Computer Science and Application 计算机科学与应用, 2018, 8(6), 860-866
Published Online June 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.86095

Hans汉斯

Personalized Movie Recommendation System Based on LDA Theme Extension

Ping Cui, Li Song, Xinkai Yang

College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai
Email: cuiping0515@163.com, xk8yang@163.com

Received: May 24th, 2018; accepted: Jun. 8th, 2018; published: Jun. 15th, 2018

Abstract

Traditional movie recommendation algorithms based on user score data have some problems, such as sparse data and false score information, which cannot really and effectively express user interest. User comments, as an effective carrier of information from users' interests and opinions, can quantify the product features through mining, analyzing and commenting information, and achieve personalized recommendation effect. Site review analysis of text information based on feature film recommended to the users in time according to the user history information analysis. User interest recommendation algorithm is proposed for expanding fusion sentiment analysis and LDA topic model features personalized selection keywords combined with TF-IDF weight item keywords and the characteristics development; the positive rate of sentiment analysis combined with topic comment expands the feature vector for item similarity calculation; the user is interested in the high similarity of the product as the recommended items list recommended. Experiments show that the proposed method improves the accuracy of recommendation.

Keywords

Recommendation System, Topic Model, Sentiment Analysis, Text Mining

基于LDA主题扩展的个性化电影推荐系统

崔　苹，宋　丽，杨新凯

上海师范大学信息与机电工程学院，上海
Email: cuiping0515@163.com, xk8yang@163.com

收稿日期：2018年5月24日；录用日期：2018年6月8日；发布日期：2018年6月15日

摘　要

依据用户评分数据的传统电影推荐算法存在数据稀疏、评分信息不能够真实有效地表达用户兴趣等问题。而

文章引用: 崔苹, 宋丽, 杨新凯. 基于 LDA 主题扩展的个性化电影推荐系统[J]. 计算机科学与应用, 2018, 8(6): 860-866.
DOI: 10.12677/csa.2018.86095

用户评论作为用户兴趣、观点反馈的信息有效载体，通过挖掘分析评论信息可以将产品特征向量化，实现个性化推荐效果。本文利用网站评论文本信息分析电影特征，在给用户推荐时可以根据用户历史评论信息分析用户兴趣，提出一种融合情感分析和LDA主题模型特征拓展的个性化推荐算法，选取主题关键词结合TF-IDF权重进行物品主题关键词特征拓展，情感分析得到的正向评论率结合主题拓展特征向量进行物品相似度计算，用户感兴趣物品相似度较高的产品作为推荐列表进行推荐。通过实验表明本文方法提高了推荐的准确度。

关键词

推荐系统，主题模型，情感分析，文本挖掘

1. 引言

评论信息已成为消费者制定商品购买决策的重要信息来源，且对消费者的商品购买决策影响非常显著[1]，然而，评论信息形式为半结构化或非结构化，商品购买决策时面临前所未有的挑战[2]。推荐系统一般使用户评分数据的推荐算法存在数据稀疏、评分信息不能够真实有效地表达用户兴趣等问题。因此以网站商品评论为物品特征描述语料，利用 word2vec 词向量模型进行评论文本情感分析，获得商品好评率，并筛选商品正向评论文本集，作为商品正向特征的描述文本集，本文使用 LDA 主题模型提取物品主题特征，在此基础上，结合 TF-IDF 提取主题关键词特征，作为物品描述文本信息的主题——关键词特征集，以增强主题粗粒度特征对物品的描述能力，提高推荐的准确率。

2. 相关研究

安悦等人[3]采取的基于内容的热门话题的个性化推荐，首先利用 TF-IDF 的文本表示方法对微博中的文本数据进行量化表示；然后利用相似度计算方法计算微博话题与用户之间的相似度，进而给出个性化的推荐列表。单京晶[4]提出一种联合 K-means 的个性化推荐方法，该方法首先对与用户感兴趣的产品特征进行聚类，将具有相似特征的产品聚到一个类别内，然后将与每个聚类中心点最近的产品推荐给用户。李峰刚等人[5]在新闻文本分类的文本分类算法中使用 LDA 进行特征降维；胡勇军等人[6]利用 LDA 模型将特征维度进行扩展，实现对短文本的分类；Chen 等学者[7]在研究文本间的相似性问题时，使用 LDA 模型计算文本相似，根据文本相似度将搜索记录数据结果集进行分类。Feng 等人[8]基于组合的概率主题模型-用户主题模型(UTM)和随机行走与重启(RWR)方法的组合推荐。UTM 通过利用用户的偏好概况和项目的内容信息来提供用户、组和项目的潜在框架，它们可以更全面地描述组兴趣和项目特征。然后将该潜在框架与 RWR 结合，通过检测综合潜在关系来预测群体对未评级项目的偏好程度。Aslanian 等人[9]提出基于协同过滤和内容推荐的混合推荐算法，介绍了一种新的提取内容特征关系矩阵的方法，然后对协同过滤推荐算法进行了改进，使得该关系矩阵能够有效地集成到算法中。与现有的算法相比，该算法能更好地解决冷起动问题。

3. 基于情感分析和 LDA 主题拓展的推荐算法

3.1. 用户评论文本情感分析

Word2vec 是 Google 公司在 2013 年提出的一款以深度学习算法思想为基础将词语表征转化为实数值

词向量的高效开源工具。主要是利用深度学习方法训练文本内容将词语转化为词向量形式，从而词语语义相似度可以通过向量空间中词向量相似度表示，也可以对文本处理简化为向量空间中词向量运算。

Word2vec 算法主要包含两种语言模型：CBOW (Continuous Bag of Words)和 Skip-gram (Continuous Skip-gram Model)。这两种模型都包括输入层(input)、映射层(projection)、和输出层(output)。CBOW 根据上下文语境预测目标词语，Skip-Gram 根据当前单词预测上下文语境窗口内的词语。

主要步骤：

1) 爬取评论信息经分词、去停用词等预处理；

2) 训练生成 word2vec 词向量模型；

3) 获取每条评论文本词语的词向量，计算每条评论文本词向量均值，做为每条评论文本的文本向量；

4) 将有标记的评论文本向量按 4:1 比例分为训练集和测试集，采用 SVM 分类器，筛选出物品的正向评论、获取物品正向评论率。物品 i 正向评论率 $posr_i$ 为物品 i 正向评论条数 $count_i$，与物品 i 总评论条数 $comments_i$ 比值。

3.2. LDA 主题模型

隐含狄列克雷分配主题模型(LDA: Latent Dirichlet Allocation)其实是基于“文本－主题－词”的三层贝叶斯产生式模型，用 LDA 主题模型对产品的评论内容进行主题词的抽取，用主题维度来代替原来的词项维度，可以较好地降低文本表示的特征维度。本文通过使用主题模型，文本就被投影到 k 个主题上。

矩阵 $T=\begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,K} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N,1} & p_{N,2} & \cdots & p_{N,K} \end{bmatrix}$，其中 $p_{n,k}$ 表示文本 D_n 在主题 T_k 上的概率。

LDA 的图模型结构如上图 1 所示，在 LDA 模型中，一篇文档生成的方式如下：

1) 文档规模的大小服从 Poisson 分布，记作 $N \sim Possion(\xi)$；

2) 文档 D_m 主题分布参数的生成 $\boldsymbol{\theta}_m \sim Dir(\alpha)$，即狄利克雷分布生成文档 D_m 的 K 维主题向量 $\boldsymbol{\theta}_m$，其中狄利克雷分布参数用 α 表示。这个操作需要重复 N 次，生成所有文档主题随机分布；

3) 生成文档特征词 $w_{m,n}$：以根据文档的主题向量 $\boldsymbol{\theta}_m$ 的多项式分布 $Mulitinomial(\theta)$ 选择该词对应的某一隐藏主题 $z_{m,n}$，接着以多项式概率分布 $Mulitinomial(\varphi(z))$ 即 $\boldsymbol{\varphi}_k$，从主题 z 中择某一特征词 $w_{m,n}$。

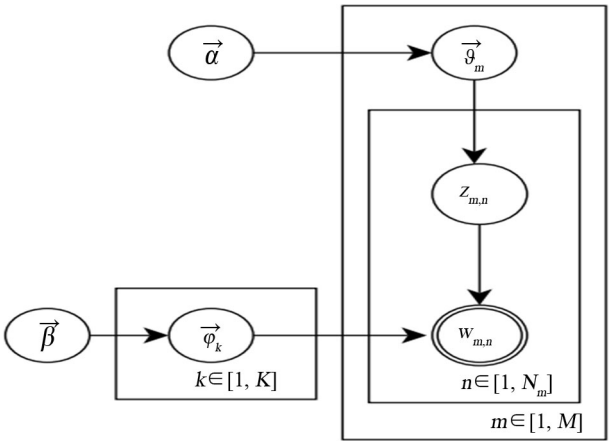

Figure 1. Three layer Bayesian network model diagram of LDA
图 1. LDA 的三层贝叶斯网络模型图

3.3. LDA 主题模型拓展

LDA 主题模型对特征维度较高的评论词进行降维处理，对于评论文本信息，在使用 LDA 主题模型进行降维之后，可以获得每个文本关于主题的概率分布。记 $D_i\left(t_1,t_2,\cdots,t_k\right)$ 为电影 i 评论文本信息主题的概率分布。

如果仅仅使用 LDA 主题模型，虽然对类似的词语进行归类合并，但也只能获取数量有限的 K 个主题来实现对文本内容的概述。这样的特征较为宽泛，粒度较为粗糙。即 $D_i\left(t_1,t_2,\cdots,t_k\right)$ 为电影 i 评论文本信息关于主题的粗粒度特征。如果在对文本进行预处理后，将每个特征词都作为文本的特征，那么就会导致粒度太细致而具有稀疏性。对此，本文通过粗粒度主题特征选取每个主题下 m 个主题关键词作为细粒度特征，即 $D_i\left(w_1,w_2,\cdots,w_{mk}\right)$ 为物品 i 评论文本信息关于主题关键词的细粒度特征。将粗、细粒度特征的结合综合特征记为 $D_i\left(t_1,t_2,\cdots,t_K,w_1,w_2,\cdots,w_{mk}\right)$，可增强文本描述能力。

对评论文本信息 $D=\{d_1,d_2,\cdots,d_n\}$ 进行 TF-IDF 值的计算选取细粒度特征 $D_i\left(w_1,w_2,\cdots,w_{mk}\right)$ 定义一种运算：

$$A\otimes B=R \tag{1}$$

其中 A 是 $m\times n$ 的矩阵，B 为 $n\times1$ 矩阵，R 为 $1\times mn$ 的矩阵。$A_{m,n}*B_n=R_{mn}$。

矩阵 $T_{D_i}=\begin{bmatrix}p_1 & p_2 & \cdots & p_K\end{bmatrix}$，矩阵 $Q_{D_i}=\begin{bmatrix}f_{1,1} & f_{1,2} & \cdots & f_{1,K}\\ f_{2,1} & f_{2,2} & \cdots & f_{2,K}\\ \vdots & \vdots & \ddots & \vdots\\ f_{M,1} & f_{M,2} & \cdots & f_{M,K}\end{bmatrix}$。

其中 $f_{m,k}$ 表示词语 $W_{m,k}$ 在文本 D_i 上的 $tf-idf$ 值。矩阵 $R_{D_i}=\begin{bmatrix}r_1 & r_2 & \cdots & r_{MK}\end{bmatrix}$，其中 $r_{mk}=f_{m,k}*p_k$，则 $Q_{D_i}\otimes T_{D_i}^{\mathrm{T}}=R_{D_i}$，$R_{D_i}$ 即为文本 D_i 的细粒度拓展主题关键词特征向量。

3.4. 个性化产品推荐列表

在用户的诸多信息中，评分是用户最简单、量化的偏好，因此在大量的推荐系统研究中均使用评分矩阵来分析用户对物品的偏好，但是不同用户的评价标准不同，只有在研究特定某一用户时，同一用户对不同物品的评分信息能够直接、有效的反映该用户对不同类型物品的喜好程度。

在给具体某用户推荐物品时还需要考虑该用户历史评分，把用户 $\boldsymbol{u}_i$ 对物品 j 的评分用 $s_{i,j}$ 进行表示，用户历史记录中有打分行为的物品数目为 n，用户 $\boldsymbol{u}_i$ 对物品 j 的评分偏好记为 $w_{i,j}$，则计算公式如下：

$$w_{i,j}=\frac{s_{i,j}}{\sum_{k=1}^{k=n}s_{i,k}} \tag{2}$$

除了以数值形式呈现的评分数据，用户的历史评论数据中也包含用户的偏爱信息，在分析具体某个用户的兴趣偏爱时，由于用户的历史评论数据量一般并不是很庞大，很难精准的分析出用户对物品某个特征的喜爱程度，存在数据稀疏的问题，因此本文分析用户对主题的偏爱，主题包含一类相似的特征，用户主题偏爱能够反映用户对同一主题下物品相似特征的喜爱程度，同时能够有效的解决数据稀疏问题。

用户主题偏爱度的获取方法是将每个用户历史评论集合作为该用户的评论文档，如用户 $\boldsymbol{u}_i$ 的评论文档记作 DU_i。使用 LDA 主题模型将所有用户的评论文档进行降维，可以获得用户评论文档关于主题的 K 维分布向量 $DTU_i=\begin{bmatrix}ut_1,ut_2,\cdots,ut_K\end{bmatrix}$。如果用户无历史评论信息或者评论文本数据过于稀疏的情况下，ut_k 均设置为 1。在 LDA 主题扩展时是将同一物品的评论作为同一文档，研究物品评论主题分布，用户兴趣分析时是将同一用户的评论文本作为同一文档，研究用户评论主题分布，但原始语料是相同的，描述物品也相同，所以主题维度 K 的选择应该等同 3.3 节物品评论的主题维度 K。

推荐系统中物品特征提取、降向量维规格化后，需要计算物品之间相似度，计算不同产品之间相似

性的方法有很多种。如余弦相似性、皮尔森相关系数和欧氏距离等，欧氏距离是直观且被广泛使用，它表示 N 维欧氏空间中两个点之间的距离，本文利用欧氏距离计算项目的相似性。

$$sim(x,y)=\frac{1}{1+\sqrt{\sum(x_i-y_i)^2}} \tag{3}$$

在个性化推荐中目标用户 $\boldsymbol{u}_i$ 历史记录中包含物品 j，其评论文本 D_j 的主题拓展特征向量可以表示为 $\boldsymbol{D}_j=\left(T_{D_j}\middle|R_{D_j}\right)=\left(p_{j,1},p_{j,2},\cdots,p_{j,K},r_{j,1},r_{j,2},\cdots,r_{j,MK}\right)$，同理对目标用户 $\boldsymbol{u}_i$ 历史记录中不包含的物品 i，其评论文本 D_i 的主题拓展特征向量也可以表示为 $\boldsymbol{D}_i=\left(T_{D_i}\middle|R_{D_i}\right)=\left(p_{i,1},p_{i,2},\cdots,p_{i,K},r_{i,1},r_{i,2},\cdots,r_{i,MK}\right)$，物品 i 和物品 j 评论文本主题拓展特征向量的欧几里德距离相似度：

$$sim_{ij}=\frac{1}{1+d\left(\boldsymbol{D}_i,\boldsymbol{D}_j\right)}=\frac{1}{1+\sqrt{\sum_{s=1}^{s=K}\left(p_{i,s}-p_{j,s}\right)^2+\sum_{s=1}^{s=MK}\left(r_{i,s}-r_{j,s}\right)^2}} \tag{4}$$

已知用户 $\boldsymbol{u}_i$ 的主题偏爱向量为 $DTU_i=[ut_1,ut_2,\cdots,ut_K]$，用户 $\boldsymbol{u}_i$ 对物品 j 的评分偏好记为 $w_{i,j}$，在为用户 $\boldsymbol{u}_i$ 进行个性化推荐时，在物品评论文本相似度计算中加入用户主题偏爱和评分偏好权重后的物品相似度计算公式为：

$$sim_{ij}=w_{i,j}\frac{1}{1+\sqrt{\sum_{s=1}^{s=K}ut_s\left(p_{i,s}-p_{j,s}\right)^2+\sum_{s=1}^{s=MK}\left(r_{i,s}-r_{j,s}\right)^2}}\cdot posr_i \tag{5}$$

其中 $posr_i$ 是物品 i 的正向评论率，是物品 i 正向评论条数与其总评论条数比值。

4. 实验及结论

实验数据采用爬虫技术从豆瓣电影网站(https://movie.douban.com)采集共 1000 部电影，8,233,422 条评论信息，并获取用户对电影评分数据。其中部分电影及用户评论信息分别如表 1、表 2 所示。实验训练测试样本 4:1，为 300 名目标用户推荐 N 部电影，与评测数据对比，计算推荐准确率、召回率信息。

Table 1. Part of films' information
表 1. 部分电影信息

ID	Movie_id	Movie_name	Comments	Rate
1	26260853	速度与激情 8/狂野时速 8(港)/玩命关头 8(台)	353,869	7.4
2	1292213	仙履奇缘/齐天大圣西游记	262,669	9.2
3	10574468	北京遇上西雅图/美丽有缘/情定西雅图	129,491	7.2

Table 2. Part of users' information
表 2. 部分用户评论信息表

ID	Username	Movie_id	Comment	Rate
1	布兰切特	26260853	喜爱大场面的千万不要错过这部汽车版的釜山行。	5
2	掉线	26260853	速度与激情系列狂热追捧	4
3	Empty	1292213	有内涵有深度的电影！	5
4	中国愤青	1292213	每一副画面都是经典	5
5	岚色 p 番茄	10574468	很适合和闺蜜或情人在电影院看。电影院全场笑	5
6	北漂青年	10574468	很温暖的爱情，平淡却又真实	5

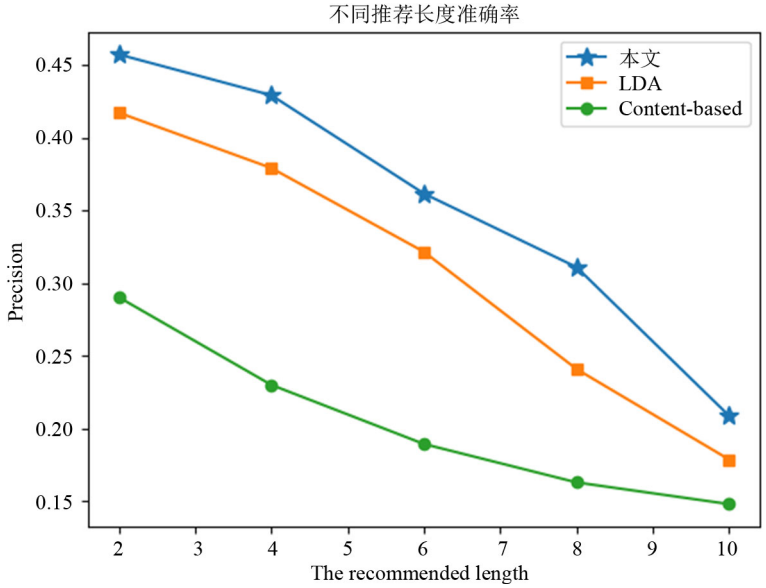

Figure 2. Recommended precision under different recommended lengths
图 2. 不同推荐长度下推荐准确率

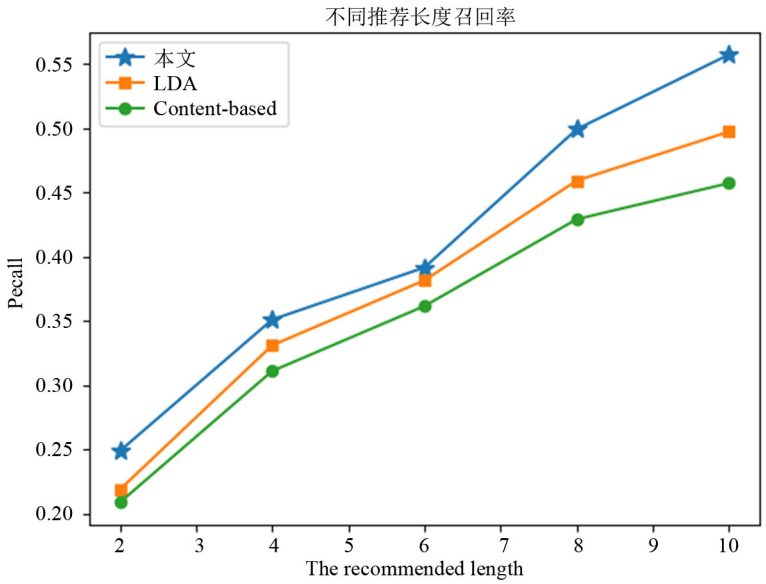

Figure 3. Recommended recall under different recommended lengths
图 3. 不同推荐长度下推荐召回率

本文算法与基于内容的推荐算法和基于 LDA 主题模型的推荐算法做比较。Top-N 不同推荐长度下推荐准确率、召回率对比分别如图 2、图 3 所示，在比较不同算法在相同的推荐长度下，采用本文基于评论文本 LDA 主题扩展的个性化推荐算法推荐相对基于评论 LDA 主题推荐和基于内容的推荐算法在推荐长度相同的情况下，准确率和召回率均有所提高。可见用户评论文本中包含有效的用户观点，同时基于 LDA 推荐算法和 LDA 主题扩展推荐算法之间的比较也能说明主题扩展特征词加强了文本的描述力，提高了推荐的准确性。

参考文献

[1] Fang, B., Qiang, Y., Kucukusta, D., *et al.* (2016) Analysis of the Perceived Value of Online Tourism Reviews: Influ-

ence of Readability and Reviewer Characteristics. *Tourism Management*, **52**, 498-506. https://doi.org/10.1016/j.tourman.2015.07.018

[2] Lee, Y.J., Hosanagar, K. and Tan, Y. (2015) Do I Follow My Friends or the Crowd? Information Cascades in Online Movie Ratings. *Management Science*, **61**, 2241-2258. https://doi.org/10.1287/mnsc.2014.2082

[3] 安悦, 李兵, 杨瑞泰, 胡沥丹. 基于内容的热门微话题个性化推荐研究[J]. 情报杂志, 2014(2): 155-160.

[4] 单京晶. 基于内容的个性化推荐系统研究[D]: [硕士学位论文]. 长春: 东北师范大学, 2015.

[5] 李锋刚, 梁钰, Gao, X., 等. 基于 LDA-WSVM 模型的文本分类研究[J]. 计算机应用研究, 2015, 32(1): 21-25.

[6] 胡勇军, 江嘉欣, 常会友. 基于 LDA 高频词扩展的中文短文本分类[J]. 现代图书情报技术, 2013(6): 42-48.

[7] Chen, M., Jin, X. and Shen, D. (2011) Short Text Classification Improved by Learning Multi-Granularity Topics. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Barcelona, Catalonia, 16-22 July 2011, 1776-1781.

[8] Feng, S., Cao, J., Wang, J., *et al.* (2016) Group Recommendations Based on Comprehensive Latent Relationship Discovery. *IEEE International Conference on Web Services*, San Francisco, CA, 27 June-2 July 2016, 9-16.

[9] Aslanian, E., Radmanesh, M. and Jalili, M. (2016) Hybrid Recommender Systems Based on Content Feature Relationship. *IEEE Transactions on Industrial Informatics*, 1-10.