文章编号:1000-5641(2015)03-0080-11

基于评论分析的评分预测与推荐

高祎璠, 余文喆, 晁平复, 郑芷凌, 张 蓉
(华东师范大学 数据科学与工程研究院 上海高可信计算重点实验室,上海 200062)

摘要: 推荐系统广泛地应用在网络平台中,推荐模型需要预测用户的喜好,帮助用户找到适合的电影、书籍、音乐等商品.通过对用户评分和评论信息的分析,可以发现用户关注的商品特征,并根据商品的特征,推测用户对该商品的喜好程度.本文提出将评论中隐含的语义内容与评分相结合,设计并实现了一种新颖的商品推荐模型.首先利用主题模型挖掘评论文本中隐含的主题分布,用主题分布刻画用户偏好和商品画像,在逻辑回归模型上训练主题与打分的关系,最终评分可以被视为是对用户偏好和商品画像的相似程度的量化表示.最后,本文在真实数据上进行了大量对比实验,结果证明该模型比对比系统性能优越且稳定.

关键词: 推荐; 潜在主题; LDA; 回归模型; 评论分析
中图分类号: TP391 **文献标识码**: A **DOI**:10.3969/j.issn.1000-5641.2015.03.010

Analyzing reviews for rating prediction and item recommendation

GAO Yi-fan, YU Wen-zhe, CHAO Ping-fu, ZHENG Zhi-ling, ZHANG Rong
(*Institute for Data Science and Engineering, Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai* 200062, *China*)

Abstract: Recommender systems are widely deployed in Web applications that need to predict the preferences of users to items. They are popular in helping users find movies, books, music, and products in general. In this work, we design a method for item recommendation based on a novel model that captures correlations between hidden aspects in reviews and numeric ratings. It is motivated by the observation that a user's preference against an item is affected by different aspects discussed in reviews. Our method first explores topic modeling to discover hidden aspects from review text. Profiles are then created for users and items separately based on aspects discovered in their reviews. Finally, we utilize logistic regression to model the user-item relationship and the rating is modeled as the similarity between user and item profiles. Experiments over real world reviews demonstrate the advantage of our proposal over state-of-the-art solution.

Key words: recommendation; hidden aspect; LDA; regression model; review analysis

收稿日期:2014-12
基金项目:国家自然科学基金(61103039,61402177);国家自然科学基金重点项目(61232002)
第一作者:高祎璠,女,硕士研究生. E-mail: yfgao@ecnu.edu.cn.
通信作者:张蓉,女,博士,副教授,主要研究方向为数据挖掘、信息检索. E-mail: rzhang@sei.ecnu.edu.cn.

1　简　　介

推荐系统广泛地应用在网络平台中,如在线广告,在线购物等.它能够有效地预测用户的喜好,帮助用户找到适合的电影、书籍、音乐等产品.目前的研究重点是如何准确地发现用户偏好和商品画像,以提高推荐性能.总体来说,推荐方法可以分为两大类[1]:协同过滤方法和基于内容的推荐.传统的协同过滤方法是根据用户的历史行为,例如:评分或评论过的电影;曾经购买过的商品等;对用户 - 商品关系进行建模,认为具有相似喜好的用户在选择产品时具有相同的偏好.另一方面,基于内容的推荐是挖掘具有相同或相似属性的商品,从而进行推荐.然而,基于内容的推荐会产生推荐商品过于单一化的问题[2].随着 Web 的流行,消费者的反馈信息即评论(包括评分)对电子商务特别是推荐性能起到积极的影响.用户的评论相较于评分包含了更丰富的对商品的意见和观点,这为生成用户偏好和商品画像提供了更多的信息[3].反过来看,利用评论产生的用户偏好和商品画像则能够更好地解释消费行为和评分.本文在协同过滤方法的基础上,试图通过结合用户的评分和评论信息构建混合的分析模型,刻画用户 - 商品关系,实现更为准确的推荐.

本文利用评分和评论内容,设计并实现了一种新颖的评分预测模型.与传统的只考虑评分的方法不同,本文通过提取评论文本中的潜在主题特征来构建用户偏好和商品画像.首先利用主题模型发现评论中潜在的主题分布,然后采用回归模型,训练出每个潜在主题对评分产生的影响,即发现潜在主题与真实评分的关系.因此,当已知用户偏好和商品画像,本文提出的方法可以做出较准确的评分预测,然后将评分高的商品推荐给用户.

传统模型由于只考虑评分,常常会发生冷启动(cold start)的问题,即当用户没有或只有较少历史记录时,系统将无法进行有效的推荐.对评论中的评分特征进行提取之后,可以为生成用户偏好和商品画像提供更丰富的信息,从而缓解推荐冷启动问题.此外,潜在的主题分布也可用于对评论进行排序,选出具有代表性的评论优先展示给用户.

因此,本文的主要贡献总结是:①基于概率主题模型的评论分析,发现评论主题分布有助于对用户和商品的准确描述,也有利于解释消费行为和评分;②结合潜在主题与回归模型,从语义分析和数字打分两个方面同时分析各潜在主题对评分的影响,实现准确的评分预测;③本文提出的模型可缓解冷启动问题,尤其是对由于评分较少导致的冷启动问题;④对评论的选择和展示提出了新的解决方案;⑤最后,在真实数据上的大量对比实验证明,该模型具有非常良好的性能.

2　相关工作

现有的协同过滤技术[4-6]通过分析用户历史评分行为,预测用户对为评价过的商品的感兴趣程度,而不考虑评论文本.另一方面,已出现很多主题发现[7-9]、情感分析[10]和意见挖掘[11-12]等方向的评论文本分析的工作.然而,这些推荐研究都未将评分和评论相联系[13-14].文献[13]提出的评论打分模型结合了文本分析;文献[14]利用意见打包的形式对评论进行建模,相较于一元和多元模型更具表现力.文献将评分视为是由一系列预定义的主题获得的综合衡量,而主题是基于评论发现的.而以上模型未考虑用户与商品间的关系,故需要给出评论内容才能做出评分预测,这样就不能直接地应用到推荐系统中.此外,文献[13]中的方法是给出一些总结性的意见,来传达一个特定商品的一些信息.

与本文设计思路最相似的研究是,结合了评分和评论分析的 HFT(Hidden Factors as Topics)商品推荐模型[3]. HFT 模型将评分中的隐藏因素和评论中的隐藏主题相融合,生成模型用户/商品画像,映射到 SVD(Singular Value Decomposition)模型[6]中做出评分预测. 然而,在 HFT 模型中,每条评论文本只能属于两个维度中的一个,也就是说,这条评论或是从商品角度来分析(属于商品的评论分类),或是从用户角度来分析(属于用户的评论分类). 这就意味着,发现的潜在主题只能从一个方面来反映评分,而另一个维度必须与潜在因素空间对齐. 为解决这个问题,本文提出的模型将两个维度映射到相同的潜在主题空间,支持分别从用户和商品的角度分析评论文本,达到更好的效果.

3 模 型

3.1 背景知识——LDA 模型[15]

与概率潜在语义分析模型(pLSA)相类似,潜在狄利克雷分布模型(LDA)是一种概率生成模型[15]. LDA 认为每一篇文档 d 服从 K 维主题分布 θ,而文档中的每一个词有概率 φ 的可能性属于主题 k. LDA 模型通过在文本和词之间引入主题维度,对向量空间进行降维. 本文中 LDA 模型用于对用户评论信息的分析,发现评论中潜在的主题.

如图 1 所示,一篇文档可视为 N 个单词组成的有序序列,一个文档集包含 M 篇文档. α 和 β 分别是文本中主题的分布 θ 和主题中词的分布 φ 的超参数,服从先验狄利克雷分布,其中 z 表示主题. 文档集的处理和参数的选择是应用 LDA 模型的关键,在实验部分会进行讨论.

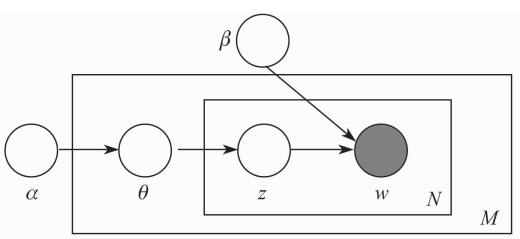

图 1 LDA 模型的图形表示

Fig. 1 Graphical model representation of LDA

3.2 评分预测流程

基于评论的评分预测实现步骤见图 2(文章中使用的符号标记说明见表 1),模型中具有两大功能模块:画像生成和评分预测.

画像生成:输入商品评论集$\{d_{ui}\}$,分别生成用户 u 的偏好 p_u 和商品 i 的画像 q_i.

评分预测:输入用户 u 和商品 i,模型预测用户 u 对商品 i 的评分.

模型的输入是评论集合$\{d_{ui}\}$和评分矩阵;然后使用 LDA 模型发现评论文本中的潜在主题(K 维)分布 θ_{ui},在 K 维潜在主题上分别生成用户偏好 p_u 和商品画像 q_i. 将用户偏好和商品画像与评分矩阵相结合,基于回归模型进行主题权重分布训练. 训练后的模型支持对用户没有评价过的商品进行评分预测.

3.3 画像生成

非结构化的评论文本包含了用户对商品不同主题维度的喜好和意见,这些潜在的主题

能够很好地反映用户评分的潜在因素.通过对评论文本的分析,分别将用户和商品映射到相同的空间 S.为了潜在主题发现和映射空间的构建,将公认度较高的文本分析和特征提取工具——LDA 模型应用到评论文本分析中.

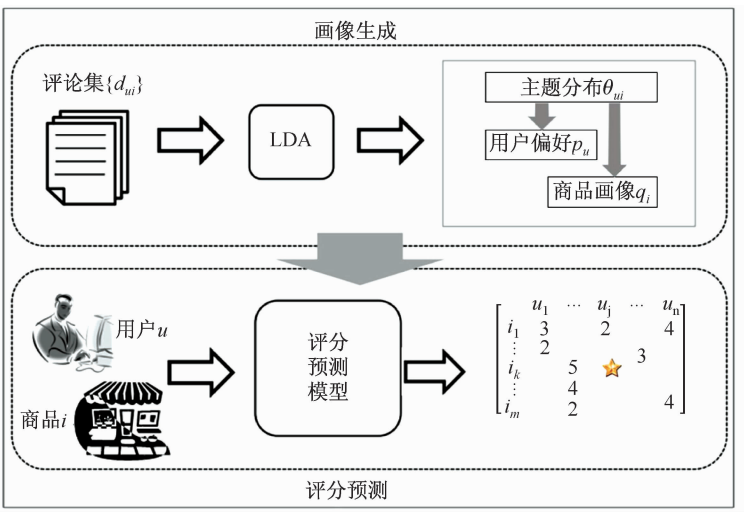

图 2　评分预测流程

Fig. 2　Rating prediction flowchart

不同于 HFT 方法,本文将每一条用户 u 对商品 i 给出的评论 d_{ui} 视为一篇文档.对评论集{d_{ui}}应用 LDA 模型,θ_{ui} 表示 d_{ui} 生成的 K 维主题分布.用户 u 所有的评论集合定义为 D_u,D_i 是商品 i 获得的所有评论的集合.每一个用户 u(或商品 i)对应于偏好 p_u(或画像 q_i).给定一个用户 u,定义用户偏好 p_u 为

$$p'_{uj}=\frac{\Sigma_i\theta_{uij}}{|D_u|},\ p_{uj}=\frac{p'_{uj}}{\Sigma_j p'_{uj}},\ j\in\{1,\cdots,k\}.$$

其中 $p_u=(p_{u1},p_{u2},\cdots,p_{uk})$,$p_{uj}$ 是用户 u 在第 j 个主题上的分布,θ_{uij} 是评论 d_{ui} 在第 j 个主题上的分布.类似的方法定义商品 i 的画像 q_i 为

$$q'_{ij}=\frac{\Sigma_u\theta_{uij}}{|D_i|},\ q_{ij}=\frac{q'_{ij}}{\Sigma_j q'_{ij}},\ j\in\{1,\cdots,k\}.$$

简单来说,p_u 和 q_i 分别是用户 u 和商品 i 所有评论主题分布的正规化结果.

表 1　符号表

Tab. 1　Table of notations

符号	描述	取值范围
K	潜在主题维度	正整数
r_{ui}	用户 u 对商品 i 的评分	正整数
$\hat{r}_{ui}$	用户 u 对商品 i 的预测评分	正整数
d_{ui}	用户 u 对商品 i 的评论	文本
q_i	商品 i 的画像	$[0,1]^K$
p_u	用户 u 的偏好	$[0,1]^K$
θ_{ui}	d_{ui} 的主题分布向量	$[0,1]^K$

3.4　评分预测

给定用户 u 和商品 i,预测用户 u 对商品 i 可能给出的评分 $\hat{r}_{ui}$,根据 $\hat{r}_{ui}$ 向用户推荐其未

给出过评价的商品. 预测评分时,将评论文本中发现的主题维度视为影响评分的潜在因素. 评分预测模型利用线性回归和逻辑斯蒂回归模型建立评分$\hat{r}_{ui}$与评论 d_{ui} 主题分布 θ_{ui} 的关系.

线性回归作为标准的回归分析模型广泛地应用在实践应用中[16]. 假设因变量与自变量存在线性关系,参数严格但易于拟合. 多项线性回归方法的评分预测函数定义是

$$\hat{r}_{ui} = W^{\mathrm{T}}\theta_{ui} + \varepsilon_{ui}.$$

其中 $W = (W_1, \cdots, W_K)$,W_j 是第 j 个主题的权重,ε_{ui} 是误差变量.

逻辑斯蒂回归是一种用于分类的概率统计模型. 本文中的多项逻辑斯蒂回归通过将概率得分作为因变量的值,从而衡量一个绝对因变量和 K 个自变量间的关系[17]. 换言之,逻辑斯蒂回归就是建立预测评分与 K 维主题分布的关系. 假设评分$\hat{r}_{ui} \in \{1, 2, \cdots, N\}$,建立多项式逻辑斯蒂回归

$$P_r(\hat{r}_{ui} = n) = \frac{\mathrm{e}^{\beta_n^{\mathrm{T}}\theta_{ui}}}{1 + \sum_{n'=1}^{N-1} \mathrm{e}^{\beta_{n'}^{\mathrm{T}}\theta_{ui}}}, P_r(\hat{r}_{ui} = N) = \frac{1}{1 + \sum_{n'=1}^{N-1} \mathrm{e}^{\beta_{n'}^{\mathrm{T}}\theta_{ui}}}.$$

其中 $n = 1, 2, \cdots, N-1$,$\beta_n = (\beta_{n1}, \beta_{n2}, \cdots, \beta_{nk})$为权重集.

两种模型均采用最大后验概率(MAP)估计权重向量(W 或者 β_n).

评分预测给定用户 u 和未评论过的商品 i,基于用户偏好 p_u 和商品画像 q_i,估计主题分布$\hat{\theta}_{ui}$是

$$\theta'_{uij} = p_{uj} q_{ij}, \hat{\theta}_{uij} = \frac{\theta'_{uij}}{\sum_j \theta'_{uij}},\ j \in \{1, \cdots, k\}.$$

3.5 代表性评论选择

本文的工作除了能够支持评分预测之外,也可以辅助解决代表性评论选择的问题. 一件商品(特别是热门商品)可能包含成百上千条评论,这为用户浏览增加了困难. 目前的评论网站一般按照时间顺序组织评论内容,或者提供简单的评论关键字搜索,但是这并不能满足用户从大量评论数据中获取有用信息的需求. 本文提出基于主题分布的代表性评论选择方案. 评论代表性能力定义为评论商品特征的潜在主题分布与商品画像主题分布的"相近"程度,旨在选择最能体现商品特征的评论. 一条评论 r_{ui} 与商品 i"相近"程度的定义为

$$d(r_{ui}, i) = \| \theta_{ui} - q_i \|_2^2,$$

选出具有较小 $d(r_{ui}, i)$的评论展示给用户,作为商品 i 的代表性评论.

4 实 验

实验环境:四核 8 线程 i7 处理器,1TB 硬盘,8GB 内存台式机一台.

编程语言:Java.

本章节中,R-Linear 和 R-Logistic 分别代表本文提出的分析模型与线性回归和逻辑斯蒂回归结合进行评分预测的方法.

4.1 数据集

大众点评网作为国内最大的餐饮品鉴类网站,包含了丰富的评论数据,是评估本文系统性能的最佳选择. 爬取的数据集包括上海地区 47,942 家餐厅的 1,205,981 条评论,共涉及 373,021 名用户,详细的数据统计展示在表 2 中. 每条数据包含用户 ID、评论时间、评论文本和 1-5 的数值评分. 餐厅共涉及 19 个大类,表 2 中展示了其中具有代表性的 10 个类别,这些类别可作为子数据集进行性能评估,其中"其他"代表未详述的 9 类数据集合. 一名用户可

访问不同类别的餐厅,但一个餐厅只属于一种类别. 数据集中包括非常受欢迎的本帮江浙菜,也包括数量最少的贵州菜. 平均来说,每家餐厅收到 25.1 条评论,每个用户撰写了 3.2 条评论,每条评论包含 119.4 个中文词汇.

表 2　分类别数据统计

Tab. 2　Data statistics on different category

餐厅类别	用户数/人	餐厅数/家	评论数/篇
本帮江浙菜	90 733	6 214	134 302
川菜	45 776	2 537	56 941
西餐	62 581	2 017	83 143
贵州菜	2 783	56	3 906
西北菜	7 007	160	8 717
火锅	60 217	1 891	74 050
湘菜	20 815	1 142	27 027
日本菜	60 427	1 823	82 856
粤菜	47 109	1 461	59 939
韩国料理	24 571	674	25 880
其他	361 420	29 967	649 220
共计	373 021	47 942	1 205 981

将数据集按照 9∶1的比例,划分为训练集和测试集. 在测试集生成时,可采用两种方法进行划分:随机分配和时序分配. 时序分配的方法认为较早的评论内容可能会对后续的评论产生影响,该方法根据评论时间选取最新的评论作为测试集,而随机分配方法不考虑时间因素而随机产生 1/10 的数据作为测试集. 实验中共选取 100 000 条评论作为测试集.

4.2　评价指标

与 HFT 论文中方法相似,采用平均均方误差(MSE)来衡量预测评分与实际得分的误差

$$\mathrm{MSE}=\frac{1}{M}\Sigma_{u,i}(\hat{r}_{ui}-r_{ui})^2.$$

其中 M 是预测评分的总数量,$\hat{r}_{ui}$ 和 r_{ui} 分别是用户 u 对商品 i 的预测评分和实际得分.

除了 MSE,还引入了准确度(accuracy)来衡量评论预测的准确度,下式中 m 表示预测评分与实际评分一致的发生次数,即 $\hat{r}_{ui}=r_{ui}$.

$$\mathrm{ACC}=\frac{m}{M}$$

值得注意的是,实际得分只可能是 1 - 5 的整数,而线性回归方法得到的预测分数 $\hat{r}_{ui}$ 要进行取整后才能与实际得分 r_{ui} 相比较,实验中采用的是四舍五入法.

4.3　对比实现系统

正如相关工作中提到的,HFT 将评论中潜在的主题与评分维度相结合,是与本文设计思路最为相似的模型. HFT 由其作者提供的源码实现①. 协同过滤方法(CF)假设有相似喜好的用户会选择相同的产品,而传统的 CF 是结合用户主观决定(比如评分)实现过滤. Slope One[18]是目前应用较为广泛的基于商品的协同过滤方法,具有简单和高效的优势,可由开源工具 MyMediaLite 3.10[19]实现②. 实验中,会将以上两种方法作为性能测试的基准

①http://www.cseweb.ucsd.edu/~jmcauley/.

②http://www.mymedialite.net/index.html.

测试系统与本文提出的推荐策略进行对比.

4.4 LDA 参数选择

超参数 α 和 β 分别取经验值 0.2 和 0.1. 分别对主题维度 K 取值 5、10 和 20 进行实验,线性回归结果展示在表 3 中. 随着 K 数值的增长,系统性能逐步提高. 然而,当 K 值从 10 增长到 20,性能提高幅度较小. 对实验结果进行核查,从主题关键词的分布可以看出,$K=10$ 时主题划分较为清晰. 实验中选取 $K=10$ 为默认主题数目,每个主题中频繁出现的 10 个代表词汇如表 4 所示. 为 LDA 能够在评论数据上快速收敛,迭代次数设置为 100.

表 3 不同主题数目下结果对比

Tab. 3 Results with different topic numbers

K	MSE	ACC/%
5	0.53	51.4
10	0.51	52.3
20	0.51	53.3

表 4 当 $K=10$ 时,各主题中频繁出现的 10 个代表词汇

Tab. 4 Top ten words for each topic with $K=10$

小吃	川菜	团购	午餐	日本料理	环境	饮料	西餐	甜品	不满
汤	辣	团购	附近	寿司	味道	咖啡	口感	蛋糕	服务员
面	嫩	一份	每次	新鲜	环境	朋友	咖喱	奶茶	差
肉	入味	套餐	好吃	三文鱼	菜	喝	脆	面包	态度
碗	鲜	量	中午	牛排	服务	舒服	配	甜	不知道
生煎	香	性价比	一直	料	喜欢	坐	烤	奶	以后
皮	烧	少	便宜	色拉	特别	下午茶	酱	巧克力	不再
小笼	牛蛙	便宜	外卖	刺身	口味	安静	芝士	杯	客人
锅贴	咸	饭	路	日本	干净	家	披萨	奶油	等了
鸡排	炒	吃了	排队	精致	热情	适合	美味	冰	难吃
汁	烤鱼	还行	公司	包房	装修	酒吧	米	饮料	只

4.5 结果和分析

首先,实验比较了我们的方法与 Slope One 和 HFT 方法的 MSE 结果. 如图 3 所示,本文的方法表现的更加优异,线性回归方法效果最好,能够达到最小的 MSE,随机分配(时序分配)方法相较于 HFT 降低了 34%(49%).

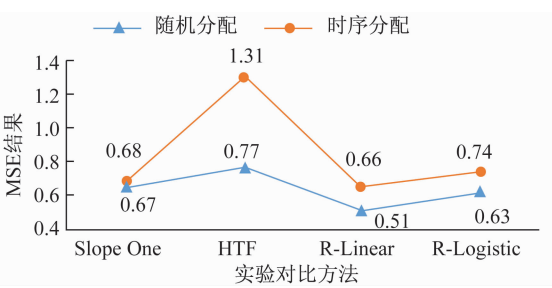

图 3 MSE 结果对比

Fig. 3 MSE values

图 4 是评分预测准确度的对比结果. 实际评分取值是 1 - 5 的整数,而 Slope One、HFT 和线性回归方法得到的预测评分均为小数,为计算 ACC 需将结果就近取整. 逻辑斯蒂回归

作为聚类模型,可以得到整数的预测评分. 从结果来看,Slope One 表现最好,本文的方法其次. 测试集随机分配情况下,线性回归与逻辑斯蒂回归方法相差无几. 整体来看,随机分配方法的评分预测性能优于时序分配方法. 以下实验均采取随机分配的数据划分方法.

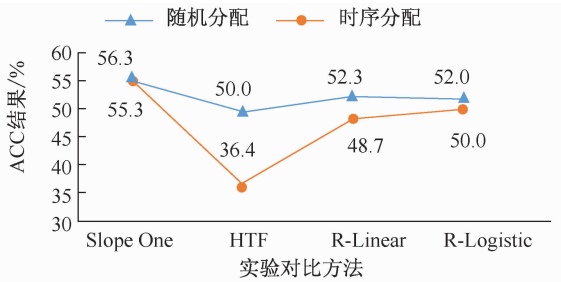

图 4　ACC 结果对比

Fig. 4　ACC values

在各种测试情况下,Slope One 均表现出较好的性能,但是作为标准的 CF(协同过滤)方法,它会受到数据稀疏性带来的严重影响. 当数据稀疏时,CF 方法预测结果的准确度难以得到保证. 在表 2 中的 10 个类别的子数据集上,对各种方法进行实验,结果如图 5 和图 6 所示. 可以明显地看出,线性回归方法在所有情况下都表现的最好,但是当数据稀疏时,如贵州菜,Slope One 的效果发生了巨大的波动. 实验证明,本文提出的方法相较于 CF 方法具有更好的稳定性,在数据稀疏的情况下亦能得到较好的 MSE 和 ACC 结果.

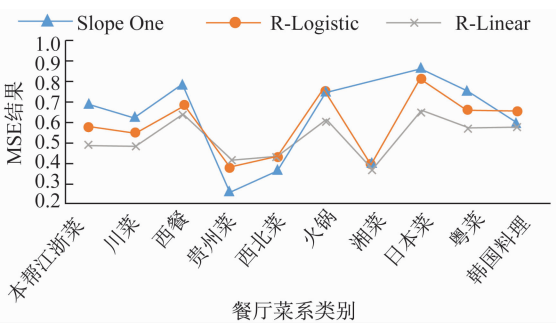

图 5　10 种类别的 MSE 对比

Fig. 5　MSE of 10 various categories

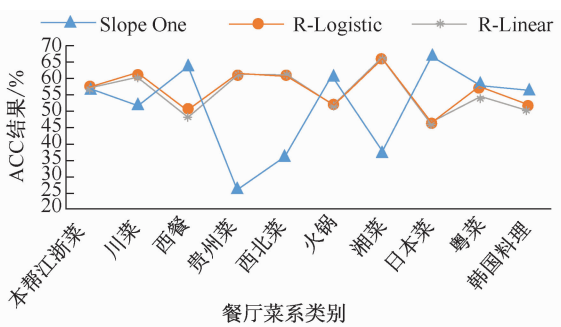

图 6　10 种类别的 ACC 对比

Fig. 6　ACC of 10 various categories

4.6 讨论

冷启动是推荐系统常见的问题. 少量的数据导致模型难以训练,系统难以对新用户做出推荐. 引入评论内容,丰富已知信息,是冷启动问题的解决办法之一. 另一方面,对用户偏好和商品画像进行建模,相较于其他随机方法,能够更加准确地获得对象的属性,有助于模型的训练. CF 作为传统推荐方法,有较好的推荐效果,但是受到数据稀疏情况的巨大限制. 已有工作将 CF 方法与基于内容分析的方法相结合,本文提出的基于内容分析的商品和用户画像分析技术与最新的 HFT 方法相比,已经具有较大的优势,同时我们与传统的 CF 相比,拥有较好的稳定系,因此后续可以考虑把我们的方法与 CF 方法结合使用.

4.7 代表性评论选择

当系统将一件商品推荐给用户时,用户会通过阅读评论来了解该商品. 图 7 中显示的是餐厅平均评论数目的分布. 首先将餐厅按照评论数目进行降序排列,前 10%的餐厅平均拥有超过 180 条的评论,半数以上的餐厅评论数目超过 50 条. 为方便用户的浏览,代表性评论的选择是非常必要. 本文提出了基于商品画像的代表性评论选择方法. 如图 8 中案例,老板娘海鲜大排档的主题分布描述为(0.05,0.07,0.13,0.05,0.21,0.11,0.01,0.13,0.06,0.15),表中展示了本文方法选取出的具有代表性的前五条评论,评论内容是极能表现餐厅概况的. 虽然还可以从覆盖度和多样性[20]等方面对代表性评论做出选择,但该方法为评论选择提供了新的参考因素.

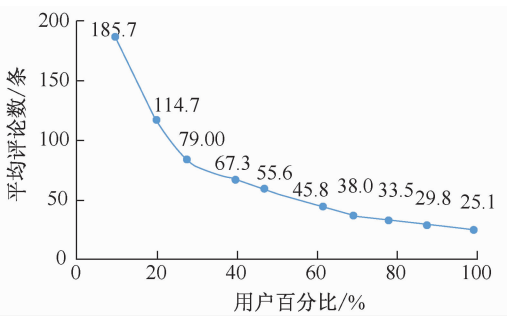

图 7 餐厅评论分布

Fig. 7 Restaurant review distribution

5 总 结

本文提出了一种同时利用评分和评论分析的商品推荐模型. 利用评论文本中的潜在主题分布,将用户和商品映射到统一的空间中,分别生成了用户偏好和商品画像. 评分预测模型建立在回归模型之上,将主题分布与预测评分相拟合. 根据预测评分做出商品推荐. 并且本文提出的方法能从语义上支持代表性评论选择的实现. 实验证明,本文提出的评分预测方法表现出稳定良好的性能,特别是针对稀疏性数据. 未来的工作中将会考虑把本方法与协同过滤方法相结合,实现更为优良的性能.

餐厅名称：老板娘海鲜大排档（共 120 条评论）	
前 5 条代表性评论	$d(r_{ui},i)$
(0.01, 0.04, 0.01, 0.04, 0.71, 0.01, 0.01, 0.16, 0.01, 0.01) 环境是不怎么样的，因为人多，坐了个包房，相对来说稍微清净一点. 时不时有人进来问要不要唱歌，二十块一首. 难听得要命啊.... 点了很多很多海鲜，毛蚶啊龙虾啊花蛤啊扇贝啊螃蟹啊蛏子啊，味道还可以的，就是咖喱蟹好像不是很新鲜. 结账人均 100.	0.021
(0.003, 0.25, 0.31, 0.01, 0.16, 0.08, 0.01, 0.003, 0.003, 0.17) 今天和朋友一起去的，朋友大概是那里的常客，直奔那就去了. 味道一般，点菜的时候感觉放在那的海鲜卖相不怎么样，味道嘛也就这样，普通. 好怀念去厦门吃的海鲜！ 点了濑尿虾、海瓜子（太小吃起来麻烦）、海肠（点的时候看着有点影响食欲，不过吃起来脆脆的还不错）、花蟹、生鱼片（卖相不好，放的芥末不够味）、赤贝（吃不惯）、海螺肉、虎头鱼（炸得挺香的），其余的不记得了. 一般吃吃	0.037
(0.002, 0.02, 0.21, 0.01, 0.10, 0.02, 0.02, 0.57, 0.04, 0.01) 昨天与朋友及其同事三人一起慕名而来！ 口味确实不错，三人吃得很爽！ 只是可惜人少，不能将想吃的都点上，只能待下次再去. 菜色就不多说，去了自己选！ 一句话，值得一去！	0.038
(0.01, 0.04, 0.02, 0.02, 0.45, 0.01, 0.01, 0.03, 0.01, 0.42) 潜水了很多年了，第一次为一家饭店跑出来写点评. 实在是太差差差了，价格比下面几位同学说的便宜，但是质量是差到不能再差了. 本来虽然环境差，味道差，海鲜不新鲜，也不会上来写点评，主要是我和我妈回去以后都闹肚子了，折腾到半夜才睡，这才跑来给大家提个醒. 点了六七个菜，濑尿虾里面肉很少，肉是松软的，有一点点膏也不是硬的；油蛤很多是闭口的；龙头烤里面太软了感觉都没有熟；小鲍鱼和扇贝一股腥味，铺满了蒜蓉都遮不掉那个味道，小鲍鱼只有两个一元硬币这么大，而且老的要命. 关键是油都一股 ho 味，椒盐也是质量很差的那种. 昨天唯一的好处是因为我觉得太难吃都没吃所以扔下老公他们买单我去路口买吃的，结果居然发现警察在贴条，赶紧在贴之前回到车上，省了 200. 否则吃得这么扫兴再被罚钱，真是要郁闷死了.	0.044
(0.01, 0.02, 0.01, 0.03, 0.13, 0.5, 0.01, 0.01, 0.01, 0.28) 去通北路吃海鲜基本都去他们家，老板娘人很好，会给打个折，还会送啤酒和小食，哈哈. 必点的就是葱油黄泥螺，超爱的，新鲜. 生蚝也是必吃的. 其他的都是每次换着吃，希望能把所有的海鲜都吃一遍	0.055

图 8　代表性评论选择实例

Fig. 8　Representative review example

[参　考　文　献]

[1] RAJARAMAN A, ULLMAN J D. Mining of Massive Datasets[M]. London: Cambridge University Press, 2011.

[2] BLANCO-FERNÁNDEZ Y, PAZOS-ARIAS J J, GIL-SOLLA A, et al. A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems[J]. Knowledge-Based Systems, 2008, 21(4): 305-320.

[3] MCAULEY J, LESKOVEC J. Hidden factors and hidden topics: understanding rating dimensions with review text [C]//Proceedings of the 7th ACM conference on Recommender systems. ACM, 2013: 165-172.

[4] SARWAR B, KARYPIS G, KONSTAN J, et al. Item-based collaborative filtering recommendation algorithms [C]//Proceedings of the 10th international conference on World Wide Web. ACM, 2001: 285-295.

[5] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

[6] KOREN Y, BELL R. Advances in collaborative filtering[M]//KANTOR P B, RICCI F, ROKACH L, et al. Recommender Systems Handbook. New York: Springer, 2010: 145-186.

[7] BRODY S, EIHADAD N. An unsupervised aspect-sentiment model for online reviews[C]//Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 804-812.

[8] JO Y, OH A H. Aspect and sentiment unification model for online review analysis[C]//Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011: 815-824.

[9] 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.

[10] TITOV I, MCDONALD R T. A Joint Model of Text and Aspect Ratings for Sentiment Summarization[C]//ACL, 2008(8): 308-316.

[11] POPESCU A M, ETZIONI O. Extracting product features and opinions from reviews[M]//KAO A, POTEET S R. Natural Language Processing and Text Mining. London: Springer, 2007: 9-28.

[12] PANG B, LEE L. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information Retrieval, 2008, 2(1-2): 1-135.

[13] QU L, IFRIM G, WEIKUM G. The bag-of-opinions method for review rating prediction from sparse text patterns [C]//Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics. 2010: 913-921.

[14] GANU G, ELHADAD N, MARIAN A. Beyond the stars: Improving rating predictions using Review text content [C]//WebDB, 2009.

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

[16] BAMMANN K. Statistical models: theory and practice[J]. Biometrics, 2006(62): 943.

[17] BISHOP C M. Pattern recognition and machine learning[M]. New York: Springer, 2006.

[18] LEMIRE D, MACLACHLAN A. Slope one predictors for online rating-based collaborative filtering[C]//SDM, 2005, 5: 1-5.

[19] GANTNER Z, RENDLE S, FREUDENTHALER C, et al. MyMediaLite: A free recommender system library [C]//Proceedings of the fifth ACM conference on Recommender systems. ACM, 2011: 305-308.

[20] TSAPARAS P, NTOULAS A, TERZI E. Selecting a comprehensive set of reviews[C]//Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011: 168-176.

(责任编辑 李 艺)