

一种引入加权异构信息的改进协同过滤推荐算法

张海霞^{1,2}, 吕振^{1,2}, 张传亭^{1,2}, 袁东风^{1,2}

(1. 山东大学信息科学与工程学院 济南 250100; 2. 山东省中国虹计划协同创新中心 济南 250100)

【摘要】协同过滤作为当前应用最成功的推荐技术之一,其推荐质量在很大程度上取决于近邻用户选取的准确性,而数据的稀疏性问题(sparsity)和相似度度量方式(similarity metrics)严重影响着最近邻的选择。该文提出了一种引入加权异构信息的改进协同过滤算法。首先利用异构网络中丰富的语义信息和边属性信息,得到用户之间基于不同元路径的相似度;然后将相似度分别应用到典型的基于用户的协同过滤推荐算法中,得到基于每个相似度的用户评分值;最后采用监督学习算法为每个打分值分配不同的权重,融合为用户最终评分。在扩展MovieLens经典数据集上的实验结果表明,本文所提算法在精确度上较传统算法有显著提高。

关键词 协同过滤; 元路径; 推荐系统; 相似度; 加权异构信息

中图分类号 TP391 **文献标志码** A **doi**:10.3969/j.issn.1001-0548.2018.01.017

An Improved Collaborative Filtering Recommendation Algorithm with Weighted Heterogeneous Information

ZHANG Hai-xia^{1,2}, LÜ Zhen^{1,2}, ZHANG Chuan-ting^{1,2}, and YUAN Dong-feng^{1,2}

(1. School of Information Science and Engineering, Shandong University Jinan 250100;

2. Collaborative Innovation Center of China Rainbow Project Jinan 250100)

Abstract Collaborative filtering is one of the most successful recommendation technologies, and the quality of collaborative filtering is determined by the accuracy of the nearest neighbors. Data sparsity problem and similarity metrics seriously affect the choice of the nearest neighbors. Different from traditional recommendation tasks, in this paper, we propose an improved meta path-based collaborative filtering algorithm for weighted heterogeneous information networks. Firstly, we calculate the similarity among users based on different meta path by utilizing the rich semantic information and attribute information in weighted heterogeneous networks. Then we apply the similarity to user-based collaborative filtering algorithm and get multiple predicted rating scores based on different similarity. Finally we calculate the final predicted scores by combining various meta path information using supervised machine learning algorithms. The method is evaluated with the extended MovieLens dataset and experimental results show that our approach outperforms several traditional algorithms and make the result of recommendation more accurate in terms of accuracy.

Key words collaborative filtering; meta path; recommendation system; similarity; weighted heterogeneous information

协同过滤(collaborative filtering)^[1]是推荐系统发展史上发展最快、应用最广的一类算法,其基本思想是相似的用户对商品的选取也是相似的,根据与目标用户最相似的 K 个邻居对目标项目的评分来进行推荐。但是传统的相似度度量方式只考虑了用户评分这一个因素,评分矩阵是极其稀疏的,这种情况下得到的用户相似性是不准确的,最终推荐精度自然不高。

近年来,社交网络朝异质性方向发展:网络中

包含多种实体(entity),实体间存在多种关系(relation),这种网络被称为异构信息网络^[2](heterogeneous information network, HIN)。大数据时代的HIN中包含丰富的语义信息,对其进行深度挖掘分析能够得到非常有意义的知识^[3-4],而HIN中包含的更多的有效信息会带来更好的推荐效果^[5-6],所以近年来开始兴起在HIN上做推荐问题的研究。HIN中包含多种类型的实体和多种多样的边信息。为了更好地利用异构网络中所蕴含的多样性内容,可以

收稿日期:2017-01-11; 修回日期:2017-10-09

基金项目:山东省自主创新及成果转化重大专项(2013ZHXC0102, 2014ZZCX03401)

作者简介:张海霞(1979-),女,博士,教授,主要从事无线通信网络和新一代信息技术方面的研究。

用元路径(meta path)表示不同的关系^[7]。两实体之间的语义信息路径为一条元路径, 两节点间不同的元路径代表不同的语义联系。基于元路径开展了大量的工作, 相似度度量是其中最基础和最重要的一个方向, 基于对称元路径的PathSim^[7]和基于任意元路径的PCRW^[8]与HeteSim^[9]等算法相继被提出。但这3种算法都是建立在无权异构网络基础上, 忽略了HIN中的边属性信息, 因此都不能被直接应用到加权HIN上的推荐工作中。利用元路径的概念可以灵活地运用HIN中丰富的信息来做推荐^[9]。

目前基于HIN的推荐工作还处于起步阶段, 文献[10]结合用户评分和多种实体相似度度量方式提出了一种新的矩阵分解推荐框架Hete-MF; 文献[11]引入异构网络信息提出了一种基于协同过滤的社会化推荐方法Hete-CF, 该方法将社会网络中的异构关系融合进了协同过滤推荐算法中, 改善了原算法中数据稀疏和冷启动问题; 文献[12]利用用户的隐式反馈信息结合用户不同的异构关系提出了一种新的个性化推荐方法Hete-Rec。一方面, 上述方法都旨在融合异构网络中的多元信息并且只考虑了HIN中的部分信息; 另一方面, 上述方法并没有考虑网络中的边属性问题, 没有关注由用户两极化评分造成的本质差别。以电影推荐网络为例, 5分表示喜欢, 1分表示讨厌; 同样, 一部电影被贴上某种标签的次数越多代表该电影越偏向于此类型。如果在推荐过程中不考虑此边属性问题, 很可能会使推荐结果有所偏差。文献[6]提出了加权HIN的概念, 通过区分网络中边上的不同属性值来探索更全面的元路径语义信息以实现更准确的推荐, 但该方法并没有提出新的相似度度量方法, 而是将有权元路径分解为有确定属性值限定的原子元路径, 利用的还是原有的相似度度量方法。本文引入加权HIN的概念, 充分考虑HIN中的节点类型信息和多种边属性信息, 并且对用户评分进行了两极化映射处理, 提出了一种引入加权异构信息的改进协同过滤算法。

1 问题定义

1.2 加权异构信息网络

定义 1 加权异构信息网络 (weighted heterogeneous information network, WHIN)^[7,9]: 可以用一个有向图 $G=(V, E, \mathbb{W})$ 表示, 且节点类型映射函数 $\mathcal{O}:V \rightarrow B$, 表示每一个 $v \in V$ 属于节点类型集合 B , 即 $\mathcal{O}(v) \in B$; ψ 表示边类型映射函数, $\psi(e) \in R$ 表示每一个 $e \in E$ 属于边类型集合 R ; 映射

函数 $\varphi:\mathbb{W} \rightarrow W$ 表示每条边的权重 $w \in \mathbb{W}$ 属于集合 W , 即 $\varphi(w) \in W$ 。只要 $|W|>0$ 且满足 $|B|>1$ 或 $|R|>0$, 该信息网络即为WHIN。图1为一个典型的WHIN, 网络中包含7种节点类型和12种实体间的关系。

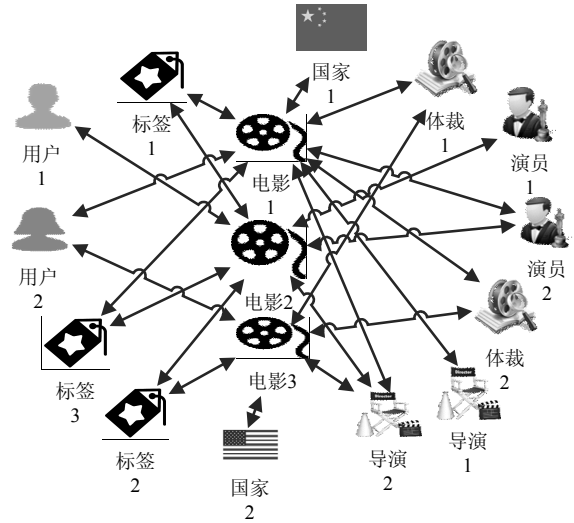


图1 加权HIN示例

1.2 加权网络模式

定义 2 加权网络模式: WHIN可以用加权网络模式^[7,9] $S=(B, R, W)$ 来简化表示, 其中 B 表示网络中的节点类型, R 表示网络中的边类型, W 表示网络中边的权重。加权网络模式对网络中节点和边的类型进行了限制, 正是这种限制导致了网络的半结构化, 指导了网络的语义内涵探索。

对于图1中的网络, 其加权网络模式如图2所示, 该网络模式清晰地表现了图1中WHIN中所有的节点类型和节点间带权重的关系。通过此网络模式, 研究者可以清楚地分辨复杂的HIN中本质上存在的节点类型以及各节点类型间可能存在的关系。

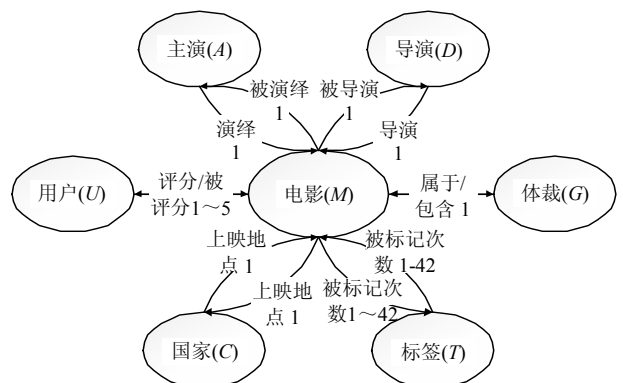


图2 加权异构信息网络的加权网络模式图

1.3 加权元路径

定义 3 加权元路径: 加权元路径^[7,9] P 定义在加权网络模式 $S=(B, R, W)$ 上, 在不引起歧义的情况

下,可以直接用节点类型和节点间边上的权重来表示加权元路径 $P = B_1(W_1)B_2(W_2)\dots(W_{l-1})B_l$ 。如元路径 $U \xrightarrow{\text{Rate5}} M \xrightarrow{\text{Rate2 by}} U$ 表示观看同一部电影(M)的两个用户(U),用元路径可直接表示为 $U(5)M(2)U$ 。否则,加权元路径表示为:

$P = B_1 \xrightarrow{R_1} B_2 \xrightarrow{R_2} B_3 \dots \xrightarrow{R_l} B_{l+1}$,表示从实体类型 B_i 到实体类型 B_{i+1} 之间的复杂关系 $R = R_1 \circ R_2 \circ \dots \circ R_l$,其中, \circ 代表关系上的连接操作,加权元路径 P 的长度为 R 的个数。

2 改进的协同过滤算法

本文引入加权异构信息计算用户间的相似度,并提出了一种新的改进协同过滤算法。接下来说明如何综合考虑该网络中用户在电影体裁、主演、导演、电影标签类型等多方面的偏好度,从而相对全面地度量两用户间的相似度。

定义 4 用户对某影响因素的偏好度。加权异构电影网络中,基于元路径 $P = B_1(W_1)B_2(W_2)\dots(W_{l-1})B_l$, U 为用户类型, B_i 为导演、演员、电影体裁等某一类型,用户 u_i 对影响因素 $y \in B_l$ 的偏好度定义为:

$$t(u_i, y) = \sum \left\{ p_{u_i \rightarrow y}^w : p_{u_i \rightarrow y}^w \in P \right\}$$

式中, u_i 表示用户; $p_{u_i \rightarrow y}^w$ 为从 u_i 到 y 的一个加权路

径实例; $\left| p_{u_i \rightarrow y}^w \right|$ 为 u_i 到 y 的所有路径实例上的权重值; $t(u_i, y)$ 为 u_i 到 y 的所有路径实例上的权重值之和。

用该公式可以推算出所有用户对所有演员基于该元路径的偏好度。如果两用户对所有演员的偏好度越接近,则两用户的喜好越相似。

定义 5 扩展交换矩阵。若一个 WHIN 为 $G = (V, E, W)$, 其网络模式为 $S = (B, R, W)$, 对于加权元路径 $P = B_1(W_1)B_2(W_2)\dots(W_{l-1})B_l$ 的扩展交换矩阵 \mathcal{M} 定义为:

$$\mathcal{M} = \mathcal{W}_{B_1(W_1)B_2} \mathcal{W}_{B_2(W_2)B_3} \dots \mathcal{W}_{B_{l-1}(W_{l-1})B_l}$$

式中, $\mathcal{W}_{B_i(W_i)B_j}$ 为节点类型 B_i 到类型 B_j 的扩展邻接矩阵。其中,扩展邻接矩阵 $\mathcal{W}_{B_i(W_i)B_j}$ 定义为:

$$\mathcal{W}_{B_i(W_i)B_j} = \begin{cases} w_{ij} & \text{节点 } B_i \text{ 与 } B_j \text{ 间边上有权重且权重为 } w_{ij} \\ 1 & \text{节点 } B_i \text{ 与 } B_j \text{ 间有连边但边上无权重} \\ 0 & \text{节点 } B_i \text{ 与 } B_j \text{ 间无连边} \end{cases}$$

所以, $\mathcal{M}_{i,j} = \mathcal{M}(i, j)$ 表示 $x_i \in B_i, y_j \in B_j$ 在元

路径 $P = B_1(W_1)B_2(W_2)\dots(W_{l-1})B_l$ 下 x_i 到 y_j 的所有路径实例的权重之和,即 x_i 对 y_j 的偏好度。

$\mathcal{M}_{U(W_1)M(W_2)A}$ 的每一行元素为该行对应的用户对网络中所有演员的偏好度。行归一化后,每一行数据代表此用户在电影主演方面的偏好,可得任意两用户 u_i 与 u_j 间在演员方面的相似度:

$$\text{sim}(u_i, u_j) = \cos(u_i, u_j) = \frac{u_i u_j}{|u_i| |u_j|} = \frac{\mathcal{M}_{i,:} \mathcal{M}_{j,:}}{|\mathcal{M}_{i,:}| |\mathcal{M}_{j,:}|}$$

式中, $\text{sim}(u_i, u_j)$ 可能为负值,用公式 $\text{sim}(u_i, u_j) = 0.5 + 0.5\text{sim}(u_i, u_j)$ 修正,使其值位于 $0 \sim 1$ 之间,值越大表明两用户越相似。如表1所示,不同的元路径蕴含不同的语义信息,某用户喜欢一部电影可能是因为喜欢该电影的主演或导演,也可能是因为喜欢电影体裁。

表1 电影推荐网络中典型的元路径及其语义信息

元路径	语义信息
P1 = UMU	与目标用户(U)看过同一部电影(M)的用户(U)
P2 = UMAMU	与目标用户(U)看过同一个主演(A)演过的电影(M)的用户(U)
P3 = UMCMU	与目标用户(U)看过在同一个国家(C)上映的电影(M)的用户(U)
P4 = UMDMU	与目标用户(U)看过由同一个导演(D)执导的电影(M)的用户(U)
P5 = UMGMU	与目标用户(U)看过属于同一个电影体裁(G)的电影(M)的用户(U)
P6 = UMTMU	与目标用户(U)看过被贴上同一个类型标签(T)的电影(M)的用户(U)

基于表1中的6条元路径求得两用户在导演、体裁、标签等方面的6个相似度值,并将各相似度应用到基于用户的协同过滤推荐算法中。在基于用户的协同过滤推荐系统中用户 u 对项目 i 的评分^[13]为:

$$r_{u,i} = \bar{r}_u + k \sum_{u \in U} \text{sim}(u, u) (r_{u,i} - \bar{r}_u)$$

式中, $r_{u,i}$ 为用户 u 对项目 i 的评分; U 为与用户 u 最相似的用户集合; $\bar{r}_{u,i}$ 为用户 u 的平均打分值,可消除用户打分偏好对最后结果的影响; k 为标准化因子, $k = 1 / \sum_{u \in U} |\text{sim}(u, u)|$ 。

得到6个受不同影响因素即不同元路径影响的用户对电影的评分值。然后采用线性回归方法为每一个基于单一元路径的预测评分值赋予不同的权重,融合为最终的预测评分。

3 仿真实验

实验环境具体配置如下: Intel® Core(TM)i7-

4710MQ CPU, 2.50 GHz, 8 GB 内存。实验中算法均用Python语言实现。

3.1 数据集

本文所用数据集为MovieLens10M扩展数据集。数据集中包含7种不同的节点类型, 如表2所示, 其加权网络模式如图2所示。

表2 MovieLens扩展数据集简介

实体类型	数目
用户 U	2 113
电影 M	10 197
演员 A	21 185
导演 D	4 060
国家 C	72
体裁 G	20
标签 T	13 222

3.1.1 数据集稀疏性分析

如果以用户和电影间已有的评分关系占所有可能存在的评分关系的比例来衡量系统的稀疏性, 那么扩展MovieLens数据集的稀疏度是3.97%。如图3所示, 仅有不超过7%的用户对10%以上的电影有评价关系, 而超过一半的用户评价的电影数目低于总电影数量的4%。传统的协同过滤算法仅利用评分信息计算用户间的相似度来确定最近邻, 在评分信息如此稀疏的情况下必然使得推荐准确度不高。

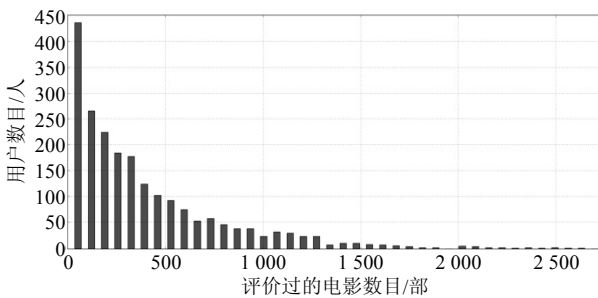


图3 用户评价过的电影数目与用户数目间的关系

3.2 评价指标

评价指标选用评分预测准确度中最经典的两种评价指标: 平均绝对误差(mean absolute error, MAE)和均方根误差(root mean square error, RMSE)。

1) MAE

MAE^[14]即为所有单个预测值与真实值的偏差的绝对值的平均数。在推荐系统中通过计算预测的用户评分与实际的用户评分之间的误差来度量, 为:

$$MAE = \frac{\sum_{(u,i) \in D} |r_{ui} - \hat{r}_{ui}|}{|D|}$$

式中, u 代表测试集中的用户; i 代表测试集中的项目; r_{ui} 代表用户对项目的实际评分数据; \hat{r}_{ui} 代表系统预测的用户对项目的评分; D 代表训练集; $|D|$ 代表训练集中 (u, i) 对个数。

2) RMSE

RMSE的平方值为所有单个预测值与真实值的偏差的平方的平均数, 为:

$$RMSE = \sqrt{\frac{\sum_{(u,i) \in D} (r_{ui} - \hat{r}_{ui})^2}{|D|}}$$

式中各符号的意义与MAE相同。MAE和RMSE都是用系统预测打分值与用户实际打分值的误差来表征一个推荐系统的准确率, 两个值都在0~1之间。若为0, 则代表该推荐系统预测用户评分准确率为100%, 反之, 值越大准确率越差。

3.3 实验结果及分析

传统的基于用户的协同过滤推荐算法中最常用的相似度度量方式为Cosine相似度和Pearson相关系数, 在HIN中最经典的计算两节点间相似度的方式是PathSim和HeteSim算法。

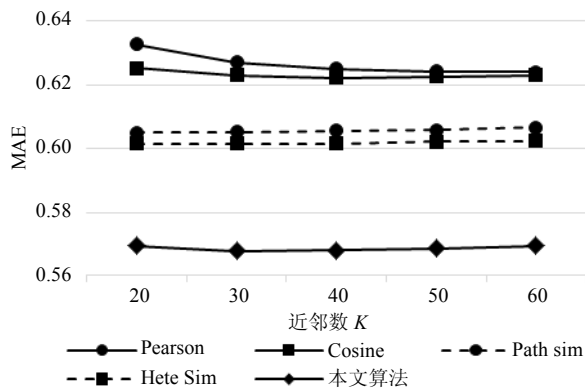
接下来将比较本文提出的算法和基于上述4种相似度度量方式的协同过滤算法在不同近邻数 K 影响下的MAE值和RMSE值。实验采用5折交叉验证方法, 最终结果为5次实验结果的平均值。

3.3.1 近邻数对算法精度的影响

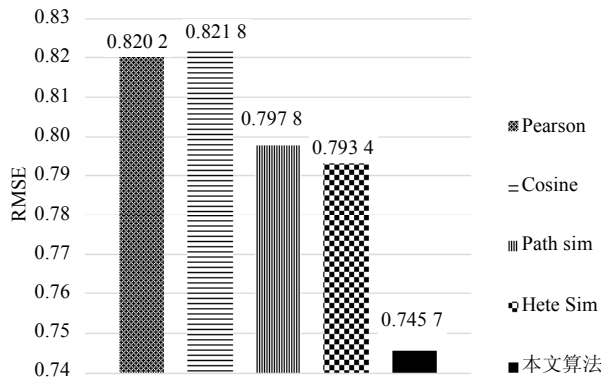
当近邻数 K 为20、30、40、50和60时, 在扩展MovieLens数据集下比较基于Cosine相似度、Pearson相关系数、PathSim和HeteSim的协同过滤算法和本文引入加权异构信息的改进协同过滤算法的MAE和RMSE大小。实验结果如图4所示。

随着邻居数的变化, 基于Cosine相似度和Pearson相关系数的协同过滤算法的MAE值始终大于其他3种算法, 推测原因可能是因为本数据集稀疏度很高, 用户的共同评分项目很少导致最终评分预测准确度低。基于Cosine相似度和Pearson相关系数的协同过滤算法的MAE值随着 K 的增加明显下降, 但是当 $K \geq 40$ 时其MAE值变化极微, 稍有下降。

基于HeteSim的协同过滤算法在 $K=30$ 时MAE略小于 $K=20$ 时的MAE, 当 $K > 30$ 时MAE变化极小, 有些许增加; 而基于PathSim的协同过滤算法的MAE随着 K 的增加变化不大, 但始终大于本文所提出的改进协同过滤算法。本文算法通过分析用户间相似度的多种因素已捕捉到相对全面的语义信息, 当 K 较小时, 随着 K 的增加, MAE几乎没有变化, 当 $K=50$ 或60时, 由于所取近邻中掺杂了与目标用户不那么相似的用户, 反而会使得MAE有所增加。如图4b所示, 随着 K 的增加, 3种算法的RMSE变化趋势与图3中MAE的变化趋势大致相同, 都是在 $K=30$ 时表现最优, 综上所述, 取 $K=30$ 。

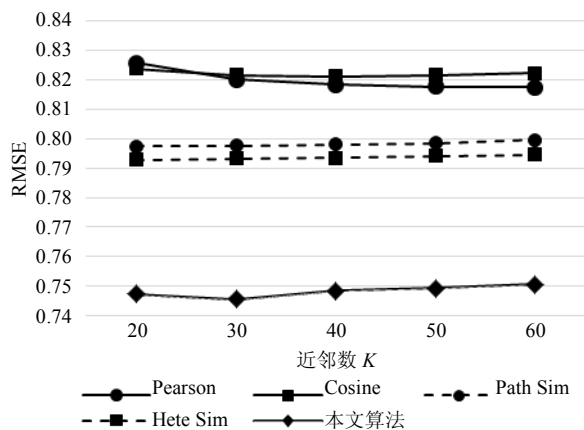


a. 近邻数K对5种算法的MAE影响



b. K=30时5种算法的RMSE比较

图5 K=30时5种算法的MAE和RMSE比较



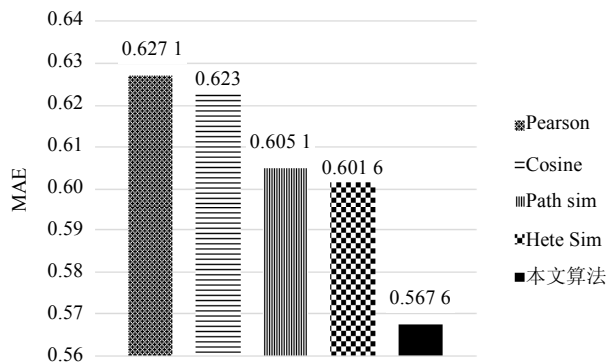
b. 近邻数K对5种算法的RMSE影响

图4 近邻数K对5种算法的MAE和RMSE影响

3.3.2 3种算法精度对比

当近邻数K=30时, 比较3种算法预测用户对电影评分数据的精度, 实验结果如图5所示。

由于融合了HIN中多条元路径携带的不同语义信息的综合影响, 并考虑了两用户间不同关系上的权重值和用户评分的两极化影响, 本文所提的改进协同过滤算法表现优于基于Cosine相似度、Pearson相关系数、PathSim和HeteSim的协同过滤算法, 其预测评分准确度在MAE和RMSE两种评价指标上均明显小于另外两种算法。



a. K=30时5种算法的MAE比较

4 结束语

本文在引入加权异构网络信息的基础上, 提出了一种基于元路径计算用户间相似度的新的协同过滤推荐算法, 全面考虑了电影推荐网络中可能影响到用户对电影评分的多种影响因素。实验结果表明本文提出的方法使得推荐算法中两大经典评价指标MAE和RMSE值明显减小, 显著地改善了预测用户评分的准确率。在今后的工作中, 可以引入用户间的社会网络信息, 考虑用户之间的朋友关系、兴趣小组关系等, 以此提升计算用户间相似性的可靠度。

参考文献

- [1] SU X, KHOSHGOFTAAR T M. A survey of collaborative filtering techniques[J]. Advances in Artificial Intelligence, 2009(12): 4.
- [2] SUN Y, HAN J, YAN X, et al. Mining knowledge from interconnected data: a heterogeneous information network analysis approach[J]. Proceedings of the VLDB Endowment, 2012, 5(12): 2022-2023.
- [3] SUN Y, HAN J. Mining heterogeneous information networks: a structural analysis approach[J]. ACM SIGKDD Explorations Newsletter, 2013, 14(2): 20-28.
- [4] SUN Y, HAN J. Mining heterogeneous information networks: Principles and methodologies[J]. Synthesis Lectures on Data Mining and Knowledge Discovery, 2012, 3(2): 1-159.
- [5] YANG B, LEI Y, LIU D, et al. Social collaborative filtering by trust[C]//Proceedings of the 23th International Joint Conference on Artificial Intelligence. [S.l.]: AAAI Press, 2013: 2747-2753.
- [6] SHI C, ZHANG Z, LUO P, et al. Semantic path based personalized recommendation on weighted heterogeneous information networks[C]//Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. [S.l.]: ACM, 2015: 453-462.

(下转第152页)

- materials[J]. Journal of the European Ceramic Society, 2004, 24(6): 1077-1081.
- [10] IMANAKA Y. Multilayered low temperature co-fired ceramics (LTCC) technology[M]. Boston: Springer Science and Business Media, 2005.
- [11] LEE W S, KIM C H, HA M S, et al. The characterizations of tape casting for low temperature sintered microwave ceramics composite[J]. Journal of the Korean Ceramic Society, 2005, 42(2): 132-139.
- [12] YUAN Y, LEE T R. Contact angle and wetting properties [M]//Surface Science Techniques. Berlin Heidelberg: Springer, 2013, 51: 3-34.
- [13] 徐自强, 杨邦朝, 石玉. LTCC异质材料匹配共烧研究[J]. 功能材料, 2011, 42(6): 1079-1082.
- XU Zi-qiang, YANG Bang-chao, SHI Yu. Research on co-fireability of different materials using LTCC technology[J]. Journal of Functional Materials, 2011, 42(6): 1079-1082.
- [14] THOMAS D, ABHILASH P, SEBASTIAN M. Casting and characterization of LiMgPO₄ glass free LTCC tape for microwave applications[J]. Journal of the European Ceramic Society, 2013, 33(1): 87-93.
- [15] LEE S H, MESSING G L, TWINAME E R, et al. Co-sintering of multilayer ceramics[J]. Key Engineering Materials, 2001, 206(1): 257-260.
- [16] BARLOW F, ELSHABINI A. Ceramic interconnect technology handbook[M]. [S.l.]: CRC Press Inc, 2007.

编辑 漆蓉

(上接第116页)

- [7] SUN Y, HAN J, YAN X, et al. Pathsim: Meta path-based top-*K* similarity search in heterogeneous information networks[J]. Proceedings of the VLDB Endowment, 2011, 4(11): 992-1003.
- [8] LAO N, COHEN W W. Fast query execution for retrieval models based on path-constrained random walks[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2010: 881-888.
- [9] SUN J, QU H, CHAKRABARTI D, et al. Relevance search and anomaly detection in bipartite graphs[J]. ACM SIGKDD Explorations Newsletter, 2005, 7(2): 48-55.
- [10] YU X, REN X, GU Q, et al. Collaborative filtering with entity similarity regularization in heterogeneous information networks[C]//International Joint Conference on Artificial Intelligence-Heterogeneous Information Network Analysis (IJCAI-HINA Workshop). Beijing: [s.n.], 2013.
- [11] LUO C, PANG W, WANG Z. Hete-cf: Social-based collaborative filtering recommendation using heterogeneous relations[C]//IEEE International Conference on Data Mining. Shenzhen: IEEE, 2014: 917-922.
- [12] YU X, REN X, SUN Y, et al. Personalized entity recommendation: a heterogeneous information network approach[C]//Proceedings of the 7th ACM International Conference on Web Search and Data Mining. [S.l.]: ACM, 2014: 283-292.
- [13] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734-749.
- [14] SHI Y, LARSON M, HANJALIC A. Mining mood-specific movie similarity with matrix factorization for context-aware recommendation[C]//Proceedings of the Workshop on Context-Aware Movie Recommendation. [S.l.]: ACM, 2010: 34-40.

编辑 叶芳