第44卷第7期 中国科学技术大学学报 Vol. 44, No. 7
2014年7月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jul. 2014

文章编号:0253-2778(2014)07-0576-06

一种融合 PageRank 的协同过滤帖子推荐方法

曹 阳[1],刘 松[1],郭剑毅[1,2],余正涛[1,2],周枫[1,2],毛存礼[1,2]

(1.昆明理工大学信息工程与自动化学院,云南昆明 650504;
2.昆明理工大学智能信息处理重点实验室;云南昆明 650504)

摘要:针对贴吧用户面临严重的信息过载问题,提出一种基于用户信息的协同过滤帖子推荐方法.分析帖子推荐的属性特点后,首先利用一个融合了用户评论行为的 PageRank 算法去判断参与一个帖子讨论中各用户的重要性,主要考虑各用户之间的回复关系以及各用户之间回复的时间关系;然后把 PageRank 得分高的用户作为聚类中心进行 k-means 聚类;最后把帖子中聚类得到的用户与推荐系统使用者通过协同过滤算法计算相似度,并结合用户的 PageRank 得分,选择与用户相关度较高的帖子作为推荐结果.实验结果表明,该模型比现在使用的热门帖子推荐有着更好的表现.

关键词:帖子推荐;PageRank;协同过滤;百度贴吧

中图分类号:TP311 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2014.07.006

引用格式:Cao Yang, Liu Song, Guo Jianyi, et al. A posts recommendation method based on the collaborative filtering and PageRank[J]. Journal of University of Science and Technology of China, 2014,44(7):576-581.

曹阳,刘松,郭剑毅,等. 一种融合 PageRank 的协同过滤帖子推荐方法[J]. 中国科学技术大学学报,2014,44(7):576-581.

A posts recommendation method based on the collaborative filtering and PageRank

CAO Yang[1], LIU Song[1], GUO Jianyi[1,2], YU Zhengtao[1,2], ZHOU Feng[1,2], MAO Cunli[1,2]

(1. *School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650504, *China*;
2. *Key Laboratory of Intelligent Information Processing Kunming University of Science and Technology, Kunming* 650504, *China*)

Abstract: In order to solve the problem of information overload in the post bar, a method of information filtration was proposed based on the user's commenting behavior. After analyzing the properties of the recommended posts, the importance of an individual user was evaluated by the PageRank algorithm, in which the weight of replies to the posts among users and the weight of reply intervals were taken into consideration. The users with a high PageRank score were then taken as a cluster center in k-means clustering. The similarity between two groups of users (one from the clustering analysis and the other from the recommending system) was calculated by a collaborative filtering algorithm. The posts with high correlations to the users were presented as the recommended results. Experimental results show that the

收稿日期:2014-03-21;**修回日期**:2014-06-15

基金项目:国家自然科学基金(61175068),云南省教育厅基金重大专项项目(KKJI201203001)资助.

作者简介:曹阳,男,1987年生,硕士生. 研究方向:数据挖掘,机器学习. E-mail:724728777@qq.com

通讯作者:郭剑毅,教授. E-mail:gjade86@hotmail.com

proposed method performs better than the recommending methods in use.

Key words: topics recommendations; PageRank; collaborative filtering; Baidu Post Bar

0　引言

百度贴吧是中国最流行的 SNS 之一，用户可以在自己的主页里面找到已加入贴吧中最新和最热门的帖子. 随着贴吧加入人数越来越多，每个贴吧的发贴量不断增加. 用户既没有时间也没有能力去把每个热门的帖子读完，即面临着严重的信息过载问题. QQ 空间、人人网、新浪微博等热门社交网站均面临着此类问题[1]，因此，急需一种为用户自动推荐有用帖子的解决方案. 本文结合百度帖吧成员关系特征，利用用户的评论行为寻找出最适宜的帖子，作为推荐结果.

当前，解决推荐问题主要有以下方法：一是基于内容(content-based)的推荐算法，例如 Somlo 等[2]以及 Zhang 等[3]等提出了利用自适应过滤技术更新用户配置文件. 利用用户的喜好信息构建配置文件，把用户的兴趣点归纳为几个主题文件，依次对比 Web 的文本内容与主题文件的相似度，把相似度较高的 Web 展示给用户并更新用户的配置文件，但该方法容易把与用户配置文件不相关的文件也引入相似度概率运算，从而加大运算量；二是基于网络结构(network structure-based)的推荐算法，例如周涛等[4-5]等和 Huang 等[6-7]等分别利用用户-产品两部分图(bipartite network)建立二者的关联关系，并据此提出了基于网络结构的推荐算法. 该方法引入两部分图上的扩散动力学部分解决了数据稀疏性问题，但并没有完全改善冷启动问题；三是协同过滤(co-llaborative filtering)推荐算法，该方法在推荐系统中应用最广泛[8]，在书籍、电影、音乐等推荐系统中有着很好的表现[9]. 针对帖子推荐问题，张新猛等[10]基于协同过滤，利用已有用户历史信息来预测用户对未读帖子的喜好概率，把概率高的帖子作为推荐结果给用户，由于用户评分数据的极端稀疏性，其相似性度量方法不能有效地计算目标用户的最近邻居，从而导致推荐质量难以保证.

本文针对帖子推荐的数据额稀疏问题，在传统的协同过滤推荐算法基础上，提出了基于用户聚类并融合 PageRank 算法的协同过滤算法，并将其应用于百度贴吧推荐. 该方法的优点在于不同聚类中的用户之间有较明显的区别，由于 k-means 聚类算法的扩展性相对较好，从而可以有效地解决用户评分数据极端稀疏情况下传统相似性度量方法存在的不足，使得计算得到目标用户的最近邻居比较准确，并且融合 PageRank 算法后使得推荐结果的准确性进一步提高.

1　基于用户评论行为的贴吧推荐

基于用户评论行为的贴吧推荐方法，首先对百度分别赋予用户回帖和回帖时间不同权重，计算每个用户的 PageRank 得分，选得分高的用户作为聚类中心进行 k-means 聚类，利用得到的聚类结果与目标用户通过 pearson correlations 计算相似度；然后结合每个用户的 PageRank 得分，找出得分高的帖子，作为结果推荐给用户. 推荐方法的具体流程如图 1 所示.

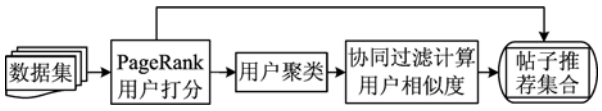

图 1　融合 PageRank 的协同过滤帖子推荐方法流程图

Fig. 1　Fusion PageRank collaborative filtering recommendation method flowchart

1.1　百度论坛贴子特征的分析

定义 1.1　贴吧的属性. 百度贴吧由一系列的主题贴吧组成，每个主题贴吧由若干个管理员与加入该贴吧的用户组成；一个主题贴吧由一系列帖子组成，帖子集合包括发帖、回帖、阅读记录. 每个帖子在主题贴吧中按照最后一次回复时间排列.

定义 1.2　帖子的属性. 每一个帖子都有一个原始的发布者，再加上不同级别的回复组成. 当回复级别超过二级时，回复的内容只显示在二级回复上，但可以通过@username 来判断他回复的对象. 每个回复都标有发表时间.

定义 1.3　用户的属性. 在百度贴吧中，存在着 2 种社会关系：朋友关系(friendship)和成员关系(membership). 这两种关系的差别在于它们基于不同的 2 个实体，用户(users)与贴吧(group)，而相对于 friendship 来说，考虑 membership 更具有优势. 因为一个用户加入一个贴吧并在上面有着很好的分数，这意味着该用户对这个贴吧有着兴趣. 而 friendship 可能是模糊的，因为两个人可能由于很

多种原因彼此加为朋友.

1.2 基于 PageRank 的用户打分

得到一个帖子,里面各参与者的重要性是一个首先需要被解决的问题.本文采用回帖人的回复关系去测量一个帖子里面每个人的重要性.这里使用 $\boldsymbol{M}^r$ 去表示帖子中的回复,$\boldsymbol{M}^r$ 是一个 $m\times m$ 的邻接矩阵.$M^r_{i,j}$ 的值代表用户 j 回复用户 i 的次数与用户 j 总回复的次数之比.这里也考虑了回复的时间对测量帖子参与人的影响.目前百度贴吧推荐系统中的热门帖子,原始发帖时间和用户看见它的时间差比较短,这就使确定每个回复的回复时间成为可能.这里定义:用户 j 回复用户 i 的最后一个帖子的时间与回复第一个帖子的时间差为用户 j 关注用户 i 的时间.这里使用 $\boldsymbol{M}^t$ 去表示回复时间关系,它也是一个 $m\times m$ 的邻接矩阵,$M^t_{i,j}$ 表示用户 j 回复用户 i 时间与用户 j 回复总时间之比.最后将 2 个矩阵线性结合起来.根据实际情况,赋予 α 和 β 一定的权值,并规定,$\alpha+\beta=1$ 则有:

$$M_{i,j}=\alpha\times M^r_{i,j}+\beta\times M^t_{i,j} \tag{1}$$

一般来说,在一个帖子中表现越重要人的回帖,会被越多的人回复,并且会有越多时间被用来回复他的帖子.于是本文采用一个融合用户评论行为的 PageRank 的算法对一个帖子参与人的重要性进行排序,用户得分表示为:

$$p(i)=\frac{d}{n}+(1-d)\times\sum_{u_j}r(u_i,u_j)\times p(j) \tag{2}$$

式中,d 为阻尼系数,且 $0<d<1$,通常取值为 0.15.n 代表着参加这个帖子的人数,$p(i)$ 表示用户 i 的 PageRank 得分,u_i 和 u_j 分别代表用户 i 和用户 j.而 $r(u_i,u_j)$ 代表这用户 i 与用户 j 的回复关系,它被定义为:

$$r(u_i,u_j)=\frac{M_{i,j}}{\sum\limits_{u_k}M_{i,k}+\lambda} \tag{3}$$

式中,λ 是一个常量,为了防止除数为 0 的情况.将式(3)带入式(2)得到:

$$p(i)=\frac{d}{n}+(1-d)\times\frac{M_{i,j}}{\sum\limits_{u_k}M_{i,k}}\times p(j) \tag{4}$$

1.3 基于用户聚类的协同过滤算法

在得到贴吧用户 PageRank 分值后,选择得分高的用户作为聚类中心进行 k-means 聚类.然后把聚类后的结果和目标用户利用 pearson correlations 去计算用户相似度,最后再结合用户的 PageRank 得分,选择与用户相关度较高的帖子作为推荐结果.

1.3.1 k-means 用户聚类

为了解决一些新发的帖子没有人评论所带来的数据稀疏和冷启动问题,并且提高系统的推荐速度,本文采用 k-means 聚类方法,选择用户对帖子的评分作为特征向量,先对用户进行聚类并生成相应的聚类中心,然后计算目标用户与各聚类中心的相似度,选出与目标用户相似度最高的 k 个聚类中心对应的聚类簇,在这 k 个聚类簇中搜索目标用户的最近邻居,从而达到在尽量少的用户空间中找到目标用户的大部分最近邻居.其具体过程如下:

(Ⅰ)从 m 个用户中选择 PageRank 得分最高的 k 个用户作为初始的聚类中心,记为 $\{w_1,w_2,\cdots,w_k\}$,其中:$W^*_j=X_{i,j}\in\{1,2,\cdots,k\}$,$i\in\{1,2,\cdots,n\}$,使每一个聚类 c_j 与聚类中心相对应.

(Ⅱ)重复每一个输入向量 $\boldsymbol{X}_i$,其中 $i\in\{1,2,\cdots,n\}$,将 $\boldsymbol{X}_i$ 分配给最近的聚类中心 W^*_j 所属的聚类 c^*_j;为每一个聚类 c_j.其中,$j\in\{1,2,\cdots,k\}$,将聚类中心更新为当前的 c_j 中所有样本的质心点,即

$$w_j=\sum_{i_j\in c_j}i_j/c_j \tag{5}$$

计算误差函数:

$$E=\sum_{j=1}^{k}\sum_{i_j\in c_j}|\ i_j-w_j\ |^2 \tag{6}$$

(Ⅲ)迭代直到改变不再明显或者聚类的成员不再变化.传统 k-means 聚类算法的初始聚类中心是随机选取的,实验过程中发现,聚类后会出现较多孤立点.这是因为协同过滤算法是在搜索最近邻居的基础上进行推荐的,无法对孤立点进行个性化推荐.研究发现,PageRank 得分高的用户可以代表一部分用户,这些用户作为聚类中心具有很好的代表性,因此,本文选择访问量高的 k 个用户作为初始聚类中心,经实验验证能较好地减少孤立点.

1.3.2 协同过滤推荐

本文针对用户的 membership 采用 pearson correlations[11] 去计算用户相似度

$$\mathrm{sim}(i,j)=\frac{\sum\limits_{u\in U}(r_{u,i}-\bar{r}_i)(r_{u,j}-\bar{r}_j)}{\sqrt{\sum\limits_{u\in U}(r_{u,i}-\bar{r}_i)^2}\ \sqrt{\sum\limits_{u\in U}(r_{u,j}-\bar{r}_j)^2}} \tag{7}$$

式中,$\mathrm{sim}(i,j)$ 表示用户 i 和用户 j 的相似性;$r_{u,i}$ 表示用户 i 对帖子 u 的评分;$u_{u,j}$ 是用户 j 对帖子 u 的评分;$\bar{r}_i$ 表示用户 i 对群组 U 帖子中的平均评分;

$\bar{r}_j$ 表示用户 j 对群组 U 帖子中的平均评分. 根据 pearson correlations 计算出目标用户与邻居用户的相似度后,则对目标用户未读过的帖子进行评分.

$$\text{Score} = \bar{r}_i + \frac{\sum \text{sim}(i,j) \times (r_{u,j} - \bar{r}_j)}{\sum \text{sim}(i,j)} \quad (8)$$

式中,Score 是帖子的分数;$\bar{r}_i$ 表示目标用户 i 对群组 U 帖子中的平均评分,$\bar{r}_j$ 表示邻居用户 j 对群组 U 帖子中的平均评分,$r_{u,j}$ 是邻居用户 j 对推荐帖子 u 的评分;$\text{sim}(i,j)$ 为目标用户 i 与用户 j 的相似度值.

1.3.3　融合 PageRank 的协同过滤算法

式(8)为传统的协同过滤算法,还没考虑帖子用户的重要性,本文结合帖子用户的重要性与帖子用户回复之间的关联度作为目标用户进行帖子推荐. 在传统的协同过滤算法基础上,结合帖子用户的 PageRank 值进行推荐. 融合思想采用加法和乘法方式,其融合速度快,实时性好. 结合 1.2 小节与 1.3.2 小节的内容,融合 PageRank 的协同过滤算法可通过下式表示:

$$\text{Score} = \bar{r}_i + p(i) + p(j) \times \frac{\sum \text{sim}(i,j) \times (r_{u,j} - \bar{r}_j)}{\sum \text{sim}(i,j)} \quad (9)$$

式中,Score 是帖子的分数;$\bar{r}_i$ 表示目标用户 i 对群组 U 帖子中的平均评分,$\bar{r}_j$ 表示邻居用户 j 对群组 U 帖子中的平均评分,$r_{u,j}$ 是邻居用户 j 对推荐帖子 u 的评分;$P(j)$ 为邻居用户 j 在帖子中所占的重要性,即用户 j 的 PageRank 得分,$P(i)$ 为目标用户 i 的 PageRank 得分;$\text{sim}(i,j)$ 为目标用户 i 与邻居用户 j 的相似度值;然后按照帖子得分的大小,在目标用户的 homepage 上面的热门帖子进行排序推荐.

2　实验和分析

2.1　实验数据获取

本文实现了一个贴吧推荐系统,并将上述提出的算法用于该系统. 系统基于 Java 语言实现,采用流行的 Web 开发框架 django,数据库采用面向文档的数据库 CouchDB;在服务器端安装了 chrome 并导入插件. 实验的数据集是由上述搭建的系统获取,通过网页爬虫抓取百度贴吧的帖子页面组成数据集. 论坛数据记录了发帖者、回复者、发帖时间、回帖标志、回帖时间等,总共包括5 261条发帖或回帖记录. 首先对数据进行预处理,提取用户数据集,包括发过帖或回过帖的 572 个用户;对全部的用户采用 PageRank 进行打分,部分记录如表 1 所示.

表 1　帖子内重要用户排序

Tab. 1　Important users sort of the posts

序号	用户	Pagerank 得分
1	雨轩觉	8.564 949
2	胖爷爱明器	7.048 351 2
3	美夕子	6.186 294
4	拉小困	5.645 916
5	拖棒棒主	5.386 274
6	白面玫瑰十一	5.145 831

2.2　实验及结果分析

为了验证提出方法的推荐效果,本文设计了 2 个实验:①在数据集中的情况下比较本文提出的算法与传统协同过滤算法的推荐效果;②在数据稀疏的情况下比较本文提出的算法与传统协同过滤算法的推荐效果.

2.2.1　数据集中情况下算法比较

本文采用准确率(precision)、召回率(recall)和 F_{measure} 作为改进算法,即融合了 PageRank 和 k-means 聚类的协同过滤算法(PageRank & k-means collaborative filtering,PKCF)推荐质量的衡量标准,并将其与传统的基于用户的协同过滤(collaborative filtering,CF)算法进行比较. 改进的协同过滤算法和传统协同过滤算法关于推荐的准确率对比如图 2 所示. 图 2 中选取了 8 个用户进行实验测试,最后一组数据是这 8 个用户数据的均值,从该组数据中不难看出,改进的算法推荐的准确率提高了约 7%.

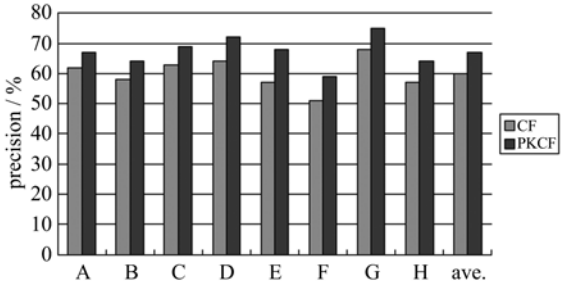

图 2　改进算法与传统协同过滤推荐准确率对比

Fig. 2　Performance comparison for the precision at different collaborative filtering algorithm

改进的协同过滤算法和传统协同过滤算法关于召回率对比分别如图 3 所示. 图 3 中也选取了 8 个用户进行实验测试,最后一组数据是这 8 个用户数

据的均值,从该组数据中不难看出,改进的算法推荐的召回率提高了约 5%.

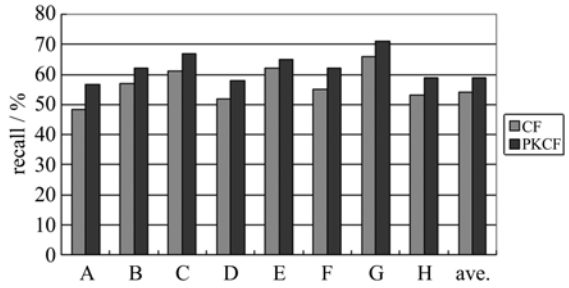

图 3 改进算法与传统协同过滤推荐召回率对比

Fig. 3 Performance comparison for the recall at different collaborative filtering algorithm

为了更好地比较算法的推荐质量,我们引入了 F_{measure} 这项评价指标. 为准确率和召回率的调和平均值,该值越高则表明推荐算法的综合性能越好.

$$F_{\text{measure}} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{10}$$

图 4 中选取了前面 8 个用户的实验准确率和召回率求得 F_{measure} 评价指标,最后一组数据是这 8 个用户数据的均值,从该组数据中不难看出,改进的算法在推荐综合性能和推荐质量上,都有较大的优势.

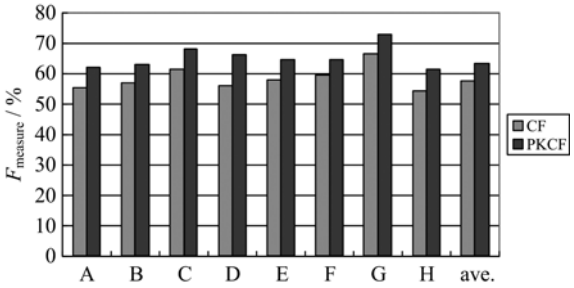

图 4 改进算法与传统协同过滤推荐 F_{measure} 对比

Fig. 4 Performance comparison for the F_{measure} at different collaborative filtering algorithm

2.2.2 数据稀疏情况下算法比较

实验中观察到,两种算法在面对稀疏数据集的情况时,其产生推荐结果的能力是不同的,图 5～7 显示了不同数据稀疏度情况下,本算法(PKCF)与传统的协同过滤算法(CF)在准确率、召回率和 F 值三项指标的对比. 图 5 显示了不同数据稀疏度情况下,PKCF 方法 CF 方法在准确率指标的对比.

从实验结果可以看到,在极端数据稀疏的情况下(密度低于 10%),传统协同过滤算法准确性较差,在密度超过 10%时逐渐提升. 而本算法在数据

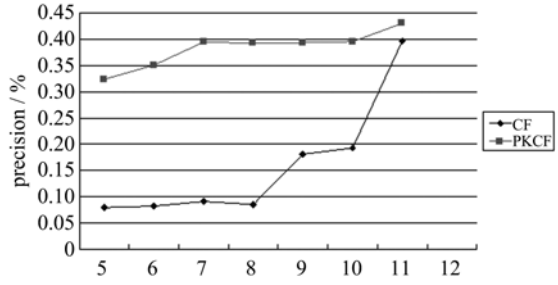

图 5 推荐准确率对比

Fig. 5 Performance comparison for the precision at different collaborative filtering algorithm

极端稀疏的情况下仍能保持较高的推荐准确率,并且随着数据密集度增大也能逐步提升,这与本算法在用户聚类过程中实现了数据的密集化有关.

图 6 显示了不同算法间召回率的对比. 可以看到,本算法在稀疏数据集中,召回率保持在稳定的较高的水平. 而 CF 的方法受数据稀疏性的影响较大,产生的推荐数量和质量都比较差. 图 7 是不同算法的 F 值比较.

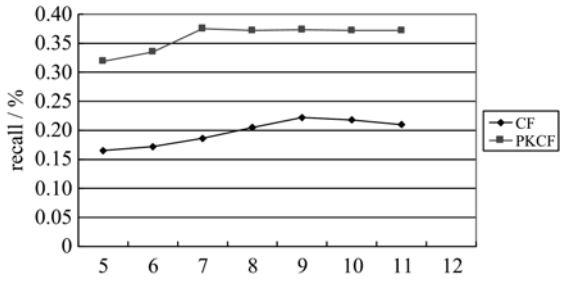

图 6 推荐准确率对比

Fig. 6 Performance comparison for the recall at different collaborative filtering algorithm

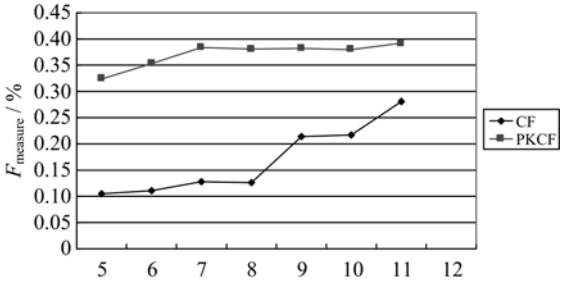

图 7 推荐准确率对比

Fig. 7 Performance comparison for the F_{measure} at different collaborative filtering algorithm

从实验结果可以看到,本算法无论在推荐综合性能和推荐质量上,都有较大的优势. 尤其在百度贴吧应用中,由于普遍存在的数据稀疏性(常常低于

10%),本算法相比传统方法具有显著的优势.

2.3　系统运行结果及分析

将本文提出的算法模型应用于百度贴吧推荐系统,得到的百度贴吧帖子的信息推荐结果如图 8,9 所示.

图 8　使用推荐系统前贴子显示状态

Fig. 8　Before using the recommended system performance of postings

图 9　用 chrome 浏览器显示推荐结果

Fig. 9　Showing results recommend using chrome browser

从图 8 可看出,未使用帖子推荐系统前,帖子信息范围杂乱,不易寻找热门及感兴趣的帖子.

从图 9 可看出,使用推荐系统后,帖子的信息较为集中,降低了信息的冗余量,解决了信息过载问题.

3　结论

本文提出一种融合了 PageRank 的协同过滤帖子推荐方法以解决百度贴吧帖子信息过载问题. 首先通过 PageRank 打分判断了每个评论参者的重要性;然后采用 k-means 聚类解决了实际情形中数据稀疏的问题,并且加快了生成推荐的运行速度. 实验数据表明,本信息推荐系统的使用,可以大大减少信息的过载及冗余量,节省用户的时间,给用户好的体验感,在公共社交网络、电子商务平台中均可发挥重要作用.

参考文献(References)

[1] Li Q, Wang J, Peter Chen Y Z, et al. User comments for news recommendation in forum-based social media[J]. Information Sciences, 2010, 180(24): 4 929-4 939.

[2] Somlo G, Howe A E. Adaptive lightweight text filtering[C]// Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis. London, UK: Springer, 2001: 319-329.

[3] Zhang Y, Callan J, Minka T. Novelty and redundancy detection in adaptive filtering[C]// Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Tampere, Finland: ACM Press, 2002: 81-88.

[4] Zhou T, Ren J, Medo M, et al. Bipartite network projection and personal recommendation[J]. Physical Review E, 2007, 76(4): 215-226.

[5] Zhou T, Jiang L L, Su R Q, et al. Effect of initial configuration on network-based recommendation[J]. Europhys Lett, 2008, 81(5): 729-736.

[6] Huang Z, Chen H, Zeng D D. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering [J]. IEEE Transactions on Information Systems, 2004, 22(1): 116-142.

[7] Huang Z, Zeng D D, Chen H. Analyzing consumer-product graphs: Empirical findings and applications in recommender systems[J]. Management Science, 2007, 53(7): 1 146-1 164.

[8] Su X Y, Khoshgoftaar T M. A survey of collaborative filtering techniques [J]. Advances in Artificial Intelligence, 2009, No. 4: 1-19.

[9] McLaughlin M R, Herlocker J L. A collaborative filtering algorithm and evaluation metric that accurately model the user experience[C]// Proceedings of 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, UK: ACM Press, 2004: 329-336.

[10] 张新猛,蒋盛益. 基于协同过滤的网络论坛个性化推荐算法[J]. 计算机工程, 2012, 38(5): 67-69.

[11] Shardanand U, Maes P. Social information filtering: Algorithms for automating "word of mouth"[C]// Proceedings of the SIGCHI conference on Human Factors in Computing Systems. Chicago USA: ACM Press, 1995: 210-217.