基于CBOW-LDA主题模型的Stack Overflow编程网站热点主题发现研究

张 景 朱国宾
(武汉大学国际软件学院 武汉 430079)

摘 要 Stack Overflow是一个热门的国外编程问答网站,通过对该网站编程提问帖的问题文本进行文本语义挖掘,能获析用户关注的编程热点。由于研究对象所代表的短文本信息具有高维性及分布不均的特点,易导致主题获取不明晰。文中提出一种基于LDA(Latent Dirichlet Allocation)主题模型的CBOW-LDA建模方法,该方法对目标语料进行相似词聚类后再完成主题建模,能有效降低文本输入维度,使主题分布更明确。采集Stack Overflow网站上2010—2015年的问题帖数据集POST,并对其进行实验,同等主题数下采用文本建模中衡量模型性能的评价指标困惑度(Perplexity)来度量算法在不同数据集容量维度下的性能。结果表明,与现有的基于词频权重的词量化主题建模TF-LDA方法相比,CBOW-LDA方法的困惑度更低,在实验语料下的困惑度降低约4.87%,证明了所提算法的性能更好。采用CBOW-LDA方法对Stack Overflow进行热点挖掘,同时使用TF-LDA方法进行对比实验,建立手工标注的标准评测集对两种方法获取的热门主题和热搜词汇进行查全率、查准率及F1值的判定,结果证实CBOW-LDA表现更佳,其热点挖掘效果较好。由实验结果可知,Java为该编程网站提问帖中最热门的主题,而C和Javascript则为该网站用户提问中被提及得最频繁的词汇。

关键词 Stack Overflow,LDA-CBOW语言模型,主题发现,热门主题,困惑度
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.04.035

Hot Topic Discovery Research of Stack Overflow Programming Website Based on CBOW-LDA Topic Model

ZHANG Jing ZHU Guo-bin
(Department of International Software,Wuhan University,Wuhan 430079,China)

Abstract Stack Overflow is a popular programming question and answer(Q&A) website,we can gather the hot programming knowledge which the developers focus on by studying the programming question text semantic mining. Owing to the high dimensionality problem which hinders processing efficiency and the topic distribution which makes topics unclear,it is difficult to detect topics from a large number of short texts in social network. To overcome these problems,this paper proposed a new LDA(Latent Dirichlet Allocation) model based topic detection method called CBOW-LDA topic modeling method. Using the model to target language and clustering similar words by vectors similarity before topic detection can decrease the dimensions of LDA output and make topics more clearly. Through the analysis of topic perplexity in the experiment dataset with different data collection capacity about the POST on Stack Overflow in 2010—2015,it is obvious that topics detected by our method has a lower perplexity,comparing with word frequency weighing based vectors named TF-LDA. In a condition of same number of topic words from the corpus,perplexity is reduced by about 4.87%,which means CBOW-LDA model performs better. When acting CBOW-LDA method in hot topic on Stack Overflow,TF-LDA method was used to be compared as well,and this paper established a manual annotation standard evaluation set and used Recall,Precision and F1 to contrast experiment results. This paper confirmed that the CBOW-LDA method has better effect because each measuring value of CBOW-LDA is better than TF-LDA,which proves that the hot spot mining effect of CBOW-LDA is good. Through ourexperiment,this paper effectively found out the hot issues of the theme and hot words in nearly six years. This paper drew the conclusion that "Java" is the hottest topic in the website,and "JavaScript" and "C" are the favorite words mentioned in questions from the users.

Keywords Stack Overflow,LDA-CBOW language model,Topic detection,Hot topic,Perplexity

到稿日期:2017-03-21 返修日期:2017-06-11 本文受国家科技支撑计划(2012BAH01F02)资助。
张 景(1992—),女,硕士生,主要研究方向为文本数据挖掘,E-mail:zhangjinghere@whu.edu.cn;**朱国宾**(1965—),博士,教授,博士生导师,主要研究方向为遥感、地理信息系统,E-mail:bgzhu@whu.edu.cn(通信作者)。

1　引言

Stack Overflow 是一个热门的国外编程问答网站,它为用户提供了一个通过发帖来寻求问题解答的平台。帖子问题涵盖大量的编程热点和前沿技术。以 Stack Overflow 问题帖的文本数据为研究对象,在数以万计的提问帖中挖掘热门问题和热搜词汇能获悉用户关注的编程热点,给用户和研究者进行相关领域的信息搜寻和热点研究带来参考和便利。

由于研究对象是该网站的海量文本型提问帖,因此其具有由社交网络短文本特征导致的文本高维性及主题分布不均等问题;并且以概率化词汇抽取为基础的 LDA(Latent Dirichlet Allocation)主题模型有着自身的局限性,导致在对该类文本进行主题挖掘时易存在文本降维难以及主题分布不明的问题。本文提出了一种基于 CBOW-LDA 的主题建模方法,即先采用基于 CBOW 词向量的方法对目标语料进行相似词聚类,再以聚类结果作为输入语料进行后续 LDA 主题模型文本的语义建模,从而有效挖掘出该编程网站提问帖中近六年的热门主题和热搜词汇。实验结果表明,与现有的基于 TF-IDF(Term Frequency-Inverse Document Frequency)的词量化 LDA 方法 TF-LDA 相比,CBOW-LDA 算法在多个不同帖子容量数据集下的对比实验中均表现出更低的困惑度,且在实验语料数据集同等主题数条件下的困惑度降低了约 4.87%,表明 CBOW-LDA 算法具有更好的性能。同时,采用 CBOW-LDA 对 Stack Overflow 进行热点挖掘,并采用 TF-LDA 方法建立对比实验。将两种方法获取到的热门主题和热点词汇与人工标注评测集进行对比,结果表明 CBOW-LDA 方法的查全率、查准率和 F1 值均更优,说明实验的热点挖掘效果合理、可靠,从而证实了 CBOW-LDA 主题模型具有良好的算法性能和实践价值。

2　国内外的研究现状

主题模型是当前文本表示研究的主要范式之一,在文本语义挖掘、机器翻译以及信息推荐等领域中都有着广泛的运用。

在相关研究工作中,Blei 等[1]在 PLSI(Probability Latent Semantic Analysis)的基础上提出了一种能够提取文本隐含主题的非监督学习模型——LDA 主题模型,其实质是一种包含词、主题、文档的三层贝叶斯概率文档主题生成模型。LDA 主题模型由于具有很好的先验概率假设,且简化了模型复杂度,加快了训练速度,因此成为目前应用范围最广的主题模型之一。在后续针对社交网络短文本进行的主题研究中,为了提高主题获取的效率和准确度,多数研究者针对研究对象的特点采取以 LDA 主题模型为基础并结合其他算法框架的方法,亦或是利用通过文本聚类降低语料输入维度的方法来改善主题建模中短文本语义稀疏及主题分布不明确的劣势。在对主题模型算法框架的创新上,Miao 等[2]提出了一种具有成本效益的系统框架,结合选定的用户行为数据与文本主题建模来检测整个主题趋势,以获取 Twitter 等微博服务的主题趋势。Lee 等[3]提出了一种结合主题模型的在线过采样主成分分析(osPCA)算法来有效识别偏离的主题数据。Wu 等[4]提出了一种学习优化 BoW 模型的新方案,将语义相关特征映射到相同词上以提升主题建模的效率。在不规则短文本聚类降维研究方面,Sharifi[5]基于降低短文本的语义维度问题,使用一个句子来概括文本所表达的主题,以使用户能快速地对热点话题进行理解。Lee[6]通过提出的一种在线文本流聚类方法有效地挖掘微博用户关注的热点主题。文献[7]提出了一种将 LDA 与 VSM(Vector Space Model)相结合的方法来研究微博话题发现,该方法基于 TF-IDF 的权重词向量,通过计算文本间的相似度先进行聚类,再进行话题发现。上述方法虽在短文本主题研究中对单一 LDA 建模方法做出了改进,降低了向量的空间维度,提升了处理效率,但在对数据维度特别高的社交网络的较大数据量文本进行主题研究时,仍然存在建模主题易分散和处理效率不高的问题。

针对 LDA 主题模型,基于概率化的单词抽取方法易导致文本主题分散混淆及处理效率不高的问题,本文在研究中采用 CBOW(Continuous Bag-of-Words)模型的词向量方法,先完成相似词的聚类,再进行后续主题建模。CBOW 语言模型是 Mikolov 等[8]于 2013 年提出的一种基于类前馈神经网络的语言模型,该模型利用文本词汇的上下文信息,通过模型训练计算构成词汇间的空间向量模型,通过词向量间的空间相似度来衡量文本语义上的相似度。采用基于 CBOW-LDA 的主题建模方法,引入 CBOW 模型先进行相似词聚类,再进行后续文本主题的建模,能在有效表达语义信息的同时降低模型处理的维度,使得获取的主题分布更准确。

3　CBOW-LDA 算法模型

3.1　Word2vec 架构简介

Word2vec 是 Mikolov 等基于神经网络模型和层次 softmax 方法优化提出的一个架构,可用来快速且有效地训练词向量。Word2vec 包含了两种训练模型,分别是图 1 所示的 CBOW 和 Skip-gram 模型。

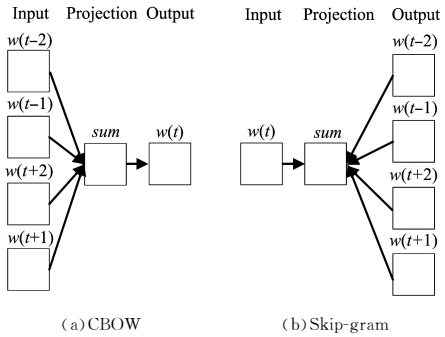

图 1　CBOW 和 Skip-gram 模型

Fig. 1　CBOW and Skip-gram model

由图 1 可知,CBOW 和 Skip-gram 模型均包含输入层、投影层和输出层。其中,CBOW 模型通过上下文来预测当前词,Skip-gram 模型则通过当前词来预测其上下文。Word2vec 提供了两套优化方法来提高词向量的训练效率,分别是梯度下降法(Hierachy Softmax)和负采样(Negative Sampling)。将训练模型和优化方法进行组合,得到 4 种训练词向量的框架,如表 1 所列。

表 1 Word2vec 词向量训练框架

Table 1 Word2vec word vector training framework

模型	CBOW	Skip-gram
Hierachy Softmax	CBOW+HS	Skip-gram+HS
Negative Sampling	CBOW+NS	Skip-gram+NS

为了探究最适合本文研究的训练框架，首先引入模型复杂度的概念[9]。模型复杂度的定义如下：

$$O=E*T*Q \tag{1}$$

其中，E 表示训练的次数，T 表示训练语料中词的个数，Q 因模型而异。T 与训练语料有关，其值越大，模型就越准确。

CBOW 模型的训练复杂度为：

$$Q=N*D+D*|V| \tag{2}$$

Skip-gram 的计算复杂度为：

$$Q=C*(D+D*|V|) \tag{3}$$

其中，N 表示 CBOW 模型输入层的窗口长度，c 表示 Skip-gram 窗口的大小，T 表示训练文本的大小，D 表示发射层维度，$|V|$ 表示训练语料的词典大小，即不同词语的个数。

比较两者复杂度的计算公式可知，Skip-gram 的计算复杂度更高，耗时更长。相比之下，CBOW 的训练速度较快，所以当训练数据很大时，使用 CBOW 技术更省时。通过比较梯度下降和负采样方法可知，CBOW+HS 和 Skip-gram+HS 框架输入层的哈夫曼树的构造过程相对复杂，直接使用梯度下降求解会导致运算复杂度非常高。CBOW+NS 和 Skip-gram+NS框架则采用了一种替代的方法，即采用相对简单的负取样来加快词向量的训练速度。因此，使用负采样的方法能实现简化求解，提高算法效率。

综上，针对本研究对象数据集容量巨大且对整体算法有较高效率要求的特点，选用 CBOW+NS 技术作为研究中 Word2vec 部分的算法框架。

3.2 CBOW 算法模型

CBOW 算法模型是在已知当前词上下文的前提下预测当前词，通过衡量词向量间的空间相似度来完成相似词的聚类。该模型以 Huffman 树为基础，构造好 Huffman 树，初始化完各个向量后就开始输入文本并对其进行训练。训练过程如图 2 所示，主要有输入层(Input)、投影层(Projection)和输出层(Output)3 个阶段。下面以样本($Context(w)$，w)为例(这里假设 $Context(w)$由词 w 的前后 c 个词构成)，对训练过程作简要说明。

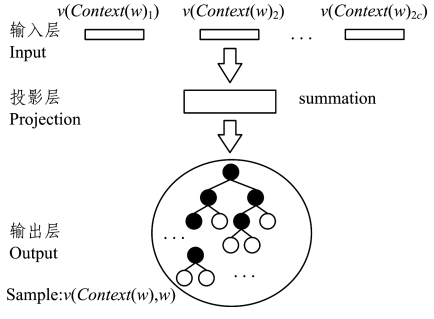

图 2 CBOW 模型的网络结构示意图

Fig. 2 Network structure schematic diagram of CBOW model

1)输入层：包含 $Context(w)$中 $2c$ 个词的词向量 $v(Context(w)_1)$，$v(Context(w)_2)$，…，$v(Context(w)_{2c})\in R^m$，$m$ 表示词向量的长度。

2)投影层：将输入层的 $2c$ 个向量累加求和，即 $X_w=\sum_{i=1}^{2c}v(Context(w)_i)\in R^m$。

3)输出层：对应一棵二叉树，它是以语料中出现过的词为叶子结点、以各词在语料中出现的次数为权值构造出来的 Huffman 树。从根节点开始，映射层的值需要沿着 Huffman 树不断地进行 logistic 分类，并且不断地修正各中间向量和词向量，具体的修正过程参照文献[10]。模型的算法流程表述如下：

```
1. e=0;
2. X_w= Σ_{u∈Context(w)} v(u);
3. FOR j=2:l^w DO
{
    3.1 q=σ(X_w^T θ_{j-1}^w);
    3.2 g=η(1-d_j^w-q);
    3.3 e:=e+gθ_{j-1}^w;
    3.4 θ_{j-1}^w:=θ_{j-1}^w+gX_w;
}
4. FOR u∈Context(w) DO
{
    v(u):=v(u)+e;
}
```

对模型的相关符号解释如下。

p^w：从根结点出发到达词 w 对应叶子结点的路径。

$l(w)$：路径 $p(w)$中包含结点的个数。

σ：sigmoid 函数。

η：学习率，值越大，对中间向量的修正跨度越大。

θ_1^w，θ_2^w，…，$\theta_{l^w}^w\in R^m$：路径 p^w 中非叶子结点对应的向量，θ_j^w 表示路径 p^w 中第 j 个非叶子结点对应的向量。

d_1^w，d_2^w，…，$d_{l^w}^w\in\{0,1\}$：词 w 的 Huffman 编码，它由 l^w-1 位编码构成，d_j^w 表示路径 p^w 中第 j 个结点对应的编码。

p_1^w，p_2^w，…，$p_{l^w}^w$：路径 p^w 中的 l^w 个结点，其中 p_1^w 表示根结点，$p_{l^w}^w$ 表示词 w 对应的结点。

CBOW 模型的优化目标函数取其对数似然函数：

$$L=\sum_{W\in C}\ln P(W(t)|Context(W(t))) \tag{4}$$

根据负采样原理，$\ln P(w_{t+j}|w_t)$模型中公式的表示为：

$$(v_{w(t+j)}^{\mathrm{T}}\cdot v_{w(t)})+\sum_{k=1}^{K}E_{w_k\sim P_v(w)}\ln(\sigma(-v_{w_k}^{\mathrm{T}}\cdot v_{w(t)})) \tag{5}$$

其中，$E_{w_k\sim P_v(w)}$表示 Huffman 树中上下文不出现某个词的期望值，$P_{v(w)}$表示整个语料中词频的分布，w_k 表示该词在 Huffman 树各层中非目标词组的节点向量和。得到 Huffman 树中路径概率最大的词向量后，通过训练整个文本词汇得到最终的词向量集。

3.3 LDA 主题模型

LDA 主题模型本质上是一个包含文档-主题-词汇的三层贝叶斯模型。该模型用来发现文档中的隐含主题，将词向量空间表达的文档约简为主题空间的低维表达。模型的有向概率图如图 3 所示。

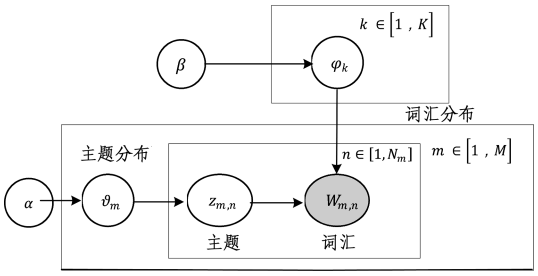

图 3　LDA 主题模型的有向概率图

Fig. 3　Directed probability diagram of LDA theme model

图 3 中,K 为主题个数;M 为文档总数;N_m 是第 m 个文档的单词总数;β 是每个主题下词的多项分布的 Dirichlet 先验参数;α 是每个文档下 Topic 的多项分布的 Dirichlet 先验参数;$z_{m,n}$ 是第 m 个文档中第 n 个词的主题;$W_{m,n}$ 是 m 个文档中的第 n 个词;两个隐含变量 ϑ_m 和 φ_k 分别表示第 m 个文档下的 Topic 分布和第 k 个 Topic 下的词分布,前者是 k 维(k 为 Topic 总数)向量,后者是 v 维向量(v 为词典中 term 总数)。LDA 概率主题模型生成文本的过程如下:

```
// topic plate
  for all topics k∈[1,K] do
  sample mixture components φk~Dir(β)
  end for
  for all documents m ∈[1,M] do
    sample mixture proportion ϑm~Dir(α)
    sample documents length Nm~Poiss(ξ)
  for all words n ∈[1,N] do
    sample topic index zm,n~Mult(ϑm)
    sample term for word wm,n~Mult(φzm,n)
  end for
end for
```

LDA 概率主题模型生成文本的相关解释如下:

1)对于主题 z,根据 Dirichelt 分布 Dir(β)得到该主题上的一个单词多项式分布向量 φ;

2)根据泊松分布 P 得到文本的单词数目 N;

3)根据 Dirichlet 分布 Dir(α)得到该文本的一个主题分布概率向量 ϑ;

4) 对于该文本的 N 个单词中的每一个单词 W_n,从 ϑ 的多项式分布 Multinomial(ϑ)中随机选择一个主题 z,从主题 z 的多项式条件概率分布 Multinomial(φ)中选择一个单词作为 W_n。

4　实验与结果分析

4.1　语料的获取

本文使用基于 sax 的 xml 解析方法,解析采集了来源于文献[11]的 StackOverflow 上 2010－2015 年用户提问帖 POST 数据集中的问题标签 tags 和问题标题 title 文本数据。数据涉及约 400 万用户在 2010 年 1 月－2015 年 6 月这 6 年内发表的约 730 万个问题帖,去除无意义文本后随机抽取 60 万个帖子文本作为实验语料。

为了近一步探究 CBOW-LDA 算法在不同样本容量维度下的性能,从实验语料中分别随机抽取 1 万、5 万、10 万、30 万个帖子文本数据作为对比数据集来进行对比实验,以验证在不同样本容量维度下 CBOW-LDA 算法的主题建模效果。

4.2　评价指标

将自然语言处理中用来衡量语言模型性能的主流评测标准——困惑度(Perplexity)作为评价指标[1]。困惑度取值越小,表示语言模型的吻合度越好,模型性能越好。困惑度的定义公式如(6)所示:

$$Perplexity(W)=\exp\{-\frac{\sum_{m=1}^{M}\ln(p(w_m))}{\sum_{m=1}^{M}N_m}\} \tag{6}$$

其中,W 为测试集,w_m 为测试集文档 m 中可观测到的单词,$p(w_m)$为模型产生文本 w_m 的概率,N_m 为文档 m 的单词数。

4.3　实验设置

4.3.1 实验流程

实验流程如图 4 所示。首先,将解析后的数据集进行分词、去停用词、提取词干等文本预处理;然后,将预处理后的数据集作为语料输入 CBOW-LDA 模型进行文本主题建模;最后通过 Python 和 Matlab 等整合数据,挖掘出热点主题和热搜词汇。

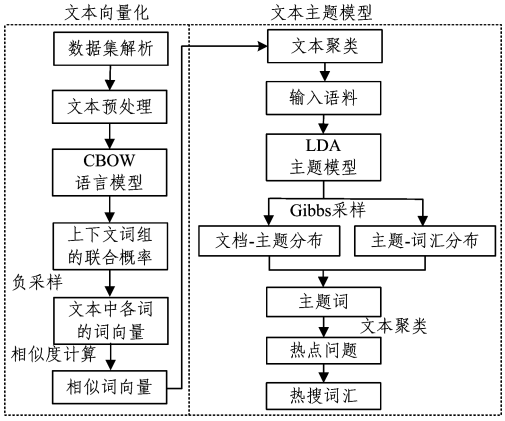

图 4　实验流程图

Fig. 4　Experimental flowchart

4.3.2　CBOW 的参数设置及降维效果

利用 CBOW 语言模型进行文本向量化处理时,设置相关参数如表 2 所列,其中 $CBOW=1$,$NS=1$,$HS=0$ 表示采取的是负采样简化求解方法。

表 2　CBOW 的参数设置

Table 2　Parameter setting of CBOW

参数	含义	数值
Window	上下文窗口大小	5
Sample	高频词亚采样阈值	0
CBOW	是否使用 Cbow 算法	1
NS	是否使用 Negative Sampling 方法	1
HsS	是否采用 Softmax 体系	0

CBOW 算法降维的实质是利用神经网络的方法将原有的大容量文本数据信息转化为向量空间信息后以高维空间点的形式展现出来,再采用基于词向量余弦相似度的算法,通过计算词向量间的余弦距离来判断词语相似度,从而完成相似词的聚类以形成新的簇。将整个文本降维后的输出语料作为 LDA 主题建模的输入语料。

在对词向量进行聚类时,参考文献[12],把相似度的阈值设定为 0.75。分别在实验语料集和不同样本容量的对比数

据集上进行实验，将文本预处理后的文本维数和通过 CBOW 算法训练处理和降维后的文本空间维数进行对比，结果如图 5 所示。

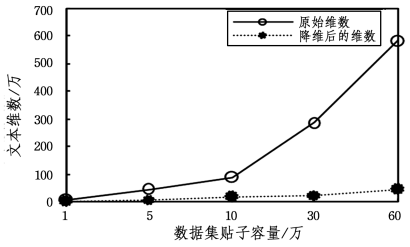

图 5 文本空间维度的变化

Fig. 5 Change of text dimension

由图 5 可知，不同样本容量的文本集在通过 CBOW 模型训练处理之后，文本空间维数均有了大幅下降，获得了较好的降维效果。将包含 60 万帖子的实验文本数据集反复聚类合并，再将每类文本合并，以形成一个新语料集，把这个降维后的新语料集输入 LDA 主题模型进行文本主题建模。

4.4 结果分析

4.4.1 模型的处理效果

将模型参数中的 α 和 β 分别设置为 0.2 和 0.01，Gibbs 抽样迭代次数设为 500，K 为隐含主题词数。选取容量为 60 万个帖子的实验语料集和容量分别为 1 万、5 万、10 万、30 万帖子的对比数据集，在不同数据集容量维度下进行多组对比实验。设置主题数 K 分别取值 5，10，15，20，25，30，35，40，45，50。对比算法选择文献[13]中的基于 TF-IDF 的权重词向量 LDA 方法(简写为 TF-LDA 算法)，在不同样本容量维度和不同主题数条件下，对比两种算法的困惑度，以度量模型的处理效果。CBOW-LDA 和 TF-LDA 算法的处理效果分别如图 6 和图 7 所示。

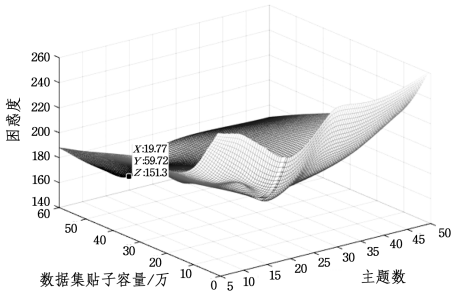

图 6 CBOW-LDA 的困惑度变化

Fig. 6 Perplexity changes of CBOW-LDA

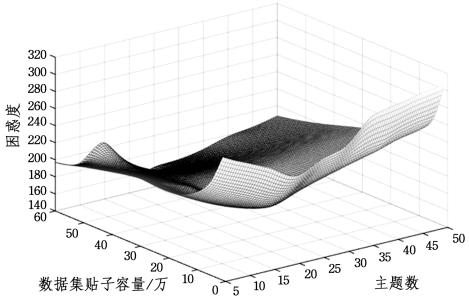

图 7 TF-LDA 的困惑度变化

Fig. 7 Perplexity changes of TF-LDA

由图 6 和图 7 可知，在实验过程中，CBOW-LDA 和 TF-LDA 算法的困惑度随样本容量的扩大而逐步降低，随主题数的增加呈现先降低后增大的变化趋势。其中，在相同的数据集容量和主题数维度变化范围内，CBOW-LDA 算法的困惑度为 151～260，而 TF-LDA 算法的困惑度为 162～302，因此 CBOW-LDA 的困惑度的浮动范围更小；并且，在不同数据集容量和相同主题数条件下，CBOW-LDA 的困惑度皆低于 TF-LDA 算法。综上证明，在不同样本容量和维度下，CBOW-LDA 算法的主题建模性能更佳。

而在最佳主题数的选择上，在相同实验数据集容量和参数设置下，通过分析 CBOW-LDA 模型在不同主题数下的困惑度来确定最佳主题数[1]。在实验语料为 60 万帖子容量的条件下，由图 6 中数据拟合的实验结果可知，CBOW-LDA 主题建模算法的困惑度在 X 取值为 19.77 时取得最小极值 151.3。因主题数需取整，故在本实验条件下的最佳主题数为 20。

为近一步探究在实验语料下 CBOW-LDA 与 TF-LDA 算法的具体性能差异，在帖子容量取 60 万的条件下，比较两种方法的困惑度随主题数目变化的具体情况(见图 8 和表 3)。

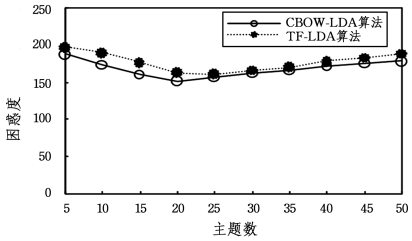

图 8 不同主题数下困惑度变化的比较

Fig. 8 Comparison of perplexity changes in different topic numbers

表 3 两种方法的困惑度

Table 3 Perplexity of CBOW-LDA and TF-LDA

主题数	5	10	15	20	25	30	35	40	45	50
CBOW-LDA	188	173	160	151	156	162	166	171	175	178
TF-LDA	197	189	176	162	160	165	170	178	182	188

通过将二者差值的百分数取平均值得出，在该主题数的范围(5～50)内，相比于 TF-LDA 方法，本文采取的 CBOW-LDA 方法的困惑度降低了约 4.87%。

在相同参数下，将主题数 K 设为最优值 20，将 Gibbs 抽样迭代次数分别设置为 1，100，200，300，400，500，观察困惑度随迭代次数变化的情况，如图 9 所示。由实验结果可知，所提 CBOW-LDA 方法虽然采用了比 TF-LDA 更为复杂的向量化方法，但是收敛速度并没有随迭代次数的增加而减慢，反而表现出较好的响应能力。

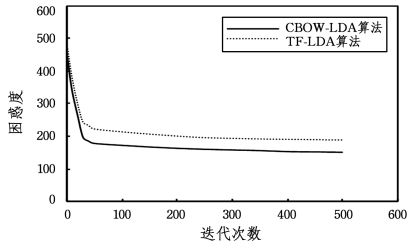

图 9 困惑度随迭代次数变化的情况

Fig. 9 Variation of perplexity when iteration number changes

综合上述多组对比实验可知,在不同样本容量和主题数维度下,CBOW-LDA 方法均展现出比 TF-LDA 更低的困惑度和更好的响应能力,从而证明 CBOW-LDA 算法的性能更佳。

4.4.2　Stack Overflow 热点挖掘

1)语料获取

在主题数 K 取最佳值 20 的设置下,对帖子容量为 60 万的实验语料数据进行基于 CBOW-LDA 主题模型的文本主题建模,得到对应的文档-主题分布和主题-词汇分布。获取包含 20 个不同主题的主题分布,标记前十个出现概率最大的主题和词汇作为热门主题和热搜词汇,并将其数据作为实验评测集。同时,采用基于 TF-LDA 的算法对实验语料进行主题建模,挖掘获取该方法下的 10 个热门主题和 10 个热搜词汇,将该结果作为实验对照集。

采用人工标注的经验验证法建立标准评测集,邀请专家进行手工标注。标注者浏览 2010－2015 年内各时期的相关话题,选取前 20 个热门的主题和热搜词汇进行标注,搜集专家标注结果,并综合整理、平滑数据后进行排序选择,再建立 10 个热门主题和 10 个热搜词汇的标准评测集。人工标注建立的热门主题和热搜词汇的标准评测集如图 10 所示。

热门主题:
Java Ios Android Sql C Python
Javascript 算法设计 编程问题 预处理
热搜词汇:
Java C Javascript Sql Php
Ios Android Data Error Python

图 10　人工标注评测集

Fig. 10　Manual annotation evaluation set

2)评价指标

以信息检索和统计学分类领域广泛采用的度量值-查全率 r(Recall)、查准率 p(Precision)和 $F1$ 值作为评估指标来评价热点挖掘的实际效果。评估参数所涉及到的系数如表 4 所列。

表 4　评估系数释义

Table 4　Definition of evaluation coefficient

	实际属于标准评测集的主题数	实际不属于标准评测集的主题数
判断为属于标准热门主题的主题数	a	b
判断为不属于标准热门主题的主题数	c	d

采用查全率 r 、查准率 p 及 $F1$ 值对此次实验获取的热门主题的效果进行评价,其计算公式如下。

$$r=\frac{a}{a+c} \tag{7}$$

$$p=\frac{a}{a+b} \tag{8}$$

$$F1=\frac{2rp}{r+p} \tag{9}$$

3)结果与分析

采用基于 CBOW-LDA 的主题建模方法对 Stack Overflow 上 2010－2015 年编程提问贴的热门主题和热搜词汇进行热点挖掘。实验获取的前十的热门主题和热搜词汇结果如表 5 所列。

表 5　2010－1015 年 Stack Overflow 的热门问题和热搜词汇

Table 5　Hot questions and words of Stack Overflow from 2010 to 2015

热门主题	主题概率	热搜词	出现次数
Java	0.06101	C	25754
Sql	0.05281	Javascript	24381
编程术语	0.05173	Sql	18969
Android	0.05153	Php	17526
界面操作	0.05134	Android	14716
算法	0.05115	Phone	13916
Ios	0.05104	Data	9928
操作指令	0.05008	File	9462
文本预处理	0.04972	Erro	9336
Java 算法	0.04923	Jquery	5651

采用 CBOW-LDA 和 TF-LDA 算法对实验语料进行热点挖掘,将获取的热门主题和热搜词汇结果与人工标注评测集进行对比,计算两种方法的查全率、查准率和 $F1$ 值,结果如表 6 所列。

表 6　相关评价指的标数值

Table 6　Values of related evaluation indexes

算法	热点挖掘对象	查全率	查准率	$F1$ 值
CBOW-LDA	热门主题	0.75	0.86	0.80
	热搜词汇	0.86	0.86	0.86
TF-LDA	热门主题	0.57	0.67	0.62
	热搜词汇	0.60	0.43	0.50

将实验中两种算法得出的热点挖掘结果与人工标准评测集的结果进行对比,计算得出两种算法的各项评测度。由表 6 可知,CBOW-LDA 方法得到的评价指标中查全率、查准率和 $F1$ 值都高于 TF-LDA 方法,表现出较好的热点挖掘效果,证明实验采取的 CBOW-LDA 方法对 Stack Overflow 热点挖掘的效果较好,具有很好的参考意义。

同时实验结果还表明,Stack Overflow 上 2010－2015 年最热门的问题主题是 Java,问及最频繁的词汇是 C 和 JavaScript。由此可知,2010－2015 年 Stack Overflow 上用户最关注的编程热点为 Java,提问帖中 C 和 Javascript 的出现频率最高。

结束语　本文提出并实现了基于 CBOW-LDA 主题模型的 Stack Overflow 编程网站热点主题发现研究。针对社交网络中短文本信息具有高维性及因分布不均易导致主题获取不明晰的问题,提出采用 CBOW-LDA 模型将文本表示模型的词向量化方法与 LDA 主题模型结合起来进行话题发现和热点挖掘。本文先对模型的输入语料进行词向量化和相似词聚类,再进行主题建模,使得话题抽取模糊度更低,表达更加明确。通过在不同数据容量维度下与 TF-LDA 方法的多组对比实验,验证了本文的 CBOW-LDA 模型具有更好的精准度和稳定性;同时成功挖掘了 Stack Overflow 编程网站上的热门主题和热搜词汇等讯息。最后,将所提方法并与 TF-LDA 方法建立对比实验,并将结果与人工标准评测集进行对比,证明了 CBOW-LDA 方法的查全率、查准率及 $F1$ 值更优,其热点挖掘效果良好。实验结果表明,Java 为该网站提问帖中最热门的主题,C 和 Javascript 为该网站用户提问中最爱提及的词汇。未来将从广度和深度两方面进一步研究和优化模型,细化时间序列,加强话题发现的效果和实用性。

参 考 文 献

[1] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[2] MIAO Z, CHEN K, FANG Y, et al. Cost-Effective Online Trending Topic Detection and Popularity Prediction in Microblogging [J]. Acm Transactions on Information Systems, 2016, 35(3): 18.

[3] LEE Y J, YEH Y R, WANG Y C F. Anomaly Detection via Online Oversampling Principal Component Analysis [J]. IEEE Transactions on Knowledge & Data Engineering, 2013, 25(7): 1460-1470.

[4] WU L, HOI S C H, YU N. Semantics-preserving bag-of-words models and applications [J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2010, 19(7): 1908-1920.

[5] RAMAGE D, DUMAIS S T, LIEBLING D J. Characterizing microblogs with topic models [C] // Fourth International Conference on Weblogs and Social Media. Menlo Park: AAAI Press, 2010: 130-137.

[6] LEE C H, CHIEN T F. Leveraging microblogging big data with a modified density-based clustering approach for event awareness and topic ranking [J]. Journal of Information Science, 2013, 39(4): 523-543.

[7] MIKOLOV T, Language Modeling for Speech Recognition [D]. Brno: Brno University of Technology, 2007.

[8] MIKOLOV T, KOPECYK J, BURGRT L, et al. Neural network based language models for highly inflective languages [C] // IEEE International Conference on Acoustics, Speech and Signal Processing. Taipei: IEEE, 2009: 4725-4728.

[9] TOMAS M, CHEN K, CORRADO G. Efficient estimation of word representations in vector space [EB/OL]. (2013-08-18) [2013-09-07]. http://arxiv.org /abs /1301.3781.

[10] PEGHOTY. word2vec 中的数学原理 [EB/OL]. http://blog.csdn.net/itplus/article/details/37969979.

[11] POST 数据集来源网址 [EB/OL]. http://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede.

[12] GUO L T, LI Y, MU D J. A Method of Topic Discovery Based on LDA Theme Model [J]. Journal of Northwestern Polytechnical University, 2016, 4(1): 698-702.

[13] HUANG B, YANG Y, MAHMOOD A, et al. Microblog Topic Detection Based on LDA Model and Single-Pass Clustering [M] // Rough Sets and Current Trends in Computing. Springer Berlin Heidelberg, 2012: 166-171.

[14] GUPTA M, KUMAR P, BHASKER B. Clustering of users on microblogging social media: A rough set based approach [C] // International Conference on Data Science and Engineering. IEEE, 2017: 1-6.

(上接第 207 页)

[4] LU Y R, WANG F, DENG B. Combined modeling approach for Web application testing [J]. Journal of PLA University of Science and Technology (Natural Science Edition), 2013(6): 617-622.

[5] THUMMALA S, OFFUTT J. An Evaluation of the Effectiveness of the Atomic Section Model [M] // Model-Driven Engineering Languages and Systems. 2014: 35-49.

[6] MARCHETTO A, TONELLA P, RICCA F. State-Based Testing of Ajax Web Applications [C] // International Conference on Software Testing, Verification, and Validation. IEEE Computer Society, 2008: 3-12.

[7] ARORA A, SINHA M. Dynamic content testing of Web Application using user session based state testing [C] // Confluence 2013: the Next Generation Information Technology Summit. 2013: 22-28.

[8] BHANU K, PRASANTH S, MOHAN G K. A Bot Driven Framework for Testing Web Applications [J]. Asian Journal of Information Technology, 2016, 15(20): 3905-3911.

[9] MANE S S, RAJMANE B A. Automatic Testing of AJAX Applications through Dynamic Analysis of User Interface State Change [J]. International Journal of Computer Applications, 2014, 95(11): 12-16.

[10] WANG L N, LI H, ZHAO L. Ajax Web automatic testing model based on simulation of users [J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2016, 44(3): 1-5. (in Chinese)
王丽娜, 李怀, 赵磊. 基于模拟用户的 Ajax Web 自动化测试模型 [J]. 华中科技大学学报(自然科学版), 2016, 44(3): 1-5.

[11] HE T, MIAO H K, QIAN Z S. Modeling and Test Case Generation for Ajax-based Web Application [J]. Computer Science, 2014, 41(8): 219-223. (in Chinese)
贺涛, 缪淮扣, 钱忠胜. 基于 Ajax 技术的 Web 应用的建模与测试用例生成 [J]. 计算机科学, 2014, 41(8): 219-223.

[12] GUO J X, GAO C, XU N S, et al. User Behavior Analysis Based on Web Browsing Log [J], Computer Science, 2014, 41(3): 110-115. (in Chinese)
郭俊霞, 高城, 许南山, 等. 基于网页浏览日志的用户行为分析 [J]. 计算机科学, 2014, 41(3): 110-115.

[13] MESBAH A, VAN DEURSEN A. Migrating Multi-page Web Applications to Single-page AJAX Interfaces [C] // Euromicro Conference on Software Maintenance & Reengineering. 2007: 181-190.