# 一种基于标签的 Top-N 个性化推荐算法

马闻锴 李 贵 李征宇 韩子扬 曹科研
(沈阳建筑大学信息与控制工程学院 沈阳 110168)

**摘 要** 随着 Web2.0 的发展,UGC 标签系统受到越来越多的关注,标签既能反映用户的兴趣又能描述物品的本身特征。现有的标签推荐算法没有考虑用户的连续行为所产生的影响,而传统的基于马尔可夫链(Markov Chain)的推荐算法虽然侧重于研究用户的连续行为来产生推荐,但它是直接作用于用户与物品的二维关系,并不适用于基于 UGC 的标签推荐。因此,通过结合马尔可夫链和协同过滤的思想,提出了一种基于标签的个性化推荐算法。该算法将〈用户-标签-物品〉的三维关系拆分为〈用户-标签〉和〈标签-物品〉两个二维关系。首先通过马尔可夫链模型计算用户对标签的兴趣度,再通过推荐标签集来匹配与其相对应的物品。为了提高推荐的精准率,该算法利用标签之间的影响,并基于匹配物品中所含标签间存在的关联关系对物品进行满意度建模,该模型是一种概率模型。在计算用户-标签和用户-物品之间的兴趣度和满意度时使用了协同过滤的思想来补全稀疏值。在公开的数据集中,与现有算法相比,该算法在精准率、召回率上均有明显提高。

**关键词** 推荐系统,标签,马尔可夫链(MC),满意度模型,协同过滤(CF)

**中图法分类号** TP301.6 **文献标识码** A

## Top-N Personalized Recommendation Algorithm Based on Tag

MA Wen-kai LI Gui LI Zheng-yu HAN Zi-yang CAO Ke-yan
(Faculty of Information & Control Engineering, Shenyang Jianzhu University, Shenyang 110168, China)

**Abstract** With the development of Web2.0, UGC tag system is receiving more and more attention. Tag can not only reflect users' interests, but also it can describe the innate character of item. Available tag recommendation algorithm does not considerae the influence of continuous behaviors of users. Although traditional recommendation algorithm based on Markov Chain produces recommendation through the emphasis on the research of continuous behaviors of users, it can not be appliedy to the tag recommendation of UCG due to its direct function on the two-dimensional relationships between user and item. Therefore, according to the thoughts of Markov Chain and Collaborative Filtering, an individual recommendation algorithm based on the tag could be applied. The algorithm splits three-dimensional relationships of 〈user-tag-item〉 into two two-dimensional relationships of 〈user-tag〉 and 〈tag-user〉. Firstly, the interest degree is calculated through the application of Markov Chain. Then correspondent item matched through the recommendation of tags. To raise the accuracy rating of recommendation, modeling of satisfaction is established by this tag according to the influence of tags and associated relationships among tags of items . This model is a kind of probabilistic model. At the same time of calculating the interest degree and satisfaction degree of user-tag and user-item, the thought of Collaborative Filtering is also used to complement sparse data. Compared with available algorithm, this algorithm is improved a lot in the aspects of precision and recall rate on the open data set.

**Keywords** Recommended system, Tag, Markov chain, Satisfaction model, Collaborative filtering

## 1 引言

随着互联网的发展,推荐系统已成为解决信息过载的重要工具,其目的在于从海量的数据中挖掘出用户感兴趣的信息,对于用户来说它可以为其推荐可能感兴趣的物品,对于网站来说因其能增加客户的点击率或客户的满意度,所以能增加其电子商务的营收额。

目前,推荐算法的主流有 3 种:基于协同过滤的算法(Collaborative Filter Algorithm, CF)、基于内容的算法(Content-Based Algorithm, CB)以及基于标签的算法(Tag-Based Algorithm, TB)[1]。这些算法的使命就是为用户和物品建立连接,实现的方式是:事先找出那些隐藏的连接并呈现给用户。从本质上讲,这是一个预测问题。从达成的连接目标角度可以分为:1)评分预测;2)行为预测。本文采用用户历史行

本文受国家自然科学基金项目(61602323),辽宁省博士启动基金项目(201601209),住建部科学技术项目(2017-K8-038)资助。

**马闻锴**(1995—),男,硕士生,主要研究方向为 Web 数据挖掘和推荐系统;**李 贵**(1964—),男,博士,教授,主要研究方向为 Web 数据挖掘与集成、大数据分析技术,E-mail:Ligui21c@sina.com(通信作者)。

为记录数据来实现个性化推荐，属于行为预测。通常这些行为数据可以用用户-物品二分图来表示，即它们之间存在二元关系，目前为止，处理这种关系最常用的技术是协同过滤算法，即通过用户之间或者物品之间的相似度来挖掘相似的用户或物品，以达到推荐用户偏好的物品的目的。此外，还可以通过辅助信息（特征）来连接用户与物品之间的关系，如用户-标签-物品，可以用三部图来表示三者之间存在的三元关系[13]。特征也有不同的表达方式，如隐语义向量（latentfactorvector）或者物品的属性集合。

本文使用基于标签的推荐方法，根据维基百科的定义，标签是一种无层次化结构的、用来描述信息的关键词，它可以用来描述物品的语义[2]。根据标注标签的人的差异，标签通常分为两种：1）为作者或专家依据物品的属性特征打标签；2）普通用户标注物品标签，即用户生成的内容（User Generated Contend，UGC）。本文使用第二种方法所生成的标签数据，即用户标注物品标签数据。

传统的基于马尔可夫链（MC）模型的推荐系统可以通过带有时间序列属性的连续行为数据，根据用户最近一次行为来预测用户的下一次行为[3,15]，通过用户偏好物品的转移矩阵来模拟用户的连续行为。而本文使用马尔可夫链模型通过用户的偏好转移矩阵来模拟用户的连续行为所包含的信息，即标签，从而根据最近一次行为所包含的标签集来预测用户下一次行为所偏好的标签集。

通过马尔可夫链可得用户下一次行为所偏好的标签集合，再分析含有用户偏好标签集合的物品对用户的吸引力。可通过历史信息所得出的标签频率来计算用户对某一物品中每个标签的吸引力，将其聚合为用户对该物品的满意度。但是，该方法仅考虑单个标签却忽略了标签间的相互影响。从用户角度来看，用户对某个标签的满意度也受用户对其他标签满意度的影响，因此在一定程度上标签之间可相互加强以提高对用户的吸引力。如果仅考虑独立的标签而忽略标签之间的影响，则会影响推荐效果。因此，在进行 Top-N 推荐时，有必要考虑标签间的影响。对此，本文提出了一种估算标签间相互作用的模型，以此作为 Top-N 推荐的依据。

## 2　相关工作

随着 Web2.0 的发展，现今采取的标签数据大多是基于大众分类法的社会化标签。最近基于社会化标签的推荐算法通常分为 3 类：1）基于张量；2）基于图论；3）基于主题。文献[4]提出了一种基于标签主题的协同过滤推荐算法，利用 Dirichlet 分布建模挖掘潜在语义主题，消除标签中存在的语义模糊问题以提高推荐效果。文献[16]通过改进基于“用户-物品-标签”三部图的物质扩散推荐算法，提出基于“用户-项目-用户兴趣标签图”的混合协同过滤推荐算法，提高了 Top-N 推荐的精确度。但这些算法都没有关注用户的连续行为对推荐效果所造成的影响。

因此，本文采用马尔可夫链模型处理连续行为数据以提高推荐效果。近年来，关于马尔可夫链的研究都限于用户-物品之间的二维关系数据，如 Rendle 等[3]通过马尔可夫链建立转移矩阵来预测用户下一次购买的物品。文献[5]提出了一种基于马尔可夫决策过程的推荐算法，利用启发式算法改进马尔可夫链状态转移的最大似然估计。本文则通过转移矩阵来计算用户下一次行为所含的偏好标签集合，其优点为：1）有效提高了用户的个性化推荐效果，基于标签的推荐使用当前用户提供的历史行为来构建用户偏好的个性化转移矩阵，而协同过滤仅需要用户静态的评分数据来发掘用户的近邻；2）进一步加强推荐解释，显示地列出推荐列表中的物品内容特征或描述，以提高用户对推荐结果的信任。

在以标签为基础进行物品排序时，受 Pham 等[8]关于地域间的兴趣点的建模以及兴趣点之间的关联关系的启发，基于标签间的关联关系构建用户对物品的满意度模型，降低不同标签间的相互影响对推荐效果所造成的影响。

综上，为构建用户对物品的满意度模型，首先将用户-标签-物品的三维关系拆分成用户-标签、标签-物品两个二维关系，在计算用户到标签的偏好时引入马尔可夫链模型，在针对标签到物品间的关联时构建满意度模型并计算用户对物品的满意度，以此对用户进行 Top-N 物品推荐。本文的贡献如下：

（1）通过马尔可夫链模型学习物品所含标签的转移矩阵以模拟用户的连续行为，已知用户最近一次行为所含标签集合，计算用户的下一步行为偏好的标签集合。

（2）构建用户对物品的满意度模型，计算用户对所含标签集合的物品的满意度，并通过标签的历史信息以及标签之间的相互影响构建用户对物品的满意度模型，并进行优化。根据用户对物品的满意度进行 Top-N 推荐。

（3）由于数据存在稀疏性，因此在构建用户对标签的兴趣度模型和基于标签的用户对物品的满意度模型时分别使用了标签的相似度及物品的相似度来补全相关稀疏数据[9-11]。

## 3　基于马尔可夫链的用户到标签的预测算法

首先给出〈用户-标签-物品〉三维关系图以及〈用户-标签〉、〈标签-物品〉的二维关系图，如图 1 所示。定义数据集 $D=\langle U,I,T,A\rangle$，其中用户集 $U=\{u_1,\cdots,u_m\}$，$|U|=m$；物品集 $I=\{i_1,\cdots,i_n\}$，$|I|=n$；用户标注形成的标签集 $T=\{t_1,\cdots,t_l\}$，$|T|=l$。对于每个用户来说，同一物品可以标注多个标签，一组用户-标签-物品的三维关系可以表示为：$A\subseteq\{(u,i,t):u\in U,i\in I,t\in T\}$。

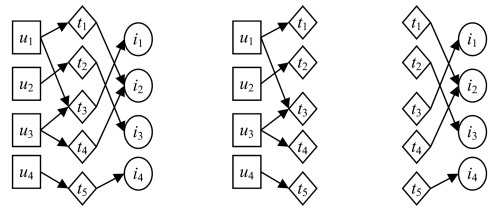



图 1　用户-标签-物品关联图

结合马尔可夫链模型给出如下定义：对于每个用户 $u$，其历史行为所含的信息，即标签集合为 $\mathscr{B}^u=(B_1^u,\cdots,B_t^u)$，且 $B_t^u\subseteq T$，表示用户 $u$ 在 $t$ 时刻查询或浏览物品的标签集合。所有用户的历史标签为 $\mathscr{B}:=\{\mathscr{B}^{u_1},\cdots,\mathscr{B}^{u_m}\}$。图 2 给出了 4 个用户的连续历史标签记录，其中圆角框代表用户直接通过查询标签来寻找物品的行为，直角框代表用户查询物品所含的标签信息，算法根据用户最近一次查询所含的标签集进行预测。

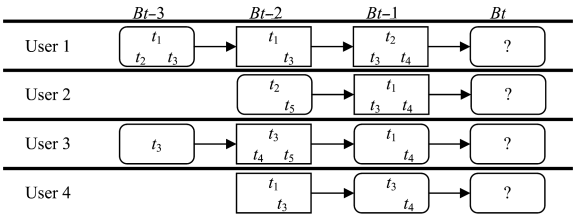



图 2 4 个用户的连续的标签集合记录

基于用户历史标签标注的行为记录,本节的目的是在用户下次展开查询或浏览时向用户推荐具有目标标签的物品。本文所处理的时间点不是基于绝对时间(例如 2018 年 1 月 1 日),而是使用事件序列关系来处理与用户相关的时间点,例如用户的第一次、第二次历史查询记录等。

### 3.1 基于标签的个性化马尔可夫链转移矩阵估计

#### 3.1.1 基于标签集的非个性化马尔可夫链

根据图 2 中的数据,基于非个性化(4 个用户同时计算)马尔可夫链模型计算的转移矩阵如图 3 所示。

| from \ to | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ |
|---|---|---|---|---|---|
| $t_1$ | 0.5 | 0.5 | 1 | 0 | 0 |
| $t_2$ | 0.5 | 1 | 0.5 | 0 | 0 |
| $t_3$ | 0.3 | 0.7 | 0.3 | 0 | 0.3 |
| $t_4$ | 0 | 0 | 1 | 0 | 1 |
| $t_5$ | 0 | 0 | 0 | 0 | 1 |

图 3 非个性化转移矩阵

例如,对于用户 4,概率为:

$p(t_1\in B_t\mid\{t_3,t_4\})=0.5(0.50+0.50)=0.500$

$p(t_2\in B_t\mid\{t_3,t_4\})=0.5(0.25+0.00)=0.125$

$p(t_3\in B_t\mid\{t_3,t_4\})=0.5(0.75+0.50)=0.625$

$p(t_4\in B_t\mid\{t_3,t_4\})=0.5(0.75+1.00)=0.875$

$p(t_5\in B_t\mid\{t_3,t_4\})=0.5(0.25+0.00)=0.125$

通过观察发现,已知用户最近一次查询的记录所含的标签集,结合上述非个性化转移矩阵,最终生成最佳推荐,与基于相似用户所生成的最佳推荐结果相差甚远。因此,下节将为每个用户生成个性化标签转移矩阵。

#### 3.1.2 个性化马尔可夫链转移矩阵估计

通常,步数为 $k$ 的马尔可夫链定义为:

$$p(X_t=x_t\mid X_{t-1}=x_{t-1},X_{t-2}=x_{t-2},\cdots X_{t-k}=x_{t-k}) \quad (1)$$

其中,$X_t,\cdots,X_{t-k}$ 是随机变量,$x_{t-k}$ 是其具体的实例。为了简化转移概率矩阵,将马尔可夫链的步数设为 1,即 $k=1$。简化后马尔可夫链表示为:

$$p(B_t^u\mid B_{t-1}^u) \quad (2)$$

同样,通过标签之间的转移矩阵表达每个用户的马氏链,但是这里的马氏链是个性化的:

$$a_{u,t_l,t_i}:=p(t_i\in B_t^u\mid t_l\in B_{t-1}^u) \quad (3)$$

依据用户最后状态给出的标签信息,挖掘出用户偏好某一标签的可能性,这可以理解为用户上一次查询的标签到查询这个标签的转移矩阵中的平均值:

$$p(t_i\in B_t^u\mid B_{t-1}^u):=\frac{1}{|B_{t-1}^u|}\sum_{t_l\in B_{t-1}^u}p(t_i\in B_t^u\mid t_l\in B_{t-1}^u) \quad (4)$$

这意味着对于每个用户 $u$,都要建立一个个性化转移矩阵 $A^u\in\mathbb{R}^{m\times m}$[14]。设标签 $t_l$ 转移到标签 $t_i$ 为一个状态对$(t_l,t_i)$,且 $t_l\in B_{t-1}^u$,$t_i\in B_t^u$,$C_{t_l,t_i}^u$ 表示用户 $u$ 的历史标签集合 $B^u$ 中状态对$(t_l,t_i)$的出现次数。对于每一个用户 $u$,标签 $t_l$ 到标签 $t_i$ 的转移概率公式为:

$$A_{t_l,t_i}^u=\begin{cases}C_{t_l,t_i}^u/\sum\limits_{t_i}^{l}C_{t_l,t_i}^u, & \sum\limits_{t_i}^{l}C_{t_l,t_i}^u\neq 0\\ 0, & \sum\limits_{t_i}^{l}C_{t_l,t_i}^u=0\end{cases} \quad (5)$$

根据图 2 中的数据,基于个性化马尔可夫链模型给出的转移矩阵如图 4 所示。

| User1 | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ |
|---|---|---|---|---|---|
| $t_1$ | 0.5 | 0.5 | 1 | 0.5 | 0 |
| $t_2$ | 1 | 0 | 1 | 0 | 0 |
| $t_3$ | 0.5 | 0.5 | 1 | 0.5 | 0 |
| $t_4$ | ? | ? | ? | ? | ? |
| $t_5$ | ? | ? | ? | ? | ? |

| User2 | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ |
|---|---|---|---|---|---|
| $t_1$ | ? | ? | ? | ? | ? |
| $t_2$ | 1 | 0 | 1 | 1 | 0 |
| $t_3$ | ? | ? | ? | ? | ? |
| $t_4$ | ? | ? | ? | ? | ? |
| $t_5$ | 1 | 0 | 1 | 1 | 0 |

| User3 | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ |
|---|---|---|---|---|---|
| $t_1$ | ? | ? | ? | ? | ? |
| $t_2$ | ? | ? | ? | ? | ? |
| $t_3$ | 0.5 | 0 | 0.5 | 1 | 0.5 |
| $t_4$ | 1 | 0 | 0 | 1 | 0 |
| $t_5$ | 1 | 0 | 0 | 1 | 0 |

| User4 | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ |
|---|---|---|---|---|---|
| $t_1$ | ? | ? | ? | ? | ? |
| $t_2$ | ? | ? | ? | ? | ? |
| $t_3$ | ? | ? | ? | ? | ? |
| $t_4$ | 0 | 0 | 1 | 1 | 0 |
| $t_5$ | ? | ? | ? | ? | ? |

图 4 个性化转移矩阵

### 3.2 用户对标签的兴趣度模型

#### 3.2.1 构建用户对标签的兴趣度

本节在研究用户、标签二维关系时,基于马尔可夫链模型按照已知用户最近一次行为所含标签集合和用户个性化标签转移矩阵 $A^u$,根据式(4)来构造用户下一次行为对标签的偏好概率矩阵 $I(u,t)$,即用户对标签的兴趣度矩阵。图 5 给出了根据图 4 生成的用户对标签的兴趣度矩阵。

| | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ |
|---|---|---|---|---|---|
| $u_1$ | 1/2 | 1/6 | 2/3 | 1/6 | 0 |
| $u_2$ | 0 | 0 | 0 | 0 | 0 |
| $u_3$ | 1/2 | 0 | 0 | 1/2 | 0 |
| $u_4$ | 0 | 0 | 1/2 | 1/2 | 0 |

图 5 用户对标签的兴趣度矩阵

#### 3.2.2 利用协同过滤补全用户对标签的兴趣度

由于用户历史行为数据存在稀疏性,3.2.1 节所得的矩阵 $I(u,t)$ 也有数据稀疏性。因此,本节引入协同过滤的思想,结合标签的相似度以补全用户对标签的兴趣度矩阵,如果数据不存在稀疏性,则不执行该操作。

标签相似度定义:若认为同一物品上的不同标签具有某种程度相似,那么当两个标签同时出现在很多物品的标签集合中时,就可以认为这两个标签具有较大的相似度。同时考虑到标注标签的次数,本节采用 Cosine 系数来进行计算,计算公式如下:

$$cossim(t,t')=\frac{\sum\limits_{i\in N(t)\cap N(t')}n_{t,i}n_{t',i}}{\sqrt{\sum\limits_{i\in N(t)}n_{t,i}^2\sum\limits_{i\in N(t')}n_{t',i}^2}} \quad (6)$$

其中,$n_{t,i}$,$n_{t',i}$ 表示物品 $i$ 被打上标签 $t$ 以及标签 $t'$ 的次数,$N(t)$ 和 $N(t')$ 表示标注标签 $t$ 以及标签 $t'$ 的物品集合。

综上,计算出补全稀疏值后用户对标签的兴趣度为:

$$I'(u,t_j)=\sum_{i=1}^{l} I(u,t_i)*cossim(t_i,t_j) \tag{7}$$

本节的目的是通过马尔可夫链和基于标签的协同过滤思想来发掘用户偏好的标签集合,标签集的大小 $T$ 应根据实际情况进行相应调整。

## 4　基于概率模型的标签到物品的排序算法

通过个性化马尔可夫链,已知用户下一次查询的物品具有的 $T$ 个标签。通过发现的标签集合,匹配对应的 $n$ 件物品,再进行 Top-N 推荐,其中 $n\geqslant N$。在标签匹配物品的过程中,会出现物品的标签数多于推荐的 $T$ 个标签的情况。如根据图 2 的转移矩阵,设 $T=2$,系统会推荐 $t_3$,$t_4$ 两个标签,假设物品 $i_1$ 包含了 3 个标签,即 $i_1=\{t_1,t_3,t_4\}$,物品 $i_2$ 包含了 4 个标签,即 $i_1=\{t_2,t_3,t_4,t_5\}$,此时,本文的目标转化为根据已知标签集构建用户 $u$ 对物品 $i$ 的满意度模型,以此进行排序,最终生成用户对物品的 Top-N 推荐。

### 4.1　Top-N 推荐

本节将构建用户基于标签对于物品的满意度模型,该模型是基于用户对物品标注标签的次数进行构建的。但基于标签的物品推荐问题中,物品不单只含一个标签,而是包含一组标签。现有的排序算法通常假设物品标签集所含标签之间的关系是相互独立的。但事实上,标签之间的关系是相互影响的。比如:某一用户购房,用户对标签 $t_1$(该标签表示为价格)的满意度为 $S_1$,对标签 $t_2$(该标签表示为地铁房)的满意度为 $S_2$,但如果某房源 $i_1$ 同时含有标签 $t_1$ 和 $t_2$,则该用户对房源 $i_1$ 的满意度为 $S_3$,且 $S_3\geqslant S_1+S_2$。

首先要考虑包含多个标签的物品 $I_r:=\{t_{i1},\cdots,t_{ik}\}$,$S_{i1\cdots ik}$ 表示用户 $u$ 对标有标签 $t_{i1},\cdots,t_{ik}$ 的物品 $I_r$ 的满意度。令 $p_u(t_{i1},\cdots,t_{ik}|k)$为用户 $u$ 查询标签 $t_{i1},\cdots,t_{ik}$ 的概率,用户将查询 $k$ 个不同的标签,$k\in[1,l]$。为方便起见,删除符号 $u$,即表示为 $p(t_i1,\cdots,t_ik|k)$。当用户 $u$ 按照标签集合进行查询时,用户 $u$ 的总预期满意度可表示为 $S_{\text{total}}$,定义如下:

$$\begin{aligned}S_{\text{total}}=&\Big(\sum_{t_{i1}\in T}S_{i1}\,p(t_{i1}|k=1)\Big)p(k=1)+\Big(\sum_{t_{i1},t_{i2}\in I_r}S_{i1i2}\,p(t_{i1},\\&t_{i2}|k=2)\Big)p(k=2)+\cdots+\Big(\sum_{t_{i1},\cdots,t_{i|I_r|}\in I_r}S_{i1\cdots i|I_r|}\\&p(t_{i1},\cdots,t_{i|I_r|}|k=|I_r|)\Big)p(k=|I_r|)\end{aligned} \tag{8}$$

由于用户主动使用标签查询物品时,使用的标签数量大多只有 1 个或 2 个,因此本文仅考虑 $S_{\text{total}}$ 中的第一项和第二项并忽略其他的高阶项。此时,$S_{\text{total}}$ 表示如下:

$$\begin{aligned}S_{\text{total}}\approx&\Big(\sum_{t_{i1}\in I_r}S_{i1}\,p(t_{i1}|k=1)\Big)p(k=1)+\Big(\sum_{t_{i1},t_{i2}\in I_r}S_{i1i2}\,p(t_{i1},\\&t_{i2}|k=2)\Big)p(k=2)\end{aligned} \tag{9}$$

本文将定义两个概率:$p(t_{i1}|k=1)$ 和 $p(t_{i1},t_{i2}|k=2)$。概率 $p(t_{i1}|k=1)$表示用户 $u$ 查询 $t_{i1}$ 的概率,即 $p(t_{i1}|k=1)=p(t_{i1}|u)$,其可以通过使用基于记忆的协同过滤[12]或矩阵分解等推荐模型来计算。对于概率 $p(t_{i1},t_{i2}|k=2)$,假设用户查询不同的标签是相互独立的,一种简单的方法是将这个概率定义为 $p(t_{i1},t_{i2}|k=2)=p(t_{i1}|u)\cdot p(t_{i2}|u)$。然而,如上所述,用户对标签的兴趣可能受到其他标签的影响,这意味着标签之间的影响不能被忽略。因此,在本文的模型中,通过等概率来定义其联合概率为:

$$\begin{aligned}p(t_{i1},t_{i2}|k=2)=&\frac{1}{2}\big(p(t_{i2}|t_{i1})\cdot p(t_{i1}|u)+p(t_{i1}|t_{i2})\cdot\\&p(t_{i2}|u)\big)\end{aligned}$$

其中,$p(t_{i1}|t_{i2})$是从 $t_{i2}$ 到 $t_{i1}$ 的转移概率,可以从历史行为数据中计算,公式如下:

$$p(t_{i1}|t_{i2})=\frac{\sharp\text{ users who queried both }t_{i1}\text{ and }t_{i2}}{\sharp\text{ users who queried }t_{i2}} \tag{10}$$

概率 $p(t_{i1}|k=1)$和 $p(t_{i1},t_{i2}|k=2)$仅根据用户的历史行为数据来定义。但是,通过整合附加信息可以给出这些概率的更复杂的定义。例如,如果知道它们之间的关系是彼此互补(或可替代)的,就可以提高(或减少)查询两个标签的概率。此外,可以利用其他方法(例如分类或文本信息)来计算查询概率,给出满意度 $S_{i1i2\cdots ik}$ 的定义。通常,用户查询的标签越多,用户获得的满意度就越高。因此,本文在项目中设置 $S_{i1}=1$,$S_{i1i2}=2$,$\cdots$,$S_{i1i2\cdots ik}=k$,其他满意度函数可以根据不同的应用需求来定义。因此,预期的满意度 $S_{\text{total}}$ 可以重新表示为:

$$\begin{aligned}S_{\text{total}}\approx&\Big(\beta\cdot\sum_{t_i\in I_r}p(t_i|u)+\sum_{\langle t_i,t_j\rangle\in O_{I_r}}p(t_i|t_j)\cdot p(t_j|u)\Big)\\&p(k=2)\end{aligned} \tag{11}$$

其中,$O_{I_r}$ 代表 $I_r$ 中所有有序对$\langle t_i,t_j\rangle$,$\beta=\dfrac{P(k=1)}{P(k=2)}$ 是用户的相关参数。但事实上,为每个用户设置 $\beta$ 是十分困难的,因此本文将 $\beta$ 设置为 1,因为它能为满意度模型提供最佳的效果,从而提高了推荐的精准度。

### 4.2　补全 Top-N 推荐

根据上文可以匹配到具有 $T$ 个标签的物品,此时如果数据存在稀疏性,则有时可能会匹配不到或者只能匹配到很少的具有 $T$ 个标签的物品,因此当 $T<N$ 时,本节需要以协同过滤的方式,通过用户对物品的兴趣度来补全推荐给用户的 $N$-$T$ 满意度。

本节使用标签来计算物品之间的相似度,主要依据物品 $i$ 和物品 $i'$被用户标注同一标签的次数越多,物品 $i$ 和物品 $i'$ 就越相似的原则,同时考虑到只关注标签是否标注,$n_{t,i}$ 和 $n_{t,i'}$ 表示物品 $i$ 以及物品 $i'$ 是否被标注标签 $t$,被标注取值为 1,否则为 0,$N(i)$ 和 $N(i')$ 分别表示物品 $i$ 和物品 $i'$ 的标签集合,因此采用 Jaccard 系数来进行计算,计算公式如下:

$$Jaccardsim(i,i')=\frac{\sum\limits_{t\in N(i)\cap N(i')}n_{t,i}n_{t,i'}}{\sqrt{\sum\limits_{t\in N(i)}n_{t,i}^2}+\sqrt{\sum\limits_{t\in N(i)'}n_{t,i'}^2}-\sum\limits_{t\in N(i)\cap N(i')}n_{t,i}n_{t,i'}} \tag{12}$$

补全用户对物品的满意度,根据 $S_{at}(u,i)$获得用户 $u$ 对于推荐过的物品 $i$ 的满意度,计算推荐过的物品 $i$ 与待推荐新物品 $i'$的相似度,求得用户 $u$ 对于待推荐新物品 $i'$的满意度,表示为:

$$S'_{at}(u,i')=\sum_{i=1}^{n}S_{at}(u,i)*Jaccardsim(i,i') \tag{13}$$

最后使用用户对物品的满意度 $S'_{at}(u,r)$来补全用户对物品的 Top-N 推荐。

## 5　算法设计

本文提出的基于马尔可夫链和协同过滤的个性化标签推

荐算法，主要考虑将标签作为用户和物品之间的一种重要纽带资源，通过建立用户对标签的马尔可夫链模型，以及通过 $T$ 个标签构建用户对物品的兴趣度模型之后，利用标签之间的关联关系构建用户对物品的满意度排序模型。依据物品评估标签间的相似度以及依据标签评估物品间的相似度，以此生成 Top-N 推荐列表。具体算法描述如算法 1 所示。

**算法 1** MCTS

输入：训练集，测试集，T 值和 K 值

输出：目标用户的 Top-N 推荐

第 1 步 将用户-标签-物品三维关系分解为用户-标签、标签-物品两个二维关系。在用户-标签关系中按照式(5)计算用户的个性化标签转移矩阵，再按照式(4)计算用户下一步行为对标签的兴趣度。

第 2 步 计算标签的相似度以补全用户对标签的兴趣模型 $I'(u,t)$。依据两个不同的标签同时被标注在很多不同物品上的次数来评估标签之间的相似度，按照式(6)采用 Cosine 系数计算得出标签的相似度，依据式(7)得到补全稀疏值后用户对于标签的兴趣度。

第 3 步 构建用户对物品的满意度模型。依据用户偏好标签集 $|T|$ 匹配物品，将匹配到的物品按照式(11)构建用户对物品的满意度模型 $S_{at}(u,i)$。

第 4 步 计算物品的相似度，根据不同物品被同一标签标注的次数评估物品之间的相似度，采用 Jaccard 系数按照式(12)计算得到物品的相似度。

第 5 步 预测目标用户对待推荐新物品的满意度。根据构建的用户对物品的满意度模型，以及待推荐新物品与已知标注物品的相似度 Jaccadsim(x,y)，按照式(13)计算用户对新物品的满意度。

第 6 步 将目标用户对新物品的预测满意度以从大到小的方式排序，取出排名前 N 的物品作为 Top-N 推荐列表并输出。

## 6 实验结果

### 6.1 数据集

为了评估算法的性能，本文采用真实的数据集 Movielens[1)] 和房谱网(House-Book)进行实验验证。两个数据集中的用户、物品、标签行为记录数如表 1 所列。

表 1 实验数据集 Movielens 和 House-Book

| 数据集 | 用户数 | 物品数 | 标签\行为记录 |
|---|---|---|---|
| Movielens | 18052 | 25308 | 753171 |
| HouseBook | 4200 | 1203 | 123896 |

Movielens 是历史最悠久的推荐系统，允许用户用标签标注电影。本文使用该数据集提供的 ml-latest 数据集，该数据集有 18052 个用户、25308 部电影的标签和 75 万条用户标注标签记录。房谱网数据集是 4200 个用户、1203 个楼盘和 12 万条用户标注标签记录。

### 6.2 实验设置

在传统的经典评估指标[2]中，Top-N 推荐一般使用准确率(Precision)和召回率(Recall)。实验中，由于准确率和召回率之间存在权衡问题，因此本文还采用 F-measure 同时评估准确率和召回率。准确率表示用户对最终生成的推荐列表感兴趣的概率，召回率表示用户真实喜欢的物品包含在最终生成的推荐列表中的概率。本文将数据集随机分成 90% 和 10% 两部分，前者作为训练集，后者作为测试集。令 $R(u)$ 表示根据用户在训练集上的行为给用户做出的 Top-N 推荐列表，$T(u)$ 表示系统向用户推荐标签后，用户实际选择的物品集。评估指标准确率和召回率以及 F-measure 的定义如下：

$$precision=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|R(u)|} \tag{14}$$

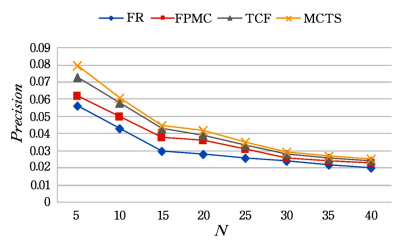
$$recall=\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|T(u)|} \tag{15}$$

$$F\text{-measure}=\frac{2*Precision*Recall}{Precision}+Recall \tag{16}$$

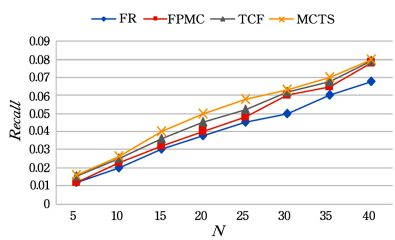
### 6.3 实验结果

本小节将本文提出的 MCTS 算法同 3 种经典的算法(基于标准张量分解的算法[3](FPMC)和 FolkRank 算法[7](定义为 FR)以及基于标签的协同过滤算法[10-11](定义为 TCF))进行实验比较，得到各算法的准确率、召回率以及 F-measure。

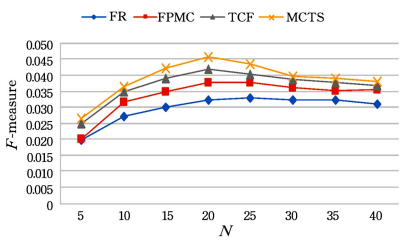
从图 6、图 7 可以看出，随着推荐物品数量的增加，MCTS 算法、FPMC 算法、FR 算法以及 TCF 算法推荐结果的准确率降低，召回率有明显提高。当 $F$ 值最高时，说明此时推荐的综合效果最佳。相较现有的推荐算法 FPMC，FR 以及 TCF，融合了 MC 模型、满意度模型以及协同过滤思想的 MCTS 算法的准确率、召回率和 $F$ 值都有所提高，说明本文提出的算法不仅能够有效利用用户的历史连续行为记录，还能够改善数据稀疏的情况，进一步论证了标签可以表达用户的兴趣和物品的特征，能有效地提高推荐的准确度。




(a) Precision



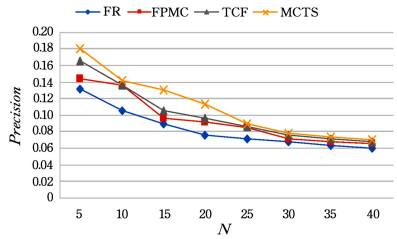

(b) Recall



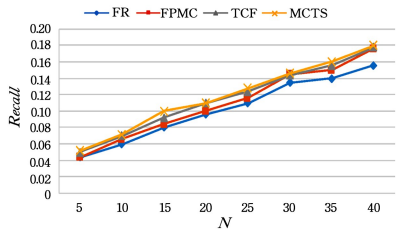

(c) F-measure

图 6 各算法基于 Movielens 数据集的准确率、召回率和 F-measure 的比较

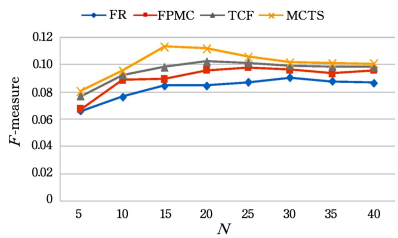
1) http://movielens.org




(a) Precision




(b) Recall




(c) F-measure

图 7　各算法基于房谱网数据集准确率、召回率和 F-measure 的比较

**结束语**　针对现有的标签推荐算法依赖用户使用标签的频率进行推荐带来的准确率低、数据稀疏等问题，通过引入马尔可夫链模型，设计并实现了基于标签和用户连续行为的个性化推荐算法。该算法将用标签作为中间纽带建立的〈用户-标签-物品〉三维关系进行拆分，分别获得用户对标签的兴趣度模型和用户对物品的满意度模型。一方面，相比基于标签的协同过滤算法，本文算法的准确率更高，同时引入协同过滤的思想降低数据稀疏性；另一方面，相比传统的基于马尔可夫链模型的推荐算法来说，使用标签更能反映出用户的兴趣和物品的特征，从而提高推荐的准确率，且标签也具备更好的解释说明性。同 3 种经典的推荐算法进行实验比较，应用两个真实的数据集 Movielens 和 HouseBook 作为实验平台。实验结果表明，本推荐算法在准确率和召回率上均有一定程度的提高。此外，如何对标签间多重影响的关系进行有效建模，以更好地体现不同权重的标签对 Top-N 推荐结果的影响，将是我们未来的研究工作。

## 参 考 文 献

[1] CHEN W, HSU W, LEE M L. A unified framework for recommendations based on quaternary semantic analysis[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011: 1023-1032.

[2] 项亮，陈义，王益. 推荐系统实践[M]. 河北：人民邮电出版社，2012：39-43.

[3] RENDLE S, FREUDENTHALER C, SCHMIDTTHIEME L. Factorizing personalized Markov chains for next-basket recommendation[C]// International Conference on World Wide Web. ACM, 2010.

[4] 文俊浩，袁培雷，曾骏，等. 基于标签主题的协同过滤推荐算法研究[J]. 计算机工程，2017(1).

[5] SHANI G, HECKERMAN D, BRAFMAN R I. An MDP-Based Recommender System [J]. Journal of Machine Learning Research, 2005, 6(1): 1265-1295.

[6] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine[C] // International Conference on World Wide Web. Elsevier Science Publishers B. V., 1998: 107-117.

[7] HOTHO A, JÄSCHKE R, SCHMITZ C, et al. Information Retrieval in Folksonomies: Search and Ranking[J]. Semantic Web Research & Applications, 2006, 4011: 411-426.

[8] PHAM T A N, LI X, CONG G. A General Model for Out-of-town Region Recommendation[C] // International Conference. 2017.

[9] PIRASTEH P, JUNG J J, HWANG D. Item-Based Collaborative Filtering with Attribute Correlation: A Case Study on Movie Recommendation[M] // Intelligent Information and Database Systems. Springer International Publishing, 2014.

[10] 刘健，张琨，陈旋. 基于标签和协同过滤的个性化推荐算法[J]. 计算机与现代化，2016(2)：62-65.

[11] 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.

[12] YE M, YIN P F, LEE W C, et al. Exploiting Geographical Influence for Collaborative Point-of-interest Recommendation [C]// International AcmSigir Conference on Research & Development in Information Retrieval. ACM, 2011.

[13] 李贵，王爽，李征宇，等. 基于张量分解的个性化标签推荐算法[J]. 计算机科学，2015，42(2)：267-273.

[14] 李贵，陈召新，韩子扬，等. 基于谱聚类群组发现和马尔可夫链的个性化推荐算法[J]. 计算机科学，2014，40(10)：44-48.

[15] 李贵，吴炎，孙平，等. 基于个性化马尔可夫链的推荐算法[J]. 计算机科学，2013，40(10)：319-322.

[16] 陈洁敏，李建国，汤非易，等. 融合“用户-项目-用户兴趣标签图”的协同好友推荐算法[J]. 计算机科学与探索，2018.