DOI: 10.3785/j.issn.1008-973X.2016.05.026

基于社交关系拓扑结构的冷启动推荐方法

张亚楠[1], 曲明成[2], 刘宇鹏[1]
(1.哈尔滨理工大学 软件学院,黑龙江 哈尔滨 150040;2.哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

摘 要: 针对冷启动用户仅有很少行为信息,很难为冷启动用户给出推荐的问题,提出基于比较社交网络中用户间社交关系拓扑结构的冷启动推荐方法.社交网络中包含多种可以反映用户偏好的社交关系,然而现有基于社交网络的冷启动推荐研究仅利用一种或者很少的社交关系,没有充分利用社交网络中的多种社交关系,很少考虑融合相异的社交关系,限制了在实际环境中对冷启动用户的推荐效果.由于社交关系在社交网络中的权重越大在推荐中的影响越大,为了给出准确的冷启动推荐,提出基于社交关系拓扑的相似用户发现方法(STSUM),基于最大熵原理融合社交网络中多种相异的社交关系,基于图形模式匹配为冷启动用户发现相似用户,给出推荐.在真实的网站中提取社交关系和用户数据,实验结果表明,STSUM 可以有效地提高对冷启动用户的推荐效果且需要较少的训练集.

关键词: 冷启动推荐;社交网络;图形模式匹配;最大熵
中图分类号: TP 391 **文献标志码**: A **文章编号**: 1008-973X(2016)05-01001-08

Recommendation method based on social topology for cold-start users

ZHANG Ya-nan[1], QU Ming-cheng[2], LIU Yu-peng[1]
(1. *Software school, Harbin University of Science and Technology, Harbin* 150040, *China*;
2. *School of computer science and Technology, Harbin Institute of Technology, Harbin* 150001, *China*)

Abstract: It is very difficult to give recommendations for cold-start user who usually has very sparse historical behavior records. A cold-start recommendation method was proposed based on comparison of topology of social relationships in social networks to improve recommendation effectiveness for cold-start user. Social network contains many social relationships which could reflect user's preference. However, most of existing social network based recommendation methods use only one or a few social relationships of social network, which do not make full use of multiple social relationships; rarely consider how to merge dissimilar social relationships, and could not give satisfactory recommendation in actual environment. In social network the higher weight a kind of social relationship takes, the greater right of recommendations it will have. In order to give accurate recommendations for cold-start user, a social topology based similar user matching method (STSUM) was proposed, Maximum entropy principle was introduced to merge multiple social relationships, and graph pattern matching was used to find similar users for cold-start user. Then recommendations were given according to similar users' records. Social relationship and user data from a real website to show the recommendation effectiveness of STSUM. The experimental results show that STSUM con give accurate recommendations for cold-start user and needs a few training set.

Key words: cold-start recommendation; social network; graph pattern matching; maximum entropy

收稿日期: 2015-12-04. **浙江大学学报(工学版)网址**: www.journals.zju.edu.cn/eng
基金项目: 国家自然科学基金青年基金资助项目(61300115).
作者简介: 张亚楠(1981—),男,讲师,从事社交网络推荐等研究. ORCID: 0000-0002-0633-826X E-mail: ynzhang_1981@163.com

新用户在登录电子商务网站的初期,通常没有或者仅有很少的购买、评价等行为信息,称这类用户为冷启动用户.为冷启动用户的推荐称为冷启动推荐.由于冷启动推荐是针对历史行为信息稀少的用户的推荐,很难根据冷启动用户的行为信息得到推荐的依据.在社交网络中,存在多种类型的社交关系,如兴趣组、评论转发关系、微博关注关系等,称为用户间多种逻辑的社交关系,这些社交关系从多个维度描述用户的特征.

由于冷启动推荐的难点是冷启动用户的历史行为信息稀少,因此对冷启动推荐的研究一直围绕着挖掘、扩充冷启动用户的信息以及为冷启动用户发现相似用户[1-6].国内外学者对冷启动推荐的研究,主要包括基于协同过滤的推荐和基于用户间信任关系的推荐.

1)基于协同过滤的推荐.基于协同过滤的推荐根据用户之间的相似程度或者项目之间的相似程度给出推荐[7].基于协同过滤的推荐效果受用户数据的稀疏程度影响.在描述用户对项目评价的用户项目矩阵中,绝大多数用户仅有很少的数据或者没有数据,很难判断哪些用户相似.为解决用户数据稀疏的问题,Jamali等[8]提出将原用户项目矩阵分解为低秩矩阵,用低秩矩阵的乘积估计用户项目矩阵中的未知值. Ma等[9]提出基于概率矩阵分解(probabilistic matrix factorization, PMF)的推荐,引入(nonnegative matrix factorization,NMF)的预测值与真实评价值的误差符合正态分布的限制条件,为冷启动用户给出推荐. Wu等[10]提出在矩阵分解的过程中引入影响用户喜好的特征向量可以提高推荐的准确度. Koren[11]提出时间敏感的协同过滤推荐算法,在用户、项目特征向量中引入时间特征,较好地解决用户兴趣漂移问题. Ren[12]提出平衡评分预测机制的协同过滤推荐方法,为个性化评分与全局评分的动态权重调整提供了一种新思路.基于协同过滤的推荐可以避免由于不完全或不精确特征抽取而产生的不准确推荐,并且可以给出较为新颖的推荐,但是推荐效果受用户历史行为信息数量的影响非常明显.

2)基于信任的推荐.信任关系是一种稳定的社交关系,由信任用户给出的推荐更可靠.准确地发现用户间信任关系是基于信任推荐的关键.对信任关系的研究主要包括信任关系的传递策略[13-14]、验证信任网络具有小世界特征[15],基于社交网络的小世界特性和用户间弱关系构建信任网络[16].如Liu等[17]提出传播过程中扩散的相对量与绝对量的权重博弈. Guha等[18]提出用户间的信任关系可通过由用户给出少量不信任或信任实例预测.印桂生等[19]提出将长尾分布与受限信任关系融合的推荐方法,为很少被关注的冷门商品给出一种推荐的途径,以及基于受限信任关系和概率分解矩阵的推荐[20],孙光福等[21]提出基于时序行为的协同过滤推荐算法,将时序因素融入推荐中,较好的解决了推荐结果偏移的问题。然而基于信任的推荐对用户历史行为信息不敏感,并且信任关系仅仅是单一维度的社交关系,不能全面描述用户的特征.本文研究如何基于社交网络的拓扑,融合社交网络中多种社交关系为冷启动用户发现相似用户,进而根据相似用户的行为记录为冷启动用户给出推荐.

1 冷启动推荐问题的定义

随着社交网络的发展,越来越多的用户加入到社交网络中,社交网络中丰富的用户社交信息可以用来反映用户在现实社会中的偏好,社交网络中存在部分用户有大量的购物记录和评价信息,也存在相当数量的冷启动用户,由于社交网络中用户间的社交关系反映了用户的行为及交友偏好,因此可以通过挖掘冷启动用户的社交关系,进而发现与其具有相似社交关系且历史评价信息充足的用户,根据这些用户的偏好给冷启动用户推荐.用户商品矩阵描述用户对商品的评价值.用户商品矩阵是以用户ID号为行,商品ID为列组成的二维矩阵.矩阵中的元素表示用户对商品的评价值,评价值越高,表示用户对商品越满意,用户选择该商品的概率要高于评价值低的商品.推荐算法的实质是预测用户项目矩阵 $\boldsymbol{R}=[r_{u,i}]_{m\times s}$ 中的未知元素,即用户对未知商品的评价值,依据评价值的排序,将Top-N评价值对应的商品推荐给用户.在用户项目矩阵中 $r_{u,i}$ 为用户 u 对商品 i 的评价值,m 为用户个数,s 为商品个数.

本文提出的相似用户发现方法((social topology based similar user matching method,STSUM)方法可以分为2步,第1步发现与目标用户存在相似社交关系的用户.第2步根据社交关系的相似程度预测用户项目矩阵中的未知项.

2 冷启动用户的社交关系拓扑

社交关系拓扑是社交网络中用户间社交关系的抽象.社交网络是用户在现实社会中交往关系在网络中的真实映射,在社交网络中用户间的社交关系

体现了用户的偏好和特征. 例如，用户加入探险俱乐部说明他很可能非常喜欢探险，加入车友会则预示他很喜欢驾车出游，因此通过用户在社交网络中的社交关系可以预测出用户的偏好. 虽然冷启动用户的购买、评价等信息少，但是在社交网络中的部分冷启动用户具有多种类型的社交关系，可以通过社交关系推测冷启动用户的偏好特征. 冷启动用户的社交关系种类越多，反映其偏好特征的信息越多，充分挖掘冷启动用户的社交关系，并为其发现具有相似社交关系的用户，根据与冷启动用户相似用户的评价或购买信息为冷启动用户给出推荐.

为发现冷启动用户的相似用户，需要从社交网络的拓扑结构中寻找与冷启动用户的社交关系的拓扑相似的用户. 社交网络中用户间的社交关系可以抽象为以用户为节点，用户间的社交关系为有向边的拓扑. 社交网络是由表示用户的节点以及表示用户间社交关系的边组成的有向图. 其中边的权重值可以表示社交关系的紧密程度，对边标记类型可以区分不同种类的社交关系. 以代表冷启动用户的节点和其连接的节点共同构成冷启动用户社交关系的拓扑.

社交关系的拓扑：令 u 表示冷启动用户，u' 表示与冷启动用户存在某种社交关系的用户，用户 u 和 u' 之间边的权重值（即社交关系紧密程度），用 u 和 u' 之间路径长度 $f_e(u, u')$ 表示，路径长度越短则对应的社交关系越紧密，路径长度越长则对应的社交关系越疏远. $f_e(u, u')=\infty$ 表示用户 u 和 u' 之间不存在社交关系. 在社交网络中，对边标记类型可以区分不同种类的社交关系，任意用户 u 和 u' 间的社交关系类型用 $f_L(u, u')$ 表示，其中 f_L 为社交关系类型的谓词. 通过社交关系类型和社交关系的路径长度可以描述用户在社交网络中的社交关系拓扑. 用户 u 的社交关系拓扑中包括与 u 直接存在社交关系的用户 u'，以及通过 u' 间接与 u 存在社交关系的用户 u". 令 $\mathrm{Topo}(u)$ 表示用户 u 的拓扑，则 $\mathrm{Topo}(u)=\sum_{u'\leftrightarrow u} f_e(u,u')\otimes f_L(u,u')$，其中 u' 为在社交网络中与 u 存在直接或者间接社交关系的用户.

3　基于社交关系拓扑结构发现冷启动用户的相似用户

基于社交网络中用户间的社交关系为冷启动用户发现相似用户，进而根据相似用户的购买或者评价等行为信息为冷启动用户给出推荐. 首先用户间的社交关系表示为以用户为节点，和带有权重的边组成的拓扑. 通过比较用户社交关系的拓扑，为冷启动用户发现相似用户. 相似的社交关系拓扑指用户间的社交关系类型一致或者相似，并且用户间该社交关系的紧密程度相似. 其中，用户间某种社交关系的紧密程度指在社交网络中通过该社交关系连接的用户节点间的距离. 例如，社交网络中的用户 a 参加摄影团队，且就职于医疗机构，与用户 a 具有相似的社交关系的用户，需要同时具备相似的社交关系类型以及相近的社交关系紧密程度. 如果存在用户 b 参加过摄影比赛，并且就职于医院，则可得出用户 a 与用户 b 比较相似. 如果存在用户 c 参加过摄影比赛，并且就读于医科院校，则可得出用户 a 与用户 c 具有一定的相似性.

表示冷启动用户的节点和与其连接的节点共同构成冷启动用户的社交关系拓扑. 为冷启动用户发现相似用户的步骤可以分为：1）标记出冷启动用户的社交关系拓扑结构，即找到与冷启动用户存在某种社交关系的节点，标记节点间的社交关系类型 f_L 和紧密程度 f_e. 2）在社交网络中发现与冷启动用户拓扑结构相似的用户. 3）基于相似用户的历史购物记录或评价记录作为冷启动用户的推荐依据. 将表示冷启动用户社交关系的拓扑结构作为匹配模式. 为冷启动用户发现相似用户的过程，可等价于在社交网络中发现与匹配模式相匹配的拓扑的过程. 为冷启动用户 u 发现相似用户的示意图如图 1 所示，图 1 中左侧虚线内为冷启动用户的社交关系拓扑，中间虚线内为整个社交网络拓扑的示意图. 实线和虚线分别表示 2 种不同类型的社交关系. 在表示社交网络的有向图中比较社交关系的类型和社交关系的紧密程度，发现与冷启动用户的社交关系相匹配的用户为图 1 中右侧虚线中的 d_1 和 f_2.

用户间社交关系的拓扑结构匹配可以借鉴图形模式匹配的思想，现有的图形模式匹配主要包括子图同形和图形模拟. 子图同形匹配条件是有向图与

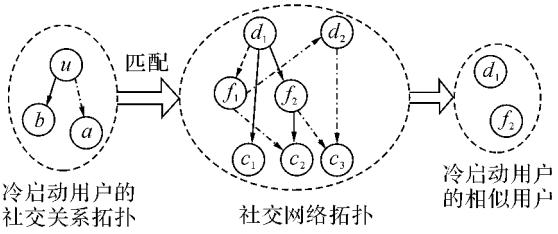

图 1　基于有界图形模式匹配的相似用户发现示意图

Fig. 1　Schematic diagram of finding similar users for cold-start users by Bounded graphic pattern matching

匹配模式中的节点存在双射关系,在社交网络中双射关系意味着用户间的社交关系是完全对称的,以微博关注关系为例,双射关系要求2个用户彼此关注,这种限制非常严格,在微博关注这种社交关系中更多的情况是众多的用户关注少数微博博主.图形模拟是一种边到边映射,边到边的映射可以比较2条边的类型、权重信息,可以在社交网络中比较存在直接社交关系的用户节点,却不能比较存在间接社交关系的用户是否相似,因此不能为冷启动用户匹配社交网络中间接的社交关系,在社交网络中,直接的社交关系数量有限,不能充分为冷启动用户发现相似用户,间接的社交关系中蕴含着范围更广阔的潜在相似用户.因此充分挖掘间接社交关系并比较用户间的间接社交关系的相似程度,可以为冷启动用户发现大量的相似用户.本文提出一种在社交网络中匹配间接社交关系的方法,为冷启动用户发现相似用户,进而为冷启动用户给出推荐.

3.1 基于有界图形模式匹配的相似用户发现方法

冷启动用户的社交关系拓扑作为匹配模式,用有向图表示整个社交网络中用户间的社交关系,在有向图中发现所有与匹配模式相匹配的节点集合,即与冷启动用户具有相似社交关系的用户.令 u 表示冷启动用户,u' 表示与冷启动用户存在社交关系的用户,$f_e(u, u')$ 表示 u 和 u' 之间的路径长度,$f_L(u,u')$ 表示 u 和 u' 间的社交关系类型,v 和 v' 表示社交网络中任意用户,$f_C(v, v')$ 表示社交网络中 (v, v') 的社交关系类型,V_Q 表示匹配模式中的节点集合,V 表示有向图中所有节点集合,令 $S \subseteq V_Q \times V$,基于有界图形模式匹配的相似用户条件:a)用户 v 的属性与用户 u 的属性相似.b)对于 (u, u'),有向图中存在一个非空且长度不超过 $f_e(u, u')$ 的路径 $v/\cdots/v'$,且存在 $(u,v) \in S$.c)对于 (u, u'),G 中存在路径 $v/\cdots/v'$,使 $f_C(v,v_1)\cdots f_C(v_n, v')$ 满足 (u, u') 上的关系 $f_L(u, u')$,其中路径 $v/\cdots/v'$ 中节点的顺序为 $(v, v_1, \cdots, v_n, v')$.设匹配模式 $P = (V_Q, E_Q, f_c, f_e)$,用户数据图 $G = (V, E, f_A)$.基于有界图形模式匹配的相似用户发现算法如算法1所示.在模式 P 和用户数据图 G 的匹配过程中,将匹配结果按照匹配程度的降序排列.

算法1 基于有界图形模式匹配的相似用户发现

输入:模式 $P = (V_p, E_p, f_c, f_e)$,用户数据图 $G = (V, E, f_A)$.

输出:当 $P \lhd G$ 的最大匹配 S

1. 计算 G 的距离矩阵 $\boldsymbol{M}$;
2. for each $(u', u) \in E_p, x \in V$ do
3. 计算 anc$(f_e(u', u), f_v(u'), x)$, desc $(f_e(u', u), f_v(u'), x)$;
4. for each $u \in V_p$ do
5. mat $(u) := \{x \mid x \in V, f_A(x)$符合 $f_v(u)$,且 out-degree $(a) \neq 0$, out-degree $(x) \neq 0\}$;
6. premv $(u) := \{x \mid x \in V$, out-degree (x) if out-degree $(a) \neq 0$, 并且 $\exists (u', u) \in E_p (x' \in$ mat $(u), f_A(x)$满足 $f_v(u')$,并且 len $(x/\ldots/x') \leqslant f_e(u', u))\}$;
7. while $(u \in V_p$ with premv $(u) \neq \varnothing)$ do
8. for (each $(u', u) \in E_p, z \in$ premv $(u) \cap$ mat $(u'))$ do
9. mat $(u') :=$ mat $(u') \setminus \{z\}$;
10. if (mat $(u') = \varnothing$) then return $\varnothing$;
11. for each u'' with $(u'', u') \in E_p$ do
12. for $(z' \in$ anc$(f_e(u'', u'), f_v(u''), z) \wedge z' \in$ premv $(u'))$ do
13. if (desc $(f_e(u'', u'), f_v(u'), z') \cap$ mat $(u') = \varnothing)$
14. then premv $(u') :=$ premv $(u') \cup \{z'\}$;
15. premv $(u) := \varnothing$;
16. $S := \varnothing$;
17. for $(u \in V_p$ and $x \in$ mat $(u))$ do $S := S \cup \{(u, x)\}$;
18. return S

对任意模式 $P = (V_p, E_p, f_c, f_e)$,待匹配用户数据图 $G = (V, E, f_A)$,基于有界图形模式匹配的相似用户发现算法的时间复杂度与模式 P 和用户数据图 G 中的节点个数以及边的个数的乘积正相关,基于有界图形模式匹配的相似用户发现的时间复杂度 $O((|V| + |V_p|)(|E| + |E_p|))$,将 $(|V| + |V_p|)(|E| + |E_p|))$ 展开得 $(|V||E| + |V||E_p| + |E||V_p| + |V_p||E_p|)$,通常模式 P 的节点和有向边规模要远小于用户数据图 G,且在用户数据图中 $|E|$ 与 $|V|^2$ 是近似相等的,因此可化简为 $O(|V||E| + |E_p||V| + |V_p||V|^2)$.

3.2 相异逻辑的社交关系赋权重值

在社交网络中,用户间具有多种逻辑的社交关系,如兴趣组、评论转发关系、微博关注关系等,这些社交关系从多个维度描述用户的特征,相似的特征映射着用户间的购买或评价行为也相似.利用多种逻辑的社交关系发现相似用户,需要为社交网络中相异逻辑的社交关系给出合理的权重值,即融合多种相异逻辑的社交关系,以得到最准确的推荐.社交网络中不同社交关系的权重是很难直接设定的,对某种社交关系给予不同的权重值,得到的冷启动用户的相似用户也不相同.各种社交关系的权重分配可以有多种组合分布,其中有一种社交关系权重的分布可以得到最大的熵.选用这种具有最大熵的分布作为社交关系组合的权重分布,是在未完整掌握社交关系权重分布时,选取符合已知条件且熵值最

大的概率分布策略. 这种推断就是符合已知条件最不确定或最随机的推断，任何其他的社交关系权重都意味着增加了有偏向性且多余的约束. 基于最大熵准则为相异逻辑的社交关系赋权值，使得社交关系的权重值的熵最大，并且使相似用户预测的评价值与真实评价值的差值与社交关系权重的乘积之和最小，其数学模型如下：

$$\left.\begin{aligned}&\hat{w}_q = \arg\max_{w_q}\{-\sum_{q=1}^{m}\ln w_q\},\\&\hat{w}_q = \arg\min_{w_q}\sum_{q=1}^{m}\sum_{i,j}w_q(r_{i,j}-\hat{r}_{i,j})^2,\\&\sum_{q=1}^{m}w_q = 1.\end{aligned}\right\} \tag{1}$$

式中：w_q 为社交关系 q 在所有社交关系中所占的权重，w_q 的权重之和为 1. $\hat{w}_q$ 为同时满足使预测评价值与真实评价值差的和取得最小值，以及社交关系的权重值的熵取得最大值时的社交关系 q 权重值. $\hat{r}_{i,j}$ 为由相似用户预测的评价值，$r_{i,j}$ 为冷启动用户真实评价值.

3.3　基于相似用户的冷启动推荐

STSUM 方法基于有界图形模式匹配和相异逻辑的社交关系赋权重值可以为冷启动用户发现相似用户，基于这些相似用户的已有评价或者购买记录为冷启动用户给出推荐. 在基于有界图形模式匹配的过程中，将模式 P 和用户数据图 G 匹配，将匹配结果按照匹配程度的降序排列，根据社交关系的相似程度预测用户项目矩阵中的未知项，其中未知项的值可由式(2)求得，为冷启动用户 a 给出推荐的形式化描述如式(3)所示：

$$p_{a,i} = \overline{r_a} + \frac{\sum_{u=1}^{M} w_{a,u}(r_{u,i}-\overline{r_u})}{\sum_{u=1}^{M} w_{a,u}}, \tag{2}$$

$$\mathrm{top}_N(a,I) := \max_{i\in I} p_{a,i}. \tag{3}$$

式中：$p_{a,i}$ 为预测的当前用户 a 对商品 i 的评价值，$w_{a,u}$ 为用户 a 对 u 的社交关系相似程度，$w_{a,u}$ 的值越大，用户的推荐权重越大，$\overline{r_a}$ 为用户 a 的平均评价值，$\overline{r_u}$ 为用户 u 的平均评价值. M 为商品的评价数量，即评价过商品 i 的用户数量. 式(3)中 $I=\{i_1, i_2, \cdots, i_s\}$ 表示商品集合. $\mathrm{top}_N(a, I)$ 为按照评价值降序排列的前 N 个推荐结果.

4　实验结果与分析

4.1　实验数据及测评方法

实验数据集来自 Epinions 网站，实验数据集包括 trust 和 rating 表，trust 表记录每个用户信任的用户 ID，rating 表记录用户对商品的评价值，其中 1 表示不推荐，5 表示非常推荐. 有 49 289 个用户对 139 544 个不同商品的评价，评价总数达到 586 361 条. 在 Epinions 网站中的社区功能中提取用户的社交关系，将用户社交关系按照社交关系类型和紧密程度标注为社交关系拓扑结构图. 为了验证 STSUM 方法的效果，选取具有冷启动用户特征的数据. 数据集中有 73.8% 的用户至多评价过 3 个商品，17.93% 的用户至多评价过 1 个商品. 测试集中的大多数用户的评价个数在 3 以下，冷启动用户数量占 91.73%.

验证实验结果的方法：均方根误差 (root mean square error，RMSE) 是评估算法计算的预测值与真实值之间差距的指标，可用于衡量推荐效果[17]，RMSE 的定义如式(4)所示，用户均方根误差 (user root mean square error，URMSE) 的定义如式(5)所示：

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(r_{u,i}-\hat{r}_{u,i})^2}{N}}. \tag{4}$$

$$\mathrm{URMSE} = \sum\nolimits_{u_K\in U}\sqrt{\frac{\sum_{i=1}^{n_K}(r_{u,i}-\hat{r}_{u,i})^2}{n_K}}\Big/|U|. \tag{5}$$

式中：$\hat{r}_{u,i}$ 为由算法预测的用户 u 对商品 i 的评价值，N 为用户评价的商品数目，U 为用户集，$|U|$ 为用户个数，u_K 为第 K 个用户，n_K 为用户 u_K 的评价数目，其取值范围在 $[0,5]$.

分别采用 RMSE 和 URMSE 评价 STSUM 的推荐效果. RMSE 和 URMSE 值越小，则算法的推荐效果越好. 选择近期在冷启动推荐取得较好效果的方法做对比实验，Bobadilla 等[1]基于神经学习提出一种优化的相似度量方法，进而为冷启动用户给出推荐. Lika 等[2]提出基于分类算法的协同过滤方法发现相似用户，进而为冷启动用户给出推荐. Ling 等[4]提出通过矢量余弦法获取用户相似矩阵，并将用户分组，使用 top-N 方法为各组推荐. 冷启动推荐的初始参数需要一定量的数据训练，将整个数据集分成训练集和测试集，取训练集依次占整个数据集的 5%、10%、20%、50%、80%.

4.2　实验结果

4.2.1　STSUM 参数选择　在确定 STSUM 中的社交关系的个数 k 时，首先，将不同类型的社交关系按照其所覆盖的用户数量降序排列，覆盖用户数量多的社交关系排在前列. 其次确定 STSUM 中社交

关系的最优个数 k. 由于不同个数的社交关系所包含的用户信息数量不同,通常包含的社交关系个数越多,能够表征的用户信息越多. 通过实验给出推荐效果和社交关系个数 k 的对应关系. 实验中,分别取不同个数的社交关系,通过基于最大熵原理的数学模型给出每个社交关系的权重,然后得到社交关系 k 的个数对 STSUM 推荐效果的影响.

不同 k 取值对 STSUM 推荐效果如图 2 所示. 其中 RMSE 值越小表明预测值与真实值的差距越小,说明算法的推荐结果越准确. 实验中当社交关系个数 k 取 1 时,STSUM 的 RMSE 值最大,表明此时给出的推荐结果与真实情况偏离最大,这是因为过少的社交关系不能完整地描述用户特征. 逐渐增加 k 值,RMSE 值降低,当 k 值取 3 时,STSUM 的 RMSE 值最小,继续增加 k 值,STSUM 的 RMSE 值不断缓慢增加,没有下降的趋势,说明继续增加社交关系的个数不能提高推荐的准确度. 由于 RMSE 是对整个测试集中用户的推荐误差求平方加和,可以衡量测试集中用户整体的推荐效果. 为了能够有效地反映 k 值对每个用户的推荐效果的影响,通过 URMSE 比较不同 k 值下对用户的推荐效果,实验结果如图 3 所示. 其中 URMSE 值越小表明预测值与真实值的差距越小. 当 k 取 1 时,STSUM 的 URMSE 值最大,表明此时给出的推荐结果与真实情况偏离最大,URMSE 实验结果与 RMSE 实验结果一致. 逐渐增加 k,其 URMSE 值降低,当 k 取 3 时,STSUM 的 URMSE 值最小,继续增加 k 值,其 URMSE 值不断增加,说明继续增加社交关系的个数不能提高推荐的准确度. 当 k 取 3 时,STSUM 的 RMSE 和 URMSE 都取得最小,并且当 k 取值继续增大的情况下,RMSE 和 URMSE 的值都缓慢上升,可知最优社交关系数量为 3.

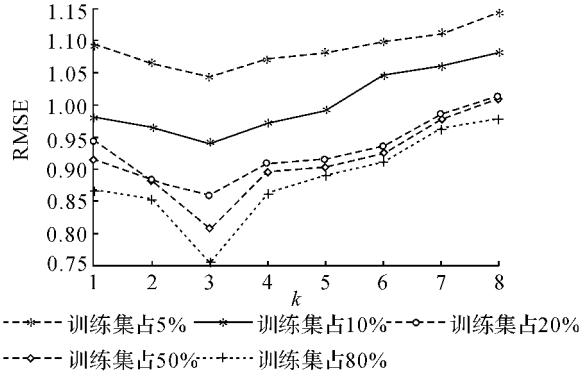

图 2 STSUM 的参数 k 与 RMSE 值关系

Fig. 2 Experimental results for different k of STSUM against RMSE

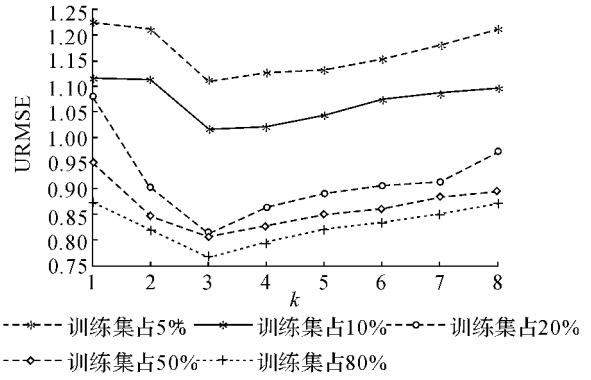

图 3 STSUM 的参数 k 与 URMSE 值关系

Fig. 3 Experimental results for different k of STSUM against URMSE

4.2.2 **对比试验结果** 为了验证基于最大熵原理给出的相异逻辑的社交关系权重值是合理的,采用对比实验验证效果. 在 k 取 3 的情况下,分别由式(1)~(3)计算社交关系权重,与各种社交关系赋予相同权重的情况做对比. 实验结果如图 4 所示,其中 t 为训练集占的百分比,采用 STSUM 得出的推荐结果在不同测试集的情况下都优于采用相同权重社交关系的推荐,为了能够有效地比较对每个用户 STSUM 和采用相同权重社交关系推荐的效果,通过 URMSE 比较推荐效果,实验结果如图 5 所示,比较图 4 与图 5 所示的结果得出 STSUM 的推荐结果在不同测试集的情况下都优于采用相同权重社交关系的推荐.

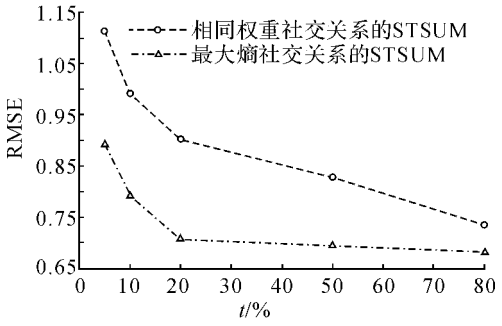

图 4 相同权重与最大熵社交关系 STSUM 的 RMSE 值比较

Fig. 4 Comparison of different social relationship weight assignment against RMSE

为了验证 STSUM 对冷启动用户的推荐效果,通过和 Bobadilla,Lika,Ling 的方法比较,以 RMSE 衡量推荐的效果,对比试验的结果如图 6 所示,按照 RMSE 值从大到小的排列上述算法,依次为 Bobadilla,Lika,Ling,STSUM. 当训练集少于 20% 时,STSUM 的 RMSE 值远小于其他方法,当训练集超过 20% 时,继续增加训练集百分比,STSUM 的

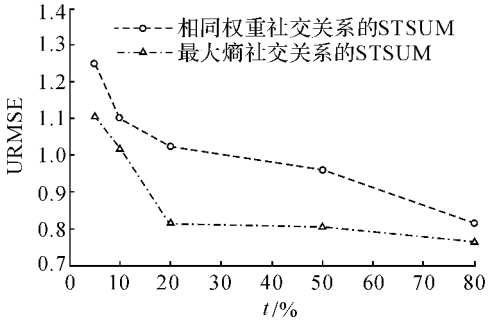

图 5　相同权重与最大熵社交关系 STSUM 的 URMSE 值比较

Fig. 5　Comparison of different social relationship weight assignment against URMSE

RMSE 值变化很小，说明 STSUM 在训练集占 20% 时就可以得到很好的推荐效果. 如图 7 所示，将上述算法按照 URMSE 值从大到小排列，依次为 Bobadilla，Lika，Ling，STSUM，与图 6 的结果一致. 当训练集小于 20%时，STSUM 的 URMSE 值小于其他算法，当训练集超过 20%时，继续增加训练集的百分比，STSUM 的 URMSE 值变化很小，说明 STSUM 在训练集占 20%时就可以得到很好的推荐效果. 在包含大量冷门商品的数据集的实验结果表明 STSUM 的推荐效果优于 Bobadilla，Lika，Ling 的方法，并且需要较少的训练集.

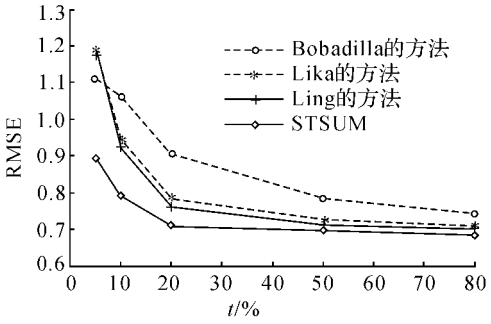

图 6　STSUM 算法与其他推荐算法的 RMSE 值比较

Fig. 6　Comparison with state-of-art methods against RMSE

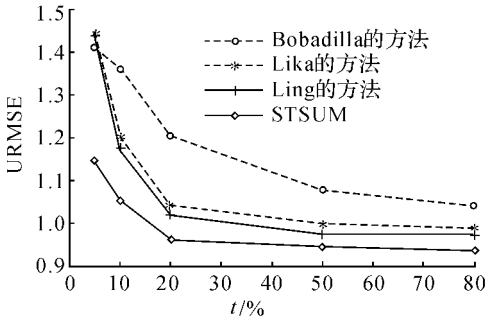

图 7　STSUM 算法与其他推荐算法的 URMSE 值比较

Fig. 7　Comparison with state-of －art methods against URMSE

5　结　语

在世界上电子商务系统中，存在部分仅有很少购买、评价等行为信息的冷启动用户. 由于缺乏冷启动用户购买、评价等信息，为冷启动用户推荐一直是推荐领域的难点之一. 为冷启动用户推荐喜欢的商品，则可以吸引更多的用户，提高用户对系统的黏滞度，因此冷启动推荐是众多网络应用的核心支撑技术之一. 社交网络是用户现实社交关系的映射，充分利用社交网络中多种逻辑的社交关系，融合相异逻辑的社交关系，可以在更广阔的范围内为冷启动用户发现相似用户，进而为冷启动用户给出推荐. 本文基于图形模式匹配、最大熵准则，充分利用用户间多种逻辑的社交关系，提出基于最大熵原理融合社交网络中多种相异逻辑的社交关系，最大限度的为冷启动用户发现相似用户，进而为冷启动用户给出合理的推荐结果. 在包含大量冷启动用户的实验结果表明，STSUM 在训练集较小的情况下，有效地提高了对冷启动用户的推荐效果.

参考文献(Reference)：

[1] BOBADILLA J S, ORTEGA F, HERNANDO A, et al. A collaborative filtering approach to mitigate the new user cold start problem [J]. **Knowledge-Based Systems**, 2012, 26(1): 225-238.

[2] LIKA B, KOLOMVATSOS K, HADJIEFTHYMIADES S. Facing the cold start problem in recommender systems [J]. **Expert Systems with Applications**, 2014, 41(4): 2065-2073.

[3] REN Y L, LI G, ZHOU W L. **PRICAI 2012: Trends in Artificial Intelligence**[M]. Berlin Heidelberg: Springer, 2012: 887-890.

[4] LING Y X, GUO D K, CAI F, et al. User-based Clustering with Top-N Recommendation on Cold-Start Problem[C]// **Proceedings of the 2013 3rd international conference on intelligent system design and engineering applications**. Hong Kong: IEEE Computer Society, 2013: 1585-1589.

[5] LOPS P, DE GEMMIS M, SEMERARO G. **Recommender systems handbook** [M]. Berlin Heidelberg: Springer, 2011: 73-105.

[6] YIN H, CUI B, CHEN L, et al. A temporal context-aware model for user behavior modeling in social media systems[C]// **Proceedings of the 2014 ACM SIGMOD international conference on Management of data**. Snowbird, USA: ACM, 2014: 1543-1554.

[7] WANG J, DE VRIES A P, REINDERS M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion[C]// **Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval.** Washington, USA: ACM, 2006: 501-508.

[8] JAMALI M, ESTER M. A matrix factorization technique with trust propagation for recommendation in social networks[C]// **Proceedings of the 4th ACM conference on Recommender systems.** Barcelona, Spain: ACM, 2010: 135-142.

[9] MA H, YANG H, LYU M R, et al. Sorec: social recommendation using probabilistic matrix factorization[C]// **Proceedings of the 17th ACM conference on Information and knowledge management.** Napa Valley, USA: ACM, 2008: 931-940.

[10] WU L, CHEN E H, LIU Q, et al. Leveraging tagging for neighborhood-aware probabilistic matrix factorization[C]// **Proceedings of the 21st ACM international conference on Information and knowledge management.** Maui Hawaii, USA: ACM, 2012: 1854-1858.

[11] KOREN Y. Collaborative filtering with temporal dynamics[J]. **Communications of the ACM**, 2010, 53(4): 89-97.

[12] REN L, GU J Z, XIA W W. An item-based collaborative filtering approach based on balanced rating prediction[C]// **Proceedings of 2011 International Conference on Multimedia Technology.** Hangzhou, China: IEEE, 2011: 3405-3408.

[13] MA H, KING I, LYU M R. Learning to recommend with social trust ensemble[C]// **Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.** Gold Coast, Australia: ACM, 2009: 203-210.

[14] KIM Y A, SONG H S. Strategies for predicting local trust based on trust propagation in social networks[J]. **Knowledge-Based Systems**, 2011, 24(8): 1360-1371.

[15] YUAN W W, GUAN D H, LEE Y K, et al. Improved trust-aware recommender system using small-worldness of trust networks[J]. **Knowledge-Based Systems**, 2010, 23(3): 232-238.

[16] JIANG W J, WANG G J, WU J. Generating trusted graphs for trust evaluation in online social networks[J]. **Future Generation Computer Systems**, 2014, 31(1): 48-58.

[17] LIU R R, LIU J G, JIA C X, et al. Personal recommendation via unequal resource allocation on bipartite networks[J]. **Physica A: Statistical Mechanics and its Applications**, 2010, 389(16): 3282-3289.

[18] GUHA R, KUMAR R, RAGHAVAN P, et al. Propagation of trust and distrust[C]// **Proceedings Of The 13th International Conference On World Wide Web.** New York, USA: ACM, 2004: 403-412.

[19] **印桂生, 张亚楠, 董红斌, 等. 一种由长尾分布约束的推荐方法**[J]. **计算机研究与发展**, 2013, 50(9): 1814-1824.
YIN Gui-sheng, ZHANG Ya-nan, DONG Hong-bin, et al. A long tail distribution constrained recommendation method [J]. **Journal of computer research and development**, 2013, 50(9): 1814-1824.

[20] **印桂生, 张亚楠, 董宇欣, 等. 基于受限信任关系和概率分解矩阵的推荐**[J]. **电子学报**, 2013, 42(5): 904-911.
YIN Gui-sheng, ZHANG Ya-nan, DONG Yu-xin, et al. A Constrained trust recommendation using probabilistic matrix factorization [J]. **Acta Electronica Sinica**, 2013, 42(5): 904-911.

[21] **孙光福, 吴乐, 刘淇, 等. 基于时序行为的协同过滤推荐算法**[J]. **软件学报**, 2013, 24(11): 2721-2733.
SUN Guang-fu, WU Le, LIU Qi, et al. Recommendations based on collaborative filtering by exploiting sequential behaviors [J]. **Journal Of Software**, 2013, 24(11): 2721-2733.