文章编号: 1007-5321(2014)增-0120-05

网络资源中基于 *K*-Means 聚类的个性化推荐

王 鑫， 黄忠义
(潍坊学院 计算机工程学院，山东 潍坊 261061)

摘要: 为了实现在网络资源中为网络用户提供针对兴趣爱好的推荐项目，提出了一种基于 *K*-means 聚类的应用于动态多维社会网络的个性化推荐算法. 首先根据用户评分数据对用户进行建模，并根据评分数据集构建多维用户网络，再加入局域世界演化理论形成动态多维网络; 然后根据改进的 *K*-means 算法对用户聚类; 最后根据最近邻居得到目标用户的预测评分作出推荐，从而形成一种应用于动态多维社会网络中的个性化推荐算法. 实验表明，相比协同过滤个性化推荐系统，新推荐策略的预测值和真实值之间的误差较小，个性化推荐水平得到了一定程度的提高.

关 键 词: 个性化推荐; *K*-means 聚类算法; 动态多维网络
中图分类号: TP393 **文献标志码:** A

Network Resource Personalized Recommendation Based on *K*-Means Clustering

WANG Xin， HUANG Zhong-yi
(School of Computer Engineering, Weifang University, Shandong Weifang 261061, China)

Abstract: A network resource personalized recommendation method based on *K*-means clustering algorithm is presented for dynamic multidimensional social network. Firstly, the user is modeled according to the user rating data, and a multidimensional network is constructed by collecting all the users' rating data, and then a dynamic multidimensional network could be formed with the help of local world evolving network model. Secondly, the network users are clustered by using the improved *K*-means algorithm. Finally, the objective user's rating could be forecasted and obtained by referring the nearest neighbors, and the personalized recommendations could be made. So far, a network resource personalized recommendation method suitable for dynamic multidimensional social network is formed. The experimental results show that the new recommendation method could reduce the error between the prediction value and the true value by comparing with the collaborative filtering recommendation system, and hereby, the new recommendation method could achieve the improved personalized recommendations.

Key words: personalized recommendations; *K*-means clustering algorithm; dynamic multidimensional network

Internet 使人们通过网络得到快捷综合信息服务功能的同时，由于海量信息的无组织呈现，人们可利用的信息往往是无序的，所以大多时候并不能充分地使用 Internet 上的信息资源，反而使得用户对

收稿日期: 2014-01-02
基金项目: 国家星火计划项目(2013GA740109)
作者简介: 王 鑫(1969—)，男，副教授，E-mail: wangxin@wfu.edu.cn.

网络资源的利用率降低. 鉴于此,个性化推荐被认为是解决当前信息过载问题的最有效的策略之一[1]. Internet 就是一个非常典型的复杂网络[1],具有大量复杂网络的构成特征. 而且如果把每类网络看成一个子网络,这些子网络的聚集就构成了多维网络系统. 如果将复杂网络理论和多维网络理论结合起来为用户做个性化推荐,可以为用户做出针对性的推荐需求[1].

目前存在的个性化推荐算法主要有协同推荐算法、基于内容的推荐算法、混合推荐算法[2-4]和最近兴起的基于网络结构的推荐算法[5]. 其中个性化推荐的关键技术是用户描述文件的表示问题,资源描述文件的表示以及推荐方法[6]. 笔者研究的是用户在多个网络活动中的兴趣趋向,并且用户能够参与资源评分.

1　算法描述及分析

1.1　用户描述文件表示

用户描述文件从内容上可以划分为基于兴趣和基于行为的 2 种类型[7],笔者提出的算法应用基于兴趣的用户描述文件并采用加权矢量模型来表示用户模型. 用户兴趣文件(profile)可由该用户对已知信息项的评估值组成,即 $P(r_1,r_2,\cdots,r_i)\ (i\leqslant m)$,其中 m 为该用户所评估过的项的总数,r_i 为用户对文档的评估值. 同时,用户输入的评分数据可以形成一个 mn 的用户 - 项目评分矩阵. 在本算法中,这个评分矩阵将作为建立动态多维网络模型的输入.

1.2　动态多维网络模型的构造

在本文的网络关系理论中,由 2 种节点组成的网络称为双模网络,由多种节点组成的网络为多模网络,而由多模网络向其中一种节点投影而形成的网络称为映射网络. 把其中一种节点定义为资源节点,另一种定义为用户节点,而用户之间表现的兴趣趋向关系是借助用户与资源的选择关系来表现的. 而相同节点的关系可以由双模式网络导出,图 1 展示了在双模式网络中用户和资源以及用户和用户之间的关系.

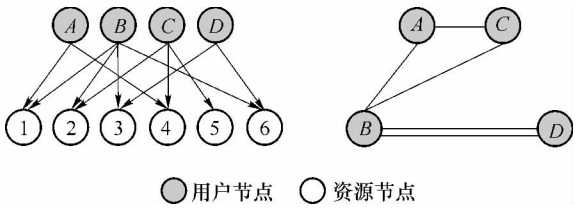

图 1　用户和资源节点关系图

将这种节点之间的关系理论扩展到多个社会网络,那么针对特定的用户群,由于用户参与到多个网络的活动中,那么每个用户就会形成多维度的连边,用户和产品之间的选择关系会形成多模网络. 具体地说,假设图 1 中节点 1 ~ 节点 6 是电影资源,用户在其中做了评价; 同时用户还会进行网页、音乐作品等其他产品的选择,这样每个用户会形成 3 种网络活动的连边,而整个网络也会形成多模网络,多模网络对用户节点进行投影后形成映射网络. 但在形成映射网络的过程中,极有可能出现两个用户之间由于评价了多个相同的资源而形成的多个连边. 不过基于评分的个性化推荐系统,不同用户即使多次对相同资源评分,并不代表他们之间的相似度较高,他们之间的评分差距有可能很大. 所以本文算法基于上述考虑,在由多模网络形成映射网络时,将用户之间出现的所有重边当作一条边处理. 这样,在多模网络对用户节点投影之后,会生成以特定用户集合为节点集,用户之间映射关系集合为边集的多维网络模型.

构建用户网络模型时,考虑动态环境下节点的加入和离开. 由于算法中的多维网络模型具有复杂网络的特性,所以文献[7]中的局域世界演化网络模型提出一个适合多维社会网络模型的演化规则.

模型的构造步骤如下: 1) 增长. 网络初始时有 m_0 个节点和 e_0 条边,每加入一个新的节点,都要与网络中现有的 $m(m\leqslant m_0$,常数) 个节点建立起连接关系[8]. 2) 定义局域世界. 随机地从网络的节点中选取 M 个已有节点,并且 $M\geqslant m$,认为是新添加并建立连接节点的局域世界 LW. 3) 优先连接. 新添加的节点以概率 p 连接到局域世界且以概率 $1-p$ 连接到局域世界外的节点[8]. 新节点的连接概率为

$$\prod(k_i)=p\prod_{\text{Local}}(k_i)+(1-p)\prod_{\text{Nonlocal}}(k_i)\quad(1)$$

其中 k_i 是节点的度. 局域世界内的优先连接概率是

$$\prod_{\text{Local}}k_i=\prod(i\in LW)\frac{k_i+r_{ij}}{\sum_{\text{Local}}(k_j+r_{ij})}=\frac{M}{m_0+t}\frac{k_i+r_{ij}}{\sum_{\text{Local}}(k_j+r_{ij})}\quad(2)$$

其中: r_{ij}是节点之间的相似度值,公式为

$$r_{i,j}=\frac{\sum_{d\in I_{ij}}(R_{i,d}-\overline{R_i})(R_{j,d}-\overline{R_j})}{\sqrt{\sum_{d\in I_i}(R_{i,d}-\overline{R_i})^2}\sqrt{\sum_{d\in I_j}(R_{j,d}-\overline{R_j})^2}}\quad(3)$$

其中: d 是用户 i 和用户 j 的共同邻居, $R_{i,d}$ 是用户 d 的评分, $\overline{R_i}$ 是用户 i 的平均评分.

I_{ij} 是用户 i 和用户 j 一起评过分的集合. 局域世界之外的随机连接概率为[8]

$$\prod_{\text{Nonlocal}}(k_i) = \prod_{i \notin \text{localworld}} \frac{1}{(m_{0+t} - M)} \tag{4}$$

其中 t 是时间间隔.

多维社会网络在以上演化规则下,形成动态多维网络模型. 这样,当有新的用户节点加入时,可以不用再由多模网络向用户节点投影更新多维网络模型,而是直接形成动态多维网络[8].

1.3 个性化推荐算法

相似性度量是一些推荐算法中的重要问题. 在寻找目标用户的邻居用户时,利用相似度公式来计算相似性. 笔者在构建动态多维网络的基础上,在复杂网络中使用 K-means 聚类找到邻居用户,进而生成最近邻用户,求得预测评分做出推荐.

笔者建立的动态多维网络没有考虑权重因素,但在用户聚类时,借鉴了加权复杂网络中关于加权度、加权聚集系数的定义来计算网络节点的综合特征值,并运用 K-means 聚类算法对用户节点聚类. 网络边的权重定义为节点的相似度,可以根据修正的余弦相似性计算. 对于用户节点集 $U = \{u_1, u_2, \cdots, u_n\}$,节点间的相似度公式为

$$\text{sim}(i,j) = \frac{\sum_{k \in I_{ij}} (R_{i,k} - \overline{R_i})(R_{j,k} - \overline{R_j})}{\sqrt{\sum_{k \in I_j} (R_{i,k} - \overline{R_i})^2} \sqrt{\sum_{k \in I_j} (R_{j,k} - \overline{R_j})^2}} \tag{5}$$

其中: I_i 和 I_j 分别表示用户 i 和 j 分别评分过的项目集合, $r_{i,j}$ 是用户 i 和用户 j 共同评分过的集合. 这样,定义用户节点 u_i 的加权度 D_i 是

$$D_i = \sum_{(u_i,u_j) \in E} \text{sim}(i,j) \tag{6}$$

其中: E 是网络的边集,定义节点 u_i 的加权聚集度为

$$K_i = \sum_{(u_j,u_k) \in R} \text{sim}(j,k) \tag{7}$$

其中: R 是由节点对 (u_j, u_k) 组成的集合, u_i、u_j、u_k 均为用户节点, u_i 与 u_j 之间, u_i 与 u_k 之间有连边.

考虑到 K-means 算法对孤立点和噪声数据敏感,所以在进行聚类之前,根据复杂网络节点的综合特征值来选取初始聚类中心. 算法中节点的综合特征值 CFV_i 的计算公式为

$$\text{CFV}_i = \frac{\omega D_i}{N} + \frac{(1-\omega) K_i}{\binom{U_i}{2}} \tag{8}$$

其中 U_i 是节点 i 的邻居节点的集合.

这样,由改进的 K-means 算法寻找目标用户的邻居用户的算法步骤是:

1) 计算用户集合中对应的各个节点的加权度 D_i 和加权聚集度 K_i,并由此得到每个节点的综合特征值 CFV_i;

2) 以综合特征值对所有节点进行排序,形成从大到小的有序序列;

3) 依次从有序序列中选取前 w 个具有较大特征值的节点作为初始聚类中心,并且保证各聚类中心之间没有连边;

4) 以所选的 w 个初始聚类中心为开始,进行 K-means 算法,得到 w 个用户聚类集 $C(c_1, c_2, \cdots, c_w)$.

为了提高效率和质量,在个性化推荐推荐之前,需要在总的邻居用户集合里根据相似性寻找出目标用户的最近相邻用户集 NNU_d.

在得到了目标用户的最近邻居之后,下一步就是要产生相应的推荐. 算法通过计算目标用户对产品的预测评分值做出推荐. 对于某一用户 u_d,找出其未评价过的资源,推荐资源的预测值由其对评价资源的平均评分和最近邻用户的评分组成. 公式为

$$P_{d,j} = \overline{R_d} + \frac{\sum_{n=1}^{k} \text{sim}(d,n)(R_{n,j} - \overline{R_n})}{\sum_{n=1}^{k} |\text{sim}(d,n)|} \tag{9}$$

其中: $\text{sim}(d,n)$ 是目标用户 d 与邻居用户 n 的相似度, $\overline{R_n}$ 是邻居用户 d 的平均评分, $\sum_{n=1}^{k} |\text{sim}(d,n)|$ 是一个规范化系数. 通过上述方法预测用户对所有未评分项的评分,然后选择预测评分最高的前若干个项作为推荐结果反馈给用户.

2 实验结果分析

将所设计的算法应用到动态多维社会网络的个性化推荐系统中验证其效率. 以在 3 种社交类型网站上为用户推荐相应网站资源为例,所有用户均参与到 3 个网站的产品推荐中. 在得到推荐网站提供的经过预先处理的部分数据后,在其中随机抽取了 56 个用户,监测到 42 个用户给出了显性评分,评价

网页一共是 980 个.

有两组对比实验: 第 1 组实验采用协同过滤的推荐系统(CFRS, collaborative filtering recommendation system) 的基于用户和基于项目的个性化推荐, 分别用 U-CFRS 和 I-CFRS 表示[1], 与 DMRS (dynamic multidimensional recommendation system) 进行结果比较.

采用统计准确性标准中的平均绝对误差 MAE 来衡量预测值和真实值之间的平均绝对偏离. 其中 MAE 定义为 $\sum_{i=1}^{n}\frac{|p_i - q_i|}{n}$, 那么当 MAE 的值越小, 即预测值和真实值之间的误差越小时, 预测的精度越高. 实验结果如图 2 所示.

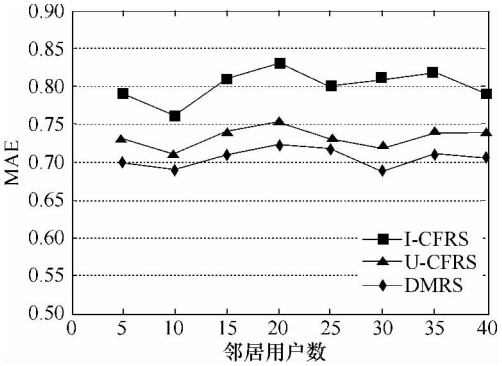

图 2　推荐结果质量比较

实验结果表明, 运用动态多维网络的推荐算法的个性化推荐系统相比较协同过滤推荐系统的两种方式, 具有更小的 MAE 值, 即推荐精度有一定程度的提高.

第 2 组实验采用基于内容的推荐系统(CBRS, content based recommendation systems) , 使用召回率和准确率为评价推荐系统优劣的指标参数[8], 在比较时, 利用 Top-*N* 推荐质量的分类准确度评价标准, 使用 10 个用户.

对比 2 个系统的召回率和准确率, 得到实验结果如图 3 和图 4 所示. 根据两组实验表明, 基于 *K*-means动态多维社会网络的推荐系统 DMRS 的召回率和准确率要高于传统的基于内容的推荐系统.

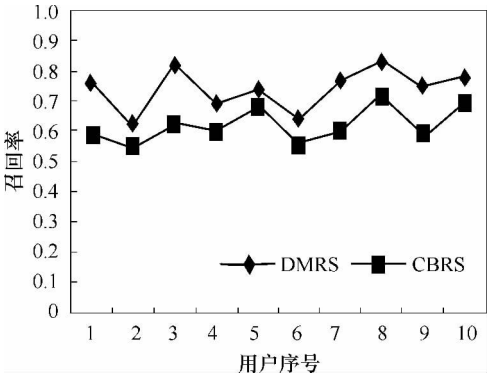

图 3　CBRS 与 DMRS 召回率比较

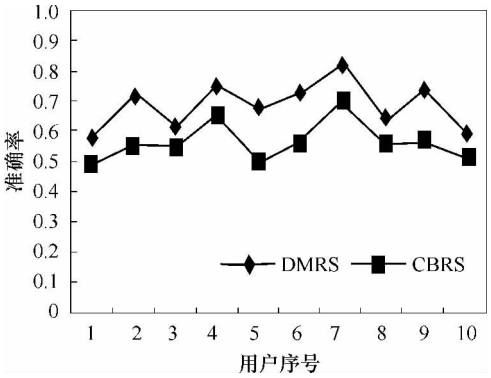

图 4　CBRS 与 DMRS 准确率比较

3　结束语

在建立用户的动态多维网络的基础上提出了一种个性化的推荐算法, 并在典型的推荐系统中做了具体应用. 通过真实的数据集对算法做了相关验证, 经实验分析显示, 该方法提高了个性化的推荐质量. 以后将对如何建立动态多维网络的维度分析和有效地确定算法中的各参数上进一步拓展, 期望进一步提高该推荐策略的个性化推荐效率.

参考文献:

[1] 张华青, 王红, 滕兆明, 等. 多维加权社会网络中的个性化推荐算法[J]. 计算机应用, 2011: 2408-2411.
Zhang Huaqing, Wang Hong, Teng Zhaoming, et al. Personalized recommendation algorithm of multidimensional weighted social network [J]. Computer Application, 2011: 2408-2411.

[2] Newman M J. The structure and function of complex network [J]. SI-AM Review, 2003, 45(2): 167-256.

[3] Liu R R, Jia C X, Zhou T, et al. Personal recommendation via modified collaborative filtering [J]. Physica A, 2009(388): 462-468.

[4] Yoshii K, Goto M, Komatani K, et al. An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model [J]. IEEE Transactions on Audio Speech and Language Processing, 2008, 16(2): 435-447.

[5] Zhou T, Jiang L L, Su R Q, et al. Effect of initial con-

figuration on network-based recommendation [J]. Europhys Lett, 2008(81):58004.

[6] Zeng C, Xing C X, Zhou L Z. A survey of personalization technology [J]. Journal of Software, 2002,13(10): 1952-1961.

[7] Wu Y H, Chen Y C, Chen A L P. Enabling personalized recommendation on the web based on user interests and behaviors [C]//Klas W. Proceedings of the 11th International Workshop on Research Issues in Data Engineering. Los Alamitos, CA: IEEE CS Press, 2001: 17-24.

[8] 张华青. 动态多维社会网络中个性化推荐方法研究[D]. 济南: 山东师范大学, 2012.
Zhang Huaqing. Personalized recommendation method in dynamic and multidimensional social networks [D]. Jinan: Shandong Normal University, 2012.