计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2014.20130342
51(10):2270-2276,2014

基于用户-标签-项目语义挖掘的个性化音乐推荐

李瑞敏　林鸿飞　闫　俊
(大连理工大学计算机科学与技术学院　辽宁大连　116024)
(hflin@dlut.edu.cn)

Mining Latent Semantic on User-Tag-Item for Personalized Music Recommendation

Li Ruimin, Lin Hongfei, and Yan Jun
(*School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning* 116024)

Abstract　Personalized recommender systems are confronting great challenges of accuracy, diversification and novelty, especially when the data set is sparse and lacks of accessorial information, such as user profiles, item attributes and explicit ratings. Collaborative tags contain abundant information about personalized preferences and item contents, and are therefore potential to help providing better recommendations. In this paper, we analyze the information on the famous music social network, Last.fm. Bipartite graph is established between users, items and tags while random walk with restart is used to analyze the relationship between the nodes discussed before and get the neighboring relations between songs or tags. After that, musicrecommended list and indirect related music collection, thus, can be obtained. At last, personalized music recommendation algorithm can be implemented by fusing and reranking the recommended list using the algorithm proposed in this paper. Experiments show that, in the same corpus, the music recommendation algorithmin this paper performs better than the ordinary method such as collaborative filtering and bipartite based algorithm. Our method built on Last.fm, therefore, satisfies the personalized requirement for users to music. Furthermore, with the development of Web2.0, our method will show its advantage as the amount of tags become more and more enormous.

Key words　social tagging; collaborative filtering; bipartite graph; music recommendation; personalized

摘　要　个性化推荐系统面临的难题是推荐的准确性、多样性以及新颖性,同时其数据集存在稀疏、信息缺失(如用户描述、项目属性以及明确的评分)等问题.协同标注中的标签包含丰富的个性化描述信息以及项目内容信息,因此可以用来帮助提供更好的推荐.算法以二部图节点结构相似与重启型随机游走为基础,分析音乐社交网络 Last.fm 中用户、项目、标签两两之间的联系,首先构建音乐间及标签间的相邻关系,初步得到音乐推荐列表和间接关联音乐集合,然后按所提算法融合结果,重新排序,得到最终推荐列表,从而实现个性化音乐推荐算法.实验表明,在该数据集上,所提方法能够满足用户对音乐的个性化需求.

关键词　社会化标注;协同过滤;二部图;音乐推荐;个性化

中图法分类号　TP391; TP181

收稿日期:2013-03-27;**修回日期**:2013-07-05
基金项目:国家自然科学基金项目(60673039,60973068);教育部高等学校博士学科点专项科研基金项目(20090041111002)

随着信息技术和互联网的发展,人们逐渐从信息匮乏的时代走入信息过载的时代.在这个时代,无论是信息消费者还是信息生产者都遇到了很大的挑战:对于信息消费者,从大量信息中找到自己感兴趣的信息是一件非常困难的事情;对于信息生产者,让自己生产的信息脱颖而出,受到广大用户的关注,也是一件非常困难的事情.推荐系统就是解决这一矛盾的重要工具.推荐算法的本质是通过一定的方式将用户和物品联系起来,不同的推荐系统利用不同的方式.

根据使用技术的不同,音乐推荐系统大致可以分为3类:基于内容的推荐算法、协同过滤推荐算法和混合推荐算法[1].基于内容的推荐算法根据项目的属性联系、项目所处的位置、项目元信息(描述项目的关键词,对于音乐,有专辑、流派、艺人名称、歌词、音频等等[2])以及用户的历史信息[3]来选择合适的项目进行推荐.但是,用户所用的关键词和项目描述标签并不能很好地对应,而将音频信息转换为数字信息会导致计算量增大、响应时间延长等问题.这种基于内容的方法还忽略了不同用户之间兴趣相似的情况,因此,并不能很好地适应社区化网络.

随着Web2.0的发展,协同过滤系统成为推荐系统的主流.协同过滤系统同样分为2类:基于内存(memory-based)的推荐算法和基于模型(model-based)的推荐算法.基于内存的推荐算法[4]是一种启发式的方法,根据用户以往的所有评价进行推荐.该方法一般是先为每个用户寻找相似的K个用户,再根据相似用户对给定项的评分以及用户相似度对项进行排序.其中,基于内存的推荐算法存在一个问题,即需要自行设定一些系数,例如:需要根据数据集或其他因素限制自行设定目标用户的"邻居"个数.基于模型的推荐算法首先利用已有的用户评价数据建立一个模型,然后根据该模型进行评价预测,例如基于贝叶斯网络的方法[5]和最大熵的方法[6].但通常基于模型算法的模型建立和更新非常耗时,且模型不能像基于内存算法一样覆盖所有的用户.

协同过滤推荐算法和基于内容的推荐算法各有侧重点,也各有不足.混合推荐算法的主要思想是将上述2种推荐方法相结合,以便充分地利用用户与资源的信息.其中最具影响力的系统是Stanford大学推出的Fab[7].国内在这方面也有很多的研究成果[8-9].

最近几年,也有一些学者提出基于图结构的推荐算法.Nick和Martin[10]通过研究点击次数在图结构中的应用达到项目推荐的目的,但是并没有有效利用用户间的社会化关系.Hilmi和Mukkai[11]提出利用基于图结构的随机游走算法来计算项目-项目间的相似度,并在MovieLens的数据集上实现了该算法.尽管随机游走算法在推荐系统中表现出了优越的性能,但是,社会标注对计算项目-项目间的相似度的贡献,并没有得到有效地验证和利用.

考虑到用户对项目赋予的标签有很大用途,一些学者也研究标签的推荐.Yang和Lu[12]从机器学习的视角,提出一种为社会化推荐系统自动推荐标签的算法.Marcus[13]认为用户提交的查询中所包含的标签,明确表明了用户的意图,并能确切描述一个项目.因此,他提出一种新的形成标签的方法,即抽取用户的意图,在社会化网络中形成面向目标项目的检索.上述学者所作的研究均表明了社会化标注的重要性,这也是本文研究的基础.

为了实现给用户推荐其可能感兴趣的音乐,本文尽可能利用社会化网络中提供的信息,挖掘用户、标签、项目三者之间的内在联系.首先,利用用户-歌曲、标签-歌曲的2个二部图分别建立项目以及标签的邻接矩阵;然后,用给定的用户向量分别在上述两个邻接矩阵上进行重启型随机游走,从而得到该用户的相关歌曲列表和相关标签列表;最后,选取相关标签列表的前N个标签,由标签-歌曲关系挖掘间接关联音乐集合,根据间接关联音乐集合调整用户相关歌曲列表的排序,推荐得分高的项目给用户.在Last.fm收集的语料集上进行实验发现,本文提出的方法比协同过滤算法表现出更好的效果.

1 推荐算法相关理论

在传统的协同推荐系统中,要求用户给出明确的评分来表明自己对该项目的喜爱程度,这些评分一般是有限且离散的,系统根据用户以往对项目的评分来预测用户对新项目的喜爱程度.Memory-based算法又可以分为基于用户的协同过滤推荐算法和基于项目的协同过滤推荐算法.

基于用户的协同过滤推荐算法基于这样一个假设,即如果用户对一些项目的评分比较相似,则他们对其他项目的评分也比较相似.算法根据目标用户的最近邻居(最相似的若干用户)对某个项目的评分逼近目标用户对该项目的评分.定义目标用户a及其未评过分的项目i,然后,预测a对i的评分:

$$p_{a,i}=\bar{r}_a+\frac{\sum_{u=1}^{K}(r_{u,i}-\bar{r}_u)\omega_{a,u}}{\sum_{u=1}^{N}\omega_{a,u}},\tag{1}$$

其中,$r_{u,i}$表示用户 u 对项目 i 的评分,$\bar{r}_a$ 和 $\bar{r}_u$ 分别表示用户 a 和用户 u 的平均打分,$\omega_{a,u}$表示用户 u 和用户 a 的相似度.

而基于项目的协同过滤算法认为,用户对不同项目的评分存在相似性,当需要估计用户对某个项目的评分时,可以用用户对该项目的若干相似项目的评分进行估计,如式(2)所示:

$$p_{a,i}=\bar{r}_i+\frac{\sum_{k=1}^{M}(r_{a,k}-\bar{r}_k)\omega_{i,k}}{\sum_{k=1}^{M}\omega_{i,k}},\tag{2}$$

其中,$\bar{r}_i$ 表示项目 i 的平均得分,$\omega_{i,k}$表示项目 i 和项目 k 的相似度. 在实际的商业应用中,基于用户的协同过滤算法要比基于项目的更有效率. 本文所用的语料中,歌曲的数量要远远大于用户数量,为了提高效率,这里采用基于用户的协同过滤算法作为本文的对比实验.

无论哪种方法,预测得分时都要计算项目-项目之间的相似度或者用户-用户之间的相似度. 相似度的计算方法有很多,本文采用最流行的 $Pearson$ 相关系数,计算方法如式(3)所示:

$$\omega_{a,u}=\frac{\sum_{i=1}^{M}(r_{a,i}-\bar{r}_a)(r_{u,i}-\bar{r}_u)}{\sqrt{\sum_{i\in I_a\cap I_u}(r_{a,i}-\bar{r}_a)^2}\sqrt{\sum_{i\in I_a\cap I_u}(r_{u,i}-\bar{r}_u)^2}}.\tag{3}$$

显然,在这种算法中,用户共同评分项目越多其相似度越高. 但是,假设 2 个用户都只对同一个项目评过分,那么这种方法计算出的 2 用户相似度就很大,这是很不合理的. 为了减少这种情况的发生,设定如果 2 用户共同评分项目个数 n 小于阈值 Tr,那么将其相似度乘以系数 n/Tr. 另外,本文采用 k-近邻法来选取目标用户的相似用户.

本文所采用的数据集是以用户听取某个歌曲的次数来表示用户对某一歌曲的喜爱程度,而不是用户根据自己的喜爱程度对项目的明确评分. 为了融合标签信息,首先由用户听取歌曲的次数及用户的标签,分别计算用户间的相似度 $\omega(UTr)_{a,u}$ 及 $\omega(UTg)_{a,u}$,用二者之和作为 $\omega_{a,u}$,如式(4)所示:

$$\omega_{a,u}=\alpha\omega(UTr)_{a,u}+\beta\omega(UTg)_{a,u},\tag{4}$$

其中,$\alpha+\beta=1$.

2 基于用户-标签-项目的音乐推荐

我们从二部图开始介绍基于用户-标签-项目中潜在语义挖掘的个性化音乐推荐算法,然后是三部图上的潜在语义挖掘算法.

2.1 二部图的关联矩阵

用户和歌曲之间的关系可以表示为一个二部图 $G1=\langle U,E1\rangle$,其中顶点集 U 表示推荐系统中的用户集合,如果用户 u_i 听过音乐 Tr_j,则为 u_i 与 Tr_j 建立一条边,听的次数为用户对该音乐感兴趣的程度. 同理,用户和标签之间的关系也可以表示为一个二部图 $G2=\langle U,E2\rangle$,若用户 u_i 与标签 Tg_j 间有联系,则为二者之间建立一条边. 然后将上述 2 个二部图分别投影到音乐和标签的维度上,投影后节点 i 与节点 j 间边的权重表示音乐(标签)i 与音乐(标签)j 的相似度. 本文选择余弦法计算节点相似度,如式(5)所示:

$$\omega_{i,j}=\frac{|\Gamma_i\cap\Gamma_j|}{\sqrt{|\Gamma_i|\times|\Gamma_j|}},\tag{5}$$

其中,Γ_i 为投影前节点 i 的邻居节点集合.

这里得到的 2 个投影图分别是由音乐项和标签项组成的 2 个不同关联图,然后,就可以分别得到音乐项和标签项对应的矩阵 ***TR0***,***TG0***. 对上述 2 个矩阵均作如下处理:当 $i=j$ 时,设置 $\boldsymbol{TR0}_{i,j}=0$,$\boldsymbol{TG0}_{i,j}=0$;否则 $\boldsymbol{TR0}_{i,j}=\omega_{i,j}$,$\boldsymbol{TG0}_{p,q}=\omega_{p,q}$. 不难看出,矩阵 ***TR0*** 和 ***TG0*** 是对称矩阵. 理论上对于图中任意一个顶点,其到其他所有顶点的概率总和应该为 1,因此需要对矩阵 ***TR0***,***TG0*** 分别进行按列归一化,得到矩阵 ***TR***,***TG***,即为关联图对应的关联矩阵. 关联矩阵中 $TR_{i,j}$表示音乐节点 i 与节点 j 的关联系数,$TG_{p,q}$表示标签节点 p 与节点 q 的关联系数. 这样,关联图中所有音乐节点及标签节点对间的关联程度都可以用 ***TR*** 或 ***TG*** 中相应的元素表示.

2.2 基于二部图的推荐算法

基于二部图的推荐算法通过排序估计用户和项之间的关系. 算法中用向量表示一个用户节点,向量中的每一维都代表关联图中的一个项目,本文中其值记录该用户听取该音乐节点的次数或有无使用过该标签,即为用户对歌曲或标签感兴趣的程度. 算法采用重启型随机游走模型(random walk with restart model, RWR)预测用户节点 $\boldsymbol{U}_i$ 对节点 Tr_j 或 Tg_j 的感兴趣程度. 从用户节点 $\boldsymbol{U}_i$ 开始随机游走,遍历整个图. 对任意节点,遍历者以概率 $1-a$ 游

走到该节点的邻居节点,以概率 a 返回节点 U_i 重新开始走步.每次游走后得出一个概率分布,该分布刻画了图中每一个顶点被访问到的概率.用这个概率分布作为下一次游走的输入,并反复迭代这一过程.迭代过程中,当前后两次概率分布相同或基本相近时,这个点的概率分布趋于收敛.收敛后即可得到一个稳定的概率分布,该概率分布就代表了用户节点 U_i 与项目的亲疏程度.基于二部图的推荐算法流程如下:

算法 1. 基于二部图的推荐算法.

输入:表示用户与项之间关系的二部图 $G1$、待推荐用户节点 U_i、重启概率 a;

输出:U_i 的相关音乐列表.

① 根据式(5)得到 G 关于音乐项目的关联图;

② 令关联图对应的矩阵 $TR0$ 对角元素为 0,以列为单位将其归一化,得到音乐项目的关联矩阵 TR;

③ 由式(6)得到用户节点 U_i,并对其归一化得到 q_i;

④ $t=0$ 时,初始化 $S_i^{tr}=q_i$;

⑤ while (S_i^{tr} 不收敛 $\|$ $t<MAXLOOP$);

⑥ {

$S_i^{tr}=(1-a)\times TR\times S_i^{tr}+a\times q_i$;

$t++$;

};

⑦ 将 S_i^{tr} 中的元素按降序排序;

⑧ 若 $S_i^{tr}[j]$ 排名在 top N,且 $q_i[j]=0$,则将 $S_i^{tr}[j]$ 对应的音乐项目推荐给用户 u_i,得到 u_i 的推荐序列.

对于用户-音乐所构成的二部图,定义 U_i^{tr} 是一个用户查询向量,根据用户 u_i 在训练集中听取歌曲的记录而建立,用 $playcount$ 表示用户 u_i 听第 j 首歌曲的次数,向量 q_i 是 U_i^{tr} 的归一化向量.U_i^{tr} 的每一维元素定义如式(6)所示:

$$U_i^{tr}=\begin{cases}playcount, & \text{用户 } u_i \text{ 听过歌曲 } j,\\ 0, & \text{否则.}\end{cases} \tag{6}$$

算法的目的是得到与该用户关系最紧密的音乐项目.若关联图由 N 个音乐节点组成,则 u_i 对应的稳态概率向量 $S_i^{tr}=[S_i^{tr}(1),S_i^{tr}(2),\cdots,S_i^{tr}(N)]$ 即为所求.经过实验验证,将迭代次数 $t=10$ 时,S_i^{tr} 已达到收敛.

2.3 基于用户-标签-项目语义挖掘的音乐推荐

在算法 1 中,只考虑到音乐与音乐间的节点相似性,在用户项目的关联图上随机游走初步得到与当前用户相关的音乐列表.但是该算法忽略了同样存在于社会化网络中用户-标签以及标签-歌曲所构成的关系图,因此,推荐准确度有待提高.本文提出基于用户-标签-项目中潜在语义挖掘的个性化推荐算法,挖掘用户与标签、标签与音乐间的关联度,对上一步得到的用户相关音乐列表重新进行排序,进一步提高推荐准确率.

每个用户都有自己的兴趣爱好,并且以标签的形式展现在个人的描述页面中.当用户听音乐时,社会化媒体允许用户为该音乐项目贴标签,而这些标签也透漏了该用户对当前收听的音乐项目的认知.随着时间的推移,用户的标签会逐渐丰富趋于完善;此外,一个音乐项目会被多个用户标注.由于这些原因,社会化媒体中有这样的现象存在:不同用户会对同一个音乐项目产生不同的看法;一个音乐项目的多个标签可能隐含同一个意思.因此,我们提出下面的观点:可以通过二部图挖掘相同含义的标签;认为一个项目和一个用户拥有共同(包括含义相近)的标签越多,那么该用户与该项目的关联度就越高.基于用户-标签-音乐中潜在语义挖掘的个性化推荐系统工作流程如图 1 所示:

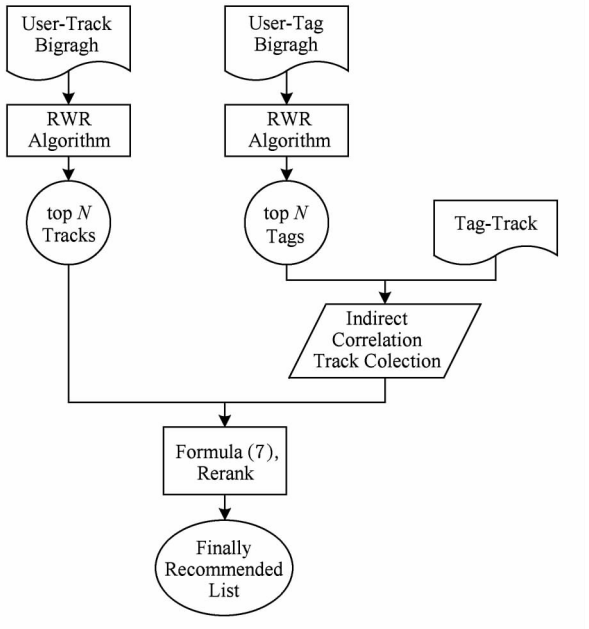

Fig. 1 Recommendation algorithm based on uset-tag-item.

图 1 基于用户-标签-音乐的推荐算法流程图

与 2.2 节中得到歌曲推荐列表的过程相似,在用户-标签构成的二部图上作同样的算法,得到与当前用户关联的标签列表 $S_i^{tg}=[S_i^{tg}(1),S_i^{tg}(2),\cdots,S_i^{tg}(M)]$.不同的是,该方法得到的标签列表中并不剔除用户本身用过的标签,因为这些标签明确指出

了用户的兴趣爱好. 然后选取与该用户关联度最高的 N 个标签,抽取这些标签对应的音乐集合,定义该集合为"间接关联音乐集合". 最后根据本次得到的歌曲集合对 2.2 节中得到的歌曲推荐列表修改权值,重新排序,得到最终推荐结果. 定义权值计算式(7)用来重新调整歌曲权值:

$$\omega'_{i,j}=\omega_{i,j}\times(1+\frac{m}{p}), \tag{7}$$

其中,m 表示不同标签对应的某相同歌曲出现的次数;p 表示该首歌曲在交集(交集是指间接关联音乐集合与原推荐列表按权重的有序相交集合)中的排序位置. 它体现了 2 个观点:1)与用户关联度较大的标签对应的音乐与用户的关联度也较大;2)与用户相关联的多个标签对应同一音乐,该音乐与用户的关联度较大. 经过权值的调整,对音乐按权值从大到小重新排序,将 top N 重新推荐给用户. 经过对数据集数据统计,平均每个用户有 10 个标签. 因此,用户的相关标签数量应该比用户本身拥有的标签数量多一些,这样才能得到扩展标签,获得更多信息. 但是,相关标签越多计算量也就越大,算法效率会下降. 根据上述 2 点,本文将选取与用户相关度较高的 30 个标签用于计算.

3 推荐算法实验结果分析

3.1 Last.fm 数据集

Last.fm 是一个音乐社交网络,它允许用户创建自己的个人页面、交友、添加标签,并记录用户听取歌曲名称及次数. 格拉斯哥大学计算机科学与技术实验室于 2008 年从音乐社区网站 last.fm 中收集并抽取了部分语料,并公开供学者们研究使用. 该语料抽取至少听过 8 首音乐的用户及其相应标签和听歌曲的情况,其中共包含 3 148 个用户、30 520 首歌曲、12 565 个标签以及 3 148 个用户中存在的 5 616 个好友关系. 本文即是在该语料集上进行的研究.

实验中对于每个用户,将语料集中所有歌曲分为 3 个部分:实验所用训练集,为用户听过歌曲的 80%;测试集,为用户听过歌曲的 20%;用户没有听过的歌曲集合.

3.2 参数选择和实验设计

大多数推荐网站是用打分机制来获得用户喜好信息,而本文所用语料收集的则是音乐的收听次数. 经过对现有语料的统计,用户收听某首音乐的次数(*playcount*)的变化范围在 1~7 939 之间. 用户会重复收听喜欢的音乐达到上千次,而对于不喜欢的歌曲听一次就不会再听.

如果当前用户的近邻听取某首音乐的次数很多,将会极大地影响推荐效果. 因此,为避免这种偏差,针对该语料特征,将协同过滤中的预测得分式(1)作如下处理:

$$p_{a,i}=\frac{\sum_{u=1}^{K}\left(\frac{r_{u,i}-\bar{r}_a}{\sigma_a}-\frac{\bar{r}_u-\bar{r}_a}{\sigma_a}\right)\times\omega_{a,u}}{\sum_{u=1}^{N}\omega_{a,u}}. \tag{8}$$

本文采用的协同过滤算法中存在 2 个参数:2 用户间听取相同歌曲的数量的阈值 Tr 和选取当前用户的相似用户的个数 K. 经过多次实验证明,这 2 个参数取值为 $Tr=20$,$K=15$ 时实验效果最佳.

在重启型随机游走算法中,重启因子需要人工设定. 为了发现应该以多大的概率返回当前待推荐用户进行重新走步,能使推荐算法达到较好的效果,设计一组关于重启概率 a 的对比实验. 设定 a 的值分别为 0.8,0.5,0.3,在由用户-音乐这个二部图投影的音乐-项目关联图上进行随机游走,图 2 为实验结果比较,可以看出 a 取值不同,$P@N$ 的表现不同.

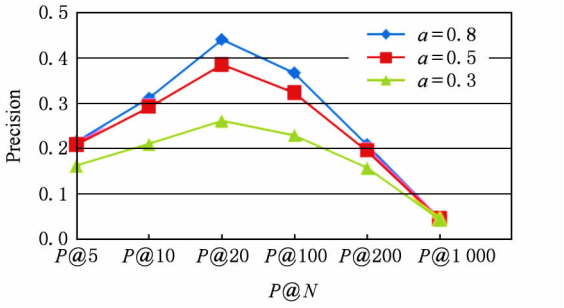

Fig. 2 Impact of restart factor values for recommendation accuracy in RWR algorithm.

图 2 RWR 算法中重启因子的取值对推荐准确率的影响

RWR 算法中重启动概率 a 常取为 0.15,本文经过实验发现,a 越大 RWR 效果越好. 出现这个现象的原因与所用关联图的半径以及数据的性质都有关联. 对于当前 Last.fm 数据集来说,用户的相似度更依赖于用户与音乐的直接路径数. 因此,下面的实验中均选用 $a=0.8$.

3.3 实验结果与分析

本文针对基于用户的协同过滤推荐算法和基于二部图的协同推荐算法进行对比实验,其中每一组实验又设计加入标签信息调整后的推荐算法与简单

模型进行对比实验.分别用 UTr,$UTrTg$ 表示利用用户-音乐关系的实验、加入用户标签信息调整后的实验.实验使用 $P@N$ 作为评价指标,针对每种方法,分别计算 $P@5$,$P@10$,$P@20$,$P@100$,$P@200$ 进行比较实验.表 1 显示出使用 Last.fm 作为数据集的不同推荐算法的效果比较.

由表 1 可以看出,在同一个数据集中,基于二部图的协同推荐算法的推荐效果明显优于基于用户的协同过滤算法,且加入用户标签的 RWR 算法达到最优,如表 1 最后一行所示.这是因为基于用户的推荐算法只考虑用户的相关性,忽略了项目之间的相关性.

Table 1 Comparison of RWR and CF

表 1 RWR 算法和 CF 算法实验结果对比

Method	Specific Methods	MAP	$P@5$	$P@10$	$P@20$	$P@100$	$P@200$
CF	*UTr*	0.030 8	0.148 0	0.128 7	0.094 1	0.049 7	0.036 5
	UTrTg	0.023 1	0.110 2	0.080 3	0.070 8	0.039 2	0.029 1
RWR	*UTr*	0.104 2	0.215 7	0.310 9	0.440 0	0.366 7	0.209 0
	UTrTg	**0.106 8**	**0.220 9**	**0.321 5**	**0.445 7**	**0.369 1**	**0.211 8**

另外,基于用户的协同过滤算法添加标签信息调整后,整个推荐效果并没有改善,推荐准确率反而下降.造成这个现象的原因是数据集中虽然有 12 565 种标签,但是根据实验统计情况,平均每个用户拥有 10 个标签.这说明用户-标签这个二部图存在的边很少,使得由该二部图得到的用户间的关联程度很小.当与由用户-音乐构成的二部图得到的用户关联矩阵进行简单的加权求和时,在一定程度上降低本来关系较为紧密的用户的关联程度,从而造成推荐准确率的下降.如果随着时间的增长,用户标签密度够大,这是一个很好的算法.总之,社会化标注仍然是研究的一个热点和重点.

而对于改进后的基于二部图的协同推荐算法,加入标签信息重新调整推荐顺序后,推荐准确率有所提升,但是效果不是很明显.用户的浏览习惯主要集中在返回列表前 20 项,之后的项目就很少会被用户注意到,因此,本算法的主要目标是提升推荐列表前 20 项的准确率.由表 1 可以看出,本文算法主要对 $P@10$ 有较明显的改善.这是所用数据集的本身性质所决定的,Last.fm 的用户数量和歌曲数量庞大,但是用户却很少为自己和歌曲标注,因而能准确定位用户和歌曲的合适标签更少.数据集中用户-标签和音乐-标签的密度分别为 4.6×10^{-4} 和 3.2×10^{-4},因此,社会化标注信息很稀疏,同时也不免掺杂一些噪音标签.这给社会化标注信息的应用增加了不少困难.尽管这样,本文提出的算法对推荐结果仍有所改善,随着时间的推移,用户信息以及社会化标注信息的丰富,利用社区网络图进行个性化推荐会更符合用户的兴趣爱好.

在计算效率方面,给定关联矩阵 ***TR*** 时,基于社会化标注与二部图的推荐算法只要经过较少次数的迭代就可以在 $O(n^2)$ 时间里得到一次推荐.但是本文算法的一个问题在于,如果关联矩阵 ***TR*** 中有一个节点发生了更新,则整个关联矩阵都需要更新,这将消耗 $O(n^2)$ 的时间,所以可以选择定时更新关联矩阵 ***TR***,同时实时更新节点向量的方法,以减少时间消耗并能得到较好的推荐效果.

4 结束语

本文通过挖掘潜在语义对基于二部图的协同推荐算法进一步改进,从而进行较为准确的音乐推荐.首先,以音乐为节点,音乐之间的标签关系及用户听取次数同时作为边,建立图模型.在图模型的基础上,利用基于二部图的推荐算法,得到与用户关联的歌曲列表.然后,将用户向量中包含的节点过滤掉,对剩余节点进行排序调整,初步得到用户的音乐推荐列表.最后,同样利用图模型获取与用户关联度较高的 N 个标签对应的歌曲集合,即间接关联音乐集合,对初次得到的推荐列表调整排序,得到最终推荐结果.通过在同一数据集上与基于用户的协同过滤算法及基于二部图的协同推荐算法的对比实验结果表明,该方法不失为个性化推荐的一种良策,尤其是随着 Web2.0 的发展及标签的增多,该方法会表现出较大的优势.

尽管本文提出的算法是基于社会化标注的,构建了用户-项目-标签的三维网络,并在推荐效果上表现出一些优势.但是作为一个典型的推荐算法,该

方法仍没有完全利用社交网络中的所有有效信息，基于社交网络图的推荐技术还是有很多地方值得继续研究. 比如标签间的相似度、用户交友关系以及歌曲固有信息间的相似度等，都可以充分合理地应用到音乐推荐技术中去，这也将是继续努力的方向.

参 考 文 献

[1] Ioannis K, Vassilios S. On social networks and collaborative recommendation [C] //Proc of the 32nd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2009: 195-202

[2] Mei Fang, Lin Hongfei. Social tag-based mobile music search [C] //Proc of the 5th China Information Retrieval Conf. Shanghai: Chinese Information Processing Society of China, 2009: 269-271 (in Chinese)
(梅放，林鸿飞. 基于社会化标签的移动音乐检索[C] //全国第五届信息检索学术会议(CCIR2009). 上海：中国中文信息学会，2009：262-271)

[3] Ferman A M, James H E, Peter van B, et al. Content-based filtering and personalization using structured metadata [C] // Proc of the 2nd ACM/IEEE-CS Joint Conf on Digital Libraries. New York: ACM, 2002: 393-393

[4] Jonathan L H, Joseph A K, Al B, et al. An algorithmic framework for performing collaborative filtering [C] //Proc of the 22nd Annual Int ACM SIGIR Conf Research and Development in Information Retrieval. New York: ACM, 1999: 230-237

[5] John S B, David H, CARL K. Empirical analysis of predictive algorithms for collaborative filtering [C] //Proc of the 14th Conf on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann, 1998: 43-52

[6] Dmitry P, David M P. A maximum entropy approach to collaborative filtering in dynamic, sparse, high-dimensional domains [C] //Proc of the 16th Annual Conf Neural Information Processing Systems. Cambridge: MIT Press 2002: 1441-1448

[7] Marko B, Yoav S. Fab: Content-based, collaborative recommendation [J]. Communications of the ACM, 1997, 40(3): 66-72

[8] Pan Yu. Design and realization of personalized coloring ring back tone recommendation system [D]. Dalian: Dalian University of Technology, 2007 (in Chinese)
(潘宇. 个性化彩铃推荐系统的设计与实现[D]. 大连：大连理工大学，2007)

[9] Wang Weiping, Wang Jinhui. Hybrid recommendation method based on tag and collaborative filtering [J]. Computer Engineering, 2011, 37(14): 34-36 (in Chinese)
(王卫平，王金辉. 基于 Tag 和协同过滤的混合推荐方法[J]. 计算机工程，2011，37(14)：34-36)

[10] Nick C, Martin S. Random walks on the click graph [C] // Proc of the 30th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2007: 239-246

[11] Hilmi Y, Mukkai S K. A random walk method for alleviating the sparsity problem in collaborative filtering [C] //Proc of the 2008 ACM Conf on Recommender Systems. New York: ACM, 2008: 131-138

[12] Yang S, Lu Z, Giles C L. Automatic tag recommendation algorithms for social recommender systems [J]. ACM Transactions on the Web, 2011, 5(1): 1-31

[13] Markus S. Purpose tagging: Capturing user intent to assist goal-oriented social search [C] //Proc of the 2008 ACM Workshop on Search in Social Media. New York: ACM, 2008: 35-42

Li Ruimin, born in 1986. Received her master's degree from Dalian University of Technology. Her main research interests includes data mining, information retrieval and recommender system.

Lin Hongfei, born in 1962. Professor and PhD supervisor in Dalian University of Technology. Member of China Computer Federation. His main research interests include information retrieval, datamining and natural language understanding.

Yan Jun, born in 1988. Received her master's degree from Dalian University of Technology. Her main research interests include data mining, information retrieval and recommender system.