Computer Science and Application 计算机科学与应用, 2019, 9(10), 1831-1838
Published Online October 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.910205

Hans汉斯

Research on Hot News Recommendation Algorithm Based on Topic Model

Ning Zhang, Zhijian Zeng, Lihong Wang, Xufei Yan

Civil Aviation University of China, Tianjin
Email: chunjingxiao@163.com

Received: Sep. 16th, 2019; accepted: Oct. 1st, 2019; published: Oct. 8th, 2019

Abstract

In the rapid development of the Internet today, network news has become the main way for people to get information, how to accurately provide personalized news recommendation for users has become an increasingly concerned problem in the industry. To solve this problem, many news recommendations based on LDA have appeared, but they only analyze the news content without considering the changes of users' interests. To solve this problem, this paper proposed a hot news recommendation algorithm based on topic interest change. Firstly, users' reading history is divided by the fixed time window size, and the probability distribution of users' interest is obtained by LDA according to users' reading history in each stage. Secondly, the time penalty weight function and the news topic distribution of users in each stage are used to predict the possible interest of users in the next stage. Finally, according to the user interest probability distribution, user-based collaborative filtering and topic distribution of news are used to complete hot news recommendation. Experiments on real data sets show that the proposed method improves the recommended performance.

Keywords

LDA, Collaborative Filtering, Topic Model, Recommendation System

基于主题模型的热点新闻推荐算法研究

张　宁，曾知涧，王利洪，燕旭飞

中国民航大学，天津
Email: chunjingxiao@163.com

收稿日期：2019年9月16日；录用日期：2019年10月1日；发布日期：2019年10月8日

文章引用：张宁，曾知涧，王利洪，燕旭飞．基于主题模型的热点新闻推荐算法研究[J]．计算机科学与应用，2019, 9(10): 1831-1838. DOI: 10.12677/csa.2019.910205

摘 要

在互联网高速发展的今天，网络新闻已成为人们获得信息的主要途径，如何准确地为用户提供个性化的新闻推荐已成为业内人士日益关注的问题。为解决这一问题，出现了很多基于LDA的新闻推荐，但它们只进行新闻内容的分析，没有考虑用户兴趣的变化。针对此问题，本文提出了一种基于主题兴趣变化的热点新闻推荐算法。首先，用固定时间窗大小划分用户的阅读历史，并在每个阶段根据用户的阅读历史，利用LDA得到用户兴趣的概率分布。其次，利用时间惩罚加权函数和用户在每个阶段的新闻主题分布预测用户下一阶段的可能兴趣。最后，根据用户兴趣概率分布利用基于用户的协同过滤和待推荐新闻的主题分布完成热点新闻推荐。通过实际数据集上的实验表明，该方法提高了推荐的性能。

关键词

LDA，协同过滤，主题模型，推荐系统

1. 引言

随着互联网的快速发展，人们了解资讯的方式和阅读的习惯都发生了很大的变化，网络新闻阅读已经成为人民日常生活不可缺少的一部分。它给人民生活带来便利，但同时由于参与用户多、资讯源头多也带来了严重资讯过载问题。如何从海量的新闻资讯或讨论话题中挖掘出用户感兴趣的内容变得越来越困难。个性化的热点新闻/话题推荐系统可以有效解决这一问题，为用户提供个性化热点新闻，提高用户的阅读和参与体验。如今日头条、一点资讯等新闻阅读产品都以自己的个性化推荐算法作为吸引用户的卖点。

个性化的热点新闻/话题推荐的关键是根据用户的阅读历史精准刻画用户的个性化阅读兴趣模型，从而为其推荐个性化阅读列表。因为新闻/话题和阅读的历史都具有一定的主题特性，因此很多学者将主题模型引入了新闻推荐过程。郭晓慧[1]在 LDA 模型的主题采样及其分布计算过程中引入平均加权值，提高了主题间的区分度。汤鲲[2]等人将 LDA 与 GRU 相结合应用于群聊会话主题挖掘，解决了传统主题模型不能解决的词语顺序问题。曹巧翔[3]等人针对 web 服务描述文本较短、缺乏足够有效信息的问题，提出了基于 Word2Vec 和 LDA 主题模型的 web 聚类方法。居亚亚[4]等人在 LDA 算法中加入了动态权重，使得在主题语义连贯性、文本分类准确率、泛化性能和精度方面比目前流行的 LDA 推理算法表现得更加优越。王丽苗[5]等人针对短视频喜好率预测面临着用户及广告的数量巨大且训练数据集高维、稀疏等问题，提出了基于 LDA-GBDT-FM 的短视频喜好率预测模型。

这些方法都注重了利用 LDA 进行新闻内容的分析，提高了推荐的准确性，但却没有考虑用户兴趣的变化。因此本文提出了一种基于主题兴趣变化的热点新闻推荐算法，它利用用户阅读历史主题的变化描述用户兴趣的变化，从而预测用户下一时刻的可能的阅读兴趣。该方法首先以一定大小的时间窗划分用户的阅读，根据每个时间段内用户阅读历史利用 LDA 模型，得到用户兴趣的主题分布。其次，利用时间惩罚加权函数和 LDA 模型得到的用户在每个阶段的新闻主题分布，得到用户兴趣。最后利用基于用户的协同过滤和待推荐新闻的主题分布完成热点新闻推荐。通过实际数据集上的实验表明，该方法提高了推

荐的性能。

2. 基于主题模型的用户兴趣变化模型的建立

本节阐述基于 LDA 概率主题分布和时间权值的用户兴趣模型生成。首先以一定大小时间窗划分用户的阅读历史，然后利用时间惩罚权值函数建模用户兴趣的变化，从而预测下个时间窗的概率主题分布以完成推荐。

2.1. 基于用户历史划分的新闻主题分布

LDA (Latent Dirichlet Allocation)是用于挖掘文本隐结构的重要的概率主题模型。本文利用 LDA 挖掘用户新闻阅读历史的隐结构作为用户兴趣[6]。

假设用户集合为$U=\{u_1,u_2,\cdots,u_n\}$，新闻集合$I=\{i_1,i_2,\cdots,i_m\}$，每个用户的以一定时间窗大小划分其新闻阅读历史后的集合为$S=\{S_{u_1},S_{u_2},\cdots,S_{u_n}\}$。对于用户$u_i$的阅读历史为$H_{u_i}$，被划分后表示成$H_{u_i}=\left\{H_{u_i}^{t_1},H_{u_i}^{t_2},\cdots,H_{u_i}^{t_{|u_i|}}\right\}$，其中$|t_{u_i}|$是用户$u_i$整阅读历史被划分的阶段数，$t_{|u_i|}$为最近交互阶段。对目标用户$u_i$的每个阶段$t_l\left(l=1,2,\cdots,|t_{u_i}|\right)$可到其阅读历史$I_{u_i}^{t_l}$，因此在阶段$t_l$用户$u_i$可看成是阅读的新闻集合$I_{u_i}^{t_l}$，每条新闻看成可观测变量$I_{u_i,j}^{t_l}\left(j=1,2,\cdots,N_{t_l}\right)$，$N_{t_l}$是$I_{u_i}^{t_l}$中新闻个数。每个新闻$I_{u_i,j}^{t_l}$从全概率角度用 LDA 可表示成式(1)。

$$P\left(I_{u_i,j}^{t_l}\right)=\sum_{k=1}^{K}P\left(I_{u_i,j}^{t_l}\middle|T_i=k\right)P\left(T_i=k\right) \tag{1}$$

其中K为主题数。

进一步定义$\Phi_{ij}^{t_l}=P\left(I_{u_i,j}^{t_l}\middle|T_i=k\right)$，$\theta_{ij}^{t_l}=P\left(T_i=k\right)$分别表示某一新闻对某一主题的重要性和某一主题对特定用户的重要性。$\Phi_{ij}^{t_l}$和$\theta_{ij}^{t_l}$通过 Gibbs 采样得到已阅读新闻产生的隐变量。因此概率$\Phi_{ij}^{t_l}$和$\theta_{ij}^{t_l}$，分别如式(2)、(3)。

$$\Phi_{ij}^{t_l}=P\left(I_i\middle|T_j,t_l\right)=\frac{C_{ij}^{NK}+\beta}{\sum_{i=1}^{N}C_{ij}^{NK}+N\beta} \tag{2}$$

$$\theta_{ij}^{t_l}=P\left(T_j\middle|u_i,t_l\right)=\frac{C_{ij}^{MK}+\alpha}{\sum_{k=1}^{K}C_{ik}^{MK}+K\alpha} \tag{3}$$

其中，C^{NK}和C^{MK}分别是$N\times K$和$M\times K$大小的矩阵，α和β是$\Phi_{ij}^{t_l}$和$\theta_{ij}^{t_l}$的超参数。α和β的默认值经常取为$50/\mathrm{K}$和 0.01。C^{NK}表示在阶段t_l主题T_j分配给新闻I_i的次数，C_{ij}^{MK}表示在阶段t_l主题T_j分配给用户u_i的次数。

因此，利用 LDA 得到表示用户u_i在t_l阶段的兴趣主题分布$P_{u_i}^{t_l}$，进而得到用户u_i阅读兴趣分布$P_{u_i}=\left\{P_{u_i}^{t_1},P_{u_i}^{t_2},\cdots,P_{u_i}^{t_{|u_i|}}\right\}$。

2.2. 基于时间惩罚权值的用户阅读兴趣变化模型

因为离当前时刻越近的阅读兴趣可能对兴趣预测的重要性越强，而离当前时刻时间越长，影响越小。因此定义了一个取值范围在[0, 1]范围内的指数函数来描述这种重要性变化的递减，如式(4)。

$$f_{u_i}\left(t_n\right)=\mathrm{e}^{-\frac{1}{|t_{u_i}|}\left(|t_{u_i}|-n\right)} \tag{4}$$

其中，$n=1,2,\cdots,\left|t_{u_i}\right|$，$\frac{1}{\left|t_{u_i}\right|}$表示递减率。当$\frac{1}{\left|t_{u_i}\right|}$值越高，比较久的阅读历史影响越小。$\left|t_{u_i}\right|-n$表示当前时刻与历史时刻相差的阶段数，值越大，表示时间越久远，影响越小。因此得到根用户兴趣的表示$P_{u_i}=\left\{f_{u_i}\left(t_1\right)P_{u_i}^{t_1},f_{u_i}\left(t_2\right)P_{u_i}^{t_2},\cdots,f_{u_i}\left(t_{\left|t_{u_i}\right|}\right)P_{u_i}^{t_{\left|t_{u_i}\right|}}\right\}$。用户下一阶段的阅读兴趣表示，如式(5)。

$$P_{u_i}^{t_{\left|t_{u_i}\right|+1}}=f_{u_i}\left(t_1\right)P_{u_i}^{t_1}+f_{u_i}\left(t_2\right)P_{u_i}^{t_2}+\cdots+f_{u_i}\left(t_{\left|t_{u_i}\right|}\right)P_{u_i}^{t_{\left|t_{u_i}\right|}} \tag{5}$$

它是用户阅读历史主题分布的加权表示。

2.3. 热点新闻/话题推荐

得到所有用户在下一阶段的兴趣分布后并计算用户间余弦相似性，得到目标用户的近邻用户。根据近邻用户的阅读历史得到待推荐新闻集合。利用 LDA 得到待推荐新闻的概率主题分布，利用式(6)计算用户兴趣与待选新闻间的相关性，推荐与用户相关性最大的新闻[7]。

$$sim\left(P_{u_i},P_{n_j}\right)=\cos\left(P_{u_i},P_{n_j}\right)=\frac{P_{u_i}\cdot P_{n_j}}{\left|P_{u_i}\right|\left|P_{n_j}\right|} \tag{6}$$

3. 实验与结果分析

3.1. 数据集

实验数据采用的从人民网爬取的 11 万 7 千条用户阅读记录，涉及到 1 万 3 千名用户。其中，每条数据包括了用户编号，新闻编号，浏览时间，新闻标题，新闻详细内容等部分。删除用户阅读记录少于 10 条及数据属性较少或无用信息较多的用户数据。最终，得到清洗之后的数据约有 6 万 5 千条。在试验过程中将每个用户最后五条阅读记录作为测试集，其余作为训练集。

3.2. 评估标准

准确率是推荐算法中最重要的评测数据，它用来衡量算法推荐结果的准确性，表示推荐给用户的资源中有多少比例是用户所接受的物品。计算公式如式(7)：

$$\text{Precision}=\frac{\sum_u\left|R(u)\cap T(u)\right|}{\sum_u\left|R(u)\right|} \tag{7}$$

其中 $R(u)$是推荐给用户 u 的资源集合，$T(u)$是用户实际操作的资源集合。

召回率是推荐算法中另一个重要测评数据，它与准确率一起被合称为精确率。表示用户所接受的资源中有多少比例是算法推荐给用户的资源。召回率计算公式如(8)：

$$\text{Recall}=\frac{\sum_u\left|R(u)\cap T(u)\right|}{\sum_u\left|T(u)\right|} \tag{8}$$

3.3. 参数敏感性分析

在实验中，超级参数 α，β 的取值会对 LDA 模型产生巨大影响，α 在文档中的主题稀疏性中起作用。高 α 值意味着主题稀疏性的影响较小，即预期文档包含大多数主题的混合[8]，而低的 α 值意味着我们希望文档仅涵盖少数主题；β 在主题中的单词稀疏性中起作用。高 β 值意味着词稀疏性的影响较小，即我们期望每个主题将包含语料库的大部分词。首先通过对 α，β 值的研究，找到并选出两者比较合适的数值，

让实验能够获取到最大的准确率。具体情况如图 1 和图 2。

从图 1 可知，随着 α 值得增大，准确率先增大后减小，当 α 值为 0.12 时准确率达到最高，因此在试验过程中将 α 值取为了 0.12。而图 2 则说明，随着 β，值得增大，准确率是缓慢降低的，因此，β 值不宜取较大的值，在试验过程中取为 0.01。

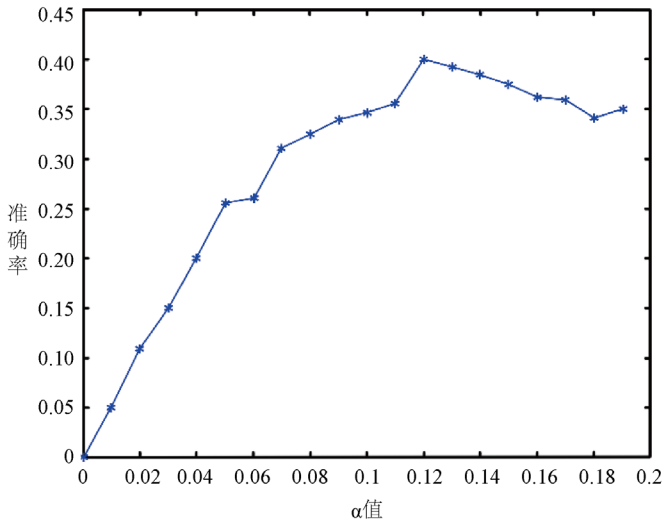

Figure 1. The effect of α on accuracy
图 1. α 值对准确率的影响

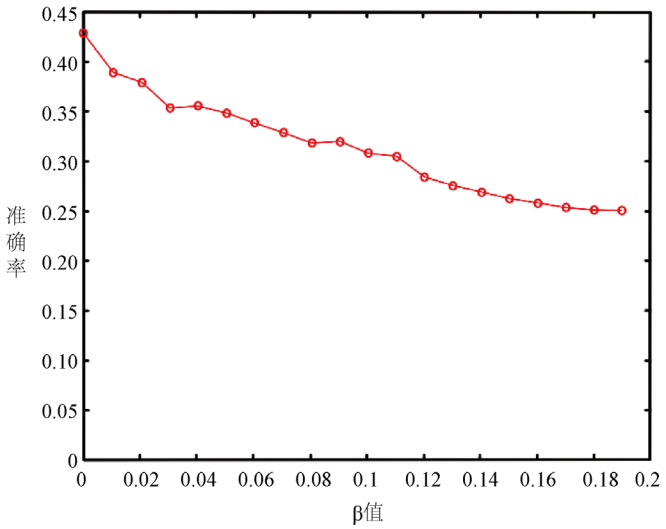

Figure 2. The effect of β on accuracy
图 2. β 值对准确率的影响

对于基于主题模型推荐算法中，主题个数也会对算法性能产生很大的影响。在试验过程中，固定 α、β 的值，讨论了主题个数对算法准确率和召回率的影响，如图 3 和图 4 所示。

图 3 表明，随着主题个数的增加，准确率先增加后减小，而从图 4 可知，在随着主题数的增加，召回率缓慢增加，因此，综合主题数对准确率和召回率存在综合影响。在试验过程中，还讨论了本文 LDA 与协同过滤想结合的算法与 LDA 算法的性能，从图 3 和图 4 可知，本文方法的准确率与召回率都优于普通的 LDA 算法。

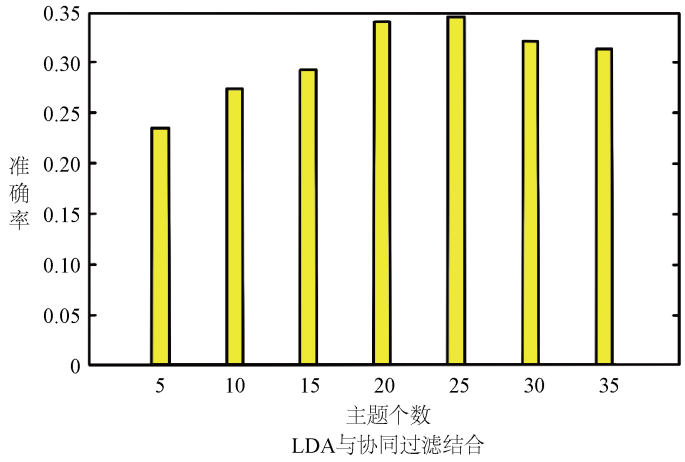

Figure 3. The influence of number of topics on accuracy
图 3. 主题个数对准确率的影响

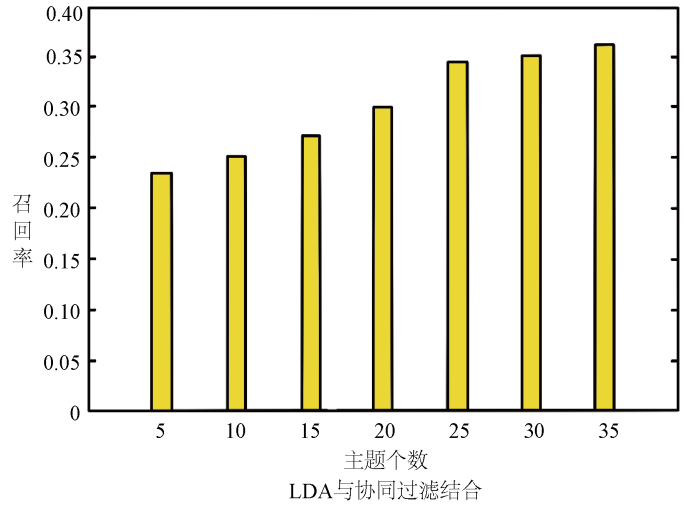

Figure 4. The influence of number of topics on recall rate
图 4. 主题个数对召回率的影响

3.4. 性能评估

为了更好的评估算法性能，实验中分别给出了本方法推荐列表长度分别位 10，20，30 的推荐结果，并与基于 LDA 的方法进行了对比，具体实验结果如表 1 所示。随着推荐列表个数的增多，无论是 LDA 还是协同过滤的 LDA，在准确率和召回率上都有提升，并且协同过滤的效果要好于 LDA 的方法。

Table 1. Comparing the accuracy rate of number of different topics between LDA and LDA-CF
表 1. LDA 与 LDA-CF 不同主题个数准确率从和召回率的对比

	准确率			召回率		
推荐列表个数	10	20	30	10	20	30
LDA	0.22	0.28	0.32	0.24	0.32	0.31
LDA-CF	0.26	0.31	0.37	0.28	0.34	0.33

3.5. 推荐案例

首先，给出一个用户 ID，通过 ID 与其浏览记录进行匹配，图 5 中以“5034018”为例，其实际浏览过的信息有 5 条，其新闻标题分别为：“失联航班乘客家属在吉隆坡机场等待消息”、“波音飞机事故史”、“马航失联航班搜救画面”、“菲律宾派飞机寻找马航失联客机”、“马航在吉隆坡国际机场召开新闻发布会”。

```
from scipy. spatial. distance import pdist
import heapq


uid = '5034018'
for idx, row in df1. iterrows ():
      if str (row ['用户编号']) == uid:
            break

dists = []
# 余弦距离
for index in range (len (df1), len (df)):
     dist = 1 - np. dot (doc_topic [idx], doc_topic [index]) / (np. linalg.norm (doc_topic [idx]) *np. linalg. norm (doc_topic [index])
     dists. append (dist)

array = np. array (dists)
sort_index = np. argsort (array) [:10]
df3=pd.read_csv ('test_data.csv', usecols=['用户编号','新闻编号','新闻标题'])
for i in range (len (sort_index)):
     print ('新闻编号：' *str (df3['新闻编号'] [sort_index [i]]))
     print ('新闻标题：' *str (df3['新闻编号'] [sort_index [i]]))
```

```
新闻编号：100648634
新闻标题：失联航班乘客家属在吉隆坡机场等待消息
新闻编号：100648717
新闻标题：菲律宾派飞机寻找马航失联客机
新闻编号：100648611
新闻标题：马航CEO举行新闻发布会
新闻编号：100648633
新闻标题：马航失联飞机上的劳务工
新闻编号：100648593
新闻标题：失联飞机上有一名孕妇
新闻编号：100651905
新闻标题：马来西亚总理主持的发布会召开
新闻编号：100648582
新闻标题：马来西亚载154名中国人航班失去联系
新闻编号：100651905
新闻标题：马来西亚总理主持的发布会召开
新闻编号：100649314
新闻标题：马航在吉隆坡国际机场召开新闻发布会
新闻编号：100649314
新闻标题：马航在吉隆坡国际机场召开新闻发布会
```

Figure 5. Recommend 10 news results to users
图 5. 向用户推荐 10 个新闻结果

对其浏览记录进行分词，计算分布等操作。最终推荐出 10 个该用户可能感兴趣的新闻信息。其中浏览过的新闻有两条，其标题分别为：“失联航班乘客家属在吉隆坡机场等待消息”、“马航在吉隆坡国际机场召开新闻发布会”

当修改代码，将推荐的新闻改为 20 个，其推荐给该用户的结果如图 6 所示：

在这 20 个新闻推荐中，用户浏览过的记录有 4 条，分别为：“失联航班乘客家属在吉隆坡机场等待消息”、“波音飞机事故史”、“马航失联航班搜救画面”、“菲律宾派飞机寻找马航失联客机”。

新闻编号：100648634
新闻标题：失联航班乘客家属在吉隆坡机场等待消息
新闻编号：100648706
新闻标题：波音飞机事故史
新闻编号：100648717
新闻标题：菲律宾派飞机寻找马航失联客机
新闻编号：100648728
新闻标题：马来西亚总理赴机场慰问失联客机乘客家属
新闻编号：100648752
新闻标题：马航事故处置组抵京召开新闻发布会
新闻编号：100648830
新闻标题：马航召开新闻发布会通报失联航班最新情况
新闻编号：100648915
新闻标题：马航失联航班搜救画面
新闻编号：100648915
新闻标题：马航失联航班搜救画面
新闻编号：100648984
新闻标题：马来西亚民航局举行新闻发布会破“舷窗”谣言
新闻编号：100648706
新闻标题：波音飞机事故史
新闻编号：100648503
新闻标题：昆明暴恐“头七”遇难者家属祭奠亲人
新闻编号：100650156
新闻标题：海军在飞机失联疑似海区打捞到救生衣和油桶
新闻编号：100650338
新闻标题：马航发布会：希望是劫机而非坠毁
新闻编号：100648611
新闻标题：马航CEO举行新闻发布会
新闻编号：100656551
新闻标题：马交通部称飞机已经跌入海底
新闻编号：100656815
新闻标题：青岛公交司机练武防身
新闻编号：100651048
新闻标题：海南一农民用摩托车发动机造飞机
新闻编号：100651215
新闻标题：马航例行记者会
新闻编号：100648598
新闻标题：消失前的马航370
新闻编号：100648915
新闻标题：马航失联航班搜救画面

Figure 6. Recommend 20 news results to users
图 6. 向用户推荐 20 个新闻结果

4. 总结

本文提出了一种基于主题兴趣变化的热点新闻推荐算法。该方法，首先用固定时间窗口大小，来划分用户的阅读记录。再利用 LDA 获取到每个阶段的用户兴趣的分布。通过使用时间惩罚加权函数，来预测用户下阶段可能感兴趣的新闻主题。最后，基于用户的协同过滤和待推荐新闻的主题分布完成热点新闻的推荐。实验结果表明，本文方法在推荐准确率和召回率上都有所提升。

参考文献

[1] 郭晓慧. 基于 LDA 主题模型的文本语料情感分类改进方法[J]. 延边大学学报(自然科学版), 2018, 44(3): 266-273.

[2] 汤鲲, 陈思思. 基于 GRU + LDA 的群聊主题挖掘[J]. 计算机与现代化, 2018(12): 72-76.

[3] 肖巧翔, 曹步清, 张祥平, 刘建勋, 李晏新闻. 基于 Word2Vec 和 LDA 主题模型的 Web 服务聚类方法[J]. 中南大学学报(自然科学版), 2018, 49(12): 2979-2985.

[4] 居亚亚, 杨璐, 严建峰. 基于动态权重的 LDA 算法[J]. 计算机科学, 2019, 46(8): 1-3.

[5] 王丽苗, 许青林, 姜文超, 符基高. 集成 FM 的短视频喜好率预测模型[J]. 计算机工程与应用, 1-5.

[6] 翟航天, 汪学明. 基于隐式反馈 LDA 模型的协同推荐算法研究[J/OL]. 计算机技术与发展, 2019(6): 1-6. http://kns.cnki.net/kcms/detail/61.1450.TP.20190306.0938.054.html

[7] 王宁, 何震, 黄泽, 周毅鹏, 武鑫良. 改进协同过滤算法在服装个性化推荐的研究[J]. 湖南工程学院学报(自然科学版), 2019, 29(1): 33-36.

[8] 张兴宇. 基于协同过滤和内容过滤的微博话题混合推荐算法[J]. 电脑编程技巧与维护, 2019(3): 52-54.