基于好友关系和标签的混合协同过滤算法

曾 安 徐小强

(广东工业大学计算机学院 广州 510006)

摘 要 针对传统推荐算法存在数据稀疏影响推荐效果的问题,考虑到社交网络中的链路预测能够综合考虑用户节点之间的拓扑结构,以及好友关系能反映用户的兴趣爱好,提出了一种融合好友关系和标签信息的推荐算法。首先,借助网络资源分配算法对社交网络的结构信息进行特征提取;然后,利用TF-IDF构建合理的社会化标签模型;最后,利用线性模型融合两方面的信息,从而实现推荐。在Last.fm和Delicious数据集上的实验表明,与传统算法相比,所提算法在推荐的召回率和准确率指标上有显著提高。

关键词 链路预测,社交关系,标签,TF-IDF,推荐算法

中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.08.042

Hybrid Collaborative Filtering Recommendation Algorithm Based on Friendships and Tag

ZENG An XU Xiao-qiang

(School of Computer, Guangdong University of Technology, Guangzhou 510006, China)

Abstract The recommendation preference of a recommendation system was greatly affected by data sparseness. In order to solve this problem, a hybird collaborative filtering recommendation algorithm based on social network and tag information was proposed in this paper. The topology similarity characteristics among user nodes can be incorporated in the prediction of links in social network, and a circle of friends can exhibit a user's interests. Thus, network resource allocation algorithm is firstly utilized to extract social network structure information. Then, the tag information is resonably extracted with the help of TF-IDF method. Finally, the recommendation is made by linearly combining both the social network structure information and the tag information. The experiment results on Last.fm and Delicious dataset suggest that the advocated alogrithm is superior to other advanced approaches in both accuracy and reliability.

Keywords Link prediction, Social relationship, Tag, TF-IDF, Recommendation algorithm

1 引言

个性化推荐系统作为缓解信息负载的有效手段,通过分析用户的历史行为数据来获取用户的特征。然而,随着用户数量的急剧增加,在有限行为数据条件下构建用户兴趣模型,进而预测用户行为,变得十分困难[1]。在Web2.0时代,用户使用社会化标签对信息进行分类,可自由组织、管理和搜索所需的资源。这些标签信息反映了用户的偏好,成为推荐系统中的一种有用的数据源[2-3]。此外,随着社交网络的发展,人们开始逐渐重视好友关系在推荐系统中的作用。由于社交网络中用户活跃,而且具有较多的用户行为(如评论好友、建立好友关系),因此其有丰富的信息可供挖掘[4-5]。

传统的协同过滤算法倾向于依赖用户对项目的评分矩阵,但是实际应用中用户数量和商品数量十分庞大,而用户给予评分的商品项目却十分稀少。因此,存在一些亟需解决的问题,如数据稀疏问题、可解释问题等。针对这些问题,已经有一些研究学者从用户标签、用户特征、用户的隐式交互行为等数据源着手寻求解决方案,但是这些方法都忽略了用户之间的关系,将用户之间的关系等同对待。而事实上,在现实生活中,来自朋友的推荐往往具有比较高的可信度,而这种信任推荐可以从社交关系数据源中充分挖掘得到。

目前已经有一些研究学者通过融合用户朋友关系和标签信息进行推荐,如丁小焕等人[6]对用户、项目、标签三元组进行高阶奇异值分解,然后再结合用户朋友关系修正张量分解结果。但是采用张量分解方法模型进行训练存在复杂耗时、模型调优困难及可扩展性不强等缺点。因此,本文尝试着从其他角度融合朋友关系和标签信息:1)从社交网络分析中的链路预测角度出发,引入社交网络的链路预测指标来融合用户朋友关系,获得用户之间的相似度;2)对于标签数据源,从信息检索和数据挖掘技术角度出发,采用TF-IDF思想挖掘用户对标签的偏爱程度,进而获得用户对项目的偏好程度。

因此,本文从社交网络中的链路预测角度出发,引入社交网络的链路预测指标[7-8],通过好友间的关系传播利用Adamic-Adar指数计算用户间的相似性;此外,从信息检索的

到稿日期:2016-07-04 返修日期:2016-08-18 本文受国家自然科学基金项目(61300107),广东省自然科学基金项目(S2012010010212),广州市科技计划项目(201504301341059),广东省科技计划项目(2014B090901053)资助。

曾 安(1978—),女,博士,教授,主要研究方向为智能信息处理、数据挖掘,E-mail:zengan2010@126.com;徐小强(1987—),男,硕士生,主要研究方向为数据挖掘、推荐系统,E-mail:xxqcheers0614@163.com(通信作者)。

角度出发，采用 IF-IDF 思想挖掘用户对标签的偏爱程度[12]，进而获得用户对项目的偏好程度；最后，采用线性组合的方式融合这两种数据源，构建推荐模型，进而为用户推荐新项目。

2　相关性研究

为了提高推荐系统的精度并缓解评分稀疏对用户相似度计算的影响，一方面，有研究人员从用户社交关系[4-5,7]或用户之间的信任关系入手，构建社会化推荐模型；另一方面，也有研究人员从社会化标签[2-3]入手，由于显式的用户项目矩阵较为稀疏，引入标签这一概念后，传统推荐系统广泛采用的用户资源双向关联便被转变成由用户、资源和标签组成的三元关联。如果对标签的利用合适，那么标签也可以当作对内容个性化信息的一种补充，最终能够提升资源型推荐系统的效果。

2.1　融合用户社交关系

社交网络是一个以用户为中心的网络，人们在社交网络平台上可以发布和获取资源信息，并不断扩展自己的朋友关系。所谓“物以类聚，人以群分”，用户在社交网络中的关系能够反映出他们的兴趣爱好在一定程度上是相似的，相互联系越紧密的人群具有相似的兴趣爱好的可能性就越大。因此，融合社交网络中的信息源用于解决推荐系统面临的稀疏问题，变得越来越重要。

社交网络中的链路预测技术能够通过已知的网络节点以及网络结构等信息，来预测网络中尚未产生连边的两个节点之间产生链接的可能性。根据网络节点的相似性可以进行链接预测，为此把复杂网络中的网络资源分配算法（Resource Allocation，RA）用于社交网络结构信息的特征提取中，进而在用户之间社交关系的基础上获得用户间的相似度。

在基于用户社交网络的推荐系统中，可以用一张抽象的图来表示用户之间的社交网络关系，如图 1 所示，图中箭头表示用户的信任关系，数字表示信任度。

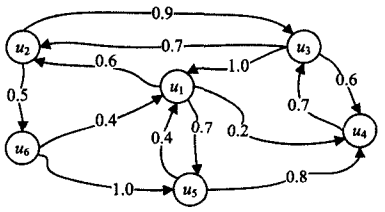

图 1　社交网络关系图

关于融合社交网络关系的研究，Yang X 等人[9]提出了一种基于社会网络的 top-k 推荐算法，该算法利用矩阵分解模型，并结合训练目标函数对未评分的数据进行调整，以提高推荐算法的性能。但是该方法并没有关注社交网络中用户节点之间的拓扑结构相似性。张燕平等人[10]针对用户历史数据和社交关系数据对用户影响力的不同，提出了全局影响力和局部影响力相结合的推荐算法，以期提升推荐效果。该方法虽然从整体和局部两个角度出发，但是并未充分利用整个网络的拓扑结构相似性。张富国[11]总结了近几年基于社交网络的推荐算法，针对社交网络中的稀疏性问题做了各方面的分析和总结，并对研究的难点和发展趋势进行了展望。因此，可以通过借助链路预测算法综合考虑用户节点之间的拓扑结构相似特征，充分挖掘用户之间的共同兴趣爱好，进而提高推荐系统的推荐精度。

2.2　融合社会化标签

社会化标签是指用户根据自己的兴趣爱好为某资源打上一个标签，以便以后重新检索或浏览。与传统标签不同，社会化标签在一个开放的平台上对标注的信息进行在线管理和存储，用户之间可以共享这些标签信息。用户可以根据自己的兴趣爱好自由建立群组，满足用户的社交需求。

在融合社会化标签的推荐算法中，蔡强等人[12]认为标签能体现用户兴趣爱好和项目的特征信息，把用户-标签-资源的三维关系转化成用户、项目和标签特征向量，并计算用户对项目的喜爱程度和项目间的相似度，然后通过用户的历史信息预测用户对其他项目的偏好值。该方法虽然能够在一定程度上缓解稀疏性的影响，但是其将用户之间的关系等同对待，并没有考虑到实际用户间的好友关系，推荐效果有待进一步提高。蔡孟松等人[13]把社交好友关系和社会化标签信息相结合，利用二部图结构建立可信用户集，最后计算社交用户的相似近邻集并作预测推荐，能够在一定程度上缓解冷启动的影响。

因此，标签作为体现用户兴趣偏好和资源特征的重要信息，将其应用到推荐算法时可提高推荐系统的性能，同时可以依据标签 TF-IDF 计算用户偏好程度和资源相似度。在此基础上，采用复杂网络领域中的网络资源分配算法 RA（Resource Allocation）对社交网络的结构信息进行特征提取[7-8]，进而在用户之间社交关系的基础上结合用户社会关系网络的拓扑结构相似性获得用户间的相似度。因此，在推荐系统中如果能有效地融合用户社交关系和社会化标签，将能获得更好的效果。

3　基于好友关系和标签的混合算法实现

本文尝试从社交网络中的链路预测角度出发，引入社交网络的链路预测指标，通过好友间的关系传播利用 Adamic-Adar 指数计算用户间的相似性。此外，对于标签数据源，本文尝试从信息检索和数据挖掘技术出发，采用 TF-IDF 思想挖掘用户对标签的偏爱程度，进而获得用户对项目的偏好程度。

3.1　用户相似性的计算

在现实生活中，人们通常喜欢从好友或者熟人那里获得推荐，社交网络中的好友关系可以反映现实中的人际关系。另外，同质性理论[14]指出具有相似特征的个体有选择彼此作为朋友的倾向；借助用户现实生活中的社交场景，在推荐解释方面会更加可靠和更具有说服力。因此，在推荐系统中不仅需要考虑用户自己的兴趣偏好，还需考虑好友们的兴趣偏好。

社交网络中的链路预测是指通过已知的网络节点以及社交网络结构等信息预测移动社交网络中尚未产生连边的两个节点之间产生链接的可能性。Adamic-Adar 指数[15]（Adamic-Adar Index）和资源分配指数（Resource Allocation Index）[16]用于计算社交网络结构图中任意两个未链接的节点之间在将来有可能会链接的概率。同时，这两项指数也经常被用于计算用户对项目的偏爱程度以及社交关系网中好友之间的信任程度。

（1）Adamic-Adar 指数：AA 指标思想是度小的共同邻居节点的贡献大于度大的共同邻居节点[8]的贡献。因此根据共

同邻居节点的度为每个节点赋予一个权重值，该权重等于该节点的度的对数的倒数。即该指数通过提升好友数量较小的相似权重来提升共同好友的权重计算，其定义如下：

$$sim^{AA}(u,v)=\sum_{z\in F(u)\cap F(v)}\frac{1}{\log_2^{K_z}} \tag{1}$$

其中，$F(u)$和 $F(v)$分别表示用户 u 和用户 v 的好友集合，K_z 表示用户 z 的好友个数。

(2)Resource Allocation 指数：该指数应用于复杂网络上进行动态资源的分配。假设两个节点 u 和 v 没有直接相连，而节点 u 可以把资源发送给 v，则它们的共同邻居节点起着传送的作用。假设每一个传送者有一个单位的资源，并且传送给每个邻居节点的概率是一样的。类似地，资源也可以从 v 传送到 u。该指数的定义如下：

$$sim^{RA}(u,v)=\sum_{Z\in F(u)\cap F(v)}\frac{1}{K_z} \tag{2}$$

显然，这样的计算方法是对称的，即 $sim(u,v)=sim(v,u)$。从这里可以看出，虽然 Adamic-Adar 指数和资源分配指数的目的不同，但是它们有着相似的结构。更进一步说，它们都降低了拥有高度数(入度和出度)的共同邻居节点的权重。Adamic-Adar 指数使用的是对数倒数的形式，而资源分配指数使用的是倒数的形式，这就说明了资源分配指数降低较高度数的共同邻居节点的力度比 Adamic-Adar 指数的力度更大。

(3)Jaccard 指数：Jaccard 指数通过计算两个节点的共同邻居数目来确定节点的相似性，也就是说两个节点如果有更多的共同邻居，则它们更倾向于连边。Jaccard 指数表示 u_a 和 u_b 的相似性由它们共同的邻居决定[18]。其计算公式如下：

$$sim^{Jad}(u,v)=\frac{|I_{u_a}\cap I_{u_b}|}{|I_{u_a}\cup I_{u_b}|} \tag{3}$$

Jaccard 指数是在共同邻居的基础上考虑两端节点度的影响，也即基于共同邻居的相似性。Adamic-Adar 指数则考虑的是两节点共同邻居的度信息，其思想是度小的共同邻居节点的贡献大于度大的共同邻居节点的贡献。Resource Allocation 指数则考虑的是网络中若没有直接相连的两个节点 x 和 y，可从 x 传递一些资源到 y，而在此过程中，它们的共同邻居就成为了传递的媒介，假设每个媒介都有一个单位的资源并且将同概率地配传给它的邻居，则 y 可以接收到的资源数就定义为节点 x 和节点 y 的相似度。

Adamic L A 等[19]指出，仅考虑节点邻居信息的若干指标中，Adamic-Adar[19]表现得最好。为此，本文算法中采用 Adamic-Adar 指数。

为了更清晰地解释文中引进的 Adamic-Adar 指数和资源分配指数、Jaccard 系数，以表 1 为例进行说明。表 1 中，行方向有 3 个用户，列方向有 5 个其他用户，如果用户间是好友则用 1 表示，否则用 0 表示。

表 1 好友关系列表

	Peter	David	Mike	Tony	Nicky	…
好友个数	15	10	4	12	14	…
Alice	1	1	1	0	1	…
Lucy	1	0	1	0	0	…
Francis	1	1	0	1	0	…
…	…	…	…	…	…	…

表 2 分别列举了 Alice 与 Lucy 和 Alice 与 Francis 在 Adamic-Adar 指数、Resource Allocation 指数和 Jaccard 指数上的计算过程。

表 2 用户间相似度的计算过程

	Sim(Alice,Lucy)	*Sim*(Alice,Francis)
Jaccard 系数	2/4=0.5	2/5=0.4
Adamic-Adar 指数	1/(log2(15))+1/(log2(4))=0.7560	1/(log2(15))+1/(log2(10))=0.5570
资源分配指数	1/15+1/4=0.3167	1/15+1/10=0.1667

3.2 用户-标签偏爱程度的计算

传统的协同过滤算法是基于用户对项目的评分来衡量用户的偏好，把与用户有着共同兴趣爱好的其他用户喜爱的项目推荐给该用户。由于推荐系统中项目数量庞大且不断增加，使得用户-项目评分矩阵变得非常巨大，同时用户给项目的评分非常少，从而导致评分数据稀疏，很难提取用户兴趣爱好。社会标签可实现对信息资源的分类，并可自由地给项目贴标注，通过分析用户使用标签的标记记录，可挖掘出用户对项目的喜爱程度。

TF-IDF 是一种统计方法，用以评估一个词对于一个文档集或者一个语料库中的某一份文档的重要程度，字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比例下降。也就是说，频数小的标签不比频数大的标签的相关性小(IDF 思想)；某一标签在某一物品的多次标注不比单一标注的相关性小(TF 思想)；热门物品不比冷门物品更受欢迎(规范化思想)。

一般地，某一个标签标注同一个物品的次数越多，就表示该标签更加符合该物品的内容。然而，许多被用户使用的热门标签也不一定能很好地描述该物品，也不一定能代表用户的兴趣。

给定用户集合 U，$U=\{u_1,u_2,u_3,\cdots,u_M\}$，$M$ 是用户个数；项目集合 R，$R=\{r_1,r_2,r_3,\cdots,r_N\}$，$N$ 是项目个数；标签集合 T，$T=\{t_1,t_2,t_3,\cdots,t_K\}$，$K$ 是标签个数；三元组 A，$A\in\{\langle u,t,r\rangle\}$，$u\in U$，$t\in T$，$r\in R$，表示用户 u 在项目 r 上标注标签 t。用户 u_i 对标签 t_k 的偏爱程度 $rel(u_i,t_k)$ 的计算公式如下：

$$rel(u_i,t_k)=TF(u_i,t_k)*IDF(t_k)=\frac{n_{u_it_k}}{n_{u_it}}*\log\frac{M}{n_{t_ku}} \tag{4}$$

其中，$n_{u_it_k}$ 表示用户 u_i 使用标签 t_k 的次数，n_{u_it} 表示用户 u_i 使用所有标签的次数，M 表示用户的个数，n_{t_ku} 表示使用标签 t_k 的用户个数，$\frac{n_{u_it_k}}{n_{u_it}}$ 表示用户 u_i 使用标签 t_k 的频率，$\log\frac{M}{n_{t_ku}}$ 表示用户 u_i 在标签 t_k 的重要程度。

3.3 用户预测评分值和算法步骤

前面通过社交网络中的好友关系分析得到了用户之间的相似性，则给定用户 u 和项目 i，基于好友关系的预测评分值 $p_u(u,i)$的计算公式如下：

$$p_u(u,i)=\sum_{v\in N_i}\frac{sim(u,v)}{IC_v} \tag{5}$$

其中，N_i 是对项目 i 有评分的用户集合，IC_v 是用户 v 评分过的项目的个数，$sim(u,v)$是用户 u 和用户 v 之间的相关性。

得到用户对标签的偏爱程度后，可以通过标签信息挖掘出用户对项目的偏爱程度。给定用户 u 和项目 i，基于标签信息的预测评分值 $p_t(u,i)$的计算公式如下：

$$p_t(u,i)=\sum_{t\in NT_i}\frac{rel(u,t)}{ITC_t} \tag{6}$$

其中，NT_i 是项目 i 的所有标签，ITC_t 是标签 t 的项目个数，$rel(u,t)$是用户 u 对标签 t 的偏爱程度。

得到基于好友关系的预测评分值和基于标签信息的预测评分值后，设定权重 α 对两种预测评分值进行线性组合，从而得到最终的预测评分值公式：

$$p(u,i)=\alpha * p_u(u,i)+(1-\alpha) * p_t(u,i) \tag{7}$$

从上述公式可以看出，当 $\alpha\neq 0$ 且 $\alpha\neq 1$ 时，该算法综合了好友关系和标签信息的预测评分值，而当 $\alpha=0$ 或者 $\alpha=1$ 时，算法都只用到了其中一个数据源预测评分值。

基于好友关系和标签的混合协同过滤算法的具体步骤如下。

输入：目标用户 u，用户好友关系矩阵 RF(m,m)，用户-标签矩阵 RT(m,k)，推荐的项目个数 N_{top}

输出：推荐给目标用户 u 的 N_{top} 个项目集 I_{top}

1. 根据式(1)，基于用户好友关系矩阵 RF(m,m)计算出用户 u 与其他用户 v 的相似性 sim(u,v)；
2. 根据式(4)，基于用户-标签矩阵 RT(m,k)计算用户 u 对标签 t 的偏爱程度 rel(u,t)；
3. 根据式(5)和步骤 1 的结果，计算用户 u 对未使用过的项目 i 基于好友关系的预测评分值 $p_u(u,i)$；
4. 根据式(6)和步骤 2 的结果，计算用户 u 对未使用过的项目 i 基于标签信息的预测评分值 $p_t(u,i)$；
5. 根据式(7)和步骤 3、步骤 4 的结果，通过调和权重 α 计算用户 u 对未使用过的项目 i 的预测评分值 p(u,i)；
6. 把预测评分值最高的前 N_{top} 个项目放入集合 I_{top} 中，并推荐给用户 u。

4　实验结果及分析

4.1　数据集与度量标准

本实验采用的数据集是 Last. fm 数据集和 Delicious 数据集。Last. fm 是 Audioscrobbler 音乐引擎设计团队的一个音乐社交平台，它允许用户在平台上创建自己的页面、交友、给音乐贴标签等，并记录所有用户听过的歌曲名称和听过的次数。Delicious 数据集来自于 Delicious. com 在线网页书签网站，用户可以对各种网页链接进行标注，用户间可以相互交流。由于这两个数据集中包括较多噪音数据，本实验筛选出部分数据作为实验数据：针对 Last. fm 数据集，要求每一个用户都至少听过 20 首歌，且每一首歌都被 20 个用户听过；针对 Delicious 数据集，选取朋友数目大于 5 的用户和标签数目大于 6 的项目。

随机选取 80%的数据集作为训练集，剩下的 20%作为测试集。实验过程中采用准确率(Precision)和召回率(Recall)作为衡量算法的标准。准确率表示用户对系统推荐资源感兴趣的概率，召回率表明一个用户喜欢的项目被推荐的概率。准确率和召回率越高，表示推荐效果越好。

4.2　实验结果分析

(1)权重值 α 对 FT-CF 算法的影响

为了衡量权重值 α 对算法的影响，分别将 α 的取值设为 0，0.1，0.2，0.3，…，1 进行实验，实验结果如图 2、表 3、图 3 所示。图 2 和表 3 表示在 Last. fm 数据集上，不同 α 值对 FT-CF 算法推荐性能的影响；图 3 示出 Delicious 数据集上不同 α 值对 FT-CF 算法推荐性能的影响。为了消除推荐项目的数量对 α 值的影响，实验过程中依次将推荐数量设置为 5，10，15，…，50，共进行 10 次实验。

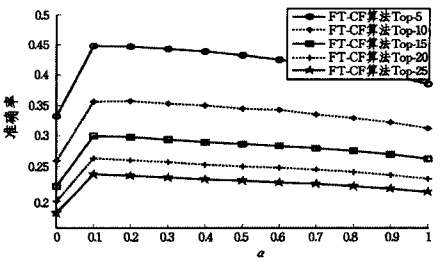

图 2　Last. fm 数据集上 α 值对 FT-CF 算法准确率的影响

表 3　不同 α 值对 FT-CF 算法准确率的影响

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Top-5	0.3331	0.4485	0.4471	0.4436	0.4395	0.4327	0.4262	0.4177	0.4092	0.4005	0.3858
Top-10	0.2588	0.3567	0.3572	0.3536	0.3503	0.3454	0.3425	0.3362	0.3300	0.3225	0.3126
Top-15	0.2168	0.3000	0.2979	0.2942	0.2901	0.2871	0.2837	0.2799	0.2752	0.2692	0.2617
Top-20	0.1925	0.2626	0.2599	0.2567	0.2530	0.2501	0.2476	0.2442	0.2406	0.2356	0.2295
Top-25	0.1734	0.2367	0.2342	0.2312	0.2288	0.2258	0.2234	0.2207	0.2173	0.2134	0.2081
Top-30	0.1594	0.2162	0.2147	0.2116	0.2089	0.2070	0.2043	0.2021	0.1993	0.1960	0.1913
Top-35	0.1485	0.2007	0.1991	0.1965	0.1935	0.1913	0.1892	0.1872	0.1850	0.1822	0.1782
Top-40	0.1389	0.1876	0.1855	0.1833	0.1805	0.1786	0.1769	0.1747	0.1727	0.1703	0.1664
Top-45	0.1321	0.1765	0.1745	0.1717	0.1694	0.1675	0.1654	0.1640	0.1622	0.1600	0.1564
Top-50	0.1251	0.1665	0.1640	0.1616	0.1599	0.1579	0.1562	0.1545	0.1529	0.1508	0.1475

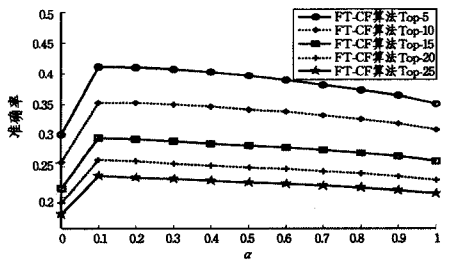

图 3　Delicious 数据集上 α 值对 FT-CF 算法准确率的影响

从图 2 和图 3 中可以看出，无论推荐项目数量是多少，针对 α 值的变化，FT-CF 算法的准确率曲线趋势基本一致，即随着 α 值增加，其开始缓慢上升，达到最大值，而后又缓慢下降。在本实验中，当 $\alpha=0.1$ 时准确率是最高的；α 在 0.1～0.5 之间时，准确率下降比较缓；α 在 0.1～0.5 之间时，准确率下降较前者快。也就是说，基于好友关系和标签的混合协同过滤算法的性能比单独基于标签信息、用户社交数据的算法的推荐性能好。

(2)稀疏性验证

为了验证所提出的基于好友关系和标签的混合协同过滤算法 FT-CF(Hybrid Collaborative Filtering Recommendation Algorithm Based on Friendships and Tag)在稀疏数据集上的推荐性能,本文将 FT-CF 算法与 PRT-CF 算法[12](Personalized Resource Recommendation Based on Tags and Collaborative Filtering)、UCTRA 算法[17](Collaborative Filtering Recommendation Algorithm Using Social and Tag Information)、Tag-CF 算法、SW-CF 算法进行了对比。PRT-CF 算法代表基于标签和协同过滤的个性化资源推荐算法;UCTRA 代表结合社交与标签信息的协同过滤推荐算法;Tag-CF 代表本文算法中只用到标签信息推荐的算法;SW-CF 代表本文算法中只用到社交网络的推荐算法。

用户好友关系矩阵和用户标签矩阵相对稀疏时,会影响用户相似性的度量,进而影响推荐系统的性能。本文用 Delicious 数据集对稀疏性问题进行验证。从 Delicious 数据集中选取用户朋友关系小于 4 的用户、标签数目小于 3 的项目进行实验,将选择的数据按 8:2 的比例随机生成训练集和测试集。实验结果如图 4 和图 5 所示,其中图 4 的评测指标是准确率,图 5 的评测指标是召回率。

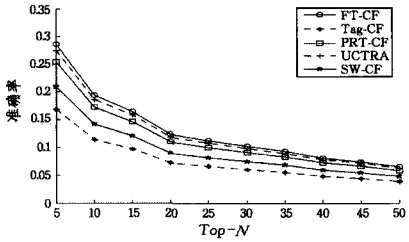

图 4 在较稀疏数据集上不同推荐算法的准确率

从图 4 可以看出,在所选取的数据集倾向于稀疏的情况下,随着推荐数目 top-N 的变化,FT-CF 算法的准确率相比于传统的方法更为优越。

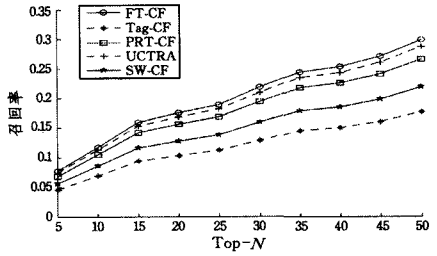

图 5 在较稀疏数据集上不同推荐算法的召回率

从图 5 可以看出,相比于传统的方法,FT-CF 算法的召回率更为优越。此外,融合好友关系和标签信息的方法比仅仅利用好友关系或者标签信息具有更好的推荐性能,这说明通过融合用户朋友关系和用户标签数据能够缓解数据稀疏时所造成的推荐精度不高的问题,这与文献[6]得出的结论也是相符的。

(3)不同算法的推荐性能的比较

为了验证所提出的基于好友关系和标签的混合协同过滤算法的性能,将本文提出的 FT-CF 算法与 PRT-CF 算法、UCTRA 算法、Tag-CF 算法、SW-CF 算法进行对比分析。

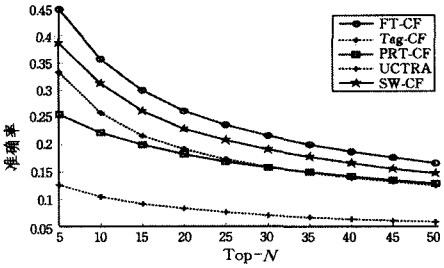

图 6 Last. fm 数据集中推荐项目个数对准确率的影响

在实验过程中,取 $\alpha=0.1$ 时,评测各个算法的准确率和召回率,在 Last. fm 数据集下各算法的准确率和召回率统计如表 4 所列。图 6 所示为推荐项目个数对准确率的影响,图 7 所示为推荐项目个数对召回率的影响。

表 4 Last. fm 数据集中推荐项目个数对准确率和召回率的影响

推荐物品数量		5	10	15	20	25	30	35	40	45	50
准确率	FT-CF	0.4485	0.3567	0.3000	0.2626	0.2367	0.2162	0.2007	0.1876	0.1765	0.1665
	Tag-CF	0.3331	0.2588	0.2168	0.1925	0.1734	0.1594	0.1485	0.1389	0.1321	0.1251
	PRT-CF	0.2561	0.2227	0.1999	0.1823	0.1684	0.1586	0.1496	0.1416	0.1351	0.1293
	UCTRA	0.1266	0.1047	0.0918	0.0837	0.0760	0.0709	0.0670	0.0636	0.0610	0.0590
	SW-CF	0.3858	0.3126	0.2617	0.2295	0.2081	0.1913	0.1782	0.1664	0.1564	0.1475
召回率	FT-CF	0.1355	0.2155	0.2719	0.3173	0.3576	0.3918	0.4245	0.4533	0.4798	0.5029
	Tag-CF	0.1006	0.1564	0.1964	0.2325	0.2619	0.2889	0.3139	0.3356	0.3592	0.3778
	PRT-CF	0.0774	0.1345	0.1812	0.2203	0.2543	0.2875	0.3163	0.3423	0.3672	0.3906
	UCTRA	0.0383	0.0633	0.0831	0.1011	0.1149	0.1285	0.1416	0.1536	0.1659	0.1782
	SW0CF	0.1165	0.1888	0.2371	0.2773	0.3143	0.3467	0.3767	0.4022	0.4251	0.4457

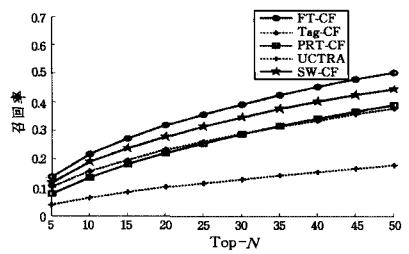

图 7 Last. fm 数据集中推荐项目个数对召回率的影响

从图 6、图 7 中可以看出,本文的 FT-CF 算法在准确率和召回率上都要优于其他算法,并且 FT-CF 算法在融合好友关系与标签信息的基础上,其推荐性能比仅仅使用标签信息(如 PRT-CF 算法和 Tag-CF 算法)或社会网络关系(如 SW-CF 算法)的算法要好。

结束语 利用标签信息和社交网络信息是推荐系统解决稀疏性问题的一种有效途径。本文从社交网络中的链路预测角度出发,引入社交网络的链路预测指标,通过好友间的关系

传播利用 Adamic-Adar 指数来计算用户间的相似性。另外,本文尝试从信息检索和数据挖掘技术出发,采用 IF-IDF 思想挖掘用户对标签的偏爱程度,进而获得用户对项目的偏好程度。采用线性组合的方式充分利用这两种数据源为用户推荐项目,并通过实验验证了较单纯使用标签信息或社会关系等算法,FT-CF 算法具有更高的准确率和召回率,而且相比其他传统算法具有更好的性能。由于标签数据代表用户对项目的语义信息,如果能够借助自然语言处理的语义分析模型以及语义分析工具将标签中的信息挖掘出来,将能更好地提高推荐精度,这也是下一步的研究目标。

参 考 文 献

[1] TAKÁCS G, PILÁSZY I, NEMETH B, et al. Investigation of various matrix factorization methods for large recommender systems[C] // IEEE International Conference on Data Mining Workshops, 2008(ICDMW'08). IEEE, 2008: 553-562.

[2] TSO-SUTTER K H L, MARINHO L B, SCHMIDT-THIEME L. Tag-aware recommender systems by fusion of collaborative filtering algorithms[C] // Proceedings of ACM Symposium on Applied Computing. New York: ACM, 2008: 1995-1999.

[3] ZHOU T C, MA H, KING I, et al. TagRec: Leveraging Tagging Wisdom for Recommendation[C] // International Conference on Computational Science and Engineering. IEEE, 2009: 194-199.

[4] JAMALI M, ESTER M. A matrix factorization technique with trust propagation for recommendation in social network[C] // Proc. of the ACM Recommender Systems Conf.. New York: ACM Press, 2010: 135-142.

[5] ZOU B Y, LI C P, TAN L W, et al. Social Recommendations Based on User Trust and Tensor Factorization[J]. Journal of Software, 2014, 25(12): 2852-2864. (in Chinese)
邹本友,李翠平,谭力文,等. 基于用户信任和张量分解的社会网络推荐[J]. 软件学报, 2014, 25(12): 2852-2864.

[6] DING X H, PENG F R, WANG Q, et al. Tensor factorization recommendation algorithm combined with social network and tag information[J]. Journal of Computer Applications, 2015, 35(7): 1979-1983. (in Chinese)
丁小焕,彭甫镕,王琼,等. 融合朋友关系和标签信息的张量分解推荐算法[J]. 计算机应用, 2015, 35(7): 1979-1983.

[7] XIANG R, NEVILLE J, ROGATI M. Modeling relationship strength in online social networks[C] // International Conference on World Wide Web. ACM, 2010: 981-990.

[8] JAVARI A, GHARIBSHAH J, JALILI M. Recommender systems based on collaborative filtering and resource allocation[J]. Social Network Analysis & Mining, 2014, 4(1): 1-11.

[9] YANG X, STECK H, GUO Y, et al. On top-k recommendation using social networks[C] // ACM Conference on Recommender Systems. ACM, 2012: 67-74.

[10] ZHANG Y P, ZHANG S, QIAN F L, et al. Local and global user influence combined social recommendation algorithms[J]. Journal of Nanjing University (Natural Sciences), 2015(4): 858-865. (in Chinese)
张燕平,张顺,钱付兰,等. 一种局部和全局用户影响力相结合的社交推荐算法[J]. 南京大学学报(自然科学版), 2015(4): 858-865.

[11] ZHANG F G. Survey of Online Social Network Based Personalized Recommendation[J]. Journal of Chinese Computer Systems, 2014, 35(7): 1470-1476. (in Chinese)
张富国. 基于社交网络的个性化推荐技术[J]. 小型微型计算机系统, 2014, 35(7): 1470-1476.

[12] CAI Q, HAN D M, LI H S, et al. Personalized Resource Recommendation Based on Tags and Collaborative Filtering[J]. Computer Science, 2014, 41(1): 69-71. (in Chinese)
蔡强,韩东梅,李海生,等. 基于标签和协同过滤的个性化资源推荐[J]. 计算机科学, 2014, 41(1): 69-71.

[13] CAI M S, LI X M, YIN Y T. Hybrid top-N recommendation method based on social user tag[J]. Application Research of Computers, 2013, 30(5): 1309-1311. (in Chinese)
蔡孟松,李学明,尹衍腾. 基于社交用户标签的混合 top-N 推荐方法[J]. 计算机应用研究, 2013, 30(5): 1309-1311.

[14] MCPHERSON M, SMITH-LOVIN L, COOK J M. Birds of a feather: Homophily in social networks[J]. Annual Review of Sociology, 2001, 27(1): 415-444.

[15] ADAMIC L A, ADAR E. Friends and neighbors on the Web[J]. Social Networks, 2003, 25(3): 211-230.

[16] ZHOU T, LII L, ZHANG Y C. Predicting missing links via local information[J]. The European Physical Journal B, 2009, 71(4): 623-630.

[17] YU H, LI J H. Collaborative Filtering Recommendation Algorithm Using Social and Tag Information[J]. Journal of Chinese Computer Systems, 2013, 34(11): 2467-2471. (in Chinese)
于洪,李俊华. 结合社交与标签信息的协同过滤推荐算法[J]. 小型微型计算机系统, 2013, 34(11): 2467-2471.

[18] JACCARD P. Etude comparative de la distribution floraledansune portion des Alpeset du Jura[J]. Bulletin De La Societe Vaudoise Des Sciences Naturelles, 1901, 37(142): 547-579.

[19] ADAMIC L A, ADAR E. Friends and neighbors on the web[J]. Social networks, 2003, 25(3): 211-230.