**Artificial Intelligence and Robotics Research** 人工智能与机器人研究, 2022, 11(4), 418-428
Published Online November 2022 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2022.114043

Hans汉斯

# 基于BERT模型和LDA主题模型的用户兴趣模型构建方法

**马海江，柴功昊**

广西科技师范学院数学与计算机科学学院，广西 来宾

收稿日期：2022年5月23日；录用日期：2022年11月2日；发布日期：2022年11月10日

## 摘　要

**用户兴趣模型构建主要结合用户的兴趣爱好信息、游览行为以及用户画像信息等综合分析用户兴趣。用户兴趣模型作为个性化信息推荐环节中的关键部分，也是个性化服务的重要部分，其质量的好坏直接影响着个性化信息推荐服务的水平。为了提高用户兴趣建模的质量，本文引入词向量模型和主题模型来准确表示用户兴趣，提出了一种基于BERT模型和LDA主题模型的用户兴趣模型构建方法。该方法将BERT模型和LDA主题模型融合，在训练过程中，模型不仅能够充分利用整个数据集的上下文信息，还能利用LDA获取隐语义信息，同时通过K-means聚类方法提取用户兴趣。实验结果表明，结合后的用户建模方法能够有效解决微博短文本的稀疏性以及上下文依赖性。与其他方法相比，提高了用户兴趣模型的质量。**

## 关键词

**用户兴趣模型，BERT模型，LDA主题模型，词向量模型**

# User Interest Model Construction Method Based on BERT Model and LDA Topic Model

**Haijiang Ma, Gonghao Chai**

School of Mathematics and Computer Science, Guangxi Science & Technology Normal University, Laibin Guangxi

Received: May 23rd, 2022; accepted: Nov. 2nd, 2022; published: Nov. 10th, 2022

## Abstract

**The construction of the user interest model mainly combines the user's hobby information,**

文章引用: 马海江, 柴功昊. 基于 BERT 模型和 LDA 主题模型的用户兴趣模型构建方法[J]. 人工智能与机器人研究, 2022, 11(4): 418-428. DOI: 10.12677/airr.2022.114043

**travel behavior and user profile information, and comprehensively analyzes the user's interest. As a key part of personalized information recommendation, user interest model is also an important part of personalized service. Its quality directly affects the level of personalized information recommendation services. In order to improve the quality of user interest modeling, this paper introduces word vector model and topic model to accurately express user interest, and proposes a user interest model construction method based on BERT model and LDA topic model. This method combines the BERT model and the LDA topic model. In the training process, the model can not only make full use of the context information of the entire data set, but also use LDA to obtain implicit semantic information, and at the same time extract user interests through the K-means clustering method. The experimental results show that the combined user modeling method can effectively solve the problems of sparsity and context dependence of microblog short texts. Compared with other methods, the quality of the user interest model is improved.**

## Keywords

**User Interest Model, BERT Model, LDA Topic Model, Word Vector Model**



## 1. 引言

随着社交网络和微博网站的迅速发展，用户兴趣挖掘以及建模在最近几年一直关注的研究课题。然而，现在的研究工作主要存在两大问题，一是它们只专注于用户的显性信息和社交网络结构来预测用户兴趣，而忽略了用户个性化信息是推断用户兴趣话题的重要来源的这一事实。二是它们使用词袋模型来表示用户的内容信息，该模型忽略了发表内容的时间顺序，因此预测的兴趣信息可能包含用户不再感兴趣的过时主题内容。微博作为一个比较流行的社交网络平台，成千上万的用户已习惯在微博平台上发表个人动态信息、个人评论信息和关注转发一些热点话题，从中可以获取、分享和传播自己感兴趣的话题信息。用户每天在微博上产生大量的微博数据，由用户产生的这些微博数据背后隐藏着巨大的商业应用价值，而如何能准确有效地从这些信息中分析并挖掘出用户兴趣，提高用户兴趣建模质量，对用户的个性化信息服务和价值利用有着十分重要的意义。

微博因其传播性、交互性和即时性等特点，在一定程度上充分反映出用户的兴趣偏好。为了充分挖掘出高质量的用户兴趣，研究者们综合考虑微博文本特点，提出了一系列的用户兴趣挖掘方法。其中最常见一种方法是根据用户所发表的微博内容以及关注点，分析并构建用户兴趣模型。文献[1]通过分析用户行为信息特征，结合用户自身因素来研究用户兴趣的变化过程，提取用户行为信息背后所隐藏的兴趣倾向，紧接着使用主题模型提取主题，构建用户兴趣模型。文献[2]为了提升用户兴趣模型的质量，在 LDA 主题模型的基础上进行改进，对微博内容等相关数据进行全面综合性的分析，获取兴趣主题。文献[3]根据 LDA 主题模型分析并挖掘用户兴趣特征，从而构建出用户感兴趣的主题。文献[4]使用改进的 LDA 主题模型检测微博热点话题信息，根据此热点抽取出用户兴趣特征。文献[5]综合分析微博信息传播方式和转发预测方法，基于结合用户兴趣特征的算法构建在线信息共享预测模型，并以新浪微博数据为例对预测模型进行测试。文献[6]基于社交媒体短文本的特征改进了词袋模型，提出了利用特征之间的语义关系

的语义表示模型，微博用户从中提取感兴趣话题的方法。

但传统的用户兴趣建模方法已无法满足当前需求，在大规模的数据上其时效性表现较差。因此本文引入词向量模型和主题模型，提出一种基于 BERT 模型和 LDA 主题模型的用户兴趣模型构建方法(记为 BERT_LDA)。BERT 是谷歌提出的预训练语言模型，其突出优点不仅在于它能够生成词向量，而且能够有效处理多义词的问题，对具有深度语义信息的文本进行分析。LDA 主题模型建立了文档－主题－词的三层概率结构，对于挖掘稀疏性较高的短文本数据中隐语义具有很好的效果。首先将预处理后的数据集通过 LDA 主题模型进行训练，提取文本主题信息；其次，使用 BERT 模型训练输入对应文本的词汇，对每个词进行向量化，生成主题词向量；再次，使用相似度函数计算文本向量与主题向量之间的相似度来描述每个文本的主题信息；最后，使用 K-means 聚类算法将用户文档集划分为几个相似的类别，进一步提取用户感兴趣的主题。通过实验对比其他模型可知，融合 BERT 模型和 LDA 主题模型的聚类方法获得较好的效果。

## 2. 相关工作

### 2.1. BERT 模型

BERT (Bidirectional Encoder Representations from Transformers) [7]模型是一种与之前自然语言表示模型不同的新模型，它主要是基于双向多层的转换编码器实现的，是通过各个层中上下文信息相互联合调节的形式实现深度双向表示[8]。具体的结构图如图 1 所示。

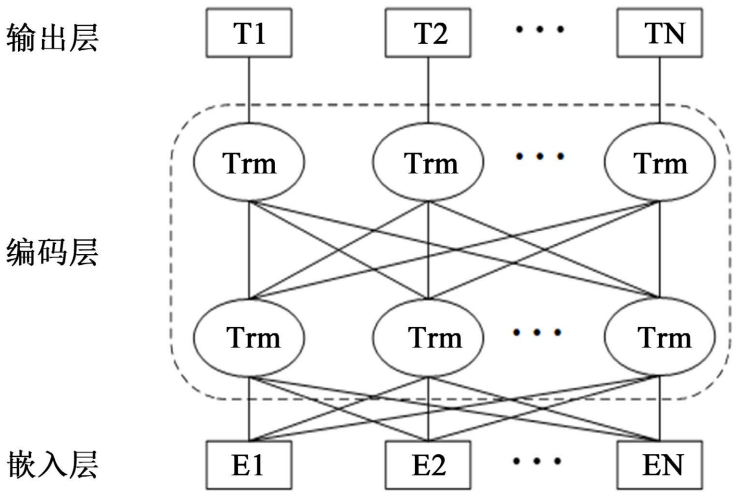


**Figure 1.** Structure of BERT model
**图 1.** BERT 模型结构图

以前的向量模型(如 Word2Vec 和 Glove)具有明显的缺点，即无法有效处理单词的多义现象。选择 BERT 语言模型的突出优点不仅在于它能够生成词向量，而且能够有效处理多义词的问题。BERT 模型的输入部分是将两个句子序列连接起来，并用符号将每个句子的开头和结尾进行标记。BERT 模型对每个单词实行了三种不同的嵌入操作(即词的位置、Word2Vec、整个句子)。将这三种嵌入结果向量进行拼接，就可以获取到 BERT 词向量[9]。其输出词向量的方式如图 2 所示。

从下图可以看出，特殊标签[CLS]和[SEP]分别添加到序列的第一个和最后一个位置，此类标签在句子或非句子对任务上的最终输出，会被视作整个句子或句子对特征表示。对于此类型以外的其他操作，即使该标签参与了序列编码，也会被最终输出结果舍弃掉。

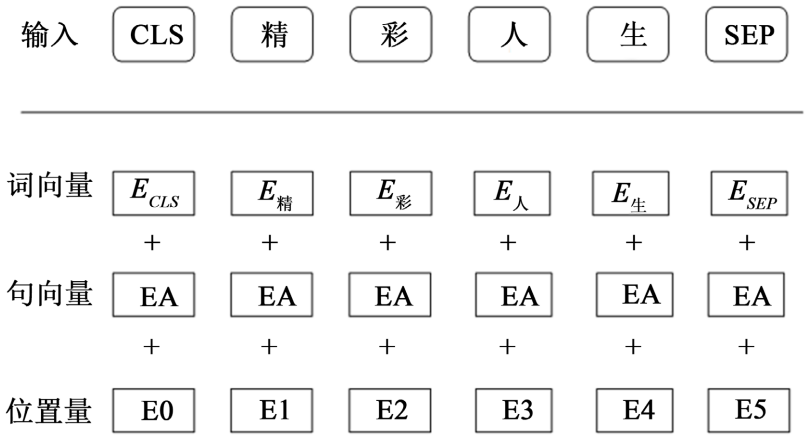


**Figure 2.** BERT model output representation
**图 2.** BERT 模型输出表示

## 2.2. LDA 主题模型

主题模型是挖掘文本中潜在主题的一种模型训练方法，主题的获取过程是根据文本中词汇的出现次数计算其概率，从概率中选择主题。LDA (Latent Dirichlet Allocation，隐含狄利克雷分布)是当前常用的一种主题获取模型，它是 Blei [10]为了提升用户兴趣主题质量而提出的。

假设语料库由 $D$ 篇文本组成，第 $m$ 个文本有 $N_m$ 个单词，这些文本一共涉及 $K$ 个主题。LDA 主题模型在预测训练集之外中的文本和主题词的分布占有很大的优势。其基本思想是：随机选择一些主题组成一个文本，并且一个词的概率分布就表示主题。LDA 主题模型将主题分布参数视为统一的隐藏 K 维随机变量，为了抑制隐藏 K 维随机变量而引入一个超参数。作为一种混合概率模型，LDA 主题模型使用单词之间的共现关系，来最大程度地提高单词在文本中查找聚类的概率。尽管 LDA 主题模型可以让文本中有多个主题出现，但它为了避免文本中包含太多的主题，就通过狄利克雷分布限制主题的占比[11]，如图 3 所示。

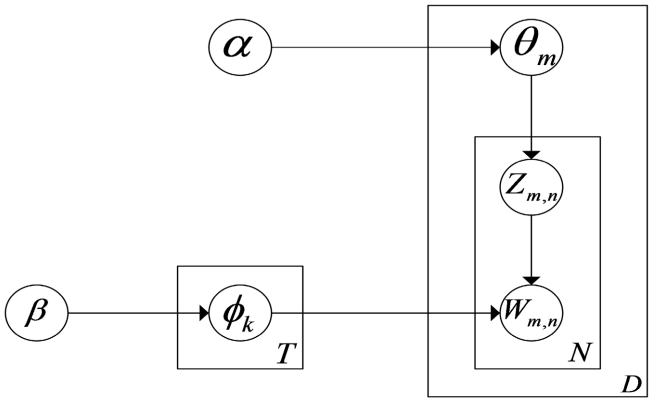


**Figure 3.** Schematic diagram of the LDA model
**图 3.** LDA 模型示意图

在 LDA 概率模型中，$\theta_m$ 为文本 $m$ 的“文本–主题”概率分布矩阵；$\phi_k$ 为主题 $k$ 的“主题–词”概率分布矩阵；$\alpha$ 和 $\beta$ 是超参数；$Z_{m,n}$、$W_{m,n}$ 为文本 $m$ 中第 $n$ 个词的主题；$T$ 为主题个数，$N$ 为文本集合 $D$ 中词的个数。

从 LDA 主题模型被提出以来，越来越多的研究者在原来的基础上，根据实际应用需求做出 LDA 算法改进及应用[12]。LDA 主题模型存在的主要缺陷是须事先设定主题个数。对于用户不太了解的一些数据，就很难指定准确的主题个数。如果指定的主题数与实际数存在很大差距的话，LDA 主题模型的性能就会受到影响，对 LDA 主题模型的使用也会受到制约。另外，当 LDA 主题模型获得文本隐藏主题的语义信息时，它也会忽略文本的词汇语义信息，其性能在一定程度上会下降。

## 3. 方法流程及关键技术

### 3.1. 方法流程

用户模型构建方法的基本流程如图 4 所示，主要包括 4 个方面：首先，对语料库进行预处理，该步骤主要针对文本中存在的特殊标点字符、表情符号、语气词、介词和没有任何意思的词等杂乱数据进行清理和停用词处理，去除存在重复的数据。对于存在缺失的完整数据进行删除或补充，删除字数较少的微博。采取常用的工具包 jieba 分词 处理语料库，将得到的文本作为词向量的输入数据。其次，通过使用 BERT 模型训练相应文本的词汇，得到词向量表示；使用 LDA 主题模型，对输入的对应文本的词进行训练，获得每个文本的潜在主题，并构建主题－词矩阵。再次，将训练好的词向量映射生成每一个主题向量和文本向量，并结合余弦相似度计算每个向量之间的相似度，来描述每一篇文本的主题信息。最后，利用 K-means 聚类对相关性较强的主题词信息进行聚类，进一步提取用户兴趣主题。

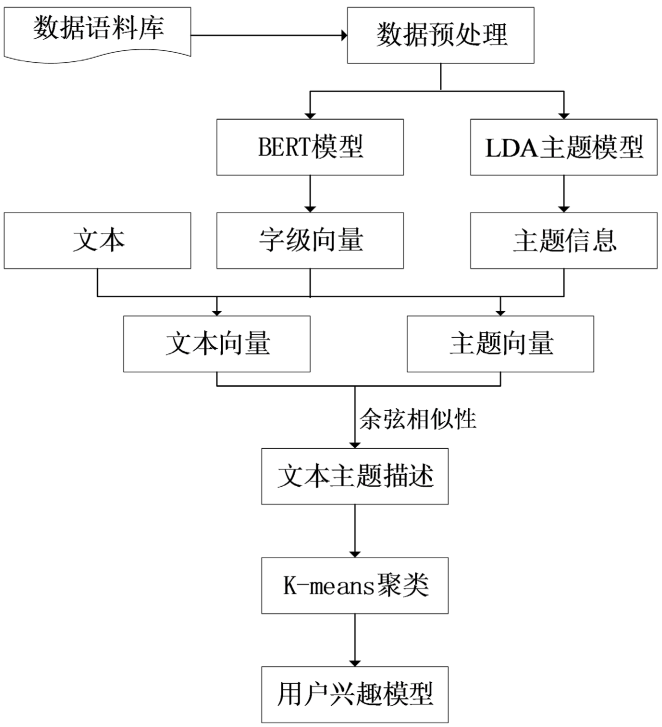


**Figure 4.** Process of user interest model construction
**图 4.** 用户兴趣模型构建方法流程

### 3.2. BERT 模型的词向量表示

为了获得更好的词向量表示，Google 提出的 BERT 预训练语言模型充分利用了词的上下文信息，可以更好地解决多义词的问题。为了在单词的左侧和右侧融合上下文信息，BERT 模型使用双向 Transformer

技术作为编码器。在模型学习过程中，将在两个方向上学习句子，以学习单词的上下文信息，从而可以更好地在不同的上下文中反映相同的单词。BERT 模型使用 Attention 机制计算句子中每个单词和其他单词之间的相关信息[13]。Word2Vec 训练的词向量表示不能很好的表示出单词的上下文信息。BERT 生成的单词表示由单词周围的单词动态表示。因此，BERT 模型在训练词向量方面较 Word2Vec 占有优势。本文亦使用 Google 提供的 BERT 模型以及中文预训练模型对文本进行向量化处理。

### 3.3. 主题相关度计算

假设给定一组文本 $X=\{x_1,x_2,\cdots,x_n\}$，是由 $N$ 个词 $\{t_1,t_2,\cdots,t_N\}$ 组成，通过 LDA 主题模型训练输出潜在主题 $\{s_1,s_2,\cdots,s_K\}$。通过 BERT 模型训练文本 $X$，并向量化每个词，生成不变长度的向量 $\{v(t_1),v(t_2),\cdots,v(t_N)\}$，主题词是从主题 $s_i$ 中前 $h$ 个词进行选取，用来生成主题向量。计算每个词占主题的权重[14]：

$$t_{i,j}=\frac{\varphi_{i,j}}{\sum_{k=1}^{h}\varphi_{i,k}} \tag{1}$$

其中，$\varphi_{i,j}$ 表示第 $i$ 个主题词中第 $j$ 个词的概率。

计算主题词向量 $v(s_i)$，即将每个词向量和它的权重的结果进行线性加权：

$$v(s_i)=\sum_{j=1}^{h}t_{i,j}v\left(t_{i_j}\right) \tag{2}$$

计算文本向量 $v(x_i)$：

$$v(x_i)=\frac{\sum_{j=1}^{c}v\left(t_{i_j}\right)}{c} \tag{3}$$

其中，$c$ 表示文本中词的个数。

对于向量 $v(x_i)$ 和 $v(s_i)$，其余弦相似度为：

$$\cos\theta=\frac{v(x_i)\cdot v(s_i)}{|v(x_i)|\times|v(s_i)|} \tag{4}$$

由于余弦夹角值范围在[−1, 1]，所以需要将其归一化到[0, 1]，即：

$$distance\left(v(x_i),v(s_i)\right)=0.5+0.5*\cos\theta \tag{5}$$

### 3.4. 基于 K-means 聚类算法的用户兴趣主题提取

通过对文本集中所有文本对应的主题描述聚类成 $k$ 个簇(cluster)，将用户兴趣的提取问题转化为对主题词描述的聚类问题。K-means 聚类算法如算法 1 所示：

**算法 1 K-means 聚类算法**

输入：聚类数目 $k$，文本主题描述
输出：$k$ 个聚类集合 $CL=\{cl_1,cl_2,\cdots,cl_k\}$

1) 设 $DT=\{dt_1,dt_2,\cdots,dt_m\}$ 为文本主题描述集合，个数为 $m$；
2) 指定簇数 $k$；
3) for $i$ in range(0,$m$+1):
$sim_{\max}=0$ //初始化簇心到样本点的距离

**Continued**

$center_{max} = -1$ //初始化簇心点

for $j$ in range(0,$k$+1):

$sim\left(v\left(dt_i\right), v\left(center_j\right)\right) = \dfrac{\boldsymbol{v}\left(dt_i\right) \cdot \boldsymbol{v}\left(center_j\right)}{\left|\boldsymbol{v}\left(center_j\right)\right|\left|\boldsymbol{v}\left(center_j\right)\right|}$ // $v\left(dt_i\right), v\left(center_j\right)$ 分别表示第 $i$ 个文本主题向量和第 $j$ 个聚类对象的向量

if $sim\left(v\left(dt_i\right), v\left(center_j\right)\right) > sim_{max}$ :

$sim_{max} = sim\left(v\left(dt_i\right), v\left(center_j\right)\right)$ //将距离相近的赋值给 $sim_{max}$

$center_{max} = center_j$ //将第 $j$ 个聚类中心对象赋值给 $center_{max}$

add $dt_i$ to $cl_j$

for $j$ in range(0,$k$+1):

$\boldsymbol{v}\left(center_j\right) = \dfrac{1}{\left|N_j\right|}\sum_{i=0}^{N_j} \boldsymbol{v}\left(dt_i\right)$

4) 重复步骤(3)，直到聚类 $cl_1, cl_2, \cdots, cl_k$ 不再发生变化。

## 4. 实验结果及分析

### 4.1. 实验语料与实验环境

为验证本章方法 BERT_LDA 的有效性，本节使用开源的用户新浪微博的数据集来作为实验数据，进一步验证用户兴趣建模方法的有效性。微博用户的数据包括初始微博，评论转发数，以及用户的基本资料和标签等信息。通过数据预处理筛选后，采用 5043 名用户的 115,736 条微博信息作为实验数据集，随机选取 100 名用户的 10,245 条微博信息作为测试集。在进行建模前，有必要对数据进行分词、去除停用词等数据预处理操作。

本文采用 Pycharm & Python 编码工具和具有丰富的 API 数值计算库的机器学习框架进行模型代码的编写，该框架被广泛地应用到各种平台上进行数值计算。本章实验的软硬件环境设置如表 1 所示：

**Table 1.** Experimental environment settings
**表 1.** 实验环境设置

| 项目 | 环境 |
|---|---|
| GPU | Nvidia Geforce GTX 1050 |
| 内存 | 16GB |
| 硬盘 | 1TB |
| 系统 | Windows10 64 |
| Python | 3.6 |
| TensorFlow | 1.7 |

### 4.2. 实验结果分析

#### 4.2.1. 评价指标

用户兴趣模型构建方法采用的指标评估方法是 $F$ 值(F-score)、召回率(Recall)和准确率(Precision)。在分类中，实例通常会被分成正类(Positive)和负类(Negative)，它会以下列四种形式出现：

1) 真正类(True Positive, TP)是指原来样本为正类时，被预测为正类。
2) 假负类(False Negative, FN)是指原样本为正类时，被预测为负类。
3) 假正类(False Positive, FP)是指原样本为负类时，被预测为正类。
4) 真负类(True Negative, TN)是指原样本为负类时，被预测为负类。

具体公式如下所示：

1) 准确率

准确率表示预测出正确的正样本数与总预测正样本中的实际个数之比，记为 $P$，其计算公式如(6)所示：

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

2) 召回率

召回率表示计算模型中预测出正确的样本与所有预测样本之比，记为 $R$，其计算公式如(7)所示：

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

3) $F$ 值

$F$ 值是对模型的综合评价指标的体现，根据准确率和召回率进行计算所得。$F$ 的计算公式如(8)所示：

$$F = \frac{2 * P * R}{P + R} \tag{8}$$

#### 4.2.2. 结果分析

为了验证 BERT_LDA 方法的有效性，分别使用本章提出的 BERT_LDA 方法与基于向量空间模型(VSM)的用户兴趣建模方法[15]和基于 LDA 主题模型的用户兴趣发现方法[16]进行实验对比。主题数 $K$ 均设置成 50，$\beta = 0.75$，先验参数 $\alpha$ 根据主题数 $K$ 进行确认，本章使用数据集中的所有文本来训练 LDA 和 BERT 来提取潜在主题和词向量，实验结果见图 5 所示。

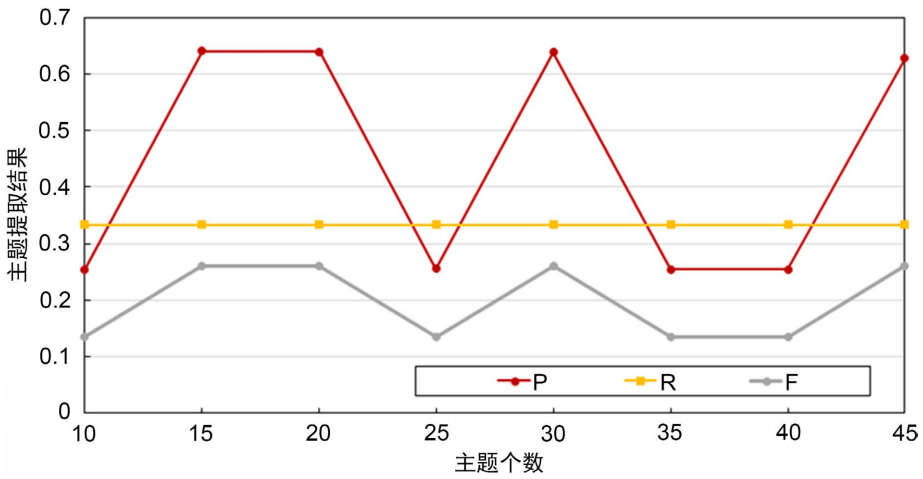


**Figure 5.** Experimental results of different subject numbers
**图 5.** 不同主题数的实验结果

图 5 展示了不同主题数对文本主题提取实验结果的影响，将 BERT 模型的向量维度设置为 128，主题数分别设置范围是 10 到 50。从图中可以得知，主题个数为 15 和 20 时其实验结果比其他情况较好，这是因为微博数据相对较为零散，过多的主题分类会引起细粒度过小。实验中通过选取不同的文档主题

个数，获得最佳的实验效果，可以清楚地看出不同主题个数对 *F* 值等指标所带来的变化。根据实验的最佳状态，确定最佳的主题个数。从表中可以清楚地看出，*F* 值随着不同主题个数的变化情况，发现在主题个数为 15 时，*F* 值达到最高。

基于上述分析，为了进一步验证本章方法对主题提取的综合效果，选取主题个数为 15，通过轮廓系数法确定聚簇数为 18。对 VSM、LDA 和 BERT_LDA 进行聚类分析，如表 2 所示。

**Table 2.** Model comparison results
**表 2.** 模型对比结果

| 模型 | 准确率(*P*) | 召回率(*R*) | *F* |
|---|---|---|---|
| VSM | 0.2981 | 0.0994 | 0.1921 |
| LDA | 0.4026 | 0.2481 | 0.2168 |
| BERT_LDA | 0.6472 | 0.3334 | 0.2620 |

表 2 展示了 VSM、LDA 及 BERT_LDA 的实验对比结果，从表 2 可以看出，*P*、*R* 和 *F* 均有一定的提升。实验对比结果中，我们可以发现，三种用户兴趣建模的方法在 *F* 值上的表现很稳定。对于仅是依赖 TF (词频)和 IDF (逆文档频率)进行词频计算，VSM 模型性能较于 LDA 模型表现的极差，主要因为它们不能够对词项本身存在的语义信息作出很好的判断，用户兴趣主题信息提取性能降低，缺乏对文本上下文的有效信息考量，具有一定的局限性。以 LDA 为对比，本章方法在用户兴趣主题提取上有良好的表现，其模型结合提取方法较为稳定。

通过对三种方法实验结果的比较，从表 2 可以看出，本章方法相对 VSM 模型，可以更好地提取很多用户兴趣主题，*F* 值有明显的提升。而相对于 LDA 主题模型，*F* 值略高于 LDA 主题模型。基于 VSM 模型和基于 LDA 主题模型的用户兴趣建模方法并没有很好地解决语义间关联性低的问题，其效果不佳。而本章方法能够很好地挖掘文本中语义信息，弥补了基于 LDA 主题模型方法的不足，也从侧面证明本章方法的有效性，其可以作为一种用户兴趣建模的有效方法，进一步提升提取用户兴趣信息的准确性。

表 3 展示了前 5 个主题及概率分布最靠前的 10 个词语。

**Table 3.** Mining results of user interest topics
**表 3.** 用户兴趣主题挖掘结果

| Topic1 | Topic2 | Topic3 | Topic4 | Topic5 |
|---|---|---|---|---|
| 住房/0.0042 | 雾霾/0.0532 | 歌手/0.0763 | 穿搭/0.0037 | 公开课/0.0653 |
| 城市/0.0039 | 两会/0.0162 | 明星/0.0457 | 前沿/0.0034 | 新东方/0.0823 |
| 市场/0.0032 | 口罩/0.0324 | 演讲/0.0356 | 时装周/0.0259 | 教育/0.0836 |
| 房产/0.0031 | 环境/0.0053 | 唱歌/0.0593 | 时尚/0.0643 | 补习班/0.0468 |
| 调控/0.0034 | 空气/0.0081 | 比赛/0.0427 | 潮流/0.0368 | 小学/0.0325 |
| 房价/0.0037 | 河北/0.0039 | 综艺/0.0323 | 风尚/0.0056 | 家长/0.0563 |
| 下降/0.0029 | 污染/0.0024 | 真人秀/0.0621 | 彩妆/0.0734 | 孩子/0.0219 |
| 上涨/0.0030 | 投资/0.0021 | 美剧/0.0623 | 卖场/0.0076 | 单词/0.0579 |
| 改革/0.0025 | PM2.5/0.0114 | 娱乐/0.0642 | 美容/0.0731 | 教室/0.0429 |
| 经济/0.0019 | 北京/0.0041 | 芒果台/0.5378 | 护肤/0.0569 | 日语/0.0376 |

从表 3 中可以很清楚地得到每个主题信息，例如，Topic5 中包含的关键词为“公开课”、“补习班”、“教育”等，从中可以很明显地读出 Topic5 讨论的是家庭教育方面的问题。

用户对主题的感兴趣程度与主题的概率分布有关，概率越高，说明用户对此项主题越感兴趣。表 4 展示了部分用户所感兴趣的主题。

**Table 4.** User-topic distribution
**表 4.** 用户－主题分布

| | | | | | |
|---|---|---|---|---|---|
| 用户 1 | 7:0.1369 | 2:0.1221 | 22:0.0852 | 18:0.0803 | 30:0.0758 |
| 用户 2 | 3:0.4193 | 13:0.0863 | 20:0.0645 | 9:0.0617 | 46:0.0524 |
| 用户 3 | 8:0.1250 | 2:0.1072 | 18:0.0893 | 43:0.0869 | 31:0.0723 |
| 用户 4 | 12:0.1410 | 3:0.1025 | 11:0.0897 | 26:0.0892 | 45:0.0618 |
| 用户 5 | 19:0.1704 | 26:0.1131 | 4:0.0743 | 11:0.0862 | 13:0.0640 |
| 用户 6 | 47:0.1481 | 37:0.0945 | 2:0.0921 | 1:0.0819 | 25:0.0581 |
| 用户 7 | 2:0.1279 | 31:0.1263 | 17:0.1046 | 27:0.0937 | 6:0.0798 |
| 用户 8 | 34:0.2823 | 18:0.0758 | 14:0.0722 | 48:0.0795 | 23:0.0587 |

表 4 展示了主题概率分布，其中随机挑选 8 个用户，选取用户比较感兴趣的 5 个主题，表中的兴趣主题主要是由主题编号和主题的概率分布。从表中可以看出，用户 2 感兴趣的主题包括 Topic3、Topic13、Topic20、Topic9 和 Topic46，根据这些信息可以对用户进行微博推荐，帮助用户更好地获取所需知识。

## 5. 结束语

本章提出了一种基于词向量和 LDA 主题模型的用户兴趣建模方法，在模型训练过程中，将 LDA 主题模型和 BERT 模型在处理上下文以及语义相关性上的优势，将其结合起来，从而能够较为有效地挖掘出用户兴趣偏好信息。实验结果表明，本章提出的建模方法的效果优于 VSM 和 LDA 主题模型，有效提高用户兴趣建模的性能。新浪微博等社交网络中包含大量结构化和非结构化数据，仅将用户发表信息作为研究是不够严谨。因此，接下来会引入知识图谱，结合用户行为信息和网络关系，全面地挖掘用户兴趣，实现个性化信息精准推荐。

## 基金项目

国家社会科学基金资助项目(19BXW110)；福建省社会科学规划项目(FJ2017B073)；广西科技师范学院科研基金项目(GXKS2022QN018)。

## 参考文献

[1] 尚燕敏, 曹亚男, 韩毅, 李阳, 张闯. 基于主题和大众影响的用户动态行为倾向预测[J]. 计算机学报, 2018, 41(7): 1431-1447.

[2] 高永兵, 许庆瑞. 基于改进 LDA 模型的微博用户兴趣挖掘研究[J]. 内蒙古科技大学学报, 2019, 38(3): 272-276.

[3] Li, H., Yan, J., Han, W., *et al*. (2014) Mining User Interest in Microblogs with a User-Topic Model. *Communications*, **11**, 131-144. https://doi.org/10.1109/CC.2014.6911095

[4] Chen, Y., Li, W., Guo, W., *et al*. (2015) Popular Topic Detection in Chinese Micro-Blog Based on the Modified LDA Model. 2015 12*th Web Information System and Application Conference* (*WISA*), Jinan, 11-13 September 2015, 37-42. https://doi.org/10.1109/WISA.2015.58

[5] 张霁雯, 周军. 一种基于用户兴趣特征的微博信息转发预测方法[J]. 辽宁工业大学学报(自然科学版), 2021, 41(3): 153-156+160.

[6] 杨仁凤, 陈端兵, 谢文波. 微博用户兴趣主题抽取方法[J]. 电子科技大学学报, 2018, 47(4): 633-640.

[7] Devlin, J., *et al*. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Volume 1, 4171-4186.

[8] Vaswani, A., *et al*. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017, 5998-6008.

[9] 李俊, 吕学强. 融合 BERT 语义加权与网络图的关键词抽取方法[J]. 计算机工程, 2020, 46(9): 89-94.

[10] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993-1022.

[11] 王丹. 基于主题模型的用户画像提取算法研究[D]: [硕士学位论文]. 北京: 北京工业大学, 2016.

[12] Huang, L.F. (2013) Optimized Event Storyline Generation Based on Mixture-Event-Aspect Model. In: *Proceedings of the* 2013 *Conference on Empirical Methods in Natural Language Processing*, ACL, Washington DC, 726-735.

[13] 石元兵. 一种基于 TextRank 的中文自动摘要方法[J]. 通信技术, 2019, 52(9): 2233-2239.

[14] Xue, M. (2019) A Text Retrieval Algorithm Based on the Hybrid LDA and Word2Vec Model. 2019 *International Conference on Intelligent Transportation, Big Data & Smart City* (*ICITBS*) *IEEE*, Changsha, 12-13 January 2019, 373-376. https://doi.org/10.1109/ICITBS.2019.00098

[15] Wang, S., Liu, S. and Liu, Z. (2011) Research of User Interest Model Based on Ordered Pair Behavior. 2011 *IEEE International Conference on Automation and Logistics* (*ICAL*), Chongqing, 15-16 August 2011, 417-421. https://doi.org/10.1109/ICAL.2011.6024754

[16] Liu, Q., Kai, N., He, Z., *et al*. (2013) Microblog User Interest Modeling Based on Feature Propagation. 2013 6*th International Symposium on Computational Intelligence and Design* (*ISCID*), Vol. 1, 383-386.