计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
50(9): 1814-1824, 2013

一种由长尾分布约束的推荐方法

印桂生　张亚楠　董红斌　董宇欣
(哈尔滨工程大学计算机科学与技术学院　哈尔滨　150001)
(yinguisheng@hrbeu.edu.cn)

A Long Tail Distribution Constrained Recommendation Method

Yin Guisheng, Zhang Yanan, Dong Hongbin, and Dong Yuxin
(*College of Computer Science and Technology, Harbin Engineering University, Harbin* 150001)

Abstract The sales of on-line shopping follow the rule of long tail distribution, therefore the total sales of unpopular goods are very large. Recommendations for unpopular goods are as important as recommendations for popular goods. However, many existing recommendation methods only focus on the recommendations for popular goods, and assign an average weight of recommendation to unpopular goods which have small number of ratings, thus it is hard to bring unpopular goods to user's attention and the sales of unpopular goods are depressed. So it is very important to improve the weight of recommendation for unpopular goods. In this paper, a long tail distribution constrained recommendation (LTDCR) method is proposed for improving the weight of recommendation for unpopular goods appropriately. The weight of recommendation in LTDCR is calculated using similarity relationship among users, where the similarity relationship is determined by the similarity of users' behaviors and is propagated under the constraint of distrust relationship. In order to improve the weight of recommendation for unpopular goods, the weight of recommendation is constrained by the long tail distribution. An accurate description of long tail distribution is also given in this paper. The experimental results in dataset containing large number of unpopular goods show that LTDCR need fewer training set to improve the effectiveness of recommendations for unpopular goods.

Key words long tail distribution; unpopular goods; weight of recommendation; similarity relationship; constraint of distrust relationship

摘　要 由于在线商品销售的长尾效应,冷门商品的总销量非常巨大,因而对冷门商品的推荐十分重要,然而由于对冷门商品的评价数量少,导致现存的推荐算法对其推荐权重接近平均推荐权重,所以很难使用户关注冷门商品,影响了冷门商品的销售,因此合理地提高冷门商品的推荐权重十分重要.提出一种由长尾分布约束的推荐方法(long tail distribution constrained recommendation method, LTDCR),由用户行为的相似度确定用户间相似关系,并应用不信任关系约束用户相似关系的传播,通过长尾分布约束由用户间相似关系计算的推荐权重,并给出一种精确描述长尾分布的方法.在包含大量冷门商品的数据集的实验结果表明,LTDCR在训练集较小的情况下,有效地提高了对冷门商品的推荐效果.

收稿日期:2013-03-30;**修回日期**:2013-06-13
基金项目:**国家自然科学基金项目**(61272186,61100007);**黑龙江省博士后基金项目**(LBH-Z12068);**哈尔滨工程大学自由探索基金项目**(HEUCF100608)
通信作者:**张亚楠**(ynzhang_1981@163.com)

关键词 长尾分布;冷门商品;推荐权重;相似关系;不信任关系约束

中图法分类号 TP391

在线购物网站的销售成本低,商品类型齐全,已经是商品销售的主要渠道之一.然而快速地从在线购物网站海量的商品中找到需要的商品却非常困难.推荐算法可以根据用户的偏好、历史浏览记录、社会关系等信息为用户推荐合适的商品.现有的推荐算法主要基于内容过滤(CBF)[1]或者协同过滤(CF)[2-9].CBF根据商品内容与用户配置文件的相关程度过滤商品,为用户给出推荐.CF可分为基于模型[2-4]和基于记忆[5-9].其中,基于模型的CF通过训练集为用户预定义一个包含对其影响最大的属性的模型,由该模型为用户筛选商品.基于记忆的CF由皮尔逊相关系数计算用户间相似度,由与当前用户相似的用户给出推荐.CF简单且有效已经被广泛应用于在线购物网站中,如Amazon[2].在线购物网站中称被评价次数超过门限值 δ 的商品为热门商品,被评价次数低于 δ 的商品为冷门商品.现有推荐算法关注对热门商品的推荐,却忽视对冷门商品的推荐,导致冷门商品很少被用户关注或购买,对用户在线购物行为的研究表明尽管单项的冷门商品的销售量非常低,但是所有冷门商品的销售量之和却非常巨大,超过某些热门商品的销售量,甚至达到总销售量的一半,研究者称这种现象为长尾效应[10-11].因此对冷门商品的推荐十分重要.

现有推荐算法通过从稀疏的用户商品矩阵中发现相似的用户[3,8-9],或者从用户的社会关系中发现并扩展用户可信任的用户范围.研究者对基于信任的推荐算法做了大量研究,并按照信任作用范围将其分为全局信任和局部信任[12-13].全局信任是由在整个互联网或者网站内所有用户对某用户信任评分的均值,是一种客观的信任.局部信任由可以直接交互的用户给对方的信任评分,是一种主观的信任.通过联合局部信任与全局信任可以得到较好的推荐效果.基于信任的推荐主要存在的问题是:信任评估过程中存在异常提高信任权重的现象,由此导致推荐出现偏差,针对这种现象孙玉星等人[14]提出了一种基于贝叶斯决策理论的推荐偏差度修正方法.基于信任的推荐的另一个问题是初始信任关系稀疏.针对这一问题,Ma等人[2]提出基于概率矩阵分解的社会推荐方法,Massa等人[15]提出通过信任传播,预测用户间潜在的信任关系.为了得到可靠的信任关系,康乐等人[16]提出通过社会化网络动态模型分析网络中的恶意信任扩张行为.用户间的不信任关系被证明可以用于约束用户间的信任关系[17],如Ma等人[18]提出一种基于不信任和信任关系混合的信任发现方法.

然而这些方法都忽视了对冷门商品的推荐,导致冷门商品很少被用户关注.针对这个问题,本文提出一种由长尾分布约束的推荐方法(long tail distribution constrained recommendation method, LTDCR),由用户行为发现用户间的相似关系,并通过不信任约束相似关系的扩展,最后由长尾分布约束商品的推荐,提高对冷门商品的推荐权重.

本文的贡献:1)提出一种基于用户行为发现用户间的相似关系的方法;2)提出基于不信任关系约束的相似关系传播;3)提出一种描述长尾分布的方法;4)提出一种提高对冷门商品的推荐权重的方法.

1 问题的定义

推荐算法的实质是预测以用户为行、商品为列的用户商品矩阵 $\boldsymbol{R}=[r_{u,i}]_{m\times s}$ 中的未知元素,即用户对未知商品的评价值,并对评价值排序,将Top-N评价值对应的商品推荐给用户.推荐算法对未知商品的推荐顺序与该未知商品的预测评价值正相关,称预测评价值为推荐权重.用户商品矩阵中 $r_{u,i}\in(0,5]$ 表示用户 u 对商品 i 的评价值,m 表示用户个数,s 表示商品个数.推荐可以分为两步,第1步计算用户间相似度:通过皮尔逊相关系数计算用户商品矩阵中代表用户对商品评价的行之间的相似度.第2步预测用户商品矩阵中的未知项,未知项的值可由

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

求得,其中 $p_{a,i}$ 表示预测的当前用户 a 对商品 i 的评价值,$p_{a,i}$ 的值越大,商品 i 的推荐权重越大,$\bar{r}_a$ 表示用户 a 的平均评价值,$\bar{r}_u$ 表示用户 u 的平均评价值,$w_{a,u}$ 表示由第1步计算出的用户 a 和 u 的相似度,$r_{u,i}$ 表示用户 u 对商品 i 的评价值.t 表示商品的评价数量,即评价过商品 i 的用户数量.

由式(1)知,对用户商品矩阵中未知项的预测

准确程度与用户相似度的准确程度正相关. 为用户 a 给出推荐的形式化描述如

$$Top(a,N):=\max_{i\in I}^{N} p_{a,i}, \tag{2}$$

其中 $I=\{i_1,i_2,\cdots,i_s\}$ 表示商品集合. 为了方便描述如何增加对冷门商品的推荐,首先给出长尾分布的定义.

定义 1. 令 $G(x)$ 表示任意一个分布的累积分布函数,其互补函数 $G^c(x)=1-G(x)$. 如果满足对任意 $\gamma>0$,当 $t\to\infty$,有 $e^{\gamma t}G^c(x)\to\infty$,则称 $G(x)$ 对应的分布为长尾分布.

由于对冷门商品的评价数量少,由式(1)预测的冷门商品的评价值接近用户的平均评价值. 为了增加对冷门商品的推荐,需要提高对冷门商品的推荐权重. 可通过构造与评价数量负相关的函数提高冷门商品的推荐权重,对在线购物网站的销售记录研究表明,商品的销量呈长尾分布[10-11],对商品的评价数量也呈长尾分布,将式(1)改写为

$$p_{a,i}=\bar{r}_a+\frac{\sum_{u=1}^{t} w_{a,u}(r_{u,i}-\bar{r}_u)}{\sum_{u=1}^{t} w_{a,u}}+\eta(t), \tag{3}$$

其中 $\eta(t)$ 与长尾分布的互补函数正相关. 为了得到合理的 $\eta(t)$,本文给出一种描述长尾分布的方法,并对该方法的效果做实例分析.

本文提出的 LTDCR 方法,由以下 3 部分组成:

1) 基于用户行为的相似关系发现;

2) 基于不信任关系约束的相似关系传播;

3) 长尾分布约束商品的推荐.

2 基于行为的相似关系发现

行为相似的用户其偏好也相似,给出的推荐更准确,本节通过用户浏览网页的相似度及对应的评价相似度计算用户间的相似度.

任意两用户的基于行为的相似度计算过程如下,设用户 a 浏览的网页记为 $page_j^a(1\leqslant j\leqslant p_1)$,用户 b 浏览的网页记为 $page_l^b(1\leqslant l\leqslant p_2)$,如图 1 所示,浅色节点代表用户浏览的网页,深色节点代表由该网页链出的网页.

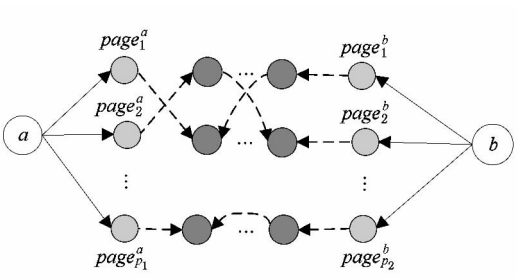

Fig. 1 An example of page similarity calculation.

图 1 网页相似度计算示意图

两个网页的相似度由其链出网页的相似度之和确定,通过 SimRank 算法[19]计算网页间的相似度,如

$$w(page_j,page_l)=\frac{c\sum_{o_1=1}^{|O(page_j)|}\sum_{o_2=1}^{|O(page_l)|} w(page_{o_1}^j,page_{o_2}^l)}{|O(page_j)||O(page_l)|}, \tag{4}$$

其中 $w(page_j,page_l)$ 代表网页 $page_j$ 与 $page_l$ 的相似度,$O(page_j)$ 代表由 $page_j$ 链出的网页集合,$|O(page_j)|$ 代表由 $page_j$ 链出的网页个数.

令 $page_{o_1}^j\in O(page_j)$, $page_{o_2}^l\in O(page_l)$, $w(page_{o_1}^j,page_{o_2}^l)$ 代表 $page_{o_1}^j$ 与 $page_{o_2}^l$ 的相似度,引入衰减因子 c 表示传播距离对网页间相似度的影响,c 取 0.8[19]. 用户 a,b 的行为的初始相似度 $w_0(a,b)$ 如

$$w_0(a,b)=\sum_{j=1}^{p_1}\sum_{l=1}^{p_2} w(page_j,page_l)\times w(rate_j^a,rate_l^b), \tag{5}$$

其中 $rate_j^a$ 表示用户 a 对网页 $page_j$ 的评价,$rate_l^b$ 表示用户 b 对网页 $page_l$ 的评价,令用户 a,b 对网页 $page_j$, $page_l$ 的评价相似度 $w(rate_j^a,rate_l^b)$ 为 $\min(rate_j^a,rate_l^b)/\max(rate_j^a,rate_l^b)$.

基于行为的相似度计算过程是递归的,记第 k 次结果为 $w_k(a,b)$,第 $k+1$ 次的计算结果如

$$w_{k+1}(a,b)=\sum_{u_1\in O(a)}\sum_{u_2\in O(b)} w_k(u_1,u_2), \tag{6}$$

其中 $O(a)$ 表示与用户 a 相似度大于阈值 ε 的用户集合,$O(b)$ 表示与用户 b 相似度大于阈值 ε 的用户集合. 5.2 节将通过实验的方法确定迭代次数 k 的值.

相似关系是以用户 ID 为行,与其相似度大于阈值 ε 的用户 ID 为列的矩阵. 令 $\boldsymbol{T}$ 表示用户行为相似关系矩阵. 提出相似耦合算法(similar coupling algorithm, SC)计算用户间相似关系,如算法 1 所示. 矩阵 $\boldsymbol{P}_{m\times p}$ 记录用户对其浏览网页的评价值,其中 m 表示用户数量,p 表示用户评价的网页数量. $\boldsymbol{Q}_{p\times p}$ 记录网页间的相似度. 算法第②至⑥步根据 $\boldsymbol{Q}_{p\times p}$ 将用户 u 所评价过的网页映射为网页 S_u. 第⑦至⑭步计算用户行为相似度,第⑮至⑯步对用户行为相似度迭代 k 次. SC 的空间复杂度为 $O(mp)$,时间复杂度为 $O(mp+kp^2)$.

算法 1. 相似耦合算法(SC).

输入:$\boldsymbol{P}_{m\times p}$,$\boldsymbol{Q}_{p\times p}$,衰减因子 c,迭代次数 k;

输出:$\boldsymbol{T}$.

① $\boldsymbol{T}=\boldsymbol{S}=\varnothing$,$tmp=0$.
② FOR($i=1$ to m)
③ FOR($j=1$ to p)
④ $S_i=P_{i,j}\times Q_{j,i}+S_i$;
⑤ ENDFOR
⑥ ENDFOR
⑦ FOR($i=1$ to m)
⑧ FOR($j=1$ to m)
⑨ 比较网页评价值 $tmp=S_i/S_j$;
⑩ IF $tmp>\varepsilon$
⑪ 标记用户相似关系 $T_{i,j}=tmp$;
⑫ ENDIF
⑬ ENDFOR
⑭ ENDFOR
⑮ $\Delta\boldsymbol{T}^k=c^k\times(\boldsymbol{T}\times\boldsymbol{T}^{\mathrm{T}})^k$;
⑯ $\boldsymbol{T}=\Delta\boldsymbol{T}^k+\boldsymbol{T}$.

3 基于不信任关系约束的相似关系传播

用户相似关系矩阵通常比较稀疏,对任意一行,以其非零元素列号为起始点的广度优先搜索,可以传播用户间相似关系,然而无约束的相似关系传播会导致相似关系的过度扩大,导致推荐结果不准确.本节提出由不信任关系约束相似关系传播.

设 $\boldsymbol{D}$ 代表不信任关系矩阵,其行向量 $\boldsymbol{D}_u$ 表示用户 u 不信任的用户集合,$\boldsymbol{D}$ 传播 r 步后记为 $\boldsymbol{D}^{(r)}$,相似关系矩阵 $\boldsymbol{T}$ 传播 q 步后记为 $\boldsymbol{T}^{(q)}$,$\boldsymbol{B}$ 代表最终的相似关系矩阵.基于不信任约束的相似关系传播包括如下 3 种方法:

1) 不信任关系直接约减相似关系,这种方法不考虑不信任关系的传播,从传播 q 步后的相似关系中直接去除不信任关系,最终的相似关系矩阵 $\boldsymbol{B}=\boldsymbol{T}^{(q)}-\boldsymbol{D}$.

2) 不信任关系约减相似关系后同步传播,这种方法从最初的相似关系中去除不信任关系后再作传播,设传播步数为 q,则最终的相似关系矩阵 $\boldsymbol{B}=(\boldsymbol{T}-\boldsymbol{D})^{(q)}$.

3) 不信任关系与相似关系异步传播后约减,当 $\boldsymbol{T}^{(q)}$ 满足与 $\boldsymbol{D}^{(r)}$ 没有交集,而 $\boldsymbol{T}^{(q+1)}$ 与 $\boldsymbol{D}^{(r)}$ 有交集,则最终的相似关系矩阵 $\boldsymbol{B}=\boldsymbol{T}^{(q)}$.

不信任关系 $\boldsymbol{D}$ 的传播如图 2 所示,如果两个用户 a,b 间存在不信任关系,而用户 b 信任用户 u,则可知用户 a 与 u 间存在不信任关系.$\boldsymbol{D}$ 传播 r 步产生的不信任关系矩阵 $\boldsymbol{D}^{(r)}=\boldsymbol{D}\cdot\boldsymbol{X}^{(r)}$,其中 $\boldsymbol{X}$ 为用户间信任关系矩阵,$\boldsymbol{X}$ 的行向量 $\boldsymbol{X}_u$ 表示用户 u 信任的用户集合.5.2 节中将通过实验确定不信任关系传播步数 r 和相似关系传播步数 q 的最优值.

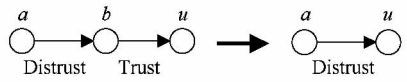

Fig. 2 An example of distrust propagation.

图 2 不信任传播示意图

方法 1 的约束效果有限,不能很好地防止相似关系过度扩大.方法 2 的约束条件过强,导致相似关系覆盖范围过小.本文采用方法 3 传播相似关系,提出一种基于列迭代的相似度传播算法(column-iteration similarity propagation, CISP),如算法 2 所示,其中相似关系矩阵 $\boldsymbol{T}_{m\times m}$,$m$ 为用户数,$\boldsymbol{T}_i$ 为 $\boldsymbol{T}_{m\times m}$ 的第 i 个列向量,$\boldsymbol{T}_i$ 传播 l 步后记为 $\boldsymbol{T}_i^{(l)}$,定义临时相似关系矩阵 $\boldsymbol{L}_{m\times m}$,$\boldsymbol{L}_j$ 为 L 的第 j 个列向量.

算法 2. 列迭代相似度传播算法(CISP).

输入:$\boldsymbol{T}$,传播步数 q;

输出:$\boldsymbol{T}^{(q)}$.

① $l=v=q$;
② WHILE($q\neq0$)
③ $l=l-q+1$;
④ FOR($i=1$ to m)
⑤ FOR($j=1$ to m)
⑥ IF($T_{i,j}\neq0$)
⑦ i 加入 $\boldsymbol{L}_j$ 中,j 加入 $\boldsymbol{T}_i^{(l)}$;
⑧ ENDIF
⑨ ENDFOR
⑩ ENDFOR
⑪ FOR($i=1$ to m)
⑫ FOR($j=1$ to $|\boldsymbol{T}_i^{(l)}|$)
⑬ IF($i\in\boldsymbol{T}_j^{(l)}$)
⑭ i 加入 $\boldsymbol{L}_j$;
⑮ ENDIF
⑯ ENDFOR
⑰ ENDFOR
⑱ $q=q-1$;
⑲ ENDWHILE
⑳ FOR($i=1$ to m)
㉑ FOR($j=1$ to m)
㉒ IF($i\in\boldsymbol{L}_j$)
㉓ j 加入 $\boldsymbol{T}_i^{(v)}$;
㉔ ENDIF

㉕ ENDFOR

㉖ ENDFOR

CISP 第④至⑩步计算用户初始 L,如果 $i\in L_k$,则把 k 加入 T_i,按列遍历矩阵 L,令初始相似关系矩阵为 $T^{(0)}$,第⑪至⑰步以列的顺序传播相似关系,第⑳至㉖步以行的顺序遍历 L,得传播 q 步后的相似关系矩阵 $T^{(q)}$. CISP 算法的时间复杂度优于广度优先传播算法(breadth-first propagation, BFP), BFP 逐一地为每个用户寻找相似用户列表,CISP 为所有用户同时寻找相似用户列表. 相似关系矩阵的维数为 m,最大传播步数 q,BFP 的空间复杂度为 $O(m^2)$,时间复杂度为 $O(qm^3)$,CISP 的空间复杂度为 $O(m^2)$,时间复杂度为 $O(qm^2)$.

4 基于长尾分布约束的推荐

本节给出式(3)中函数 $\eta(t)$ 的描述. 由于 $\eta(t)$ 与长尾分布的互补函数正相关,首先提出一种由超指数函数(hyper-exponential function, HEF)描述长尾分布的方法,并验证 HEF 描述长尾分布的效果,最后给出 $\eta(t)$ 的表达式和基于长尾分布约束的推荐过程.

4.1 长尾分布的描述

根据学者 Feldmann 等人[20]的研究表明长尾分布不适合用单一的函数描述其特征,但可以通过多个函数的叠加描述. 提出一种通过 HEF 描述长尾分布的方法.

为满足定义 1 所述的长尾分布条件,提出由 n 个底为 e 的指数函数线性组合(HEF)描述长尾分布函数,如

$$P(x)=\sum_{j=1}^{n}p_j\mathrm{e}^{-\lambda_j x}+C. \tag{7}$$

符合长尾分布的样本点数据记为 (x_i,y_i) $(i=1,\cdots,h)$,$F(x)$ 表示 HEF 与数据 (x_i,y_i) 的平方差,如:

$$F(x)=\sum_{i=1}^{h}\omega_i\Big(\sum_{j=1}^{n}p_j\mathrm{e}^{-\lambda_j x_i}+C-y_i\Big)^2, \tag{8}$$

其中 $\omega_i>0$ 为点的权系数,对长尾分布的描述可通过求使 F 最小的参数 $p_j,\lambda_j>0(j=1,\cdots,n)$ 得到.

由多元函数求极值的必要条件 $\partial F(x)/\partial\lambda_j=0$, $\partial F(x)/\partial p_j=0$,得:

$$\frac{\partial F(x)}{\partial p_j}=2\sum_{i=1}^{h}\omega_i\Big(\sum_{j=1}^{n}p_j\mathrm{e}^{-\lambda_j x_i}+C-y_i\Big)\mathrm{e}^{-\lambda_j x_i}=0, \tag{9}$$

$$\frac{\partial F(x)}{\partial\lambda_j}=2\sum_{i=1}^{h}\omega_i\Big(\sum_{j=1}^{n}p_j\mathrm{e}^{-\lambda_j x_i}+C-y_i\Big)(-x_i)p_j\mathrm{e}^{-\lambda_j x_i}=0. \tag{10}$$

令 $\varphi_j(x)=\mathrm{e}^{-\lambda_j x}(j=1,\cdots,n)$,$\varphi_l(x)=\mathrm{e}^{-\lambda_l x}(l=1,\cdots,n)$,$(\varphi_j,\varphi_l)=\sum_{i=1}^{h}\omega_i\varphi_j(x_i)\varphi_l(x_i)$,$y(x_i)=y_i-C$,$(y,\varphi_l)=\sum_{i=1}^{h}\omega_i y(x_i)\varphi_l(x_i)$,且 φ_j 与 φ_l 线性无关,则式(9)变为

$$\sum_{j=1}^{n}(\varphi_j,\varphi_l)p_j=(y,\varphi_l), \tag{11}$$

其系数矩阵如下:

$$\boldsymbol{\Phi}=\begin{bmatrix}(\varphi_1,\varphi_1) & (\varphi_1,\varphi_2) & \cdots & (\varphi_1,\varphi_n)\\ \vdots & \vdots & & \vdots\\ (\varphi_n,\varphi_1) & (\varphi_n,\varphi_2) & \cdots & (\varphi_n,\varphi_n)\end{bmatrix}. \tag{12}$$

由于存在 $n<h$,可解 n 阶线性方程组得到 $P(x)$ 的系数 p_j. 令 $\psi_j(x)=p_j^2x\mathrm{e}^{-\lambda_j(x+1)}(j=1,\cdots,n)$,$\psi_l(x)=p_l^2x\mathrm{e}^{-\lambda_l(x+1)}$ $(l=1,\cdots,n)$,$(\psi_j,\psi_l)=\sum_{i=1}^{h}\omega_i\psi_j(x_i)\psi_l(x_i)$,$(y,\psi_l)=\sum_{i=1}^{h}\omega_i y(x_i)\psi_l(x_i)$,且 ψ_j 与 ψ_l 线性无关,则式(10)变为

$$\sum_{j=1}^{n}(\psi_j,\psi_l)\mathrm{e}^{\lambda_j}=(y,\psi_j), \tag{13}$$

其系数矩阵如下:

$$\boldsymbol{\Psi}=\begin{bmatrix}(\psi_1,\psi_1) & (\psi_1,\psi_2) & \cdots & (\psi_1,\psi_n)\\ \vdots & \vdots & & \vdots\\ (\psi_n,\psi_1) & (\psi_n,\psi_2) & \cdots & (\psi_n,\psi_n)\end{bmatrix}, \tag{14}$$

解 n 阶非线性方程组可得到 $P(x)$ 的系数 λ_j. 4.2 节给出用 HEF 描述长尾分布效果的实例分析.

4.2 HEF 描述长尾分布的效果分析

本节对 HEF 描述长尾分布的效果做实例分析. 韦伯分布(Weibull distribution)的累积分布函数如:

$$f(x)=1-\mathrm{e}^{-\alpha x^{\beta}}, \tag{15}$$

其互补累积分布函数 CCDF 如:

$$f^{c}(x)=\mathrm{e}^{-\alpha x^{\beta}}, \tag{16}$$

其中,α 是尺度参数,β 是形状参数. 当 $\beta<1$ 时,韦伯分布的 CCDF 具有明显的长尾分布特征,取韦伯分布参数 $\alpha=0.10799$,$\beta=0.3$. 通过分析 HEF 拟合韦伯分布的效果,可验证 HEF 描述长尾分布的效果.

HEF 由 n 个指数函数组成,参数 n 与 HEF 拟合韦伯分布的效果关系如表 1 所示. 表 1 第 1 列表示 HEF 的项数 n,第 2~4 列是评测拟合效果的 3 个指标,分别为误差项平方和(sum of squares for error, SSE),决定系数 R-Square,均方根误差(root-

mean-square error, RMSE). SSE 测量拟合值与实际值总体的偏移程度,SSE 越接近 0 表示拟合效果越好. R-Square 表示系数对被拟合方程的解释适合度,R-Square 越接近 1,表示对被拟合函数的解释能力越好. RMSE 衡量观测值同真实值之间的偏差. 如表 1 中所示,当 HEF 的 n 值不断增加,其 SSE 和 RMSE 值先降低后升高,其 R-Square 值先升高后降低,当 $n=2$ 时,HEF 的 SSE 和 RMSE 值取最小,R-Square 值最大,说明当 $n=2$ 时 HEF 对韦伯分布的拟合效果最好.

Table 1 Effectiveness of Parameter n of HEF in Fitting Weibull

表 1 HEF 的项数与拟合韦伯分布效果的关系

n	SSE	R-Square	RMSE
1	0.106	0.887	0.046 91
2	0.001 88	0.998	0.006 39
3	0.071 4	0.924	0.040 3
4	0.035 7	0.962	0.029 1
5	0.038 8	0.959	0.031 2

为了证明 HEF 最适合描述长尾分布,选取高斯函数、傅里叶函数、多项式函数做对比实验. 由于拟合函数的项数影响拟合的效果,为了得到高斯函数、傅里叶函数、多项式函数拟合参数为 $\alpha=0.107\,99$, $\beta=0.3$ 的韦伯分布的最好效果,逐渐增加拟合函数的项数直到其取得最优拟合效果. 当高斯函数的项数为 2 时,其拟合韦伯分布 CCDF 的最好效果如图 3 所示. 当傅里叶函数的项数为 2 时,其拟合韦伯分布 CCDF 的最好效果如图 4 所示. 当 HEF 的系数 $p_1=0.467\,1$, $p_2=0.346\,8$, $\lambda_1=1.546$, $\lambda_2=0.063\,98$, $C=0.184\,9$ 时,其拟合韦伯分布 CCDF 的最好效果如图 5 所示. 当多项式函数的项数为 4 时,其拟合韦伯分布 CCDF 的最好效果如图 6 所示. 可知图 5 所示的 HEF 对韦伯分布的拟合效果最好,说明 HEF 可以有效地描述长尾分布.

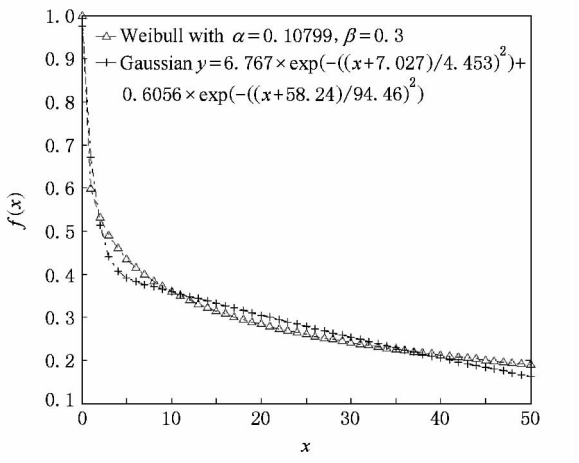

Fig. 3 Fitting CCDF of Weibull by gaussian function.

图 3 高斯函数拟合韦伯分布的 CCDF

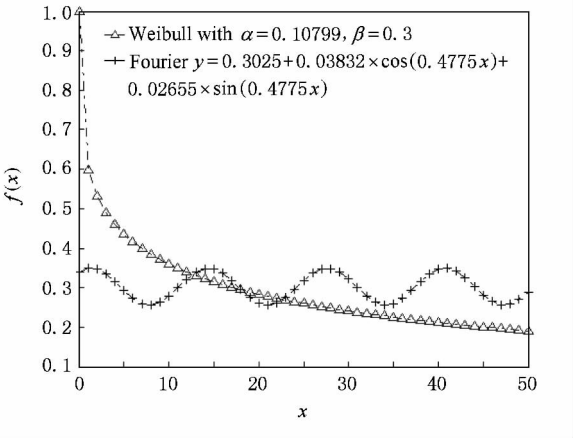

Fig. 4 Fitting CCDF of Weibull by fourier function.

图 4 傅里叶函数拟合韦伯分布的 CCDF

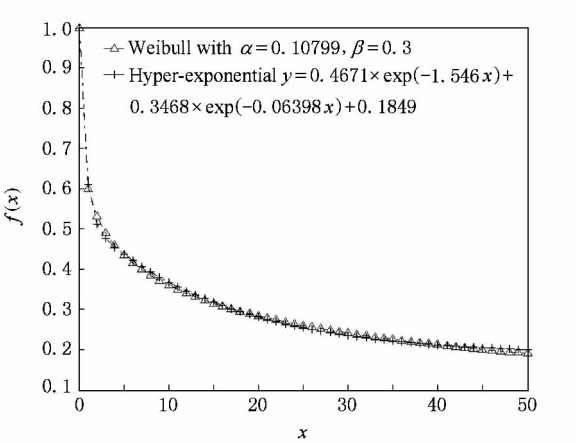

Fig. 5 Fitting CCDF of Weibull by HEF.

图 5 HEF 拟合韦伯分布的 CCDF

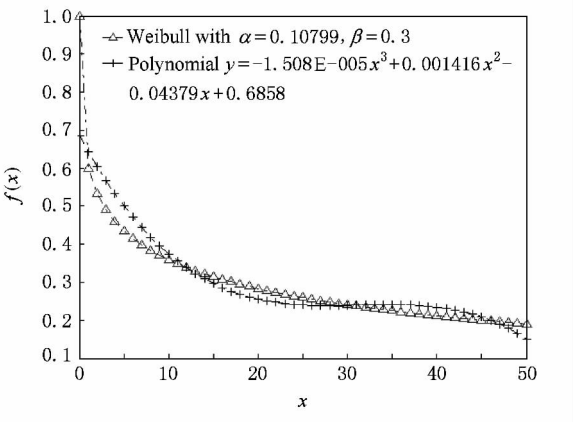

Fig. 6 Fitting CCDF of Weibull by polynomial function.

图 6 多项式函数拟合韦伯分布的 CCDF

4.3 基于长尾分布约束的推荐过程

4.2 节验证了 HEF 描述长尾分布的有效性，并且得出当其项数 $n=2$ 时，对长尾分布的描述效果最优. 首先令函数 $g(x)=1/(1+e^{-x})$ 将长尾分布的函数值 $P(t)=p_1e^{-\lambda_1 t}+p_2e^{-\lambda_2 t}+C$ 映射在[0,1]之间，其中参数 p_1，p_2 由式(11)确定，λ_1，λ_2 由式(13)确定，C 为常数，由于商品被评价的次数呈长尾分布，为了增加对冷门商品的推荐，令式(3)中 $\eta(t)$ 为：

$$\eta(t)=1-\frac{1}{1+e^{-P(t)}}, \tag{17}$$

则式(3)变为

$$p_{a,i}=\bar{r}_a+\frac{\sum_{u=1}^{t}w_{a,u}(r_{u,i}-\bar{r}_u)}{\sum_{u=1}^{t}w_{a,u}}+1-\frac{1}{1+e^{-P(t)}}. \tag{18}$$

基于长尾分布约束的推荐 LTDCR 过程如算法 3 所示. 其中计算用户行为的相似关系、相似关系传播和计算用户 a 的 $Top(a,N)$ 的空间复杂度分别为 $O(mp)$，$O(m^2)$，$O(t)$，时间复杂度分别为 $O(mp+kp^2)$，$O(qm^2)$，$O(t+m\,\mathrm{lb}\,m)$，则 LTDCR 的空间复杂度为 $O(mp+m^2+t)$，时间复杂度为 $O(mp+kp^2+qm^2+t+m\,\mathrm{lb}\,m)$.

算法 3. 长尾分布约束的推荐方法(LTDCR).

输入：$\boldsymbol{Q}$，$\boldsymbol{P}$，k，$\boldsymbol{T}$，$\boldsymbol{D}$，$\boldsymbol{X}$，q；

输出：用户 a 的 $Top(a,N)$ 对应的商品.

1) 基于用户行为相似度确定推荐关系.

① 计算用户浏览的网页相似度；

② 由用户浏览网页相似度与对应评价相似度得到用户行为相似关系；

2) 基于不信任关系约束的相似关系传播.

① 不信任约束传播 r 步产生 $\boldsymbol{D}^{(r)}$；

② 相似关系传播 q 步产生 $\boldsymbol{T}^{(q)}$；

③ 当 $\boldsymbol{D}^{(r)}\cap\boldsymbol{T}^{(q)}=\varnothing$，$\boldsymbol{D}^{(r)}\cap\boldsymbol{T}^{(q+1)}\neq\varnothing$，得相似关系为 $\boldsymbol{T}^{(q)}$.

3) 由式(18)预测用户对商品的评价值.

4) 给出用户 a 的 $Top(a,N)$ 评价值对应的商品.

5 实验结果与分析

5.1 实验数据及测评方法

实验数据集来自 Epinions 网站. Epinions 是创立于 1999 年的一个著名的社区交流网站. 用户在这个网站分享他们对商品的评价，用户评价包含 5 个等级，1 表示不推荐，5 表示非常推荐. 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. 数据集中只有 23.8% 的商品被至少 1 个用户评价过，7.93% 的商品被至少 3 个用户评价过. 冷门商品占 68.27%.

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[3]，RMSE 的定义如

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_{u,i}-\hat{r}_{u,i})^2}, \tag{19}$$

其中，$r_{u,i}$ 表示用户 u 对商品 i 的真实评价值，$\hat{r}_{u,i}$ 表示由算法预测的用户 u 对商品 i 的评价值，N 表示用户评价的商品数目. 为了能有效地反映对每个用户的推荐效果，引入用户均方根误差(user root mean square error, URMSE)，URMSE 的定义为

$$URMSE=\frac{1}{|U|}\sum_{u_K\in U}\sqrt{\frac{1}{n_K}\sum_{i=1}^{n_K}(r_{u,i}-\hat{r}_{u,i})^2}, \tag{20}$$

其中，U 表示用户集，$|U|$ 表示用户个数，u_K 表示第 K 个用户，n_K 表示用户 u_K 的评价数目. 采用 RMSE 和 URMSE 评价 LTDCR 的推荐效果. RMSE 和 URMSE 值越小，则算法的推荐效果越好.

选择如下方法做对比试验：概率矩阵分解(probabilistic matrix factorization, PMF)方法[3]；基于概率矩阵分解的社会推荐(social recommendation using probabilistic matrix factorization, SoRec)方法[2]；基于信任感知的推荐方法，其中 MT1，MT2，MT3 分别表示信任传播距离为 1，2，3 步的情况[15]；基于不信任关系和信任关系的推荐方法[18]，包括基于不信任的推荐(recommendation with distrust, RWD)和基于信任的推荐(recommendation with trust, RWT). 整个数据集被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%.

5.2 实验结果

5.2.1 LTDCR 参数选择

首先确定 LTDCR 中的参数，包括基于行为的相似关系发现中的迭代次数 k 和相似关系传播步数 q. 迭代次数 k 对 LTDCR 推荐效果的影响如图 7 所示. 当 $k=1$ 时，LTDCR 的 RMSE 值最大，逐渐增加 k 值，RMSE 值降低，当 $k=3$ 时，LTDCR 的 RMSE

值最小,继续增加 k 值,LTDCR 的 RMSE 值不断增加. k 值对 LTDCR 的 URMSE 值影响如图 8 所示. 当 $k=1$ 时,LTDCR 的 URMSE 值最大,逐渐增加 k,其 URMSE 值降低,当 $k=3$ 时,LTDCR 的 URMSE 值最小,继续增加 k 值,其 URMSE 值不断增加. $k=3$ 时,LTDCR 的 RMSE 和 URMSE 最小,可知最优迭代次数为 3.

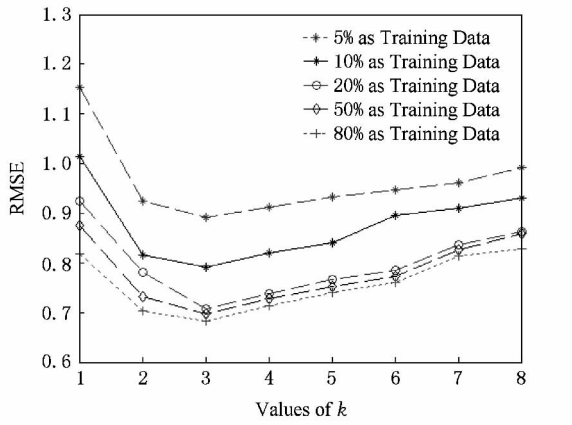

Fig. 7 Results for different k in LTDCR against RMSE.

图 7 LTDCR 的参数 k 与 RMSE 值关系

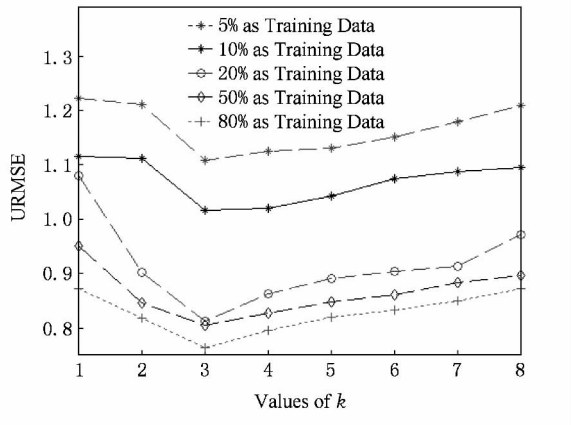

Fig. 8 Results for different k in LTDCR against URMSE.

图 8 LTDCR 的参数 k 与 URMSE 值关系

不信任关系矩阵由 rating 表中对同一商品的评价值相差超过 2 的用户构成. 当不信任关系矩阵传播 2 步时,不信任关系矩阵与初始的相似关系矩阵交集不为空,为了扩大相似关系矩阵覆盖范围,不信任关系矩阵的传播步数 $r=1$. 相似关系矩阵传播 q 步时,LTDCR 的 RMSE 值如图 9 所示,当 $q=2$ 时,RMSE 值最小,继续增加 q 值,RMSE 值不断增加. 不同 q 值情况下,LTDCR 的 URMSE 值如图 10 所示,当 $q=2$ 时对应的 URMSE 值最小,继续增加 q,URMSE 值不断增加,可知 q 的最优值为 2.

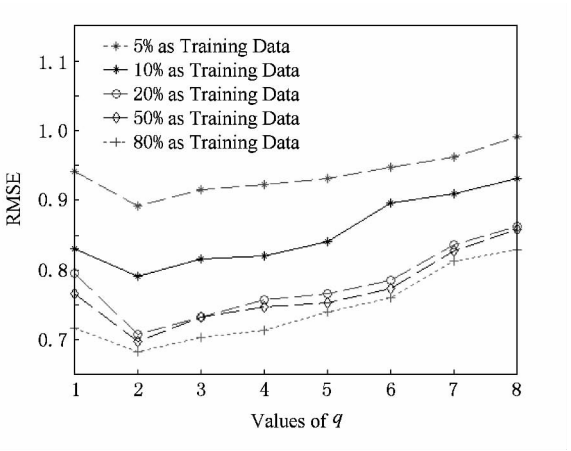

Fig. 9 Results for different values of q in LTDCR against RMSE.

图 9 LTDCR 的参数 q 与 RMSE 值关系

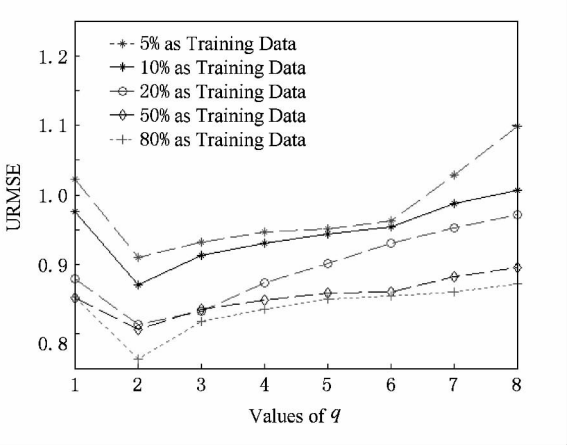

Fig. 10 Results for different values of q in LTDCR against URMSE.

图 10 LTDCR 的参数 q 与 URMSE 值关系

5.2.2 对比试验结果

为了证明长尾分布约束的效果,与无长尾分布约束的推荐(non-long tail distribution constrained recommendation, Non-LTDCR)作比较. 如图 11 所示,训练集取不同百分比的情况下,LTDCR 的 RMSE 一直小于 Non-LTDCR 的 RMSE. LTDCR 与 Non-LTDCR 的 URMSE 比较结果如图 12 所示,LTDCR 的 URMSE 一直小于 Non-LTDCR 的 URMSE,说明长尾分布约束可提高推荐效果.

LTDCR 和 PMF, MT1, MT2, MT3, SoRec, RWD, RWT 的比较结果分别如图 13,14 所示. 如图 13 所示,按照 RMSE 值从大到小的排列上述算法,依次为 PMF, SoRec, MT3, MT1, MT2, RWD, RWT, LTDCR. 当训练集少于 20% 时,LTDCR 的 RMSE 值远小于其他方法,当训练集超过 20% 时,

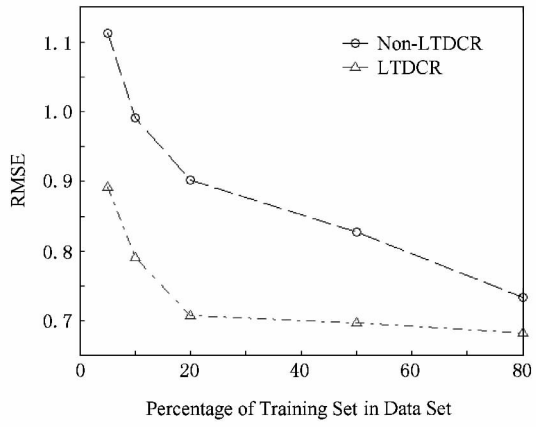

Fig. 11 Comparison of Non-LTDCR and LTDCR against RMSE.

图 11 Non-LTDCR 与 LTDCR 的 RMSE 值比较

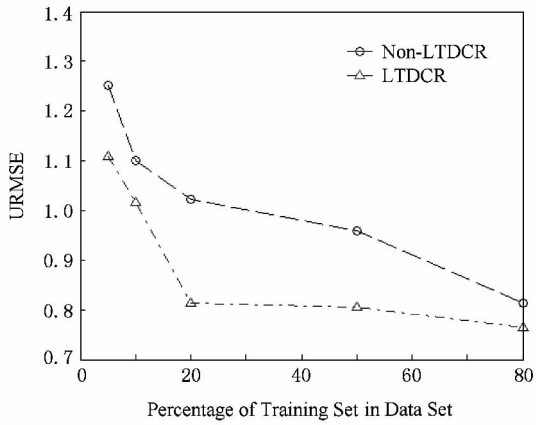

Fig. 12 Comparison of Non-LTDCR and LTDCR against URMSE.

图 12 Non-LTDCR 与 LTDCR 的 URMSE 值比较

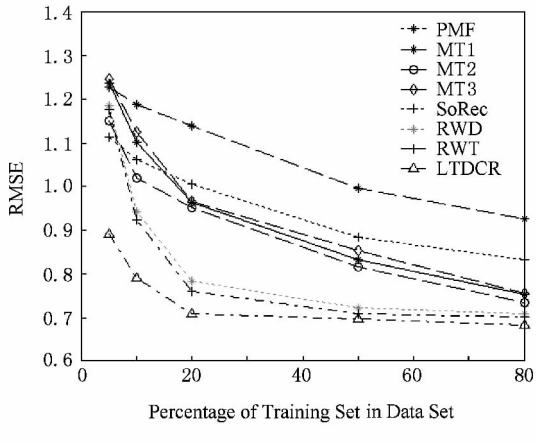

Fig. 13 Comparison of PMF, MT1, MT2, MT3, SoRec, RWD, RWT and LTDCR against RMSE.

图 13 PMF, MT1, MT2, MT3, SoRec, RWD, RWT, LTDCR 算法的 RMSE 值比较

继续增加训练集百分比，LTDCR 的 RMSE 值变化很小，说明 LTDCR 在训练集占 20% 时就可以得到很好的推荐效果. 如图 14 所示，将上述算法按照 URMSE 值从大到小排列，依次为 PMF, SoRec, MT3, MT1, MT2, RWD, RWT, LTDCR，与图 13 的结果一致. 当训练集小于 20% 时，LTDCR 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，LTDCR 的 URMSE 值变化很小，说明 LTDCR 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 LTDCR 的推荐效果优于 PMF, MT1, MT2, MT3, SoRec, RWD, RWT，并且需要较少的训练集.

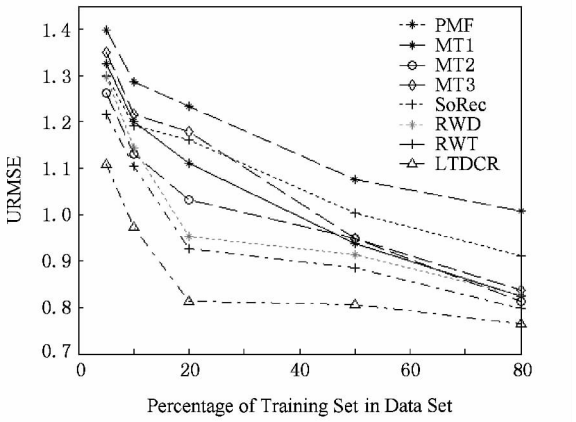

Fig. 14 Comparison of PMF, MT1, MT2, MT3, SoRec, RWD, RWT and LTDCR against URMSE.

图 14 PMF, MT1, MT2, MT3, SoRec, RWD, RWT, LTDCR 算法的 URMSE 值比较

6 结 论

本文针对现存的推荐方法给冷门商品的推荐权重过低，导致其很难被用户关注. 提出一种由长尾分布约束的推荐方法 LTDCR，在商品的预测评价值中加入评价数量因素，提高了对冷门商品的推荐权重，解决了由于冷门商品的评价数量少而很难被用户关注的问题. 通过不信任关系约束由用户行为相似度计算的相似关系的传播，既扩大了相似关系的覆盖范围又保证了其准确度. 给出一种准确描述长尾分布的方法，并由实例分析其效果. 在包含大量冷门商品的数据集的实验表明，LTDCR 可以有效地提高对冷门商品的推荐效果，并且需要较少的训练集.

参 考 文 献

[1] Balabanović M, Shoham Y. Fab: Content-based, collaborative recommendation [J]. Communications of the ACM, 1997, 40(3): 66-72

[2] Ma H, Yang H, Lyu M R, et al. Sorec: Social recommendation using probabilistic matrix factorization [C] //Proc of the 17th ACM Conf on Information and Knowledge Management. New York: ACM, 2008: 931-940

[3] Salakhutdinov R, Mnih A. Probabilistic matrix factorization [J]. Advances in Neural Information Processing Systems, 2008, 20(1): 1257-1264

[4] Jamali M, Ester M. A matrix factorization technique with trust propagation for recommendation in social networks [C] //Proc of the 4th ACM Conf on Recommender Systems. New York: ACM, 2010: 135-142

[5] Bobadilla J, Ortega F, Hernando A, et al. A collaborative filtering approach to mitigate the new user cold start problem [J]. Knowledge-Based Systems, 2012, 26(10): 225-238

[6] Zhu Rui, Wang Huaimin, Feng Dawei. Trustworthy services selection based on preference recommendation [J]. Journal of Software, 2011, 22(5): 852-864 (in Chinese)
(朱锐, 王怀民, 冯大为. 基于偏好推荐的可信服务选择[J]. 软件学报, 2011, 22(5): 852-864)

[7] Luo Xin, OuYang Yuanxin, Xiong Zhang, et al. The effect of similarity support in *K*-Nearest-Neighborhood based collaborative filtering algorithm [J]. Chinese Journal of Computers, 2010, 33(8): 1437-1445 (in Chinese)
(罗辛, 欧阳元新, 熊璋, 等. 通过相似度支持度优化基于 *K* 近邻的 CF [J]. 计算机学报, 2010, 33(8): 1437-1445)

[8] Ma H, King I, Lyu M R. Effective missing data prediction for collaborative filtering [C] //Proc of the 30th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2007: 39-46

[9] Wang J, de Vries A P, Reinders M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion [C] //Proc of the 29th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2006: 501-508

[10] Brynjolfsson E, Hu Jeffrey Y, Smith M. From niches to riches anatomy of the long tail [J]. Sloan Management Review, 2006, 47(4): 67-71

[11] Park YJ, Tuzhilin A. The long tail of recommender systems and how to leverage it [C] //Proc of the 2008 ACM Conf on Recommender Systems. New York: ACM, 2008: 11-18

[12] Wang Shouxin, Zhang Li. Evaluation approach of subjective trust based on cloud model [J]. Journal of Software, 2010, 21(6): 1341-1352 (in Chinese)
(王守信, 张莉. 一种基于云模型的主观信任评价方法[J]. 软件学报, 2010, 21(6): 1341-1352)

[13] He Lijian, Huang Houkuan, Zhang Wei. A survey of trust and reputation systems in multi-agent systems [J]. Journal of Computer Research and Development, 2008, 45(7): 1151-1160 (in Chinese)
(贺利坚, 黄厚宽, 张伟. 多 Agent 系统中信任和信誉系统研究综述[J]. 计算机研究与发展, 2008, 45(7): 1151-1160)

[14] Sun Yuxing, Huang Songhua. Bayesian decision-making based recommendation trust revision model in ad hoc networks [J]. Journal of Software, 2009, 20(9): 2574-2586 (in Chinese)
(孙玉星, 黄松华. 基于贝叶斯决策的自组网推荐信任度修正模型[J]. 软件学报, 2009, 20(9): 2574-2586)

[15] Massa P, Avesani P. Trust-aware recommender systems [C] //Proc of the 2007 ACM Conf on Recommender Systems. New York: ACM, 2007: 17-24

[16] Kang Le, Jing Jiwu, Wang Yuewu. The trust expansion and control in social network service [J]. Journal of Computer Research and Development, 2010, 47(6): 1611-1621 (in Chinese)
(康乐, 荆继武, 王跃武. 社会化网络服务中的信任扩张与控制[J]. 计算机研究与发展, 2010, 47(6): 1611-1621)

[17] Guha R, Kumar R. Propagation of trust and distrust [C] // Proc of the 13th Int Conf on World Wide Web. New York: ACM, 2004: 403-412

[18] Ma H, Lyu M. Learning to recommend with trust and distrust relationships [C] //Proc of the 3rd ACM Conf on Recommender Systems. New York: ACM, 2009: 189-196

[19] Antonellis I, Molina H G. Simrank++: Query rewriting through link analysis of the click graph [J]. Journal of the VLDB Endowment, 2008, 1(1): 408-421

[20] Feldmann A, Whitt W. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models [J]. Performance Evaluation, 1998, 31(3): 245-279

Yin Guisheng, born in 1964. Professor and PhD supervisor in Harbin Engineering University. His main research interests include database and knowledge based application system, virtual reality, internetware.

Zhang Yanan, born in 1981. PhD candidate and student member of China Computer Federation. His research interests include data mining and recommender system.

Dong Hongbin, born in 1963. Professor and PhD supervisor in Harbin Engineering University. His main research interests include natural computing, machine learning, data mining, and multi-agent systems.

Dong Yuxin, born in 1974. PhD and associate professor in Harbin Engineering University. Her main research interests include trust evolution, social networks.