基于用户评分和评论信息的协同推荐框架*

谭云志 张 敏 刘奕群 马少平

(智能技术与系统国家重点实验室 北京 100084)
(清华信息科学与技术国家实验室(筹) 北京 100084)
(清华大学 计算机科学与技术系 北京 100084)

摘 要 用户的反馈通常包含一个数值评分和一段文本形式的评论.文中利用用户评论学习商品特征在不同主题上的分布及用户对商品不同特征的偏好程度,把商品特征和用户偏好的契合度引入传统的协同过滤算法中,提出基于用户评分和评论信息的协同推荐框架.使用该框架可较方便地将用户评论信息引入到现有的协同过滤算法中.通过引入用户评论信息,可一定程度缓解传统协同过滤算法面临的数据稀疏性问题.在 22 个亚马逊的真实数据集上的实验证明文中方法的有效性.

关键词 推荐系统,协同过滤,用户评论,文本分析
中图法分类号 TP 18 **DOI** 10.16451/j.cnki.issn1003-6059.201604008
引用格式 谭云志,张 敏,刘奕群,马少平.基于用户评分和评论信息的协同推荐框架.模式识别与人工智能,2016,29(4):359-366.

Collaborative Recommendation Framework Based on Ratings and Textual Reviews

TAN Yunzhi, ZHANG Min, LIU Yiqun, MA Shaoping
(*State Key Laboratory of Intelligent Technology and Systems*, *Beijing* 100084)
(*Tsinghua National Laboratory for Information Science and Technology*, *Beijing* 100084)
(*Department of Computer Science and Technology*, *Tsinghua University*, *Beijing* 100084)

ABSTRACT

The feedback of users usually contains a numeric rating and a textual review. In this paper, textual review information is used to learn the distributions of item features on different topics and the user preference to different features of items. Then, the topic-based user preference similarity is incorporated into the traditional collaborative filtering recommendation systems. A recommendation framework based on ratings and textual reviews is proposed. With the proposed framework, review information can be easily introduced into the existing recommendation algorithms. By employing textual reviews, the problem of data sparsity in the traditional recommendation algorithms is relieved. Experiments are conducted on 22 real-world datasets from Amazon and the experimental results demonstrate the advantages and the effectiveness of the proposed framework.

* 国家重点基础研究发展计划项目(973 计划)(No. 2015CB358700)、国家自然科学基金项目(No. 61472206,61073071)资助
Supported by National Basic Research Program of China (973 Program) (No. 2015CB358700), National Natural Science Foundation of China (No. 61472206,61073071)
收稿日期: 2015-05-12; **修回日期:** 2015-08-25; **录用日期:** 2015-09-10
Manuscript received May 12,2015; revised August 25, 2015; accepted September 10, 2015

Key Words Recommendation System, Collaborative Filtering, Textual Review, Text Analysis

Citation TAN Y Z, ZHANG M, LIU Y Q, MA S P. Collaborative Recommendation Framework Based on Ratings and Textual Reviews. Pattern Recognition and Artificial Intelligence, 2016, 29(4): 259 - 366.

推荐系统通过提供个性化的推荐帮助人们克服信息过载的问题,其核心是通过个性化的算法、利用不同用户对于不同商品的反馈信息发现他们的喜好. 在所有推荐算法中,因为协同过滤算法可利用群体的智慧进行推荐[1],所以不仅在学术界受到广泛重视,还在产业界获得广泛应用.

协同过滤算法的基本假设是两个用户给很多商品评分类似或在很多商品上具有相似的行为表现(如购买、观看等),那么他们可能在其它商品上的评分或表现也会相似[2]. 传统的相似度算法基于用户的历史评分,若用户之间共同评分的商品数量较少,则这些算法得到的相似度不可靠.

为了减轻数据稀疏带来的影响,一类重要的方法是混合协同过滤(Hybrid Collaborative Filtering)算法[3-4],这类方法通常结合协同过滤(Collaborative Filtering, CF)算法和基于内容的算法,通过使用额外信息(如商品描述、浏览日志、统一资源定位器(Uniform Resource Locator, URL)信息和用户兴趣、需求信息等),在一定程度上缓解数据稀疏性问题,取得更好的推荐效果. 但这类方法使用的额外信息严重依赖于所推荐的商品类型,且有些信息难以获得. 同时这些信息具有不同的形式,需根据信息的不同呈现形式设计不同的使用方法,很难找到一个通用的表达形式充分利用信息.

随着越来越多用户评论信息的出现,把评论信息引入推荐算法的方式受到越来越多的重视[3,5-12]. 与混合协同过滤算法使用的其它信息相比,评论信息较易获得且包含大量有用的信息. 通过分析这些评论,可理解商品的特征分布和用户的个人偏好分布. 但评论信息通常以纯文本的形式出现,如何让计算机理解、分析,进而提取有用的信息是一个巨大的挑战. Goldberg 等[2]使用人工方式从评论中提取特征信息,这就需要丰富的领域知识,同时需要大量的人力成本. 文献[6]和文献[7]使用情感分析的方法自动分析评论信息,但这些方法通常依赖于自然语言处理技术,且计算复杂度较高. 同时使用评分和评论信息进行推荐的方法基于矩阵分解和主题模型[8],只能对商品或用户两者其一使用评论信息,且未综合考虑用户评论与用户对商品的情感倾向,在一定程度上影响推荐性能.

区别于传统的协同过滤算法只使用用户的历史打分,本文提出基于用户评分和评论信息的协同推荐框架(简称为 TopicCF). 该框架同时使用用户评分和评论信息进行推荐,从用户评论中发现商品的特征分布和用户的偏好分布,在跳过复杂的自然语言处理技术的同时结合评论内容与用户对应的情感倾向. 同时考虑用户在历史评分上的相似性和在个人偏好上的相似性,可更准确确定用户之间的相似度,提升参与协同推荐用户的质量. 此外,通过寻找特征分布相似的商品,提升参与协同推荐的用户数量,在一定程度上缓解数据稀疏性问题. 最后,在亚马逊的 22 个真实数据集[8]上的实验表明本文方法可较好地把评分和评论信息引入现有的推荐算法中,进而提升推荐系统的性能.

1 特征提取和相似度计算

用户评论一个商品时,商品典型的、给人印象深刻的特征易在评论中体现,同时这些特征也是该用户关注的特征,具有不同喜好的用户在评论中讨论的内容不同,通过分析这些评论,可理解商品的特征分布和用户的喜好分布.

1.1 商品特征和用户偏好分布学习

本文使用主题模型潜在狄利克雷分布(Latent Dirichlet Allocation, LDA)[10]抽取商品特征在 n(如 $n = 5$)个主题上的概率分布. LDA 是一个生成概率模型,具体到文本数据建模上,LDA 将针对每个文档生成可用以代表对应文档的主题概率分布. 对于第 i 个商品,结合它所有相关评论组成该商品的评论文档 d_i,然后使用 LDA 得到每个商品文档在 n(如 $n = 5$)个主题上的概率分布向量 f_i(n 维),f_i 也是对应商品 i 的特征分布向量.

因为用户在评论中既会讨论他们喜欢的商品特征,也会讨论他们厌恶的商品特征,所以学习用户的偏好分布时需结合用户表现的情感倾向. 为此,结合商品的特征分布和用户给商品打出 1 ~ 5 分的数值评分,分别学习用户的喜好分布和厌恶分布. 具体如下:

$$\boldsymbol{p}_u = \frac{\sum_{i \in I_u} \boldsymbol{f}_i (r_{u,i} - 1)}{\sum_{i \in I_u} (r_{u,i} - 1)}, \quad (1)$$

$$\boldsymbol{h}_u = \frac{\sum_{i \in I_u} \boldsymbol{f}_i (5 - r_{u,i})}{\sum_{i \in I_u} (5 - r_{u,i})}, \quad (2)$$

其中,$\boldsymbol{p}_u$ 和 $\boldsymbol{h}_u$ 为 2 个 n 维的向量,表示用户 u 对 n 个主题的喜好和厌恶分布; I_u 为用户 u 评分的商品集合; $\boldsymbol{f}_i$ 为商品 i 的特征分布向量; $r_{u,i}$ 为用户 u 对商品 i 给出 1 ~ 5 分的评分值. 根据式(1)可看出,用户给商品的评分越高,商品对应特征分布占其它喜好分布越大,即用户越关心该商品对应的特征. 类似地,式(2)中用户给商品的评分越低,该商品对应的特征分布占其厌恶分布越大.

1.2　用户相似度计算

为融合来自用户评分和评论的双重信息,任何两个用户的相似度被分解为评分相似度和偏好相似度,其中偏好相似度又由喜好相似度和厌恶相似度共同表示.

已有多种方法基于评分计算用户之间的相似度[1],选择其中校准的余弦相似度(Adjust Cosine Similarity, ACS)[14]计算用户之间的评分相似度 $rsim$. 对于任意两个用户 u、v,具体如下:

$$rsim_{u,v} = \frac{\sum_{i \in I_{u,v}} \left(r_{u,i} - \bar{r}_i\right) \cdot \left(r_{v,i} - \bar{r}_i\right)}{\sqrt{\sum_{i \in I_{u,v}} \left(r_{u,i} - \bar{r}_i\right)^2} \cdot \sqrt{\sum_{i \in I_{u,v}} \left(r_{v,i} - \bar{r}_i\right)^2}},$$

其中,$I_{u,v}$ 为用户 u、v 共同打分的商品集合,$\bar{r}_i$ 为商品 i 的平均得分.

基于用户的喜好和厌恶分布,使用余弦相似度(Cosine Similarity, CS)计算任意两个用户 u、v 之间的喜好相似度 $psim_{u,v}$ 和厌恶相似度 $hsim_{u,v}$,然后使用喜好相似度和厌恶相似度的几何平均值计算用户的偏好相似度 $tsim_{u,v}$,即

$$psim_{u,v} = \cos(\boldsymbol{p}_u, \boldsymbol{p}_v) = \frac{\boldsymbol{p}_u \cdot \boldsymbol{p}_v}{\|\boldsymbol{p}_u\| \cdot \|\boldsymbol{p}_v\|},$$

$$hsim_{u,v} = \cos(\boldsymbol{h}_u, \boldsymbol{h}_v) = \frac{\boldsymbol{h}_u \cdot \boldsymbol{h}_v}{\|\boldsymbol{h}_u\| \cdot \|\boldsymbol{h}_v\|},$$

$$tsim_{u,v} = \frac{2 \cdot psim_{u,v} \cdot hsim_{u,v}}{psim_{u,v} + hsim_{u,v}}.$$

使用参数 α 平衡评分相似度和偏好相似度的重要性,得到任意两个用户 u、v 的相似度:

$$usim_{u,v} = \alpha \cdot rsim_{u,v} + (1 - \alpha) \cdot tsim_{u,v}. \quad (3)$$

当 $\alpha = 1$ 时,用户相似度退化为评分相似度; 当 $\alpha = 0$ 时,用户相似度退化为偏好相似度; 当 $\alpha \in (0,1)$ 时,同时使用用户评分和评论信息衡量用户之间的相似性.

2　基于用户评分和评论信息的协同推荐框架

2.1　基本框架

为了把用户评论信息引入现有的推荐算法,提升预测准确率,提出 TopicCF. 该框架在已有推荐算法的预测结果上引入修正项,同时使用用户评分和评论信息衡量用户之间的相似度. 预测用户 u 对商品 i 的最终评分如下:

$$\hat{r}_{u,i} = b_{u,i} + \frac{\sum_{v \in U_i} \left(r_{v,i} - \bar{r}_v\right) \cdot usim_{u,v}}{\sum_{v \in U_i} |usim_{u,v}|}, \quad (4)$$

其中,$\hat{r}_{u,i}$ 为预测得到的用户 u 对商品 i 的评分,$b_{u,i}$ 为使用已有推荐算法得到的预测结果,U_i 为所有给商品 i 评过分的用户集合.

$b_{u,i}$ 为基础评分项(Basic Rating Term),基础评分项后面部分为评分修正项(Boost Term). 基础评分项可使用任何已有评分预测算法计算得到,如基于用户的协同过滤算法、奇异值分解(Singular Value Decomposition, SVD)、潜在因素模型(Latent Factor Models, LFM)等. 这些算法都仅使用评分信息进行评分预测,对于用户刻画仅停留在使用用户的历史评分层面. 本文框架在它们的预测结果上增加评分修正项,将评论信息引入已有的评分预测算法,结合评分和评论信息对用户建模,使用与待预测用户最相似的用户信息修正已有评分预测算法的预测结果.

2.2　引入相似商品信息

当数据稀疏时,U_i 含有用户数量较少,引入推荐与商品 i 相似的商品评过分的用户信息,在一定程度上缓解数据稀疏问题. 为此,基于商品的特征分布,寻找与商品 i 在特征水平上最相似的几个商品,组成扩展集,利用给扩展集中的商品评过分的用户进行协同推荐. 商品特征相似度:

$$isim_{i,j} = \frac{\boldsymbol{f}_i \cdot \boldsymbol{f}_j}{\|\boldsymbol{f}_i\| \cdot \|\boldsymbol{f}_j\|}.$$

通过引入扩展商品及用户信息,修改式(4)得

$$\hat{r}_{u,i} = b_{u,i} + \frac{\sum_{j \in E_i} \left(isim_{i,j} \cdot \sum_{v \in U_j} \left(r_{v,j} - \bar{r}_v\right) \cdot sim_{u,v}\right)}{\sum_{j \in E} \left(isim_{i,j} \cdot \sum_{v \in U_j} |sim_{u,v}|\right)},$$

$sim_{u,v} = usim_{u,v} \cdot \Omega(I_u \cap I_v \neq 0)$，

其中，E_i 为商品 i 的扩展商品集合，需说明的是 $i \in E_i$ 且 $isim_{i,i} = 1$；I_u、I_v 分别为用户 u、v 评过分的商品集合；$\Omega(\cdot)$ 为示性函数，当用户 u、v 无共同评分商品时为 0，否则为 1.

3 实验及结果分析

3.1 实验数据集

为了评价 TopicCF 的性能，基于 Amazon.com 的 22 个真实数据集开展实验. 每个数据集都是在亚马逊网站上出售的一类真实商品的集合，这些数据集的统计信息如表 1 所示. 表 1 最后一列为每个数据集中历史评分不超过 3 个用户(即“沉默”用户)的比例，对于这种评分数量很少的用户，仅使用用户评分的协同过滤算法很难给出准确预测.

表 1 实验数据集分布

Table 1 Distribution of experimental datasets

数据集	用户数	商品数	评分数	沉默用户占比/%
Arts	24070	4207	27752	98.62
Automotive	133255	47540	188392	96.71
Baby	13929	1651	16073	98.55
Beauty	167724	28805	248900	95.66
Cell Phones and Accessories	68040	7336	76656	95.66
Clothing and Accessories	128793	65688	577978	83.98
Gourmet Foods	112543	23368	153757	96.50
Health	311635	39276	421646	96.97
Home and Kitchen	644508	78364	967115	95.44
Industrial and Scientific	29589	22603	136720	95.91
Jewelry	40593	18622	58190	95.74
Movies	827806	241599	6337023	60.65
Music	1024450	486604	4109752	78.48
Musical Instruments	67006	14116	84411	97.64
Office Products	110471	14110	132752	98.34
Patio	166831	19384	203169	98.38
Software	68463	10657	82039	98.29
Sports and Outdoors	329231	67872	506352	95.71
Tools and Home Improvement	283513	50739	403709	96.11
Toys and Games	290712	52223	391088	96.95
Video Games	228569	20336	364580	94.72
Watches	62040	10279	68035	99.42

3.2 对比方法和评价指标

使用 4 种已有推荐算法预测基础评分项.

1) 用户评分平均值(User Mean，UMean). 此方法仅使用用户历史评分的平均值预测对应用户新的评分. 用户评分的平均值也是传统的基于用户协同过滤算法使用的基础评分项.

2) SVD++[15]. 此方法基于矩阵分解算法，同时包含隐含信息和明确信息. 以这个模型为核心的推荐算法赢得 2007 ~ 2009 年 Netflix 公司举办的推荐算法竞赛.

3) LFM. 此方法在传统的用户-商品评分矩阵分解的基础上为用户、商品增加评分偏置，也是目前最成功的算法之一.

4) 隐藏因素和主题模型(Hidden Factors and Hidden Item Topics，HFT(item))[8]. 基于传统的 LFM，融合隐性的评分表示和评论主题表示. 预测准确度高于 LFM，赢得 2013 年 Yelp 挑战赛的大奖，是目前同时使用评分与评论信息的最成功的推荐算法之一.

在使用 4 种方法得到基础评分后，使用 2.2 节中加入相似商品扩展的评分修正项修正推荐结果(加入评分修正项的方法标注为对应的基本方法 + Boost，如 UMean + Boost).

对于 HFT，在实验过程中使用 McAuley 等[8]公布的源代码生成基础预测评分(取 $K = 5$). 对于其它 3 种算法，使用推荐系统算法开源库 MyMediaLite 生成基础预测评分.

最后采用常见的评价指标 MSE(Mean Squared Error)评价评分预测的效果，MSE 定义为预测数值和真实数值之间平方误差的平均值：

$$MSE = \frac{1}{|D|}\sum_{(u,i)\in D}\left(r_{u,i} - \hat{r}_{u,i}\right)^2.$$

其中，D 为数据中所有评分对应的“用户-商品”对的集合，$|D|$ 为该集合的大小. MSE 越小，说明预测结果离真实值越近，预测效果越好.

3.3 评分预测性能

实验过程中，随机选取数据集的 80% 数据用于实验，10% 数据用于验证，剩下的 10% 数据用于测试. 因为在真实场景下预测的数据无评论信息，所以仅在训练过程中使用评论信息. 此外设定主题数量和扩展商品集合的大小都为 5，用网格搜索确定式(3)中参数 α 的最优值. 表 2 为 TopicCF 与其它 4 种推荐算法的性能对比，其中每个数据集上的最好预测结果使用黑体数字进行标注.

表 2　融合 TopicCF 框架的 UMean, SVD ++, LFM 和 HFT 模型的评分预测结果(MSE 值)

Table 2　Rating prediction results of UMean, SVD ++, LFM and HFT models combined with TopicCF framework (MSE value)

数据集	UMean	UMean + Boost	提升比/%	SVD ++	SVD ++ + Boost	提升比/%	LFM	LFM + Boost	提升比/%	HFT	HFT + Boost	提升比/%
Arts	1.6358	1.6331	0.1648	1.6012	1.5960	0.3297	1.4443	1.4406	0.2596	1.4239	**1.4190**	0.3494
Automotive	1.6521	1.6478	0.2614	1.5718	1.5658	0.3798	1.4276	1.4247	0.2029	1.4192	**1.4128**	0.4467
Baby	1.9842	1.9830	0.0585	1.6265	1.6263	0.0113	**1.5403**	1.5407	−0.0268	1.5494	1.5456	0.2465
Beauty	1.4647	1.4451	1.3434	1.6164	1.5869	1.8253	1.3745	1.3597	1.0727	1.3726	**1.3500**	1.6397
Cell Phones and Accessories	2.3414	2.3415	−0.0065	2.1546	2.1549	−0.0145	**2.1009**	2.1016	−0.0350	2.1305	2.1308	−0.0159
Clothing and Accessories	0.3628	0.3527	2.7724	0.7250	0.7015	3.2441	0.4037	0.3920	2.9096	0.3413	**0.3332**	2.3723
Gourmet Foods	1.6128	1.6102	0.1604	1.5082	1.5026	0.3716	1.4159	**1.4159**	0.0009	1.4444	1.4378	0.4576
Health	1.7398	1.7263	0.7748	1.6459	1.6279	1.0961	1.5040	**1.4935**	0.6973	1.5167	1.4990	1.1629
Home and Kitchen	1.8660	1.8510	0.8022	1.6730	1.6481	1.4893	1.5167	**1.5076**	0.6014	1.5333	1.5155	1.1572
Industrial and Scientific	0.4016	0.3996	0.4857	0.4363	0.4331	0.7402	0.3681	0.3659	0.5991	0.3479	**0.3452**	0.7556
Jewelry	1.2968	1.2936	0.2507	1.4417	1.4346	0.4928	1.2831	1.2776	0.4344	1.2638	**1.2513**	0.9862
Movies	0.8032	0.6687	16.7390	1.1355	0.8675	23.6020	0.6056	**0.5324**	12.0770	0.8515	0.6943	18.4520
Music	0.8246	0.7525	8.7487	0.9710	0.8605	11.3832	0.7119	**0.6832**	4.0310	0.8273	0.7563	8.5823
Musical Instruments	1.6229	1.6208	0.1300	1.4554	1.4543	0.0712	1.3921	**1.3909**	0.0869	1.4098	1.4036	0.4404
Office Products	1.9055	1.8992	0.3344	1.7538	1.7466	0.4084	1.6081	**1.6038**	0.2645	1.6173	1.6098	0.4645
Patio	2.0141	2.0109	0.1615	1.7829	1.7795	0.1928	1.6800	**1.6790**	0.0613	1.6884	1.6850	0.2042
Software	2.5957	2.5950	0.0275	2.3067	2.3070	−0.0150	**2.1557**	2.1595	−0.1737	2.1709	2.1693	0.0769
Sports and Outdoors	1.2856	1.2788	0.5300	1.3092	1.2978	0.8674	1.1316	1.1252	0.5670	1.1181	**1.1091**	0.8070
Tools and Home Improvement	1.8083	1.8005	0.4336	1.6191	1.6096	0.5884	1.5003	**1.4957**	0.3090	1.5139	1.5059	0.5254
Toys and Games	1.7295	1.7307	−0.0718	1.4240	1.4258	−0.1235	**1.3328**	1.3354	−0.1949	1.3415	1.3407	0.0575
Video Games	1.9005	1.8934	0.3719	1.5306	1.5496	−1.2414	**1.4617**	1.4877	−1.7787	1.4671	1.4887	−1.4710
Watches	1.5537	1.5514	0.1490	1.4799	1.4783	0.1073	1.4711	**1.4702**	0.0547	1.4824	1.4809	0.1017
平均 MSE	1.5637	1.5494	0.9181	1.4895	1.4661	1.5703	1.3377	**1.3310**	0.5003	1.3560	1.3402	1.1635

从表 2 可看出,通过引入修正项,推荐框架在多数数据集上提升基础预测结果的准确度. 当引入修正项后,4 种方法的表现都得以提升,最高提升达到 1.5703%. 需指出的是,针对引入修正项前后的预测结果进行成对(Pair-Wise)的 T 检验,结果表明所有提升对应的 p 值都小于 0.01,即引入修正项后性能显著提升.

3.4　沉默用户预测的效果分析

从表 1 可看出,实际系统中包含大量评分很少的沉默用户,基于评分的协同过滤算法很难对那些“沉默”用户作出准确推荐,基于很少的历史评分得到的用户相似度不可靠. 然而,通过引入评论信息和扩展相似用户,推荐框架可大幅提升沉默用户的预测精度.

为了说明本文方法在缓解数据稀疏性上的表现,使用 LFM 产生基础预测,统计 TopicCF 框架对 Movies 测试数据集中拥有不同评分数量用户(x 坐标)的 MSE 提升情况(y 坐标,等于 LFM 模型的 MSE 值减去 TopicCF 模型的 MSE 值,即 $MSE_{\mathrm{LFM}} - MSE_{\mathrm{LFM+Boost}}$),对应提升为正值表示使用修正项后可取得更好的预测效果,正值越大表示预测效果提升越明显. 实验结果如图 1 所示.

由图 1 可知,本文框架能全面提升评分预测的准确度,特别是对于历史评分数量很少的用户. 这是因为相比数值的评分,文本形式的评论包含的信息非常多. 即使通过很少的评论信息,也能大致刻画用户的偏好分布和商品的特征分布. 此外,TopicCF 对于历史评分数量大于 10 的用户预测准确度也有较大提升,这是因为随着用户评分增多变杂,在一定程度上掩盖用户真正的关注点,通过引入评论信息对用户建模,可同时从评分相似度和偏好相似度这两个角度衡量用户之间的相似性,从而更好找出与被

推荐用户相似的其他用户，使用高质量的用户参与协同推荐，提升真正相似的用户在协同推荐中所起的作用，更好提高推荐系统的准确性.

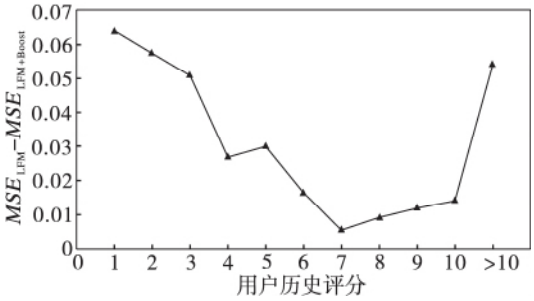
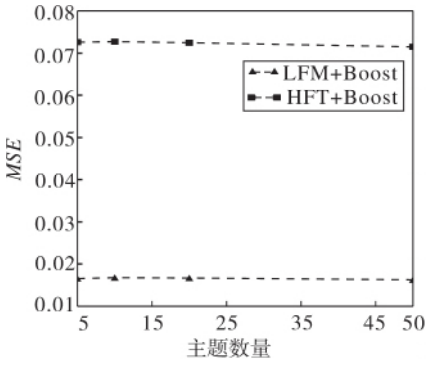


图 1 TopicCF 模型对 Movies 数据集上具有不同数量历史评分的用户的预测结果

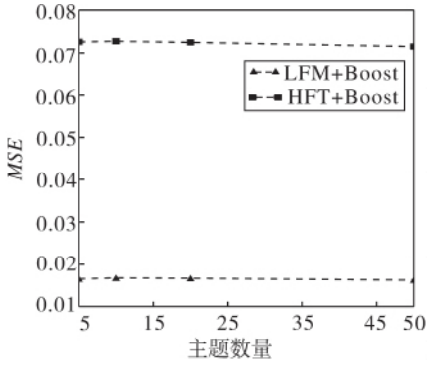
Fig. 1 Prediction result of TopicCF model on users with different number of historical ratings on Movies dataset

3.5 主题数量对推荐性能的影响

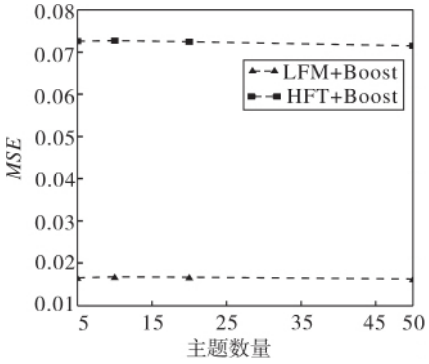
保持扩展商品数量为 5，在上述 4 个数据集分别设定主题数量为 5，10，20，50，研究主题数量对预测性能的影响，实验结果如图 2 所示.



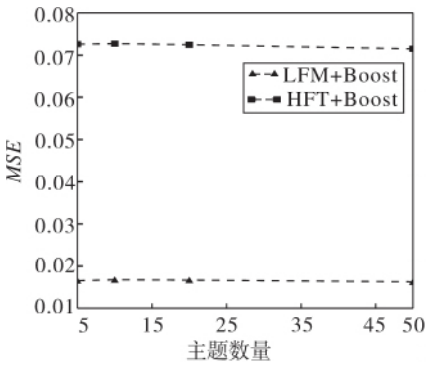
(a) Clothing and accessories



(b) Health



(c) Office products



(d) Tools and home

图 2 4 个数据集上主题数量对预测结果的影响

Fig. 2 Effect of topic number on prediction result on 4 datasets

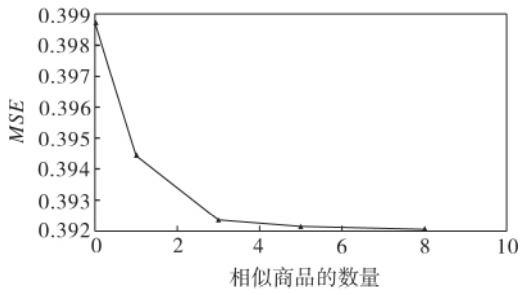
从图 2 可看出，随着主题数量的增加，模型的预测准确度提高甚微，即模型对于主题数量变化具有较好的稳定性. 这主要是因为虽然商品特征很多，但用户在评论中讨论且决定用户对该商品的态度的特征只有少数几个，所以可使用较少的主题数量得到较好的推荐效果.

3.6 相似商品数量选取对推荐性能的影响

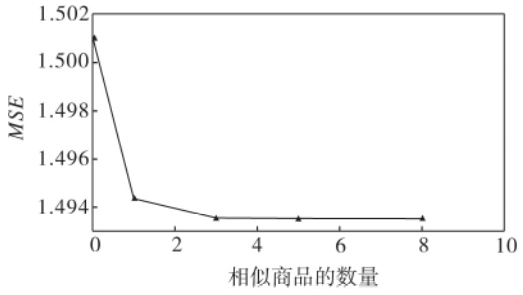
为了说明扩展相似商品对于提升推荐性能的作用，在主题数量等于 5 时，随机选取 4 个数据集研究预测准确度，选取最相似的前 0，1，3，5，8 个商品的变化情况. 当选取 0 个相似商品时，模型退化到 2.1 节中描述的基本模型. 实验结果如图 3 所示.

从图 3 可看出，随着选取相似商品的数量增加，模型的预测准确度也得到提升，特别是当相似商品的数量从 0 增到 1 时. 同时，当相似商品的数量增加到一定程度(如 5) 之后，预测性能趋于稳定. 因此框架只扩展与某一商品最相似的前5个商品，提升预

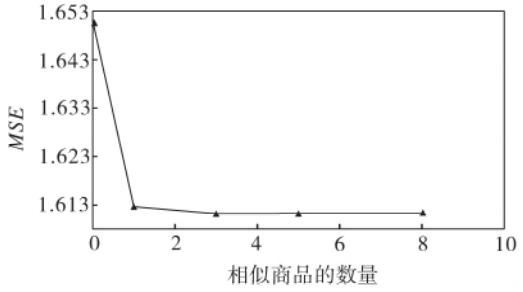
测准确度的同时减少模型的计算复杂性.



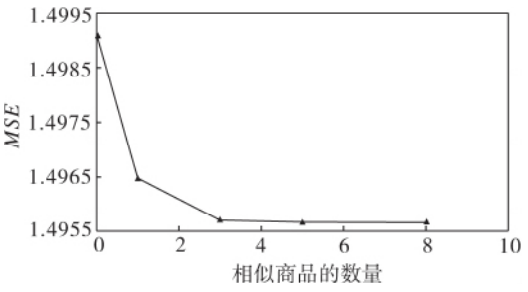
(a) Clothing and Accessories



(b) Health



(c) Office products



(d) Tools and home

图 3　4 个数据集上相似商品数量对预测结果的影响

Fig. 3　Effect of similar item number on prediction result on 4 datasets

4　结 束 语

本文提出 TopicCF 以融合评分和评论信息进行协同推荐,使用评论信息学习商品的特征分布和用户的偏好分布,将基于主题和偏好的相似度引入协同过滤框架中,通过扩展相似商品缓解数据稀疏性问题. 本文框架可方便地为现有推荐算法引入评论信息,同时提升参与协同推荐的用户质量和数量,进而提升推荐效果.

本文框架可通用地和现有的推荐算法结合,提升推荐性能; 在理解商品特征和用户偏好上具有较好性能,同时考虑用户在历史评分和个人偏好上的相似性可提升参与协同推荐用户的质量; 通过引入评论信息,扩展相似商品和相似用户,可较好缓解数据稀疏性对推荐效果的影响.

下一步将利用商品的特征分布和用户的偏好分布进行 TOP-*N* 推荐,同时拟使用商品的特征分布帮助商品生产者发现商品的优点和不足,从而更有针对性地改进商品.

参 考 文 献

[1] SU X Y, KHOSHGOFTAAR T M. A Survey of Collaborative Filtering Techniques // AGUIRRE A H, BORJA R M, GARCIA C A R, eds. Advances in Artificial Intelligence. New York, USA: Hindawi Publishing Corporation, 2009. DOI: 10.1155/2009/421425.

[2] GOLDBERG K, ROEDER T, GUPTA D, *et al*. Eigentaste: A Constant Time Collaborative Filtering Algorithm. Information Retrieval, 2001, 4(2): 133 – 151.

[3] MELVILLE P, MOONEY R J, NAGARAJAN R. Content-Boosted Collaborative Filtering for Improved Recommendations // Proc of the 18th National Conference on Artificial Intelligence. Edmonton, Canada, 2002: 187 – 192.

[4] ZIEGLER C N, LAUSEN G, SCHMIDT-THIEME L. Taxonomy-Driven Computation of Product Recommendations // Proc of the 13th ACM International Conference on Information and Knowledge Management. Washington, USA, 2004: 406 – 415.

[5] GANU G, ELHADAD N, MARIAN A. Beyond the Stars: Improving Rating Predictions Using Review Text Content [J/OL]. [2015 – 04 – 24]. http://paul.rutgers.edu/~gganu/resources/WebDB.pdf.

[6] JAKOB N, WEBER S H, MÜLLER M C, *et al*. Beyond the Stars: Exploiting Free-Text User Reviews to Improve the Accuracy of Movie Recommendations // Proc of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion. Hong Kong, China, 2009: 57 – 64.

[7] MUSAT C C, LIANG Y Z, FALTINGS B. Recommendation Using Textual Opinions // Proc of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 2684 – 2690.

[8] MCAULEY J, LESKOVEC J. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text // Proc of the 7th ACM Conference on Recommender Systems. Hong Kong, China, 2013: 165 - 172.

[9] LEUNG C W K, CHAN S C F, CHUNG F L. Integrating Collaborative Filtering and Sentiment Analysis: A Rating Inference Approach // Proc of the ECAI Workshop on Recommender Systems. Riva del Garda, Italy, 2006: 62 - 66.

[10] ZHANG Y F, LAI G K, ZHANG M, *et al*. Explicit Factor Models for Explainable Recommendation Based on Phrase-Level Sentiment Analysis // Proc of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. Gold Coast, Australia, 2014: 83 - 92.

[11] YU X H, LIU Y, HUANG J X J, *et al*. Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain. IEEE Trans on Knowledge and Data Engineering, 2012, 24(4): 720 - 734.

[12] TAN Y Z, ZHANG Y F, ZHANG M, *et al*. A Unified Framework for Emotional Elements Extraction Based on Finite State Matching Machine // Proc of the 2nd Conference on Natural Language Processing and Chinese Computing. Chongqing, China, 2013: 60 - 71.

[13] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003, 3: 993 - 1022.

[14] SARWAR B, KARYPIS G, KONSTAN J, *et al*. Item-Based Collaborative Filtering Recommendation Algorithms // Proc of the 10th International Conference on World Wide Web. Hong Kong, China, 2001: 285 - 295.

[15] KOREN Y. Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model // Proc of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2008: 426 - 434.

[16] HOFMANN T. Latent Semantic Models for Collaborative Filtering. ACM Trans on Information Systems, 2004, 22(1): 89 - 115.

作者简介

谭云志,男,1991 年生,硕士研究生,主要研究方向为机器学习、个性化推荐、情感分析. E-mail: cloudcompute09@gmail.com.

(**TAN Yunzhi**, born in 1991, master student. His research interests include machine learning, personalized recommendation and sentiment analysis.)

张 敏(通讯作者),女,1977 年生,博士,副教授,主要研究方向为信息检索与挖掘、用户行为分析、机器学习、推荐系统. E-mail: z-m@tsinghua.edu.cn.

(**ZHANG Min**(Corresponding author), born in 1977, Ph. D., associate professor. Her research interests include information retrieval and mining, user behavior analysis, machine learning and recommendation systems.)

刘奕群,男,1981 年生,博士,副教授,主要研究方向为网络搜索技术、信息检索、用户行为分析. E-mail: yiqunliu@tsinghua.edu.cn.

(**LIU Yiqun**, born in 1981, Ph. D., associate professor. His research interests include web search technology, information retrieval and user behavior analysis.)

马少平,男,1961 年生,博士,教授,主要研究方向为智能信息处理、信息检索、文本信息检索的模型与方法. E-mail: msp@tsinghua.edu.cn.

(**MA Shaoping**, born in 1961, Ph. D., professor. His research interests include intelligent information processing, information retrieval and models and methods of text information retrieval.)