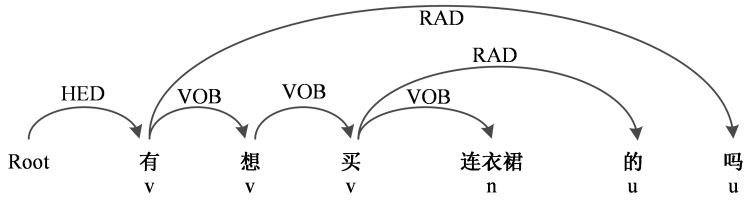

图 3　LTP 平台依存句法分析实例
Fig. 3　An example of platform dependency syntax analysis

的名词的数量; n_url_{ij}表示第i篇微博中第j个名词在所指向的URL链接中出现的次数; α,β和γ代表各自的权重。式(4)中, N表示微博数据集和京东问答数据集中博文的总和, n_j表示其中包含单词j的数量。对于每一篇微博的消费意图对象, 可以视为通过式(1)计算得到的 Score 最大的单词。

3 实验与分析

由于包含消费意图的微博非常少, 所以本文采用搜索消费意图触发词的方式, 爬取一定量的新浪微博数据。我们采用人工方式标注 10000 条无消费意图的微博文本和 10000 条具有显性消费意图的微博文本, 并在 10000 条显性消费意图的微博中挑选出 1000 条, 标注其意图对象, 作为提取意图对象的数据集。我们将一部分包含显性消费意图的句子拆成两部分: 包含意图触发词的句子和不包含意图触发词的句子, 分别放入标注的显性消费意图和隐性消费意图数据集。例如, “我的手机坏了, 最近打算入手 p30”中, “我的手机坏了”是不含触发词, 但有一定的隐性消费意图, 因此将其作为隐性消费意图识别的语料。

对于京东商品问答数据集, 我们假定在社区发表过商品问题咨询的用户均具有消费意图, 并爬取京东问答数据 200000 条。经观察, 发现其中极少文本含有消费意图触发词, 因此在预处理阶段, 将有购买触发词的文本删除, 将其他问答数据视为具有隐性消费意图的文本, 加入京东问答数据集。

3.1 评价方法

实验采用准确率P、召回率R以及F1值共3个指标评价模型的性能, 计算公式如下:

$$P=\frac{\text{TP}}{\text{TP+FP}}, \tag{5}$$

$$R=\frac{\text{TP}}{\text{TP+FN}}, \tag{6}$$

$$\text{F1}=\frac{2\times P\times R}{P+R}。 \tag{7}$$

3.2 性能分析

3.2.1 基于迁移学习的目标域标注数据集构成

先单独使用标注的微博数据集分别训练 SVM, LSTM 和 Bi-LSTM 模型, 结果如表 2 所示。可以发现, 在数据量不多的情况下, LSTM 与 SVM 模型的效果相当。

本文分别使用标注微博数据集、标注微博数据集+原始京东数据集和标注微博数据集+置信度高的京东数据集, 分别训练 LSTM 模型, 结果如表 3 所示。可以看出, 直接迁移京东数据到训练集中训练出的模型性能低于只使用微博数据集训练出的 LSTM 模型。使用迁移置信度高的京东数据训练出的模型性能比直接使用微博数据训练出的模型有一定程度的提升。本文还复现 Chen 等[12]提出的 Co-Class 算法, 并进行对比。

3.2.2 基于 Attentional-Bi-LSTM 模型的隐性消费意图识别

在使用基于迁移学习方法生成目标域标注数据集的基础上, 分别训练 LSTM, LSTM+Attention 和 Bi-LSTM 等模型, 结果如表 4 所示。可以看出, 将京东平台的问答数据迁移到微博数据集中, 可以提高对隐性消费意图识别的准确率和召回率。

3.2.3 基于 TF-IDF-VOB 的显性意图对象提取

由于目前没有公开的数据集, 因此本文通过人工挑选出具有消费意图对象的微博文本作为消费意图对象提取的数据, 共 1000 条。由于样本较少, 我

表 2 基于微博数据集的不同模型性能对比

Table 2 Performance comparison of different models based on Weibo dataset

模型	P	R	F1
SVM	0.815	0.785	0.799718
LSTM	0.844	0.763	0.801458
Bi-LSTM	0.832	0.801	0.816205

表 3 基于不同数据组合的模型性能对比

Table 3 Performance comparison of models based on different data combinations

模型	P	R	F1
LSTM (微博数据)	0.844	0.763	0.801458
LSTM (微博数据+直接迁移京东数据)	0.805	0.759	0.781323
Co-Class (微博数据+直接迁移京东数据)	0.829	0.780	0.803753
LSTM (微博数据+迁移置信度高的京东数据)	**0.929**	**0.891**	**0.909603**

说明: 粗体数字表示最优结果。

表 4　基于迁移后生成数据集的不同模型的性能对比

Table 4　Performance comparison of different models based on migrated data sets

模型	*P*	*R*	F1
LSTM	0.929	0.891	0.909603
LSTM+Attention	0.936	0.905	0.920239
Bi-LSTM	0.920	0.912	0.915982
CNN+LSTM	0.918	0.922	0.919996
Attentional-Bi-LSTM	0.939	0.920	0.929403

们采用十折交叉验证方法，将 1000 条数据平均分成 10 份，共做 10 轮实验，每轮选取不同的 1 份作为测试集，另外 9 份作为训练集，最终得到的评价指标是 10 轮实验的平均准确率。

假定 α, β 和 γ 符合 $\alpha+\beta+\gamma=1$，且 α, β 和 γ 均大于 0，通过调节各自的取值使得准确率最高。通过分别限定 α, β 和 γ 的值为 0，分析另外两个参数对准确率的影响，见表 5。可以发现，当与意图触发词构成动宾关系的名词权重较大时，算法能取得较好的准确率，与我们预期的消费意图对象通常与意图触发词构成动宾关系的结论保持一致。

本文通过适当调节得到最优参数，并与传统的基于 ATT (方法 1)、VOB (方法 2)和只使用 TF-IDF 的 (方法 3) 3 种基线方法进行比较，同时也复现了基于单词对齐的社交媒体检索方法[15](方法 4)，性能对比如图 4 所示。可以发现，本文提出方法的准确率高于其他方法。这是因为本文方法通过分配权重的方式，提高了 VOB 关系中名词的权重，同时也能考虑到其他名词的权重，在一定程度上弥补了 VOB 以及 ATT 方法的不足，也解决了当不存在 VOB 或 ATT 关系时，无法从名词中提取到意图对象的问题。

表 5　不同参数的对比

Table 5　Comparison of different parameters

限制条件	α	β	γ	*P*/%
α=0	–	0.35	0.65	50.5
β=0	0.62	–	0.38	64.3
γ=0	0.78	0.22	–	75.4

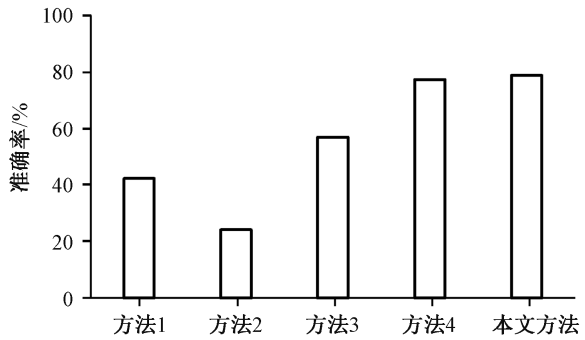

图 4　不同意图对象提取方法的性能对比

Fig. 4　Performance comparison of different intent object extraction methods

4 结论与展望

本文提出一种基于迁移学习的方法，使用京东问答平台数据作为辅助数据，在此基础上提出一种基于 Attentional-Bi-LSTM 模型识别用户的隐性消费意图。我们还提出一种基于 TF-IDF-VOB 的消费意图对象提取算法。实验结果表明，迁移京东问答平台的数据，并与微博数据相融合，可以有效地扩充训练集，训练的分类模型具有较高的准确率和召回率；融合 VOB 和 TF-IDF 的显性消费意图对象的提取方法在性能方面有较大的提升。在未来工作中，我们将尝试抽取微博中的多个意图对象，同时设计相应的推荐算法。

参考文献

[1] 付博, 刘挺. 社会媒体中用户的隐式消费意图识别. 软件学报, 2016, 27(11): 2843–2854

[2] Goldberg A B, Fillmore N, Andrzejewski D, et al. May all your wishes come true: a study of wishes and how to recognize them // Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Boulder, 2009: 263–271

[3] Kröll M, Strohmaier M. Analyzing human intentions in natural language text // International Conference on Knowledge Capture. Redondo Beach, 2009: 197–198

[4] Wang J, Cong G, Zhao W X, et al. Mining user intents in twitter: a semi-supervised approach to inferring intent categories for tweets // Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, 2015: 318–324

[5] 陈浩辰. 基于微博的消费意图挖掘[D]. 哈尔滨: 哈尔滨工业大学, 2014

[6] Ding X, Liu T, Duan J, et al. Mining User consumption intention from social media using domain adaptive convolutional neural network // AAAI. Aus-

tin, 2015, 15: 2389–2395

[7] Liu B, Lane I. Attention-based recurrent neural network models for joint intent detection and slot filling // Proceedings of INTERSPEECH. San Francisco, 2016: 685–689

[8] Ding X, Cai B, Liu T, et al. Domain adaptation via tree kernel based maximum mean discrepancy for user consumption intention identification // Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm, 2018: 4026–4032

[9] Yang M, Wang D, Feng S, et al. An Empirical study on learning based methods for user consumption intention classification // National CCF Conference on Natural Language Processing and Chinese Computing. Dalian, 2017: 910–918

[10] 钱岳, 丁效, 刘挺, 等. 聊天机器人中用户出行消费意图识别方法. 中国科学: 信息科学, 2017, 47 (8): 49–59

[11] 余慧, 冯旭鹏, 刘利军, 等. 聊天机器人中用户就医意图识别方法. 计算机应用, 2018, 38(8): 36–40

[12] Chen Z, Liu B, Hsu M, et al. Identifying intention posts in discussion forums // Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, 2013: 1041–1050

[13] Song H J, Park S B. Identifying intention posts in discussion forums using multi-instance learning and multiple sources transfer learning. Soft Computing, 2017, 22(4): 1–12

[14] Park D H, Fang Y, Liu M, et al. Mobile app retrieval for social media users via inference of implicit intent in social media text // Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. Indianapolis: ACM, 2016: 959–968

[15] 付博, 刘挺. 基于跨社交媒体检索的微博消费对象识别. 计算机科学与探索, 2015, 9(10): 1247–1255

[16] Duan J, Chen Y, Liu T, et al. Mining intention-related products on online Q&A community. Journal of Computer Science and Technology, 2015 (5): 1054–1062

[17] Wang J, Zhao W X, Wei H, et al. Mining new business opportunities: identifying trend related products by leveraging commercial intents from microblogs // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, 2013: 1337–1347