基于用户－兴趣－项目三部图的推荐算法*

张艳梅 王 璐 曹怀虎 毛国君

(中央财经大学 信息学院 北京 100081)

摘 要 目前大多数个性化推荐算法为了追求较高的推荐精度而在不同程度上受到用户兴趣过拟合问题的影响，因此提出通过挖掘用户隐含的兴趣信息进行推荐的算法. 首先利用概率主题模型抽取用户兴趣分布，并建立用户－兴趣－项目加权三部图. 然后在用户－兴趣和兴趣－项目的概率加权二部子图上依次利用物质扩散算法配置项目资源值，并根据项目资源值的高低排序产生 Top-K 推荐. 在 Movielens 数据集上的实验表明，基于用户－兴趣－项目三部图的推荐算法能缓解过拟合问题，同时可提高准确率等方面的性能.

关键词 用户兴趣，个性化推荐，三部图，物质扩散，概率主题模型

中图法分类号 TP 393 **DOI** 10.16451/j.cnki.issn1003-6059.201510006

Recommendation Algorithm Based on User-Interest-Item Tripartite Graph

ZHANG Yan-Mei, WANG Lu, CAO Huai-Hu, MAO Guo-Jun

(*School of Information, Central University of Finance and Economics, Beijing* 100081)

ABSTRACT

Since most of the existing personalized recommendation algorithms pursue a higher accuracy, their performance is affected by the problem of user interest over-specialization. An algorithm is proposed to fully mine and use the implicit user interest information for recommendation. The probabilistic topic model is adopted to extract user interest distribution, and the weighted tripartite graph of user-interest-item is generated. Then the user item resource value is allocated by material diffusion algorithm in user-interest and interest-item bipartite graphs respectively, and the Top-K recommendation list is generated according to the rank of item resource values. Experimental results on Movielens datasets show the proposed algorithm relieves the problem of user interest over-specialization. Meanwhile the recommendation accuracy is improved .

Key Words User Interest, Personalized Recommendation, Tripartite Graph, Material Diffusion, Probabilistic Topic Model

* 国家自然科学基金项目(No. 61309029,61273293)、中央财经大学学科建设基金项目(No. 2015XX04)资助
收稿日期: 2014－09－10; 修回日期: 2015－01－11
作者简介 张艳梅(通讯作者)，女，1976 年生，博士，副教授，主要研究方向为电子商务、服务计算. E-mail: jlzym0309@sina.com. 王璐，女，1989 年生，硕士研究生，主要研究方向为电子商务. 曹怀虎，男，1977 年生，博士，教授，主要研究方向为电子商务、网络计算. 毛国君，男，1966 年生，博士，教授，主要研究方向为数据挖掘.

1 引 言

个性化推荐通过分析目标用户的历史行为和兴趣偏好信息,主动提供能满足需求的产品或服务. 近年来兴起的基于图结构的推荐是个性化推荐的主流方法之一[1],其不但在推荐准确性上优于传统协同过滤算法,而且在推荐多样性上也明显高于传统的协同过滤算法[2]. 文献[3]最早提出基于二部图的物质扩散算法,取得较好的推荐准确性和多样性.

Web2.0 的发展及它的附属应用形成一种以用户为主导的互联网发展模式. 社会化标签作为网络资源增加元数据的一种主要方式,不仅反映用户的个人兴趣,也可表达商品间的语义关系. 目前,社会化标签与个性化推荐结合的研究被广泛关注,主要有 FolkRank 推荐方法[4]和基于三部图网络结构的推荐方法. FolkRank 推荐方法来源于 PageRank[5],由于它推荐给用户的不是资源而是标签,因此该算法应用存在局限性. 在三部图网络结构的推荐方法中,较经典的是 Zhang 等[6]在二部图结构的基础上引入新的标签内容,提出用户 - 商品 - 标签的三部图物质扩散算法,明显提高推荐的准确性、多样性.

然而,现在很多社会化标注系统对标签的添加无任何要求限制,用户可自由地对所选资源添加描述词语或关键词,这种随意性的行为会降低标签质量,不规范,不严密,如标签的一词多义和多词一义问题,不可避免地影响推荐性能. 许多学者以此为出发点,提出用标签聚类[7-9]缓解词语过多的问题,这种方法是当前研究的重点. 它的基本思想是根据标签之间的共现频率,利用聚类算法,将同义词合成为一个词. Durao 等[10]将项目与标签的紧密度作为标签聚类划分的依据,分别计算用户和各个聚类间的亲密度及每个项目聚类对应的主题,进而依据聚类结果为用户提供个性化推荐. 由于标签使用呈现“幂律分布”特性,标签聚类算法只能发现球状的聚类,并对噪声数据特别敏感,从而制约推荐的有效性和覆盖度. 在语义方法中,采用基于本体的原则[11-12]组织标签,构造并解释标签之间的语义联系. Baruzzo 等[13]通过领域本体和标签术语形成对同一概念的语义共识,实现为网络资源主动推荐新的标签. Fu 等[14]提出一个社会化标签语义仿真模型,该模型认为社会化标签源于一个自发的主题演化过程,可预先对资源的语义解释做出指导.

针对标签噪声问题,张量降维的方法也是重要的研究方向,Leginus等[15]提出将社会化标注系统中的用户、资源和标签信息建模成三层张量,采用高维奇异值分解 Kernel-奇异值平滑技术缩减维度,进行语义分析. 还有学者采用基于主题的方法发现潜在主题,Krestel 等[9]利用概率主题模型确定标签和资源所划归的主题,分别计算主题在标签上的概率分布及资源在主题上的概率分布,把每个主题下最高的标签推荐给用户.

综上所述,传统的基于用户 - 项目 - 标签三部图的推荐算法一般基于标签数据进行推荐,在不同程度上缩减信息量,从而导致信息损失. 而且由于网络数据形式更多的情形是用户在保存一个商品的同时并未给出任何标注,导致推荐算法在无标签数据或标签信息不完整的环境下无法应用. 大多数个性化推荐方法为追求较高的推荐精度而在不同程度上受到用户兴趣过拟合问题的困扰. 因此,本文提出基于用户 - 兴趣 - 项目三部图的推荐算法(User Interest Item-Diffusion on Tripartite Graph, UII-DTG). UII-DTG 既能发挥三部图推荐的有效性和多样性,解决用户兴趣的过拟合问题,同时又能避免标签数据本身固有的噪声问题.

2 基于用户 - 兴趣 - 项目三部图的推荐算法思想

为更清晰阐述 UII-DTG 思想,需先分析传统的基于用户 - 项目 - 标签三部图推荐算法(DTG)的特点. 如图 1(a)所示,DTG 的物质扩散方式为以项目为中心的并列扩散,该扩散方式可能带来如下 3 个问题.

1) 将用户 - 项目和标签 - 项目二部图上的物质扩散看作两个独立的过程,用户、项目间的选择关系及标签、项目间的标注关系并无关联,各自独立决定最终每个项目获得的资源情况.

2) DTG 推荐中标签直接与项目相连,更多体现项目的属性信息,并未很好起到联系用户和项目的纽带作用,忽视用户对项目偏好关系的体现. 然而在现实生活中,用户对项目的选择往往基于用户的兴趣偏好,因此,如果识别出用户的潜在兴趣需求,就可更深刻地理解用户、项目间的联系.

3) DTG 的三部图是无权的,在项目之间资源转移的过程中,将项目资源平均分配给用户,用户又将分到的资源再平均分配给项目. 但在真实的社会网络中,用户与项目之间的边权具有重要的现实意义.

(a) DTG　　(b) UII-DTG

图 1　2 种算法的物质扩散方式

Fig. 1　Diffusion modes of 2 algorithms

与 DTG 的并列扩散方式不同,UII-DTG 的扩散方式为以兴趣为中心的顺序扩散(见图 1(b)) . UII-DTG 主要可分为 2 个步骤实现.

1) 利用潜在狄利克雷分布(Latent Dirichlet Allocation, LDA) 概率主题模型抽取用户兴趣分布,从而以兴趣为中心建立用户 – 兴趣 – 项目加权三部图.

2) 在用户 – 兴趣 – 项目的二部子图上依次完成两次物质扩散算法,将用户 – 兴趣子图上的扩散结果作为兴趣 – 项目上扩散过程的输入.

LDA 是近年来机器学习领域提出的一种话题模型,是一种完全生成的贝叶斯层次主题模型,有较好的先验概率假设. UII-DTG 将 LDA 模型引入三部图结构的推荐中,将用户集合表示为兴趣上的一个概率分布,而每个兴趣又是在项目空间上的概率分布. 建立用户 – 兴趣 – 项目三部图,将兴趣视为中心节点,以用户的兴趣分布作为用户 – 兴趣二部图的边权,以兴趣在项目上的分布作为兴趣 – 项目二部图的边权重. 该三部图的兴趣节点可理解为用户对项目的需求,而对于项目而言,这是它们的某些特征. 基于这层隐含的兴趣分布,可更真实充分地解释项目被推荐给用户的原因,从而深层次地理解用户、兴趣及项目间的关系.

总之,DTG 基于用户-项目-标签数据结构,标签一般源于系统预先设定的固定标签集合,并不能根据用户的需求变化动态调整. 以 Movielens 数据集为例,假设用户初始偏好 Romance 题材的电影,随着时间推移,用户喜好可能会转变为小众类型或新型题材的电影,如 Micro_Film,但由于不具备描述这种题材的标签,因此这种类型的变化并不能在 DTG 中得以体现. 如果推荐系统不能捕获这些用户喜好的变化动态,仍根据用户过去的喜好进行推荐,易导致过拟合问题. 而在 UII-DTG 中,以用户-项目选择关系作为 LDA 的输入数据,这些输入数据能实时反映用户对项目选择关系的动态变化,UII-DTG 基于这些动态变化的输入数据抽取新的兴趣层,从而及时感知并捕捉用户兴趣变化动态,更准确预测和匹配用户的新兴趣,在理论上可缓解过拟合问题.

3　基于用户 – 兴趣 – 项目三部图的推荐算法实现

UII-DTG 将主题建模技术引入社会化推荐系统,采用 LDA 主题模型学习用户的兴趣分布,然后利用抽取的兴趣模型构建用户 – 兴趣 – 项目三部图,并将用户在兴趣上的概率分布和兴趣在项目上的概率分布作为用户与兴趣、兴趣与项目间的边权,在用户 – 兴趣和兴趣 – 项目二部图上依次执行物质扩散算法,按照节点间的边权占该节点权重和的比例分配资源,最后将两部分结果综合形成资源推荐.

3.1　用户兴趣建模

UII-DTG 将概率主题模型移植到基于图结构的推荐系统中. 首先假定用户的网络行为由一个隐式因子——兴趣决定,将用户集合表示为兴趣上的一个概率分布,而每个兴趣又是在项目空间的概率分布. 其次,UII-DTG 忽略用户网络行为中的消费时序信息,假定任两个时间点的用户行为可交换. 基于上述 2 点假设,可使用概率主题模型抽取用户的兴趣分布.

在 M 个用户、N 个项目的数据集上,UII-DTG 的概率图模型如图 2 所示.

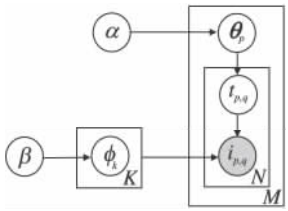


图 2　UII-DTG 的 LDA 概率图模型

Fig. 2　LDA probability graph model of UII-DTG

图 2 中只有灰色圆是可观察值,其他的空心圆都是隐式变量,箭头方向代表条件概率方向,矩形表示重复过程,大矩形表示从先验 Dirichlet 分布中为集合中的每个用户反复抽取兴趣分布,小矩形表示从兴趣分布中反复抽取产生购买项目(假设用户的网络行为是购买行为) .

对于目标用户 U_p,这个概率图可分解为 2 个主要过程.

1) $\alpha \rightarrow \boldsymbol{\theta}_p \rightarrow t_{p,q}$. 从先验分布 $\boldsymbol{\theta} \sim Dirichlet(\alpha)$ 中采样,生成用户 p 的兴趣分布向量 $\boldsymbol{\theta}_p$,这个过程仅执行一次,然后根据 $\boldsymbol{\theta}_p$ 生成该用户的第 q 个购买项

目 $i_{p,q}$ 的兴趣分布，$q \in (1,N)$，重复 N 次.

2) $\beta \rightarrow \phi_k \rightarrow i_{p,q} \mid k = t_{p,q}$. 这个过程可描述如下：根据 $Dirichlet(\beta)$ 生成兴趣在项目上的分布 ϕ_k，然后在 ϕ_k 和 $t_{p,q}$ 已确定的条件下生成用户集合中用户 p 的第 q 个项目，重复 N 次.

最终，LDA 的联合概率分布为

$$p(\boldsymbol{\theta},t,i \mid \alpha,\beta) = p(\boldsymbol{\theta} \mid \alpha) \prod_{n=1}^{N} p(t_n \mid \boldsymbol{\theta})\, p(i_n \mid t_n,\beta)),$$

其中，α、β 表示 Dirichlet 分布的参数，$\boldsymbol{\theta}$ 是用户级兴趣分布向量，$\theta_{p,k} = p(t_k \mid U_p)$.

在 LDA 主题模型建立过程中，抽取用户兴趣是一个隐变量推导过程，由观察到的项目计算隐含变量，属于 NP 难问题. 一般来说直接计算模型中的联合概率分布不现实，因此需用近似推理方法对其进行估计. 由于 Gibbs 抽样速度较快且易于实现，所以这里采用 Gibbs Sampling 间接计算模型参数 α、β、$\boldsymbol{\theta}$、ϕ.

Gibbs 的目的是构造收敛于某目标概率分布的马尔科夫链，并从链中提取接近该概率分布值的样本，算法最终得到的 2 个 Dirichlet 后验分布在贝叶斯框架下的估计值：

$$\theta_{p,k} = \frac{n_p^{(k)} + \partial_k}{\sum_{k=1}^{K} n_p^{(k)} + \partial_k},$$

$$\phi_{k,q} = \frac{n_k^{(q)} + \beta_q}{\sum_{q=1}^{N} n_k^{(q)} + \beta_q},$$

其中，$n_p^{(k)}$ 表示用户 p 选择过的项目中属于兴趣 k 的频次；$n_k^{(q)}$ 为项目 q 被分配到兴趣 k 的次数.

最终得到用户的兴趣分布及兴趣在项目空间的概率分布，将抽取的 2 个矩阵作为下一步三部图推荐的输入数据.

3.2 物质扩散过程

考虑一个由 M 个用户、N 个产品和 K 个兴趣构成的推荐系统中，如果用户 p 偏好兴趣 k，在 p 和 k 之间建立一条连边

$$\alpha_{p.k} = \theta_{p,k},\ p = 1,2,\cdots,M,\ k = 1,2,\ldots,K,$$

否则 $\alpha_{p,k} = 0$. 如果项目 q 拥有兴趣 k，在 k 和 q 之间建立一条连边

$$\alpha'_{k,q} = \phi'_{k,q},\ k = 1,2,\cdots,K,\ q = 1,2,\cdots,N,$$

否则 $\alpha'_{k,q} = 0$.

由此构造出的加权三部图如图 3 所示.

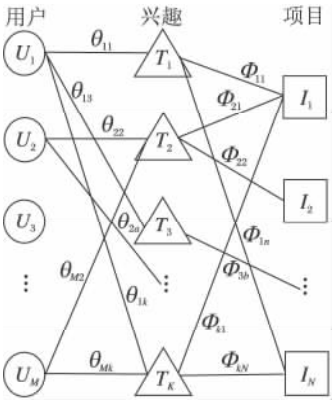


图 3 用户-兴趣-项目加权三部图

Fig. 3 Weighted tripartite graph of user-interest-item

再在概率加权的二部子图上依次执行物质扩散算法.

对于兴趣变量，用一个 K 维的 0/1 向量 $f = (f_1, f_2, \cdots, f_K)$ 表示兴趣集合的初始资源向量，即如果用户偏好兴趣 k，$f_k = 1$，否则为 0.

选定目标用户后，即可在 $\boldsymbol{\theta}$ 矩阵中得到用户的兴趣分布向量，由于该分布用 0 ~ 1 之间的概率值表示，导致某些兴趣因子上的值可能极低，但不为 0. 为此，设置阈值 $\lambda \in (0,1)$，只有大于 λ 的兴趣因子才可参与扩散过程，这些参与扩散的兴趣因子可看作用户兴趣的代表性特征. 在用户-兴趣二部加权图上进行用户-兴趣-用户-兴趣三次资源转移过程后，兴趣得到的资源向量 f' 为

$$f'_k = \sum_{l=1}^{M} \frac{\alpha_{l,k}}{N(U_l)} \sum_{s=1}^{K} \frac{\alpha_{s,l} f_s}{N(T_s)},$$

$$k = 1,2,\cdots,K,\ \alpha_{l,k} > \lambda,\ 0 < \lambda < 1,$$

其中

$$N(U_l) = \sum_{j=1}^{K} \alpha_{l,j},$$

表示用户 U_l 连接所有兴趣的边权之和，

$$N(T_s) = \sum_{r=1}^{M} \alpha_{r,s}$$

为兴趣 T_s 连接所有用户的边权和.

对于项目而言，用一个 N 维的向量 w 表示项目集合的初始资源向量，项目初始资源转移向量由兴趣资源向量和兴趣-项目二部图上的边权共同决定. 项目初始向量表示为

$$w_i = \sum_{k=1}^{K} \frac{\alpha'_{k,i} f'_k}{\sum_{j=1}^{N} \alpha'_{k,j}},\ i = 1,2,\cdots,N.$$

项目获得由兴趣传来的初始资源后,进行第二次物质扩散. 在兴趣–项目二部加权图上,经过资源二次分配后,项目最终得到的转移向量 w' 为

$$w_i' = \sum_{l=1}^{K} \frac{\alpha_{l,i}'}{N(T_l)} \sum_{s=1}^{N} \frac{\alpha_{s,l} w_s}{N(I_s)},\ i = 1,2,\cdots,N,$$

其中

$$N(T_l) = \sum_{j=1}^{N} \alpha_{l,j}',$$

为兴趣 T_l 所有项目连边的权值之和,

$$N(I_s) = \sum_{r=1}^{K} \alpha_{r,s}',$$

为项目连接所有兴趣的边权和.

w' 即为最终的资源转移向量,根据项目最终获得的资源值,将用户未选择过的项目按数值高低排序,取分数最高的前 L 个项目产生推荐.

3.3　算法代码

算法伪代码如下所示.

Input: the matrix of user ratings

λ: the proportion of interest involved in the diffusion

K: the number of interest

α, β: hyperparameter of dirichlet distribution in LDA

Output: Recommendation list L for U_0

For all interest $k \in [1, K]$, do

Sample mixture components $\phi_k \sim Dir(\beta)$

For all users $U_m, m \in [1, M]$, do

Sample mixture proportion $\theta_m \sim Dir(\alpha)$

For all items preferred by user $U_m, m \in [1, M]$, do

Sample interest index $K_{m,n} \sim Mult(\theta_m)$

Sample item for interest $K_{m,n}$

Return the matrix of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$

Initialize resource of interest owned by the target user U_0

Diffusion_from_Interest_to_User($\lambda, K, \boldsymbol{\theta}$)

Diffusion_from_User _to_ Interest ($\lambda, K, \boldsymbol{\theta}$)

Return the final resource $\boldsymbol{\gamma}$ of involved interest

Initialize resource of item according to $\boldsymbol{\gamma}$

Diffusion_from_Item_to_ Interest ($\lambda, K, \boldsymbol{\phi}$)

Diffusion_from_ Interest _to_ Item ($\lambda, K, \boldsymbol{\phi}$)

3.4　复杂度分析

UII-DTG 分为3 个步骤. 这里设 UII-DTG 的总时间复杂度为 C,3 个步骤的时间复杂度分别为 C_1、C_2 和 C_3.

1) 第一步是对于兴趣(K 个),所有的用户(M 个)、所有用户偏爱的项目(M 个) 生成两个矩阵 $\boldsymbol{\theta}$ 和 $\boldsymbol{\phi}$. 三层循环生成矩阵的时间复杂度为

$$C_1 = O(KM^2).$$

2) 第二步是根据初始化目标用户拥有的兴趣的初始资源,返回最终用户拥有的兴趣的资源. 用户–兴趣二部图物质扩散的算子:

$$f_k' = \sum_{l=1}^{M} \frac{\alpha_{l,k}}{N(U_l)} \sum_{s=1}^{K} \frac{\alpha_{s,l} f_s}{N(T_s)},$$

$$k = 1,2,\cdots,K,\ \alpha_{l,k} > \lambda,\ 0 < \lambda < 1,$$

时间复杂度 $C_2 = O(M^2K^2)$,其中

$$N(U_l) = \sum_{j=1}^{K} \alpha_{l,j}$$

的时间复杂度为 $O(K)$,

$$N(T_s) = \sum_{r=1}^{M} \alpha_{r,s}$$

的复杂度为 $O(M)$,而且还有一个是 M 项求和、一个是 K 项求和. 由于进行两次物质扩散,所以第二步的总时间复杂度 $C_2 = O(M^2K^2)$.

3) 第三步是根据第二步返回的最终资源计算初始化资源. 项目–兴趣二部图物质扩散的算子:

$$w_i = \sum_{k=1}^{K} \frac{\alpha_{k,i}' f_k'}{\sum_{j=1}^{N} \alpha_{k,j}'},\ i = 1,2,\cdots,N,$$

时间复杂度为 $O(NK)$. 而兴趣到项目的物质扩散的算子:

$$w_i' = \sum_{l=1}^{K} \frac{\alpha_{l,i}'}{N(T_l)} \sum_{s=1}^{N} \frac{\alpha_{s,l} w_s}{N(I_s)},\ i = 1,2,\cdots,N,$$

时间复杂度为 $O(K^2N^2)$. 第三步总时间复杂度为

$$C_3 = O(NK) + O(K^2N^2).$$

综上所述,UII-DTG 的时间复杂度为

$$C = O(KM^2) + O(M^2K^2) + O(NK) + O(K^2N^2).$$

从上式可看出,UII-DTG 的时间复杂度与用户数目 M、兴趣数目 K、项目数量 N 有关. 它的复杂度最高阶项是一个 M^2K^2 的四次项,是一个可接受的复杂度.

4　实验及结果分析

4.1　数据处理

实验数据来自 GroupLens 小组提供的 Movielens-10M/100K 数据集,该数据集包含电影推荐系统中用户对电影的评分数据,评分为 1 ~ 5 之间的整数,后来 Movielens 引入大众标签特性,是测试推荐算法性能的标准数据集之一.

对于用户评分数据的处理,由于本文算法属于

二元选择关系的推荐，推荐结果只有用户喜欢与不喜欢2种情形，MovieLens数据集为5级评分，通常3～5分被认为是用户喜欢，1～2分被认为是用户不喜欢[17]. 因此把评分值大于3的数据视作用户喜欢的项目，去掉用户评分小于3的项目.

对于标签数据的处理，由于Movielens系统对用户添加标签无任何语法限制，导致标注信息出现一些个性化的符号，因此，去除只出现过一次的标签，同时去除孤立节点以确保每部电影至少有2位用户评价，也至少被一个标签描述.

表1给出原始数据和预处理后的数据集统计信息. 表1中，k_u为用户平均评价的电影数目，k_i^u为一部电影被评价的平均用户数，k_i^T为一部电影拥有的平均标签数，k_T^i为一个标签指向的平均电影数.

表1 Movielens数据集基本统计信息

Table 1 Basic statistical information of Movielens dataset

数据集	Movielens-100K	Movielens-10M
用户数	943	3549
电影数	1682	6054
标签数	–	5828
评分数	82518	–
k_u	87.51	14.41
k_i^u	49.06	8.45
k_i^T	–	10
k_T^i	–	10.39

为评价算法性能，将上述预处理后的数据分为2部分：随机抽取每个用户70%的数据作为训练集，用于计算用户对未评价电影的喜欢程度，剩下的30%作为测试集，用于对照评价算法的推荐结果.

实验运行环境如下：Windows XP操作系统，CPU主频2.26 GHz，2 GB内存，SQL Server 2000数据库. 实验代码在Eclipse平台中用Python实现.

实验基准算法如下：1）二部图物质扩散推荐算法（DBG）[8]；2）三部图物质扩散推荐算法（DTG）[6]；3）基于用户的协同过滤推荐算法（User-Based Collaborative Filtering，UCF）[17]. 采用Pearson公式计算用户相似度，最终产生Top-K推荐，协同过滤算法采用开源代码库Apache Mahout实现.

4.2 评价指标

为了从多方面评价算法性能，选取5个评价指标分别测试算法的准确性、多样性和新颖性.

1）准确性. 准确率（Precision）是广泛用于信息检索领域的评价标准之一. 准确率表示用户对一个被推荐资源感兴趣的可能性，定义为推荐列表中用户喜欢的项目占所有推荐项目的比率：

$$P = \frac{N_{rs}}{N_s},$$

其中，N_{rs}为正确推荐数，N_s为算法生成的推荐总数.

相应地，召回率（Recall）表示用户喜欢的产品被推荐给该用户的概率，定义为推荐列表中用户喜欢的项目与系统中用户喜欢的所有项目的比率：

$$R = \frac{N_{rs}}{N_r},$$

其中，N_{rs}为正确推荐数，N_r为用户喜欢的项目总数.

准确率和召回率需一起使用才能全面评价算法的好坏. $F1$指标由于把准确率和召回率统一到一个指标而得到广泛应用. $F1$指标表示为准确率和召回率的调和平均值：

$$F1 = \frac{2PR}{P+R},$$

其中，P为准确率，R为召回率.

2）多样性. 为了评价不同用户推荐列表的差异，利用平均海明距离（Average Hamming Distance）衡量推荐结果中推荐列表的多样性. 2个用户U_1、U_2的海明距离：

$$Hamming_{ij} = 1 - \frac{Q_{ij}}{L},$$

其中，L为推荐列表的长度，Q_{ij}为用户U_i、U_j推荐列表中相同项目的数量.

推荐系统中所有用户推荐列表的多样性定义为H_{ij}的平均值. 海明距离的最大值为1，意味着所有用户的推荐列表完全不同；最小值为0，即所有用户的推荐列表完全相同.

3）新颖性. 一个好的推荐系统不但有较高的准确率，同时还能将长尾中鲜为人知的商品推荐给用户，新颖性（Novelty）用来衡量推荐系统向用户推荐新信息的能力，数值越小，新颖度越高，公式表示为

$$Novelty = \frac{1}{nL}\sum_{i=1}^{n}\sum_{I_r \in I_R^i} k(I_r),$$

其中，n为用户数量，L为推荐列表长度，$k(I_r)$为项目的度.

4.3 实验结果

4.3.1 算法参数选取

在UII-DTG第一阶段，利用LDA抽取用户兴趣分布时，采用LDA在处理文档集时的常用设置，$\beta = 0.01$，$\alpha = 50/K$，K为参数，表示预设的兴趣数量.

为了测试实验参数的最优值，在推荐列表长度$L=10$的条件下，计算λ和K的不同取值对准确度指标Precision、多样性指标海明距离Hamming的影

响. 实验结果如表 2 所示.

表 2　λ、K 不同取值时推荐结果对比

Table 2　Comparison of recommendation results with different λ and K

	λ	$K=20$	$K=60$	$K=100$
准确率	0.1	0.1061	0.2219	0.1931
	0.2	0.1539	0.1865	0.1756
	0.3	0.2236	0.1809	0.1686
	0.4	0.2076	0.1691	0.1691
	0.5	0.1952	0.1636	0.1628
	0.6	0.1878	0.1647	0.1564
	0.7	0.1802	0.1617	0.1516
	0.8	0.1715	0.1614	0.1604
	0.9	0.1710	0.1631	0.1579
海明距离	0.1	0.808	0.847	0.852
	0.2	0.889	0.838	0.733
	0.3	0.802	0.817	0.708
	0.4	0.843	0.714	0.612
	0.5	0.795	0.718	0.521
	0.6	0.728	0.697	0.474
	0.7	0.704	0.577	0.426
	0.8	0.716	0.508	0.365
	0.9	0.702	0.501	0.313

从表 2 可看出,UII–DTG 的推荐精度受到 λ 和 K 变化的综合影响,因为 λ 和 K 的乘积值决定最终有多少兴趣因子参与运算. 表 2 准确率部分直观反映出: λ 和 K 的乘积越大,推荐精度越差. 这说明在用户兴趣过于分散时,会导致推荐算法无法准确判断用户偏好重心,最终不能给出有效的推荐结果.

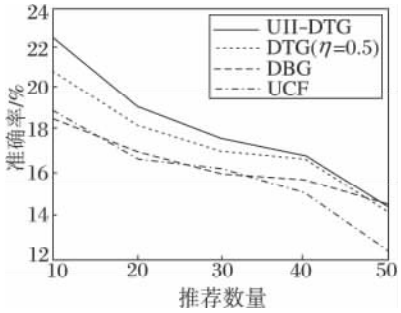
表 2 同时给出 λ 和 K 对多样性指标 Hamming 的影响. 同样地,Hamming 指标与 Precision 呈现相同的变化规律: λ 和 K 的乘积越大,推荐列表的多样性越差. 这是因为,越多的兴趣因子参与扩散,导致每个兴趣因子在扩散过程中对整个兴趣集合的影响越小,最终所有用户得到的资源向量趋同,仅在最后一步滤掉用户已选择过的项目时,才产生推荐结果的差异.

综合表 2 的结果和上述分析发现,在 $\lambda=0.3$, $K=20$时,算法性能达到最优,此时目标用户约有 6 个稳定的兴趣特征因子参与三部图的扩散过程. 同时,由于兴趣数目 K 取较小值,也有利于提高 LDA 建模速度,减少算法离线运行时间.

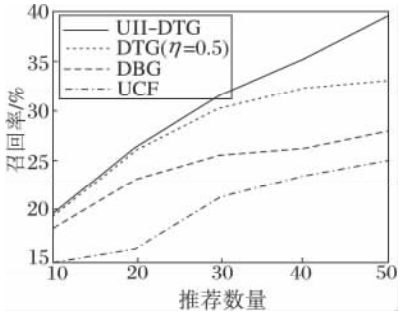
4.3.2　实验结果对比

通过 Movielens 数据集进行对比实验,其中 UII–DTG、DBG 和 UCF 采用 Movielens–100k 数据集,DTG 使用引入标签信息的 Movielens–10M 数据集,线性叠加参数 $\eta=0.5$. 下面的实验中,DTG 的推荐结果均是在 $\lambda=0.3$,$K=20$ 的条件下取得.

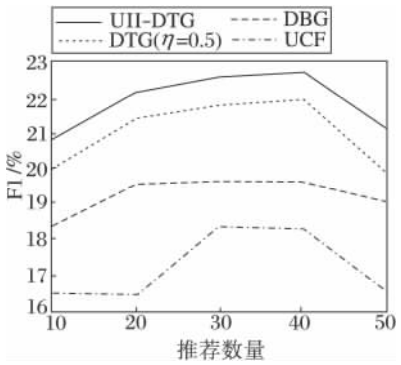
首先,为更好展示 UII–DTG 与各基准算法在准确性方面的对比,将从准确率、召回率和 $F1$ 指标 3 个角度描述推荐性能. 图 4 分别展示各算法在上述 3 个指标上随推荐数量变化的表现.



(a) *Precision*



(b) *Recall*



(c) *F*1

图 4　准确性指标随推荐数量 L 的变化

Fig. 4　Accuracy indexes changing with recommendation number L

在图 4(a) 中,将准确率从高到低排序依次是 UII–DTG、DTG、DBG 和 UCF. 3 个基于图结构的推荐算法在准确率上均优于传统的协同过滤算法,这也符合在第 2 节中所做的分析. DBG 和 UCF 均是在用户 – 项目二元数据结构下产生推荐,从图中可看出,

相比 UCF，DBG 的推荐准确度有小幅提升，在引入标签数据后，DTG 取得更好的推荐效果，特别是在推荐数量较少时. 而 UII-DTG 在无标签信息的条件下可达到高于三部图算法的推荐准确率，*Precision* 指标大约平均提升 3.6%.

图 4(b) 反映召回率与推荐数量的相互关系. 召回率可从另一个角度描述系统的推荐准确度. 与准确率指标变化规律相反的是，UII-DTG 和 DTG 恰恰在推荐数量较多时，取得更好的召回率. 这可能是因为随着推荐数量增长，推荐列表能以更高的比例将测试集中用户喜欢的电影囊括进来.

图 4(c) 中展示 $F1$ 指标随推荐数量的变化情况，$F1$ 指标综合准确率和召回率的推荐结果. 各算法的 $F1$ 值均出现先增后降的趋势，但 UII-DTG 与基准算法在推荐准确度上的排位仍无变化. 这说明，从用户对项目的选择关系中挖掘隐含兴趣层是有价值的，这个隐含层次能起到替代标签的作用，传递更多的信息.

其次，对比各算法在推荐多样性上的表现，这里采用平均海明距离衡量. 图 5 展示推荐系统多样性与推荐数量的关系.

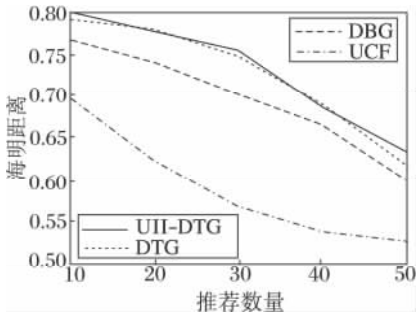


图 5 多样性指标随推荐数量 L 的变化

Fig. 5 Diversity index changing with recommendation number L

图 5 中一个明显的趋势是，基于图结构的推荐算法在用户推荐列表多样性上显著高于 UCF，至少提升约 18%，而相比 UCF，UII-DTG 更有 24.08% 的提升. 由此可见，以兴趣为中心的两次扩散过程能有效放大兴趣表征的用户个性化信息，提高用户间推荐结果的差异性. 但 UII-DTG 并不是始终优于 DTG，这可能是由于兴趣与标签均是用户个性化偏好的表现方式.

最后，对比各推荐算法向用户推荐长尾中新项目的能力. 图 6 反映新颖性指标随推荐数量的变化关系.

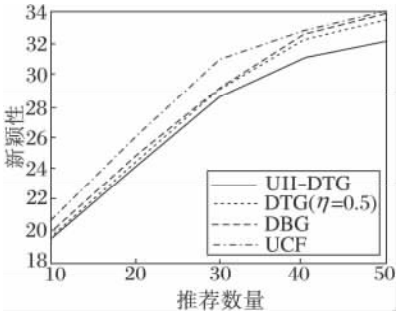


图 6 新颖性指标随推荐数量 L 的变化

Fig. 6 Novelty index changing with recommendation number L

图 6 中随着推荐数量的增加，*Novelty* 也呈现递增趋势，其中 UII-DTG 的 *Novelty* 最低且增速较缓，这也就证明利用隐含兴趣产生的推荐结果可有更多的新鲜“冷门”推荐产生.

从上述实验可看出，UII-DTG 不仅在推荐的准确性上表现较好，而且在多样性、新颖性指标上也明显优于传统的三部图物质扩散算法. 而多样性可在广度上衡量过拟合问题，其表现为不同用户间推荐列表的差异程度. 新颖性指标可在深度上体现推荐的过拟合度，防止推荐结果过度匹配用户过去的行为，导致未来的新行为得不到准确预测. 从图 5、图 6 的对比实验中不难看出，UII-DTG 在两个维度中均较好缓解推荐中的过拟合问题.

4.3.3 兴趣实例分析

UII-DTG 利用 LDA 抽取用户兴趣，而该兴趣是一个隐含的层次，为了测试隐含兴趣是否与用户真实存在的兴趣存在对应关系，在实验结果中随机抽取 2 个兴趣因子，并识别每个兴趣项下概率值最高的 Top-5 电影，如表 3 所示.

表 3 2 个隐含兴趣因子对应的 Top-5 电影列表

Table 3 List of Top-5 movies corresponding to 2 implicit interest factors

兴趣因子 1	兴趣因子 2
Chasing Amy	Mulholland Falls
Mrs. Brown (Her Majesty, Mrs. Brown)	He Walked by Night
Bridges of Madison County	Touch of Evil
Shine	M
Breakfast at Tiffany´s	Devil in a Blue Dress

分析兴趣因子 1 的 Top-5 电影，发现它们的一个共性特征是都属于 Drama、Romance 题材类的电影，一般来说，实际拥有这类兴趣的用户主要是年轻

人. 兴趣因子 2 对应的 Top-5 电影均可标记为 Thrill、Crime 及 Mystery. 在现实生活中,这类兴趣对应的用户群体主要是成年男性影迷. 因此,UII-DTG 提取的隐式兴趣可与用户的真实兴趣对应,隐含的兴趣层是以抽象的形式存在且具有实际意义的兴趣表示,只是它不能用文本直观表示. 这说明,UII-DTG 可准确挖掘用户兴趣,以抽象表示的兴趣层取代大众标签,起到联系用户和项目的"纽带"作用.

5　结 束 语

针对三部图推荐算法在无标签数据或用户标注信息不完整的环境下应用受限的问题,及大众标签固有的噪声问题,本文提出基于用户 - 兴趣 - 项目的三部图推荐算法(UII-DTG). UII-DTG 通过用户的购买、收藏行为挖掘这些行为背后的兴趣驱动原因,并将兴趣模型引入到三部图结构推荐算法中,从而将三部图算法成功移植到用户-项目的二元数据结构中,在真实的数据集上实验表明该算法能较大幅度提升推荐效果.

在接下来的研究工作中,考虑在算法中加入时间因素,准确跟踪用户兴趣变化,分阶段考虑不同时间段用户消费行为对将来行为的影响. 同时,本研究只验证 UII-DTG 在既有数据集上的推荐效果,并未考虑其在数据稀疏或是新用户冷启动问题中的应用,这部分有待于在后续工作中展开深入研究. 最后,希望能结合标签元数据及语义分析方法找到兴趣的隐含含义,更直观体现用户选择资源的原因,进一步提升用户体验.

参 考 文 献

[1] Li X, Chen H C. Recommendation as Link Prediction in Bipartite Graphs: A Graph Kernel-Based Machine Learning Approach. Decision Support Systems, 2013, 54(2): 880 - 890

[2] Liu J G, Zhou T, Wang B H. Personalized Recommendation System Research Process. Progress in Natural Science, 2009, 19(1): 1 - 15 (in Chinese)
(刘建国,周 涛,汪秉宏. 个性化推荐系统的研究进展. 自然科学进展, 2009, 19(1): 1 - 15)

[3] Zhou T, Ren J, Medo M, *et al*. Bipartite Network Projection and Personal Recommendation. Physical Review E, 2007. DOI: 10.1103/PhysRevE.76.046115

[4] Gemmell J, Schimoler T, Ramezani M, *et al*. Improving Folkrank with Item-Based Collaborative Filtering [EB/OL]. [2014 - 08 - 30]. http://ceur-ws.org/Vol-532/paper3.pdf

[5] Brin S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine [EB/OL]. [2014 - 08 - 20]. http://ilpubs.stanford.edu: 8090/361/1/1998 - 8.pdf

[6] Zhang Z K, Zhou T, Zhang Y C. Personalized Recommendation via Integrated Diffusion on User-Item-Tag Tripartite Graphs. Physica A: Statistical Mechanics and Its Applications, 2010, 389(1): 179 - 186

[7] Song Y, Zhang L, Giles C L. Automatic Tag Recommendation Algorithms for Social Recommender Systems. ACM Trans on the Web (TWEB), 2011. DOI: 10.1145/1921591.1921595

[8] Shepitsen A, Gemmell J, Mobasher B, *et al*. Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering // Proc of the ACM Conference on Recommender Systems. Lausanne, Switzerland, 2008: 259 - 266

[9] Krestel R, Fankhauser P. Personalized Topic-Based Tag Recommendation. Neurocomputing, 2012, 76(1): 61 - 70

[10] Durao F, Dolog P. A Personalized Tag-Based Recommendation in Social Web Systems // Proc of the Workshop on Adaptation and Personalization for Web 2.0. Trento, Italy, 2009: 40 - 49

[11] Symeonidis P, Nanopoulos A, Manolopoulos Y. A Unified Framework for Providing Recommendations in Social Tagging Systems Based on Ternary Semantic Analysis. IEEE Trans on Knowledge and Data Engineering, 2010, 22(2): 179 - 192

[12] García-Crespo Á, Colomo-Palacios R, Gómez-Berbís J M, *et al*. SEMO: A Framework for Customer Social Networks Analysis Based on Semantics. Journal of Information Technology, 2010, 25(2): 178 - 188

[13] Baruzzo A, Dattolo A, Pudota N, *et al*. Recommending New Tags Using Domain-Ontologies // Proc of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technologies. Milan, Italy, 2009, III: 409 - 414

[14] Fu W T, Dong W. Collaborative Indexing and Knowledge Exploration: A Social Learning Model. IEEE Intelligent Systems, 2010, 27(1): 39 - 46

[15] Leginus M, Dolog P, Žemaitis V. Improving Tensor Based Recommenders with Clustering // Proc of the 20th International Conference on User Modeling, Adaptation, and Personalization. Montreal, Canada, 2012: 151 - 163

[16] Blattner M, Zhang Y C, Maslov S. Exploring an Opinion Network for Taste Prediction: An Empirical Study. Physica A: Statistical Mechanics and Its Applications, 2007, 373: 753 - 758

[17] Goldberg D, Nichols D, Oki B M, *et al*. Using Collaborative Filtering to Weave an Information Tapestry. Communications of the ACM, 1992, 35(12): 61 - 70