

LDA-RR:一种基于评分和评论的推荐方法

王 建 黄佳进

(北京工业大学电子信息与控制工程学院 北京 100124)

摘 要 推荐系统是解决互联网信息过载问题的有效途径之一,其中具有代表性的是协同过滤推荐。传统的协同过滤推荐方法只考虑评分信息,而评论信息则包含了用户和物品更具体的特征信息。使用主题模型 LDA 并结合评分信息和评论信息,提出了一种基于用户改进的 LDA 算法。假设每个用户下隐含着主题分布,主题下隐含着物品分布,同时词语的分布由主题和物品共同决定,该算法根据潜在主题分布挖掘用户兴趣进而完成推荐。实验结果表明,改进的算法有效提升了推荐质量。

关键词 推荐系统,信息过载,协同过滤,主题模型

中图分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.02.044

LDA-RR: A Recommendation Method Based on Ratings and Reviews

WANG Jian HUANG Jia-jin

(School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China)

Abstract Recommender system is one of the effective ways to solve the problem of information overload. The collaborative filtering method is a typical method of recommender systems. The traditional collaborative filtering algorithm only takes rating information into account, while reviews contain more specific characteristic information about users and items. In this paper, we proposed an improved LDA algorithm which can combine ratings with review opinions of users. We assumed that each user has an implicit topic distribution, each topic has an implicit item distribution, and the distribution of words is determined by the topic and the item, then we used the potential topic distribution to mine user's interests and make recommendations. The experiment shows that our algorithm can effectively improve the recommendation quality.

Keywords Recommendation system, Information overload, Collaborative filtering, Topic model

1 引言

随着互联网的迅猛发展,信息在给人们的生活带来方便的同时,也产生了信息过载的问题^[1],大量的无关信息降低了人们获取信息的效率。对于电子商务系统,如何利用推荐算法解决信息过载问题并帮助用户找到喜欢的物品显得十分重要。

协同过滤算法假设相似的用户具有相似的偏好^[2],将用户划分为兴趣相投的群体,进而相互协助推荐物品。但是传统的协同过滤只是在用户评分中进行兴趣建模,而商品的特征是多元的,不同用户看重的特征也不尽相同,这就造成对同一个商品评分很高的用户的偏好可能并不相同的问题。而基于评论的推荐算法可以很好地解决这个问题^[3]。

用户的评论中包含了商品或者服务的特征信息,而基于评论的推荐算法可以在非监督的条件下对商品或者服务进行

分类,并且抽取其具体特征^[4]。主题模型 LDA (Latent Dirichlet Allocation)^[5]是广泛应用于评论推荐领域的一个非监督的三层贝叶斯文档生成模型。文献[6]将 LDA 算法引入到对评分矩阵的兴趣建模中,把物品看作词语,把每个用户购买的物品集合看作文档,构造了用户、主题、物品三级模型,实现了 iExpand 算法。文献[7]则结合 iExpand 算法和矩阵分解算法,提出了两步推荐框架 HTMMF。HTMMF 算法模拟了用户为物品评分的行为过程,分别预测了用户为一个物品评分的概率和评分的值,它们的乘积作为用户评分的期望,从而根据评分期望生成推荐列表。但是, iExpand 算法和 HTMMF 框架只考虑了评分信息,忽略了评论信息。基于此,本文在原始 LDA 模型中加入物品,并结合评分信息和用户评论提出了基于用户评论的推荐方法 LDA-RR (A Recommendation Method Based on Ratings and Reviews)。LDA-RR 假设每个用户下隐含着主题分布,主题下隐含着物品分布,同时词

到稿日期:2015-12-10 返修日期:2016-04-19 本文受国家自然科学基金(61562009)资助。

王 建(1989—),男,硕士生,主要研究方向为数据挖掘、信息检索,E-mail: wj1989@emails.bjut.edu.cn;黄佳进(1977—),男,博士,助理研究员,主要研究方向为数据挖掘、信息检索。

类别筛选。数据集包括商品编号、用户编号、购买时间、顾客评分、对商品的评论等信息,大约 520000 条记录。为了避免数据集过于稀疏,数据集保留了评论记录多于 7 条的用户以及被评论记录多于 6 条的商品。同时,数据集集中的评论去除了停用词和出现次数少于 5 的词语,以及文字评论少于 20 个词语的记录。使用 PorterStemmer 对词语进行还原处理后,数据集包含 2754 个用户和 845 个商品,评价记录大约 40000 条。详细统计信息见表 2。

表 2 数据集信息	
亚马逊数据集	
用户	2754
商品	845
评论	40179
稀疏度	99.1%

实验中数据集被随机分成 5 份,每次取其中一份作为测试集,其他 4 份作为训练集,共进行 5 次交叉实验,分别记为 Data_1,Data_2,Data_3,Data_4 和 Data_5。为了评价算法的有效性,采用两种排序预测推荐算法(包括原始 LDA 算法和基于 LDA 算法的 iExpand 算法)与本文提出的 LDA-RR 算法进行对照实验。采用获得 Top-N 推荐列表后,计算准确率、召回率和 F1 指标来评价推荐效果。准确率(P)、召回率(R)和 F1 衡量指标^[12]如式(4)所示:

$$\begin{aligned} P &= \frac{|hits|}{|recset|} \\ R &= \frac{|hits|}{|testset|} \\ F1 &= \frac{2 \times P \times R}{P + R} \end{aligned} \quad (4)$$

其中,hits 表示推荐物品命中个数,recset 表示推荐物品总数,testset 表示命中物品数在理论上的最大值。

下面将展示实验的实施过程和实验结果,具体主要包括:
1)3 种 LDA 算法中参数的设置;2)3 种算法交叉对比实验的结果展示和分析。

3.1 参数设置

原始 LDA 算法是以用户评论作为语料库进行用户兴趣建模的,语料库中主题分布 θ 和主题中词语的分布 φ 都是 Dirichlet 分布,由超参 α 和 β 决定。iExpand 算法是传统 LDA 算法的一个引申运用,是物品、主题、用户的三级模型,把用户-物品的评分表中购买同一个物品的用户的集合看作文档,将用户看作词语,所有文档的集合作为语料库,对用户的兴趣进行建模。而本文提出的 LDA-RR 算法在 LDA 算法的基础上假设由主题和物品共同决定词语的选择,实验结果也表明这种方法是可行的。在下面的参数优化中,以 LDA-RR 算法为例来介绍参数的调整过程。实验中,假设吉布斯采样过程中 3 个超参 α, β_1 和 β_2 的初始值均为 0.5,主题个数 K 的取值范围为 $[10, 110]$,吉布斯采样次数为 1000,Top-N 中 N 取 10。

实验过程中,首先在 5 次划分的数据集上针对不同主题数量进行参数优化,确定不同主题数量下的最优参数,然后用最优参数进行交叉验证,以便对推荐结果进行客观地评价。5 次交叉验证中每次对每个主题进行 3 次实验,以 15 次实验结

果的平均值作为最终结果。不同主题下推荐效果的 F1 指标如图 3 所示。可以看出,随着主题数量 K 的增加,推荐效果越来越好,当 K 大于 30 后推荐效果明显提升,当 K 为 70 时推荐效果最佳,此时 $\alpha=0.04, \beta_1=0.01, \beta_2=0.03$ 。为了比较的公平性,调整对比算法的参数得到最优结果,此时原始 LDA 算法中 $k=50, \alpha=0.2, \beta=0.003$;iExpand 算法中 $K=50, \alpha=0.03, \beta=0.1$ 。

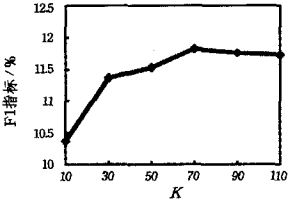


图 3 LDA-RR 指标 F1 随主题 K 的变化

3.2 实验结果对比

在选定最优参数后,一共进行 5 次交叉验证,每个数据集重复进行 3 次实验,每个实验在 Top-10 的推荐列表下进行考察,最终对每个数据集的结果取平均值。3 种算法的召回率对比结果如图 4 所示。为了更好地展示算法的准确率对比结果,在不同划分的数据集上选取原始 LDA 算法的准确率作为基准(设为 0),展示 LDA-RR 算法和 iExpand 算法在推荐准确率上与前者的相对增益,对比结果如图 5 所示。

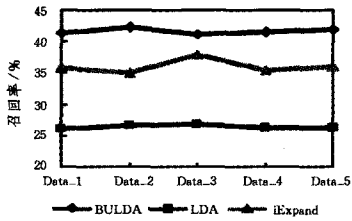


图 4 召回率

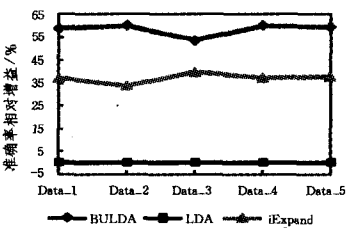


图 5 准确率相对增益

原始 LDA 算法只考虑了文本评论,而 LDA-RR 算法整合了评分和文本评论,利用用户对物品的评分作为其评论的加权。在图 4 和图 5 中,与原始 LDA 算法相比,LDA-RR 算法的召回率和准确率分别提高了 57.2% 和 58.3%,表明采用引入评分加权使得用户偏好的物品出现次数增加且用户感兴趣的物品的属性权重也增加的方法有助于提高推荐质量。同时,LDA-RR 在原始 LDA 模型中使用评分对评论进行加权,也在一定程度缓解了由于数据集稀疏造成的单个用户下评论信息少造成的影响。

iExpand 算法只是用评分对用户购买物品的出现次数进行加权。而引入评论信息的 LDA-RR 算法与 iExpand 算法

(下转第 305 页)

- sion for face recognition [J]. IEEE Trans. PAMI, 2010, 32(11): 2106-2112.
- [3] ZHANG L, YANG M, FENG X C. Sparse representation or collaborative representation which helps face recognition[C]//ICCV. 2011.
- [4] DING C X, TAO D C. A comprehensive survey on pose-invariant face recognition [R]. arXiv:1502.04383, 2015.
- [5] CHAI X J, SHAN S G, CHEN X L, et al. Locally linear regression for pose-invariant face recognition[J]. IEEE Trans. on Image Processing, 2007, 16(7): 1716-1725.
- [6] SHARMA A, JACOBS D W. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch[C]//CVPR. 2011.
- [7] LI A N, SHAN S G, GAO W. Coupled bias-variance tradeoff for cross-pose face recognition [J]. IEEE Trans. on Image Processing, 2012, 21(1): 305-315.
- [8] ZHANG H C, ZHANG Y N, HUANG T. Pose-robust face recognition via sparse representation [J]. Pattern Recognition, 2013, 46(5): 1511-1521.
- [9] TAI Y, YANG J, LUO L, et al. Orthogonal Procrustes Problem based Regression with Application to Face Recognition with Pose Variations[C]//Isdc. 2015.
- [10] GREEN B. The orthogonal approximation of an oblique simple structure in factor analysis [J]. Psychometrika, 1952, 17: 429-440.
- [11] GOWER J, DIJKSTERHUIS G. Procrustes problems [M]. Oxford University Press, 2004.
- [12] VIKLANDS T. Algorithms for the weighted orthogonal procrustes problem and other least squares problems [D]. UMEA University, 2006.
- [13] WAGNER A, WRIGHT J, GANESH A, et al. Towards a practical face recognition system: robust alignment and illumination by sparse representation [J]. IEEE Trans. PAMI, 2012, 34: 372-386.
- [14] YU L B, ZHANG M, DING C. An efficient algorithm for L1-norm principal component analysis[C]//ICASSP. 2012.
- [15] YANG A, GANESH A, ZHOU Z, et al. Fast l1-minimization algorithms and application in robust face recognition [R]. UC Berkeley, 2010.
- [16] SIM T, BAKER S, BSAT M. The cmu pose, illumination, and expression database [J]. IEEE Trans. PAMI, 2003, 25(12): 1615-1618.
- [17] GROSS R, MATTHEWS I, COHN J, et al. Multi-pie [J]. Image and Vision Computing, 2010, 28: 807-813.

(上接第269页)

相比,在召回率和准确率上分别提高了15.5%和27%。由此可见,充分挖掘评论中的信息有助于提高推荐的准确度。

结束语 在推荐系统中,充分利用用户的评论信息,挖掘用户潜在的兴趣模型,进而提高推荐的准确度,是一个值得深思的问题。针对传统的协同过滤推荐系统只考虑了评分信息的问题,本文结合商品的评论信息和用户的评分信息,提出了基于用户的改进的LDA算法,该算法对系统的正确推荐起到了积极作用。但是考虑到之前对数据集的处理工作,如何改进本文的算法使其在稀疏数据集中达到较好的推荐效果是下一步工作的重点。

参考文献

- [1] WANG H N, LU Y, ZHAI C X. Latent aspect rating analysis on review text data: a rating regression approach[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2010: 783-792.
- [2] MA H W, ZHANG G W, LI P. Survey of collaborative filtering algorithms [J]. Journal of Chinese Computer Systems, 2009, 30(7): 1282-1288. (in Chinese)
马宏伟, 张光卫, 李鹏. 协同过滤推荐算法综述[J]. 小型微型计算机系统, 2009, 30(7): 1282-1288.
- [3] YAO W, ESTER M. FLAME: A Probabilistic Model Combining Aspect Based Opinion Mining and Collaborative Filtering[C]//Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, 2015: 199-208.
- [4] TITOV I, MCDONALD R. Modeling online reviews with multi-grain topic models[C]//Proceedings of the 17th International Conference on World Wide Web. ACM, 2008: 111-120.
- [5] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993.
- [6] LIU Q, CHEN E, XIONG H, et al. Enhancing collaborative filtering by user interest expansion via personalized ranking. [J]. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society, 2012, 42(1): 218-233.
- [7] ZHAO X Y, NIU Z D, CHEN W, et al. A hybrid approach of topic model and matrix factorization based on two-step recommendation framework [J]. Journal of Intelligent Information Systems, 2015, 44: 335-353.
- [8] JINDAL N, LIU B. Opinion spam and analysis[C]//International Conference on Web Search & Data Mining. 2008: 219-230.
- [9] XU G, WANG H F. The development of topic models in natural language processing [J]. Chinese Journal of Computers, 2011, 34(8): 1422-1436. (in Chinese)
徐戈, 王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报, 2011, 34(8): 1422-1436.
- [10] KONSTANTIONS C, GREGORIS M. A topic-based recommender system for electronic marketplace platforms[J]. Expert Systems with Applications, 2013, 40(11): 4370-4379.
- [11] ZHU J, HU W B. Recent Advances in Bayesian Machine Learning [J]. Journal of Computer Research and Development, 2015, 52(1): 16-26. (in Chinese)
朱军, 胡文波. 贝叶斯机器学习前沿进展综述[J]. 计算机研究与发展, 2015, 52(1): 16-26.
- [12] JANNACH D, ZANKER M, FEFERNING A, et al. Recommender Systems: An Introduction[M]. 蒋凡, 译. 北京: 人民邮电出版社, 2013: 118.