



基于用户行为认知的在线社交网络协同推荐

亓 晋,许 斌,胡筱旋,徐匾珈,肖星琳

(南京邮电大学 南京 210003)

摘 要:近年来,在线社交网络成为人们工作、生活不可或缺的信息共享与交流工具,如何对海量庞杂、大范围时空关联的用户行为信息进行认知并据此提供个性化的推荐服务,已成为在线社交网络发展重点关注的问题。为此,提出了一种基于用户行为认知的在线社交网络协同推荐框架,在对用户特征、文本信息及兴趣偏好等行为进行认知的基础上,利用协同过滤算法,实现个性化的推荐服务。实验结果验证了提出的基于用户行为认知的协同推荐策略具有较好的稳定性和实际应用效果。

关键词:在线社交网络;行为认知;协同推荐;用户兴趣模型

doi: 10.11959/j.issn.1000-0801.2015273

Collaborative Recommendation for Online Social Networks Based on User Behavior Cognitive

Qi Jin, Xu Bin, Hu Xiaoxuan, Xu Bianjia, Xiao Xinglin

(Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Recently, online social network (OSN) has become essential tools for information sharing and communication in people's work and life. How to cognitive massive, complex, large-area and spatiotemporal association user behavior information and provide personalized recommendation services have become problems need special attention in development of OSN. Thus, a frame of collaborative recommendation for OSN based on user behavior cognitive was proposed which used collaborative filtering algorithm to provide personalized recommendation services on the basis of analysis of behavior of user characteristics, text information and interest preferences, etc. Experimental results verify that the proposed collaborative recommendation strategy based on user behavior cognitive has good stability and actual application effect.

Key words: online social network, behavior cognitive, collaborative recommendation, user interest model

收稿日期:2015-08-23;修回日期:2015-10-13

基金项目:江苏省产学研联合创新资金资助项目(No.BY2012035),南京邮电大学人才引进项目(No.NY213050, No.NY214098)

Foudation Items: Application & Research Joint Innovation Funding Product of Jiangsu Province(No.BY2012035), NUPTSF(No.NY213050, No.NY214098)

论文引用格式:亓晋,许斌,胡筱旋等.基于用户行为认知的在线社交网络协同推荐.电信科学,2015273

Qi J, Xu B, Hu X X, *et al.* Collaborative recommendation for online social networks based on user behavior cognitive. Telecommunications Science, 2015273



1 引言

随着移动互联网的飞速发展,Facebook、Twitter、微博、微信、社交网站等构成的在线社会网络^[1](online social network, OSN)成为人们生活不可或缺的信息共享与沟通工具,并逐渐成为社会舆论传播的新兴主流途径。一方面,OSN是连接现实社会和虚拟网络空间的纽带,汇聚了大量的关系、行为等可感知和可计算的人类社会的数字足迹。各种类型的OSN中产生的网络行为已经能够在很大程度上反映现实世界以及宏观舆论导向的真实状况,因而受到从政府机关到商业企业等各个领域的广泛关注;另一方面,OSN快速崛起,已发展为规模异常庞大且具有实时随机动态变化特征的行为载体,呈现出多源异构、信息传播突发涌现、用户和网络行为阵发性等特点,其上聚集的海量庞杂、多源异构、大范围时空关联的用户及网络行为信息的认知及个性化推荐正在成为社会科学和自然科学领域的研究热点。

2 相关工作

推荐系统在20世纪90年代中期成为一个独立的研究领域,被认为是应对“信息过载”的一个有效手段,现有的推荐系统可以分为3类:基于内容的推荐、协同过滤推荐(collaborative filtering recommendation, CFR)和混合推荐^[2-4],其中协同过滤推荐是应用最为广泛的推荐算法,它通过先推测出所有用户的偏好,再根据互为邻居的使用情况向其推荐相应的项目。

目前,国内外的学者和研究团队已经对面向OSN的推荐系统展开了一系列相关的研究工作^[5-6],并取得了一定成果。Sanchez等人结合协同过滤与基于内容的技术,提出了一种混合图像推荐系统,以创造一种高度适应任何类型用户的移动社交网络^[7]。Zanda等人设计和实现了一种移动社交活动推荐系统,整合Facebook社交网络移动数据和传感器数据向用户提出活动推荐^[8]。Lara等人将社交网络中的组推荐系统通过模型的实例化,应用于Facebook的电影推荐中^[9]。上述3篇参考文献分别从算法、数据来源、模型等各方面改善了OSN推荐系统,丰富了OSN推荐系统的内容。Sun Z B等人考虑到用户之间的社会关系,提出了一种社会正则化方法来提高推荐的准确性^[10]。Kardan等人基于社交网络上的信息及过去的广告和状态,提出了一种在社交网络平台上更具针对性的广告推荐框架,为用户提供

更精确的信息推荐^[11]。上述两篇参考文献利用用户的社会关系OSN上的用户信息来提高信息推荐的准确性,但总体来说,以上方法对于如何充分挖掘OSN上用户行为并进行更为有效的个性化推荐,仍存在一定的局限性。

随着对于认知科学研究的深入,将认知科学与信息科学相融合已经成为目前研究的热点之一。目前国内外研究多是集中在信息检索上^[12],形成了许多具有代表性的认知信息检索模型。如果能结合OSN用户特征、文本信息及兴趣偏好等可认知的用户行为与推荐系统相结合,并实现高效、个性化的推荐服务,将会是有意义的研究方向。因此,本文设计了一种基于用户行为认知的OSN协同推荐框架,在数据感知基础上实现微博用户特征、文本特征和用户兴趣偏好的行为认知,利用协同过滤算法,给出预测结果,实现个性化的电影推荐服务。

3 基于用户行为认知的在线社交网络协同推荐框架

基于用户行为认知的在线社交网络协同推荐框架分为3层,如图1所示。

第一层是数据感知层,负责从数据源感知相关数据,包括用户信息以及用户相关的文本信息数据,并将数据存入数据库中,方便行为认知层计算;第二层为行为认知层,是框架的核心,负责将第二层感知到的初始数据进行处理,包括用户特征分析、文本信息分析和用户兴趣建模3部分,认知结果将为上层的协同推荐提供支撑;第三层是协同推荐层,通过分析行为认知层所提交的认知结果,利用协同过滤算法,实现对用户的个性化推荐。

3.1 数据感知层

数据感知层主要负责感知数据源的特定信息,包括:用户信息,包括用户的基本信息、用户关注的好友等;用户发布的文本信息,包括用户发布的信息、用户转发的信息、用户评论和点赞等信息。常用的数据感知方式有网络爬虫以及OSN提供的API。

- 网络爬虫。最常用的数据感知方式是网络爬虫,负责从OSN获取用户及其文本信息以便后续的数据处理。以新浪微博为例,网络爬虫首先需要模拟用户的登录过程进行相关数据的爬取,登录成功后即可获取相关信息:一种是以文件的形式将用户及其微博信息进行记录;另一种是预先设定关键词,根据关键词匹配进行抓取数据。

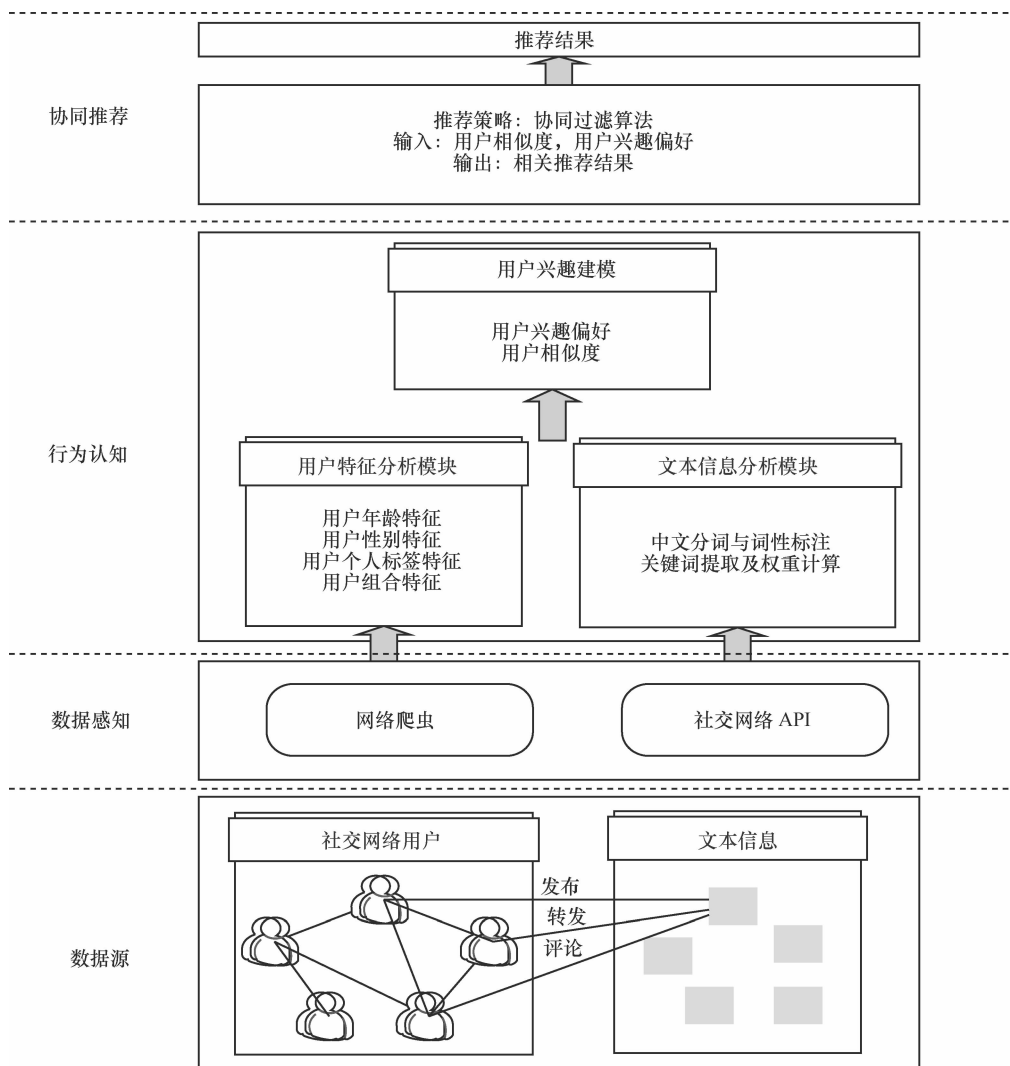


图 1 基于用户行为认知的在线社交网络协同推荐框架

- OSN 开放 API。除网络爬虫的方式外,部分 OSN 向企业及个人提供开放 API 以满足开发需求。该方法比网络爬虫更为简单高效,但对单次获取的信息量有一定的限制。

3.2 行为认知层

行为认知层主要负责将数据感知层获取的初始数据进行处理,通过对用户、文本信息分析,获取用户及文本特征,建立用户兴趣模型,为推荐层的协同推荐提供支撑。

3.2.1 用户特征

特征 1 (年龄特征)用户的年龄是一个自然数。为了避免数据稀疏的问题,将用户年龄以 15 岁为起始点(即高一学生的平均年龄)进行分组,每 6 年为一组<15~20>、<21~26>、<27~32>、<33~38>、<39~44>、<45~50>、<50 以上>。考虑个人隐私问题,将部分不填写或错误填写信息的用户

划为无年龄组,每组的编号为用户的年龄特征,即年龄特征=<15~20,21~26,27~32,33~38,39~44,45~50,50 以上,无年龄>。

特征 2 (性别特征)用户的性别特征分 3 类,即性别特征=<男性、女性、保密>。

特征 3 (个人标签特征)OSN 向用户提供自定义标签的功能或兴趣标签功能,用户可根据自己的喜好设置相关标签。本系统采用隐含狄利克雷分配 LDA 来分析用户的兴趣类别,通过计算用户潜在的兴趣向量,选择权重最好的一个作为用户的兴趣类别,指定 30 类兴趣,再加上部分用户不使用自定义标签,则共分 31 类。

特征 4 (组合特征)这是特征生成中较为常见的一种方式。本框架采用的组合特征包括<年龄×性别>、<年龄×标签>、<性别×标签>及<年龄×性别×标签>。



3.2.2 文本信息

本系统采用中国科学院开发的分词系统 ICTCLAS 系统进行文本信息分析,可有效进行中文分词、词性标注、未登录词识别等功能。

中文分词和词性标注后,需要进行关键词的提取。关键词的提取是文本处理过程中的关键步骤,选词的结果将会直接影响最终的分类结果。文本特征抽取算法各有优缺点,考虑到 OSN 用户行为认知中保留重要的词语的特点,该框架采用改进的 TF-IDF 算法^[13]。

首先,在 TF-IDF 中引入了置信度 $conf(w_j, c_m)$:

$$conf(w_j, c_m) = \frac{N(w_j, c_m)}{N(w_j, all)} \quad (1)$$

其中, $0 < conf(w_j, c_m) \leq 1$, c_m 为一个特定的类别, w_j 为特征词, $N(w_j, c_m)$ 表示 c_m 中包含特征词 w_j 的文档总数, $N(w_j, all)$ 表示的所有类别中包含特征词 w_j 的文档总数。

其次,引入特征词 w_j 的支持度 $sup(w_j)$:

$$sup(w_j) = \frac{N(w_j, all)}{N(all)} \quad (2)$$

其中, $0 < sup(w_j) \leq 1$, $N(all)$ 表示全部训练文档集中文档的总数, $N(w_j, all)$ 表示的所有类别中包含特征词 w_j 的文档总数。

由式(1)、式(2)可以得到一个主导度量:

$$dom(w_j, c_m) = f(conf(w_j, c_m), sup(w_j)) = \begin{cases} 1 & ((conf(w_j, c_m) \geq threshold \wedge sup(w_j) \geq threshold)) \\ 0 & ((conf(w_j, c_m) < threshold \wedge sup(w_j) < threshold)) \end{cases} \quad (3)$$

满足阈值 $dom(w_j, c_m) = 1$ 的特征词被当作关键词。若文档 d_j 中特征词集合包含关键词 $kw(w_j, c_m)$, 文档 d_j 就会被划分至 c_m 中。即关键词可以用于确定文档是否被分至一个特定的类别。

为了避免特征词的组合爆炸现象,该算法采取分层的方法进行选择,且能够较好地选择高频词汇作为可选择的关键词。

3.2.3 用户兴趣建模

在用户特征分析和文本特征分析的基础之上,基于分布估算法的建模方法构建用户兴趣模型^[14]。

首先找出用户“感兴趣”的项目集合 s_u , 定义 s_u 中的项目是用户 u 评分高于其平均评分的项目:

$$s_u = \{(t_1, r_{u,1}), (t_2, r_{u,2}), \dots, (t_n, r_{u,n})\} \quad (4)$$

其中, $r_{u,n}$ 表示用户 u 对项目 n 的评分, t_n 表示项目 n 的特征向量:

$$t_n = \{(k_1, w_1), (k_2, w_2), \dots, (k_j, w_j)\} \quad (5)$$

其中, k_j 是特征关键字, w_j 是 k_j 的权重, 而权重则由 TF-IDF 计算得出。

建模过程如下。

(1) 随机地从 s_u 中抽取 n 个关键词放入 K_u 中, 其中权重分别赋为 0.1、0.3、0.5、0.7、0.9:

$$K_u = \begin{Bmatrix} (k_1, 0.1) \cdots (k_1, 0.9) \\ \vdots \\ (k_n, 0.1) \cdots (k_n, 0.9) \end{Bmatrix} \quad (6)$$

(2) 初始化群体 $B_0 = \{\text{profile}_1, \text{profile}_2, \dots, \text{profile}_n\}$, 其中, $\text{profile}_n = ((k_i, w_i), (k_j, w_j), \dots, (k_k, w_k)), (k_i, w_i)$ 按照模型 $p = (((k_1, 0.1), c_{1,1}), \dots, ((k_1, 0.9), c_{1,5}), \dots, ((k_n, 0.9), c_{n,5}))$ 随机采样。其中 $c_{n,i}$ 表示 (k_i, w_i) 的选取概率。 $c_{n,i}$ 初始为 $1/N$, N 为 K_u 中元素个数。

(3) 当演化代数小于 Maxgen 时, 执行下列过程:

- 使用 $\text{Fitness}(\text{profile}_n) = \sum_{i \in s_u} \log \frac{r_{u,i}}{\text{sim}(\text{profile}_n, t_i)}$ 计算 B_0

中每个个体的适应值, 选择前 $M (M < n)$ 个个体更新 p , 用 X_s 表示选择后的优势群体。

- 使用更新后的 p 对 K_u 进行 N 次随机采样, 得到新群体, 并返回步骤(3)。

(4) 选择适应值最好的个体作为建模结果。

3.3 协同推荐层

协同推荐层通过使用协同过滤算法, 对行为认知层处理完成的数据进行分析, 利用修正余弦相似度寻找最近邻并产生相应推荐, 实现对用户的个性化推荐。

常用的相似度计算有余弦相似性、曼哈顿相似度、修正余弦相似度、Tanimoto 相似度和皮尔逊(Pearson)相似度。本框架选择修正余弦相似度来查找最近邻^[14]。

对于任意用户和, 相似度为:

$$\text{sim}(u_i, u_j) = \frac{\sum_{c \in I_{ij}} (w_{i,c} - \bar{w}_i)(w_{j,c} - \bar{w}_j)}{\sqrt{\sum_{c \in I_i} (w_{i,c} - \bar{w}_i)^2} \sqrt{\sum_{c \in I_j} (w_{j,c} - \bar{w}_j)^2}} \quad (7)$$

其中, I_{ij} 为兴趣模型中共有关键词的集合, $\bar{w}_i$ 是用户 u_i 关键词权重的平均值, $w_{i,c}$ 是用户 u_i 关键词 c 的权重。

得到最近邻后, 产生相应推荐。假设用户 u 的最近邻集合为 NN_u , 则 u 对推荐对象 i 的评分可以通过计算得出, 即:

$$PR_{u,i} = \bar{r}_u + \frac{\sum_{n \in NN_u} \text{sim}(u, n)(t_{n,i} - \bar{r}_n)}{\sum_{n \in NN_u} |\text{sim}(u, n)|} \quad (8)$$

其中, $sim(u, n)$ 指的是用户 u 和 NN_u 中的用户 n 的相似度, r_u 指的是用户 u 对推荐项目的评分, PR_{ui} 则是指预测。

4 实验结果分析

4.1 实验数据来源

本实验所用的数据来源于 GroupLens 研究小组提供的 MovieLens 公开数据集 (<http://movielens.umn.edu>), 包括 943 个用户对 1 682 个项目(影片)的 10 万条投票记录, 其中每个用户至少对 20 部电影进行过评价, 评分区间为 1~5 分。并选取其中 80 000 条作为训练集, 剩余 460 个用户的 20 000 条评价作为测试集。

4.2 实验环境

实验硬件环境为 Intel inside CORE-i5 系列 CPU、2.8 GHz 主频、4 GB 内存。实验软件环境为 Windows7 操作系统、MyEclipse、SQL Server 2008 数据库。

4.3 评价指标

为了验证本文所提出的推荐框架的有效性, 实验分别选取成功率、平均绝对误差(mean absolute error, MAE)和 Pearson 积矩相关系数(Pearson product-moment correlation coefficient, PCC)作为评价系统性能的指标, 通过分析 3 种评价指标验证框架的优越性。

(1) 成功率

成功率 r 表示系统推荐的 N 部电影与用户所有评分电影里的前 N 部的吻合程度。

(2) 平均绝对误差

mae 表示某个用户的预测评分与实际评分之间的偏差程度, 而 MAE 通过计算所有用户 mae 的均值来体现推荐系统的整体性能, mae 和 MAE 越小表明推荐系统质量越高。设预测的用户总数为 m , 用户 i 评分集为 $(p_{i1}, p_{i2}, \dots, p_{in})$, 实际的用户评分集 $(q_{i1}, q_{i2}, \dots, q_{in})$, 则 mae 的计算式为:

$$mae = \frac{\sum_{j=1}^n |p_{ij} - q_{ij}|}{n} \quad (9)$$

MAE 的计算式为:

$$MAE = \frac{\sum_{i=1}^m \sum_{j=1}^n |p_{ij} - q_{ij}|}{n \times m} \quad (10)$$

(3) Pearson 积矩相关系数

Pearson 积矩相关系数(PCC)衡量的是预测评分和真实评分的线性相关程度, 取值区间为 $[-1, 1]$, 越接近 1, 表示系统的预测值与实际值越相近, 即系统性能越好。设 r_α 表示

电影 α 的实际评分, $\bar{r}'_\alpha$ 表示预测评分, 则 PCC 计算式为:

$$PCC = \frac{\sum_{\alpha} (r'_\alpha - \bar{r}')(r_\alpha - \bar{r})}{\sqrt{\sum_{\alpha} (r'_\alpha - \bar{r}')^2} \sqrt{\sum_{\alpha} (r_\alpha - \bar{r})^2}} \quad (11)$$

4.4 实验结果及分析

4.4.1 成功率

分别对 460 个用户进行推荐 5 部电影和推荐 10 部电影的成功率分析, 从中随机选取 125 个用户成功率, 结果如图 2 所示。

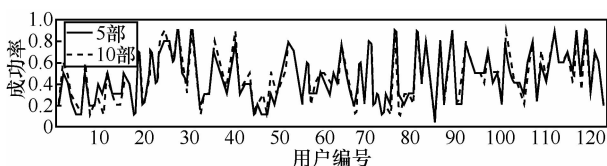


图2 推荐5部和10部电影的成功率

图2中横坐标表示随机挑选的125个用户, 纵坐标表示对某一用户推荐 N 部电影的成功率, 其中成功率为0的点表示该用户并未评价任何电影, 以致无法评价其成功与否。由图2可知, 推荐5部电影的实线与推荐10部电影的虚线大致吻合, 说明本文所提出的框架稳定性较好。经计算, 推荐5部和推荐10部的平均成功率分别为 $\bar{r}_5=37.13\%$ 、 $\bar{r}_{10}=55.84\%$, 均大于30%, 属较高水平。同时实验也选取了15部、20部、25部, 计算其平均成功率并与上述实验结果一起比较, 见表1。

表1 推荐不同部数平均成功率

推荐部数	成功率
5	37.13%
10	55.84%
15	66.23%
20	72.92%
25	77.65%

由表1可知, 推荐系统成功率与推荐部数密切相关, 推荐部数越多, 成功率越高。但如果推荐太多, 用户反而难以翻阅, 且增加系统负担, 因此结合成功率与实际情况, 推荐5部较为合适。

4.4.2 平均绝对误差

根据第4.4.1节对于成功率的分析, 计算出当推荐5部电影时, 每个用户的 mae 值如图3所示。

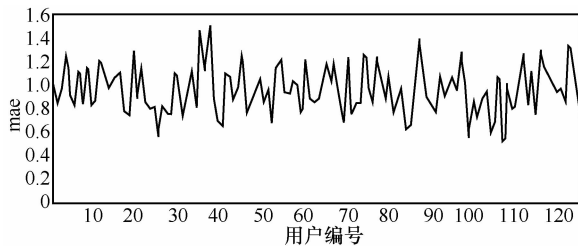


图3 每个用户 mae 分布

用户对电影的评分是 1~5 的整数,当用户未对电影评分时默认为 0。由图 3 可知,在推荐的电影数量为 5 部时,纵坐标代表各个用户的 mae 值,mae 值的大小表示用户实际评分与预测评分相差的数值。由式(10)计算出系统整体的 MAE 为 0.967,表示用户实际评分与预测评分相差为 1 分左右,说明用户的实际评分和系统的预测评分差距不大,预测值基本能表示用户对该电影的喜恶程度。

4.4.3 Pearson 积矩相关系数

利用电影的评分向量,计算每部电影的 PCC,以估计预测评分和实际评分的线性相似度,结果如图 4 所示。

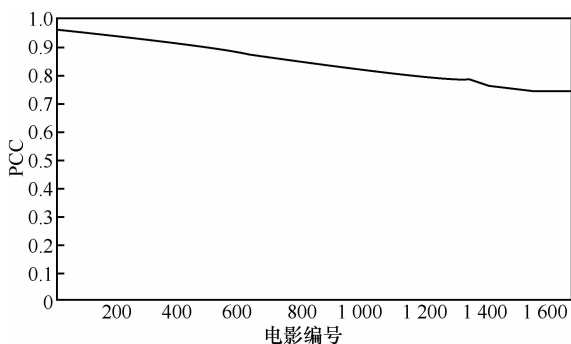


图4 1680 部电影的 PCC 值

图 4 中,PCC 指标表示越接近 1,则系统预测评分与用户实际评分越接近,说明该系统性能较好。由图 4 可知,1600 部电影对应的 PCC 值始终位于 0.7 以上,处于一个比较高的水平,说明本文提出的推荐系统的预测值与实际值较为接近,推荐系统的推荐质量较高,性能较好。

5 结束语

随着社交网络的不断发展,用户与用户之间的社交关系也愈加复杂,社交网络每天也会产生愈加庞大的数据信息,如何从中发掘有用的信息并为用户提供合适的个性化推荐有着重要的意义。实验结果表明,同时考虑用户社交特征以及用户的数据信息能够更为有效地为服务推荐提供决策支持,使得成功率以及误差都有着较

好的结果。本文构建了一个基于协同过滤算法的推荐框架,通过分析用户的社交特征以及用户的数据信息,为用户进行兴趣建模,从而为用户进行合适的推荐。实验证明,本框架具有较好的成功率和实验误差,具有良好的应用效果。

参考文献

- 徐恪,张赛,陈昊等. 在线社会网络的测量与分析. 计算机学报, 2014,36(9): 165~188
Xu K, Zhang S, Chen H, *et al.* Measurement and analysis of online social network. Chinese Journal of computers, 2014, 36(9): 165~188
- Burke R. Hybrid recommender systems: survey and experiments. User Modeling and User-Adapted Interaction, 2002, 12(4): 331~370
- Herlocker, J L, Konstan J A, Terveen K, *et al.* Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 2004, 22(1): 5~53
- Hofmann T. Latent semantic models for collaborative filtering. ACM Transactions on Information Systems, 2004, 22(1): 89~115
- Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734~749
- Resnick P, Varian H R. Recommender systems. Communications of the ACM, 1997(40): 56~58
- Sanchez G, Faustion A, Barrilero G, *et al.* Social and content hybrid image recommender system for mobile social networks. Mobile Networks and Applications, 2012, 17(6): 782~795
- Zanda A, Eibe S, Menasalvas E. SOMAR: a social mobile activity recommender. Expert Systems with Applications, 2012, 39(9): 8423~8429
- Quijano-Sanchez L, Diaz-Agudo B, Recio-Garcia J A. Development of a group recommender application in a social network. Knowledge-Based Systems, 2014(71): 72~85
- Sun Z, Han L, Huang W, *et al.* Recommender systems based on social networks. Journal of Systems Software, 2015(99): 109~119
- Kardan A A, Hooman M. Targeted advertisement in social networks using recommender systems. Proceedings of 7th International Conference on e-Commerce in Developing Countries: With Focus one-Secury(ECDC), Kish Island, Iran, 2013
- Koren J, Zhang Y, Liu X. Personalized interactive faceted search. Proceedings of World Wide Web Conference Series, Beijing, China, 2008: 477~486
- Zhang Y T, Gong L, Wang Y C. An improved TF-IDF approach for text classification. Journal of Zhejiang University - Science A: Applied Physics & Engineering, 2005, 6A(1): 49~55

- 14 梁天一, 梁永全, 樊健聪等. 基于用户兴趣模型的协同过滤推荐算法. 计算机应用与软件, 2014(11)

Liang T Y, Liang Y Q, Fan J C, *et al.* Collaborative filtering recommendation algorithm based on user interest model. Computer Applications and Software, 2014(11)

[作者简介]



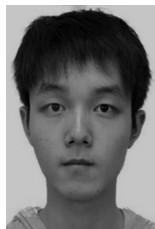
亓晋,男,博士,南京邮电大学物联网学院讲师,主要研究方向为新一代网络、大数据管理与智能计算、云计算。



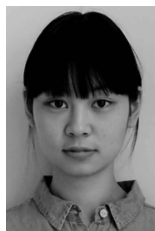
许斌,男,博士,南京邮电大学物联网学院讲师,主要研究方向为智能计算。



胡筱旋,女,南京邮电大学硕士生,主要研究方向为模式识别与智能系统。



徐匾珈,男,南京邮电大学硕士生,主要研究方向为智能优化算法。



肖星琳,女,南京邮电大学本科在读,主要研究方向为智能优化算法。