基于社交用户标签的混合 top-*N* 推荐方法*

蔡孟松，李学明，尹衍腾
（重庆大学 计算机学院，重庆 400044）

摘　要：针对协同过滤方法的冷启动问题，提出一种将社交用户标签与协同过滤相结合的混合 top-*N* 推荐方法。通过社交用户关系获得可信用户集，然后根据个性化标签采用结构上下文相似性算法（SimRank）计算社交用户相似近邻集并进行预测推荐，最后结合传统协同过滤方法进行推荐。实验结果表明，该方法能够提高在一般数据集及冷启动用户数据集下的推荐精度。

关键词：推荐系统；协同过滤；社交网络；个性化标签；冷启动
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1001-3695(2013)05-1309-03
doi:10.3969/j.issn.1001-3695.2013.05.007

Hybrid top-*N* recommendation method based on social user tag

CAI Meng-song, LI Xue-ming, YIN Yan-teng
(*College of Computer Science, Chongqing University, Chongqing* 400044, *China*)

Abstract: Aiming at the cold start user problem of tradition recommendation system, this paper proposed a hybrid top-*N* recommendation method which combined social user tag and collaborative filtering. The method obtained trustworthy user set through social relations. Based on user personalized tags, it applied a structural-context similar algorithm named SimRank to generate the similar neighbors set, which was used to generate predictive ratings. Finally, it combined the predictive ratings of traditional collaborative filtering to provide the recommendations. The experimental evaluations on all users and cold start users dataset demonstrate the proposed method outperforms the collaborative filtering approach in terms of recall and precision.

Key words: recommender system; collaborative filtering; social network; personalized tags; cold start

0　引言

互联网的普及信息量急剧增长，促进了推荐系统（recommender systems）的广泛应用。研究表明，在基于电子商务的销售行业中使用个性化推荐系统可以使销售收入提高 1% ~ 2%[1]。协同过滤是当前推荐服务中应用最广泛且最成功的方式，它根据用户的相似性来推荐资源。当为用户提供一个预测评分的列表时，通常采用 top-*N* 推荐方法[2]。

随着 Web 2.0 的发展，社交网络已经步入人们生活的各个方面，人们在现实生活中的关系逐渐延伸到互联网，Facebook、Twitter 等社交网站也因此取得了巨大的成功。由于社交网络中用户非常活跃，而且具有较多的用户行为（如评论好友、添加话题），因而包含大量可供挖掘的信息。

协同过滤只依赖于用户对项目的评分矩阵，对于各种特定的应用都有较好的适用性，但它也存在一些难以完全解决的问题，如新用户问题、数据稀疏性问题及可信问题等[3]。针对这些问题，已经有一些工作分别从用户标签[4]、用户特征[5]、用户浏览行为[6]等方面来解决。但是，这些方法都是在传统的协同过滤方法基础上予以提升，忽略了用户之间的关系，而在现实生活中，一个朋友的推荐则具有更高的可信及精确度，而这种现实朋友关系通常也体现在社交网络中。

目前已经有一些研究将用户关系结合协同过滤进行推荐，如文献[7,8]利用用户信任关系构建信任网络，采用随机游走算法结合协同过滤算法以解决协同过滤算法的新用户，但是仅从信任关系上选取用户，没有考虑到用户间的行为及偏好。文献[9]则由用户社交关系提取用户相关性结合协同过滤算法进行研究，对预测精度有很好的提升，但是忽略了对冷启动问题的解决，且没有考虑到社交用户除了关系以外的社交行为信息的影响。文献[10]利用社交关系并使用矩阵分解方法进行推荐，取得了很好的效果，但是也没有考虑社交用户的行为信息及社交关系对冷启动问题的影响。

因此，本文提出了一种基于社交用户标签的混合 top-*N* 推荐算法。本文的主要贡献如下：

a）针对协同过滤算法冷启动及精确性问题，提出了一种将社交推荐与协同过滤相结合的 top-*N* 推荐方法，并有很好的效果。

b）提出了一种社交推荐方法，利用用户在社交网络中的个性化标签及用户关系采用图的分析方法，最终形成推荐。

c）针对社交推荐及协同过滤的推荐结果的结合问题，提出了一种基于权重的组合推荐方法。

1　基于社交用户标签的 top-*N* 推荐方法

算法思想：a）使用基于用户的协同过滤算法对目标用户进行预测评分并产生推荐；b）在社交网络平台根据用户社交行为寻找用户相似近邻集合进行预测评分并产生推荐；c）将这两种方法得到的推荐结果结合产生最终的推荐项目。

收稿日期：2012-09-06；**修回日期**：2012-10-27　　**基金项目**：国家自然科学基金资助项目（61103114）

作者简介：蔡孟松（1988-），男，江苏宿迁人，硕士研究生，主要研究方向为推荐系统、数据挖掘（abscasey@gmail.com）；李学明（1967-），男，重庆人，教授，主要研究方向为数据挖掘、网格计算；尹衍腾（1987-），男，硕士研究生，主要研究方向为数据挖掘、电子商务.

1.1 基于用户的协同过滤 top-N 推荐

基于用户的协同过滤算法的 top-N 推荐算法通常包括:建立用户档案;计算用户相似性,选取近邻集合;预测评分计算并产生 top-N 推荐项目。

1)建立用户档案

根据通常的定义方法[11],定义用户集合为 $U=\{u_1,u_2,\cdots,u_m\}$,商品项目集合为 $I=\{i_1,i_2,\cdots,i_n\}$;$r_{u,i}$表示用户 u 对项目 i 的行为(如购买或评分行为),当用户无行为时则 $r_{u,i}$ 记为 0;所有用户对项目的行为可以表示为一个 $m\times n$ 的矩阵,记为 R_{mn}。

2)计算用户相似性

计算用户相似性方法常见的有余弦相似性、皮尔森相关系数等,本文采用皮尔森相关系数方法进行计算。用户 u 和用户 v 共同评价的项目集合是 $I_{u,v}$,则用户 u 和 v 之间的相似性为

$$\mathrm{sim}_{c_{u,v}}=\frac{\sum_{i\in I_{u,v}}(r_{u,i}-\bar{r}_u)(r_{u,i}-\bar{r}_v)}{\sqrt{\sum_{i\in I_{u,v}}(r_{u,i}-\bar{r}_u)^2}\sqrt{\sum_{i\in I_{u,v}}(r_{v,i}-\bar{r}_v)^2}} \tag{1}$$

其中:$r_{u,i}$是用户 u 对项目 i 的评分,$\bar{r}_v$ 是用户 v 的平均评分。选取协同过滤下最为相似用户个数 K_c,得到用户的近邻集合 U_c。

3)计算预测评分

选取评分最高的 N_c 个项目生成推荐集 R_c,对目标用户 u 按公式进行预测评分,得到预测项 $r_{c_{u,i}}$:

$$r_{c_{u,i}}=\frac{\sum_{v\in U_{c_u},i\in I_v}\mathrm{sim}_{c_{u,v}}\times r_{v,i}}{\sum_{v\in U_{c_u},i\in I_v}\mathrm{sim}_{c_{u,v}}} \tag{2}$$

其中:$r_{c_{u,i}}$是用户 u 对项目 i 的预测评分,U_{c_u}是用户 u 的近邻集,$\mathrm{sim}_{c_{u,v}}$是协同过滤下用户 u 和 v 的相似性。

1.2 基于社交网络的推荐

基于社交网络用户标签推荐的主要思想是根据社交网络中用户关系及用户个性化标签,使用图的相似性计算方法 SimRank 算法找出目标用户的可信相似近邻集,然后对商品进行预测评分。

1.2.1 相关定义

在社交平台上,用户之间以朋友的关系相连,包含了用户间的信任关系,这就构成了一个巨大的社交网络。此外,社交用户会产生一定的用户行为,如给个人标注标签或给其他人或物标注标签,这些标签可以表征用户的喜好、表达用户观点[12]等。本文则是以社交网络的用户标签及用户关系为研究出发点。

定义 1 个性化标签。表征用户身份、偏好的自我描述或他人对用户的评价、印象或用户常用于对物品的标签,可以是简单的短语,如 80 后、武侠、爱猫等。

定义 2 信任度。社交网络中某用户对其他用户的信任程度,记为 trust,取值为 0 ~ 1,用户自己对自己的信任度为 1,对陌生人信任度则为 0。

定义 3 信任递减因子。用户在社交网络中因对直接朋友及间接朋友的信任度的不同而表现出来的信任递减因子 μ $(0<\mu<1)$,如用户 a 距离用户 b 的深度为 k,则用户 a 和 b 的信任度 trust(a,b)为 μ^k。

定义 4 用户信任集。距离用户深度为 k 的可信用户的集合,记为 U_t。

1.2.2 算法描述

社交网络的推荐步骤如下:

a)构建用户信任集合 U_t。在社交网络中,从目标用户 u 出发,使用广度优先算法(breadth first searchs,BFS)寻找在距离用户 u 深度为 1 ~ k 范围内的用户构建用户信任集合 U_t。根据六度分隔理论,任何人经过至多六个人即可以与陌生人产生联系,当两个用户深度为 6 时其信任度 trust 接近 0,即 μ^6 值要很小,因此取 μ 值为 0.5。因为随着 k 的增加,搜索的用户数会急剧增加而信任度迅速递减,则会带来较大的处理成本但却对结果影响较小,综合以上因素取深度 k 值为 3。

b)计算用户近邻集合。根据用户的个性化标签,可以将用户—标签构成一个二部图,如图 1 所示。本文使用 SimRank 算法计算信任相似度。

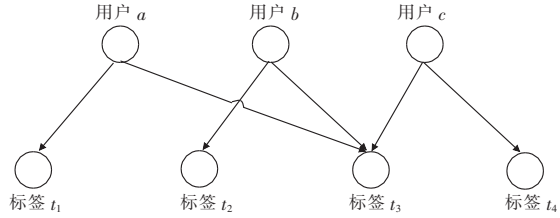

图 1 用户—标签二部图

SimRank 算法是 2002 年为了衡量结构性上下文(structural-context)的相似性而提出的[13]。SimRank 算法基于这样的假设:如果两个对象分别与其他相似的对象相似,则这两个对象也相似。

为了计算本文用户信任集上的用户相似性,将 SimRank 扩展到二部图,则有假设:如果两个用户分别具有相似的标签,则这两个用户相似;如果两个标签都标注在相似的用户上,则这两个标签相似。构造如图 1 的二部图,分为用户集合 N_U 和标签集合 N_T,如果用户 a 具有标签 t_1,则添加一条由用户 a 指向标签 t_1 的有向连线。用户及标签相似性按如下公式计算:

$$S_{m+1}(u,u')=\frac{\mathrm{trust}(u,u')\times C_1}{|O(u)||O(u')|}\sum_{i=1}^{|O(u)|}\sum_{j=1}^{|O(u')|}S_m(O_i(u),O_j(u')) \tag{3}$$

$$S_{m+1}(t,t')=\frac{C_2}{|I(t)||I(t')|}\sum_{i=1}^{|I(t)|}\sum_{j=1}^{|I(t')|}S_m(I_i(t),I_j(t')) \tag{4}$$

其中:m 是指迭代的次数;$u'\in U_t$,trust(u,u')表示用户 u 和 u'的信任度;$O(u)$表示用户 u 的出邻居(out-neighbor)集合,即由 u 出发所指向的标签集合;$|O(u)|$表示用户 u 的出邻居个数;$O_i(u)$为 u 的单个出邻居$(1\leqslant i\leqslant|O(u)|)$;$I(t)$表示标签 t 的入邻居(in-neighbor)集合,即所有指向 t 的用户集合;$|I(t)|$表示标签 t 的入邻居个数;$I_i(t)$为 t 的单个入邻居$(1\leqslant t\leqslant|I(t)|)$;$C_1,C_2$ 为常数,在 0 ~ 1 之间。

初始迭代的值记为 $S_0(a,b)$,当 $a=b$ 时 $S_0(a,b)=1$,否则为 0。根据文献[13],经过约五次迭代计算即可得到稳定的迭代结果,将最终计算得到的用户 u 和 v 间的相似性 $S_5(u,v)$记为 $\mathrm{sim}_{s_{u,v}}$。对目标用户 u 选取 K_s 个此时最相似用户,记为近邻集 U_s。

c)计算预测评分,选取评分最高的 N_s 个项目生成推荐集 R_s。对目标用户 u 按公式进行预测评分,得到预测评分为 $r_{s_{u,i}}$,将评分最高的 N_s 作为推荐结果。

$$r_{s_{u,i}}=\frac{\sum_{v\in U_{s_u},i\in I_v}\mathrm{sim}_{s_{u,v}}\times r_{v,i}}{\sum_{v\in U_{s_u},i\in I_v}\mathrm{sim}_{s_{u,v}}} \tag{5}$$

其中:$r_{s_{u,i}}$是预测评分,U_{s_u}是最近邻集合,$\bar{r}_v$ 是用户 v 的平均评

分,$\mathrm{sim}_{s_{u,v}}$是社交网络下经过 SimRank 算法迭代得到的用户 u 和 v 的相似性 $S_5(u,v)$。

1.3　混合的推荐

由基于用户协同过滤方法得到目标用户 u 的 K_c 个最近邻用户并推荐数量为 N_c 的项目集,记为 R_c。由社交网络方法得到 K_s 个最近邻并推荐出 N_s 个项目集合,记为 R_s。将这两个推荐集合组合进行推荐,最终推荐的 top-N 集合 R 表示为

$$R = R_c + R_s - R_s \cap R_c \tag{6}$$

其中:R 的个数 N 满足

$$N = \alpha \times N_c + (1-\alpha) \times N_s \tag{7}$$

α 在 0 ~ 1 间变化,即对于某个推荐个数 N 及 α,可以确定协同过滤推荐个数 N_c 及社交推荐个数 N_s。

2　实验评估

2.1　实验设计

实验使用 ACM 第 5 次推荐系统大会公开的 Last. fm 数据集。Last. fm(http://www. lastfm. com)是一个音乐社交网站,用户在收听音乐的同时可以对艺术家进行标注,产生社会标签,并可以添加好友进行社交活动。该数据集包含 1 892 名用户,音乐覆盖 17 632 名艺术家,11 946 个社会标签,1 271 条用户之间关系。平均每个用户有 13. 4 个朋友,有 98. 6 个标签,每个艺术家被 5. 3 个用户收听。

Last. fm 数据集中用户对艺术家有收听行为但没有直接评分。本文将用户对艺术家的收听次数转换为间接评分进行实验。另外,实验中选取用户对艺术家标记的标签作为用户个性化标签。实验数据集分为训练集和测试集,其中训练集占 90%,测试集占 10%。实验过程分为两个部分,首先评估本文方法在整个数据集上的推荐效果,然后在冷启动用户占 50% 的数据集上实验,其中把用户行为数少于 5 个的用户认为是冷启动用户。实验中为了取得更好的对比效果,本文选取 N 值为 50。

2.2　评估标准

对 top-N 推荐结果的评估,本文使用召回率(recall) 和准确率(precision) 来度量,召回率和准确率是信息检索领域中最常用的度量标准之一。1968 年由 Cleverdon 提出并一直使用至今[14]。

召回率表示推荐列表中相对于用户测试集中实际选择对象命中数与测试集中用户实际跟踪对象数之比,其值越高表示推荐的效果越好。

$$\mathrm{recall} = N_{rs}/N_r \tag{8}$$

准确率表示推荐列表中相对于用户测试集中实际选择对象的命中数与推荐对象数之比。

$$\mathrm{precision} = N_{rs}/N_s \tag{9}$$

2.3　实验结果及分析

2.3.1　*一般数据集效果评估*

首先,实验先选取不同的 α 值在[0,1]间并且 K_c 与 K_s 分别在(0,50)变化域上进行推荐。召回率随 α 变化曲线如图 2 所示。一般数据集下在 α 值为 0. 85 时取得最好效果,表明在一般数据集中协同过滤方法对推荐影响更大。此时协同过滤最近邻个数 K_c 为 4,社交网络推荐最近邻个数 K_s 为 3。

图 3 表示召回率随用户最近邻个数变化曲线。图中曲线分别对应基于用户协同推荐、社交推荐、在 $K_c=4$ 时 K_s 变化的混合推荐、在 $K_s=3$ 时 K_c 变化的混合推荐的召回率变化曲线。由图 3 可见,当混合推荐中最近邻个数的 K_c 或 K_s 单独变化时,其结果都要优于任意单一推荐效果。

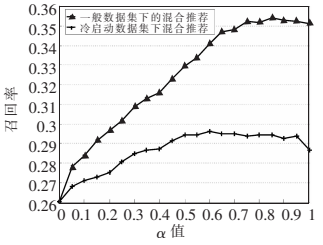

图 2　召回率随 α 值变化图

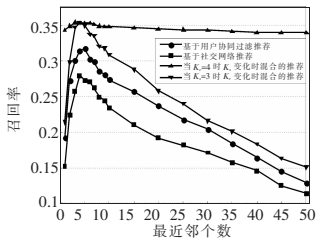

图 3　召回率随最近邻个数变化曲线

图 4 表示准确率随用户最近邻个数变化曲线。同样可以看出,混合推荐的效果优于任意单一方法。

2.3.2　*冷启动用户数据集效果评估*

按照 2. 3. 1 节方式进行实验。如图 2 所示,在冷启动数据集下,当 α 值为 0. 6 时召回率值达到最高,即表明在冷启动数据集下提高社交推荐的比例可以提高推荐效果,此时 $K_c=3$, $K_s=3$。

召回率随最近邻个数变化曲线如图 5 所示。图中曲线分别对应基于用户协同推荐、社交推荐、在 $K_s=3$ 时 K_c 变化的混合推荐、在 $K_c=3$ 时 K_s 变化的混合推荐的召回率变化曲线。由图 5 可以看出,冷启动数据集下协同过滤推荐效果受到较大影响,通过混合推荐可以有效提高推荐效果。

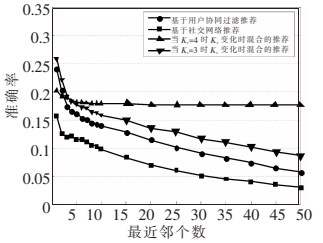

图 4　准确率随最近邻个数变化曲线

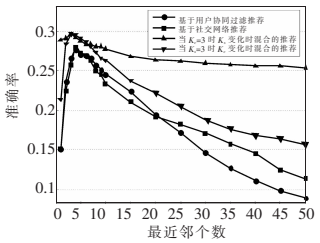

图 5　在冷启动用户数据集下召回率随最近邻个数变化曲线

准确率随最近邻个数变化曲线如图 6 所示。由图 6 可以看出,混合推荐方法在冷启动用户数据集上准确率有一定的提高。

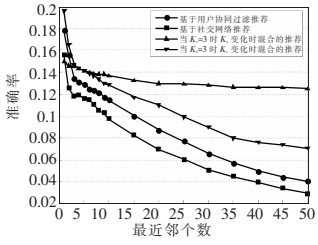

图 6　在冷启动用户数据集下准确率随最近邻个数变化曲线

3　结束语

随着社交化的迅速普及,社交网络在各方面得到了广泛的应用。针对常用的协同过滤算法的不足,本文利用社交网络中用户的关系及个性标签结合传统协同过滤算法进行混合推荐。实验结果表明,本文方法能够有效缓解冷启动问题,提高推荐质量。下一步将对社交用户行为进行研究,提取更加个性化的标签,以进一步提高推荐效果。

(下转第 1344 页)

繁抖动影响,使得任务被频繁地剥离出处理器核,造成处理器在处理的过程中会产生间隙,同时总任务数较低,也拉低了系统的整体利用率;而随着任务数的逐渐增多,可以很容易地观察到 SMD 算法优先执行被依赖的节点任务的优势得以体现,其利用率稳步上升,当任务集中的任务数达到 300 个时,其任务利用率甚至比经典 RM 算法提高了近 15 个百分点。

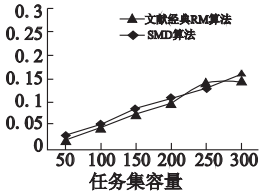

图4 两种算法执行6组任务集后的丢失率

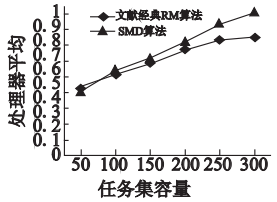

图5 两种算法执行6组任务集后的处理器利用率

4 结束语

本文通过对多核处理器环境下的特殊实时调度问题进行分析研究,从如何提高复杂实时系统中的处理器利用率及降低软实时系统的丢失率入手,提出了一种新的基于 MST 矩阵的及任务动态因子的调度算法。该算法通过对 MST 矩阵的特性进行分析,将任务划分为一个个的并行集,再通过综合考虑已执行时间、任务间的依赖关系及任务的最早截止期几个要素,以动态因子的形式对任务进行实时调度。实验结果表明,该算法能较好地对文中所提出的这种既具有依赖关系又具有周期性的实时任务集进行调度,比之经典 RM 算法,在处理器利用率及任务丢失率两项重要指标上皆有所提升。未来工作的重点在于将单行树的任务集扩展为一般树的形式,以及将一棵树的形式扩展为多棵树的形式,以适应范围更广的应用要求。

参考文献:

[1] LIU C L, LAYLAND J W. Scheduling algorithms for multiprogramming in a hard-real-time environment[J]. Journal of ACM, 1973, 20(1): 46-61.

[2] LEONTYEV H. Compositional analysis techniques for multiprocessor soft real-time scheduling[D]. Chapel Hill: the University of North Carolina, 2010.

[3] CHETTO H, CHETTO M. Some result of the earliest deadline scheduling algorithm[J]. IEEE Trans on Parallel and Distributed Systems, 1989, 15(10): 1261-1269.

[4] MOK A K. Fundamental design problems of distributed systems for the hard-real-time environment[D]. Massachusetts: Massachusetts Institute of Technology, 1993.

[5] BRANDENBURG B B, ANDERSON J H. On the implementation of global real-time schedulers[C]//Proc of the 30th IEEE Real-time Systems Symposium. Washington DC: IEEE Computer Society, 2009: 214-224.

[6] BARUAH S K, COHEN N K, PLAXTON C G, *et al*. Proportionate progress: a notion of fairness in resource allocation[J]. Algorithmica, 1996, 15(6): 600-625.

[7] 杨根科,吴智铭. 周期实时任务在多处理器下的可调度条件[J]. 上海交通大学学报, 2004, 38(9): 1597-1600.

[8] LOPEZ J M, DIAZ J L, GARCIA D F. Minimum and maximum utilization bounds for multiprocessor RM scheduling[C]//Proc of Euromicro Workshop on Real-time Systems. [S. L.]: IEEE Press, 2001: 67-75.

[9] 黄金贵,李荣珩. 独立多处理机任务静态调度问题的近似算法[J]. 软件学报, 2010, 21(12): 3211-3219.

[10] 王涛,刘大昕. 多处理器单调速率任务分配算法性能评价[J]. 计算机科学, 2007, 34(1): 272-277.

(上接第 1311 页)

参考文献:

[1] LAWRENCE R D, ALMAS G S, KOTLYAR V, *et al*. Personalization of supermarket product recommendations[J]. Data Mining and Knowledge Discovery, 2001, 5(1/2): 11-32.

[2] McLAUGHLIN M R, HERLOCKER J L. A collaborative filtering algorithm and evaluation metric that accurately model the user experience[C]//Proc of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2004: 329-336.

[3] 李春,朱珍民,叶剑,等. 个性化服务研究综述[J]. 计算机应用研究, 2009, 26(11): 4001-4005.

[4] 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): 35-36.

[5] 李改,李磊. 一种解决协同过滤系统冷启动问题的新算法[J]. 山东大学学报:工学版, 2012, 42(2): 11-17.

[6] 余小高,余小鹏. 基于隐式评分的推荐系统研究[J]. 计算机应用, 2009, 29(6): 1585-1589.

[7] JAMALI M, ESTER M. Using a trust network to improve top-*N* recommendation[C]//Proc of the 3rd ACM Conference on Recommender Systems. New York: ACM Press, 2009: 181-188.

[8] LI Rui-feng, ZHANG Yin, YU Hai-han, *et al*. A social network-aware top-*N* recommender system using GPU[C]//Proc of ACM/IEEE Joint Conference on Digital Libraries. 2011: 287-296.

[9] ELEFTHERIOS T, YANNIS M. Product recommendation and rating prediction based on multi-modal social networks[C]//Proc of the 5th ACM Conference on Recommender Systems. New York: ACM Press, 2011: 61-68.

[10] MA Hao, ZHOU Deng-yong, LIU Chao. Recommender systems with social regularization[C]//Proc of the 4th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2011: 287-296.

[11] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions[J]. IEEE Trans on Knowledge and Data Engineering, 2005, 17(6): 734-749.

[12] SEN S W. Nurturing tagging communities[D]. Minnesota: University of Minnesota, 2009.

[13] JEH G, WIDOM J. SimRank: a measure of structural context similarity[C]//Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2002: 538-543.

[14] HERLOCKER J L, KONSTAN J A, TERVEEN L G, *et al*. Evaluating collaborative filtering recommender systems[J]. ACM Trans on Information Systems, 2004, 22(1): 5-53.