一个面向新兴趣点发现的模糊兴趣挖掘算法*)

唐 灿 唐亮贵 刘 波

(重庆工商大学计算机与信息工程学院 重庆 400067)

摘 要 本文在分析现有个性化推荐算法的基础之上,针对其难以发现新的用户兴趣点的不足,提出了一种面向新兴趣点发现的协作算法,建立了包括新兴趣点的多商品模糊兴趣模型。实验表明,该模型比现有模型更优。

关键词 个性化,模糊兴趣模型,推荐系统,新兴趣点

A Fuzzy Interest Data Mining Algorithm for New Interest Oriented

TANG Can TANG Liang-Gui LIU Bo

(Department of Computer Science, Chongqing Technology and Business University, Chongqing 400067)

Abstract Based on the analysis the of current personal recommendation generation algorithms, to resolve the problem that can't find new interest, a fuzzy Interest pattern for new Interest oriented is proposed in this paper. We build a serial of recommend algorithms to create the pattern. The experiment results indicate the effect is good.

Keywords Personalization, Fuzzy interest pattern, Recommender systems, New-interest point

1 引言

个性化服务是互联网时代商家克敌制胜的法宝,电子商务网站越来越多地使用个性化推荐来吸引人群。传统的个性化推荐分为基于用户的协作推荐[1]和基于商品项的协作推荐[2]。在个性化研究的初期,基于用户兴趣模型的协作过滤推荐技术进入人们的研究视野,基于用户兴趣模型的个性化推荐具有推荐具体化的特点,对单一商品的推荐很有利,因此在虚拟商店上得到了较多的支持。

然而,当虚拟商店的生意越来越好,多元化的商品经营理念自然就会出现在商家的脑子里。如同 Amazon[4] 这样的网站,以网上书店起家,现已将商品扩大到从 CD、DVD 到家居各类商品。当虚拟商店由单一商品供应向提供琳琅满目的各类商品的商场过渡后,如何建立针对用户的多类商品的模糊兴趣集就成为摆在眼前必须面对的问题。

传统的用户兴趣模型只有用户购买商品才能建立,无法主动为用户发现新的兴趣点,无法实现跨商品类的推荐。此后,基于商品项的推荐成为第二代的推荐算法,它在多类商品推荐中大显身手,成为现有推荐算法的主流。然而,此推荐模型却难以成为用户个性化的最佳体现,难以实现针对某用户的特殊推荐。用户兴趣模型重新进入研究的视野,如何能够解决基于用户兴趣模型的协作推荐的不足,找到支持新兴趣点发现的出路呢?

2 面向新兴趣点发现的模糊兴趣模型的建立

经研究证明,模糊兴趣模型比传统用户兴趣模型具有更好的个性化表现力和推荐效果[3],但以前的模糊兴趣模型仍然没有解决新兴趣点发现的问题,无法做到更好的新类推荐,针对这一问题,本文尝试提供一种解决的思路。

要无中生"有",发掘出用户新的兴趣点,从而建立完整多类商品模糊兴趣集。其根本的思路就是根据已经有的用户兴趣模型来派生出新的用户兴趣模型。这需要一系列的步骤。

首先,先给出多类商品集需要使用的一些定义:

定义 1(互为异类商品) 设现有两个不同的商品 A、B,A 的属性集为 $\{S_{11}, S_{12}, S_{13}, \cdots, S_{1i}, \cdots, S_{1n}\}$,$B$ 的属性集为 $\{S_{21}, S_{22}, S_{23}, \cdots, S_{2j}, \cdots, S_{2n}\}$,$N_1=\{S_{11}, \cdots, S_{1i}\}$ 是商品 A 的主属性,$N_2=\{S_{21}, \cdots, S_{2j}\}$ 是商品 B_2 的主属性,如果 $N_1 \neq N_2$,则商品 A,B 是互为异类商品。记为 $A<>B$,否则为同类商品,记为 $A=B$。

定义 2(同类商品集 c) $c=\{A_1, A_2, \cdots, A_n | A_1=A_2=\cdots=A_n\}$

定义 3(多类商品集 C) $C=\{c_1, c_2, \cdots, c_n\}$。

通过以上的定义,我们可以看出,多类商品集 C 中的每一子元素所对应的主属性都互不相同。

定义 4 被选子类商品集 C_{select}:

$C_{select}=\{c_1, c_2, \cdots, c_r | c_1, c_2, \cdots, c_r \subset C; U \subset c_1, \cdots, c_r;\}$

式中 U 表示用户已选择商品集,r 表示有 r 类商品被选择。

定义 5 缺项商品集 C_x:

$C_x=\{c_1, c_2, \cdots, c_m | c_1, c_2, \cdots, c_m \subset C; c_1, c_2, \cdots, c_m \not\subset C_{select};\}$

2.1 模糊兴趣关联规则

在文[3]中探讨了如何通用用户兴趣收集来建立单类商品的模糊兴趣集,设单类商品 R 对应的模糊兴趣集为 I_r,则可以给出多类简单模糊兴趣集的定义。

定义 6 多类简单模糊兴趣集 I_s:

$I_s=\{I_1 \cup I_2 \cup \cdots \cup I_r | I_1, \cdots, I_r \subset I\}$

由于 I_s 由用户选择的多个单类模糊兴趣集直接并列而成,这样的用户兴趣集未加任何处理,故称为多类简单模糊兴趣集。

为实现从多类简单模糊兴趣集中找出新兴趣点的目标,可选择关联规则挖掘算法。关联规则挖掘算法大多是从事务集出发进行挖掘的。事务集的特性在于其中的每一个元素是一个独立的事务。一个普通的事务对于商品交易而言,通常包含一次购物,或购物车中的商品记录等动作。事务一般具有时间独立性,事务间一般具有空间独立性。

仔细分析多个用户的 I_s 可以发现:用户之间具有天然的

*)基金项目:重庆市科攻关项目(CSTC,2005AC2090),重庆市教委科技项目(KJ060704)。唐 灿 讲师,主要研究方向:电子商务,软件理论。

空间无关性，单个用户的 I_s 兴趣点之间相互隔离、互不相关，I_s 集也可以看作是用户的事务集。也可以使用通用的数据挖掘算法来寻找相应的兴趣点之间的模糊兴趣关联规则。

设 $I_{all}^{out}=\{I_{s1}^{out},I_{s2}^{out},\cdots,I_{sl}^{out},\cdots,I_{sn}^{out}\}$ 为多类简单模糊兴趣总集，它与传统关联规则挖掘中的事务集相对应，则模糊兴趣关联规则可以描述为：

$X=>Y,X\subseteq I_s^{out},Y\subseteq I_s^{out}$

其中，X 被称为此模糊兴趣规则的前件，B 被称为此模糊兴趣规则的后件。“Love $a(0.6)$→Love $B(0.3)$”就是一个模糊关联规则的例子。

跟通常的关联规则挖掘一样，模糊兴趣关联规则挖掘算法同样需要产生强模糊关联规则。这就需要定义两个函数来估计此关联规则。

通常有三种方式来定义这样两个函数[6]：

- 扩展支持度和信任度来支持模糊关联规则；
- 得用统计学中的相关工具，衡量候选模糊关联规则的有效性。

本文采用模糊逻辑中的近似推理原理来判断候选模糊兴趣关联规则。采用信任度和平蕴涵度来评价关联规则的有效性[4]。

定义 8 I_{all}^{out} 中对 X 的支持度 $I_{supp}(X)$ 为：

$$I_{supp}(X)=\frac{\sum_{i=1}^{n}I_{supp_i}(X)}{|I_f|}$$

其中 $|I_f|$ 表示 I_{all}^{out} 的数据量。$I_{supp_i}(X)$ 为数据库中第 i 条数据的支持度：

$I_{supp_i}(X)=T(x_{1i},x_{2i},\cdots,x_{pi})$

其中 T 是广义三角模，当 $p=2$ 时，可采用以下形式：$T(a,b)=\min(a,b),T(a,b)=\max(a,b),T(a,b)=ab$ 等。

定义 9 $X\Rightarrow Y$ 规则的蕴涵度为：

$$I_{imp}(X=>Y)=\frac{\sum_{i=1}^{n}I_{imp_i}(X=>Y)}{|I_f|}$$

其中，$I_{imp_i}(X=>Y)=FIO(I_{supp_i}(X),I_{supp_i}(Y))$

FIO 表示某一具体形式的模糊蕴涵算子[6]。

2.2 模糊关联规则挖掘算法

使用模糊兴趣关联规则挖掘算法可以生成模糊兴趣关联规则。为了避免用户兴趣数据太多引起的噪声干扰，需要对已有的模糊兴趣模型进行一些预处理，以达到数据清洗的目的。

① 对多类简单模糊兴趣集的剪枝

由于待处理的数据 I_s 是模糊数据，相当多的内容并非传统的“是”或“否”的关系，没有进行量化，却正好满足模糊关联规则挖掘的要求。由于直接处理此 I_s 将包括相当多的冗余，因此需要对它进行剪枝。剪枝过程将采用以下策略：

if $I_i<minHotspot$ then 剪枝

剪枝后的 I_i 记为 I_i^{out}

② 模糊兴趣关联规则的挖掘

接下来，本文将利用支持度和蕴涵度进行模糊关联规则的挖掘。

定义 8 给定最小支持度 Min_supp 和最小蕴涵度 Min_imp，如果候选模糊兴趣关联规则 $X\Rightarrow Y$ 满足：

- $I_{supp}(X\Rightarrow Y)\geqslant$Min_supp
- $I_{imp}(X\Rightarrow Y)\geqslant$Min_imp

则 $X=>Y$ 被称为强模糊兴趣关联规则。

如果一条关联规则的支持度很高，则表示这条规则在兴趣总集中出现的频率很高；如果一条关联规则的蕴含度很高，表明这条规则可以为此关联规则提供精确的预测，同样也表明这条规则很重要。

发现兴趣关联规则的任务就是发生形如 $=>XY$ 这样的规则，规则的支持度必须大于或等于给定的最小支持度，规则的蕴含度同样也应当大于或等于给定的最小置信度。发现兴趣关联规则一般分为二步走：

- 从模糊数据库中发现所有的频繁集；
- 从频繁集中产生候选关联规则，针对每一关联规则计算其支持度和蕴含度，然后用 Min_supp 和 Min_imp 取得所有强关联规则。

其算法表示如下：

```
Algorithm 1  ArrayList getStrongRules()
///输入：I：模糊数据库，FIO：模糊蕴涵算子，Min_supp 最小支持度，
        Min_imp 最小蕴涵度
///输出：StrongR 强模糊关联规则集
{
  L=GenFrequent_Set(If, Min_supp);//调用函数 GenFrequent_
  set，生成 If 的频繁集 L
  for each f as Li in L
          {//f 为候选规则
              if (Iimp (f)≥Min_imp)
                    StrongR. add(f);//插入此规则到强集中
          }
}//end getStrongRules
```

在算法 1 中我们使用了算法 2 提供的函数用于产生频繁集：

```
Algorithm 2  ArrayList GenFrequent_Set(Df, Min_supp)
///输入：If：模糊数据库，Min_supp 最小支持度
///输出：L 表示全部的频繁集
///功能：产生全部频繁集
{
  L1={frequent 1-Sets};
  for(k=2;Lk-1！=Null,K++)
  {//循环到 L 为空
      Ck=Apriori-Gen(Lk-1);//利用 Apriori 算法生成事务集
      for each t in If
      {
        Ct=subset(Ck,t);//生成 t 的子集
        for each c in Ct
    {
          for n=1 to k
          {
            c. supp+=T(t1i,t2i,…,tpi)//T 为三角模算子，我们取
  的是 min;
          }//得到 c. supp
        }
        if (c. supp>=Min_supp×|Df|) Lk. add(C);
        //插入 Lk
  return L=∪Lk//取所有的合集
}//End of Algorithm 2
```

2.3 包含新兴趣点的完整商品模糊兴趣集的建立

包含新兴趣点的完整商品模糊兴趣集的建立需要经过以下步骤：

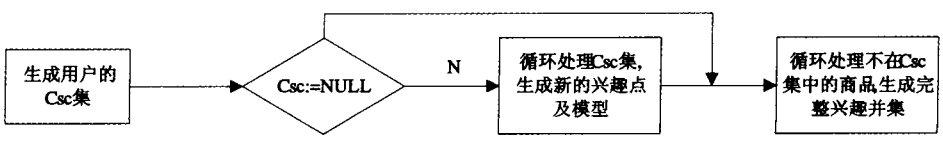

图 1 完整商品模糊兴趣集流程图

针对缺项商品集 Csc，我们可以通过循环使用算法 3 来生成一个完整的模糊兴趣集。

```
Algorithm 3  ArrayList Gen_NewFIP(u,IR)
///输入:u 缺少的商品，IR 用户的兴趣集，输出:Iu 商品类 u 的兴趣集
{   Ir= cutIntereSet(IR);//剪枝 IR
tmpN=GetStrong(StrongR ,Ir);//过滤 StrongR 规则库中不含 Ir 项的规则，
tmpN. Sorting; //排序
  if(n=1;n<=10;n++)//并取得前 10 名给 Iu
  {
    Iu. Add(tmpN(n));//加上第 n 项
  }
IR. Insert(Iu);//插入 Iu 到 IR 中
return IR;
}
```

在此算法过程中，我们选择了 Y 位于商品 u 和 X 位于 I_R 中符合 Strong 条件的 Top-10 规则 $X=>Y$，然后直接使用 X 作为新的 I_u 的生成项。这个生成项就是最后作用于 I_R 的生成集。

整个生成的过程图示如图 2 所示。

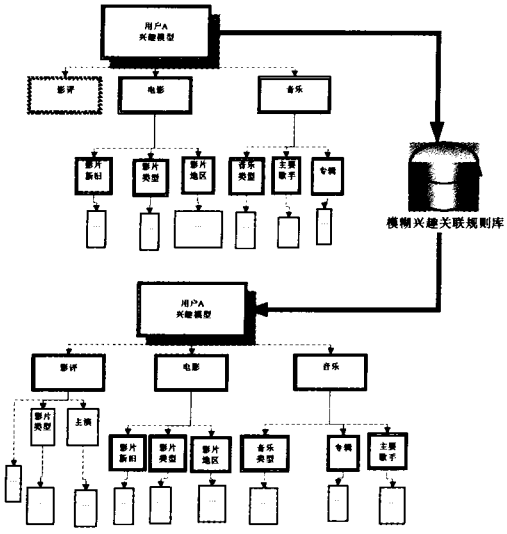

图 2　包含新兴趣点的完整模糊兴趣模型生成示例

在上图中，我们展示三个不同的兴趣类：电影、音乐和影评。其中，对于用户 A 来说，影评属于缺项集 Csc。通过挖掘所有的用户兴趣模型可以得到所有的强模糊关联规则，生成模糊关联规则库，然后对用户 A 可以派生出新的兴趣点，从而生成了一个完整的多类商品模糊兴趣模型。

3　算法原型实验及实验结果分析

本文最后生成的模型简称为 FIP 模型。为证明其有效性，我们一共使用三种不同类型的广义商品来作推荐实验。这三种不同类型的商品分别是：音乐、电影、书评。在原形实验中，针对三类广义商品，我们使用了 Item-based 和 FIP-based 推荐算法进行比较。由于 Item-based 推荐算法不存在生成用户兴趣集的问题，因此也就不存在缺项推荐。而本评测是为了了解缺项模糊用户模型推荐算法的可能性，故仍以缺项推荐为推荐的背景。

测试使用了 10 位用户，他们故意地在本系统中进行缺项选择，然后使用 FIP-based 算法生成缺项用户模型，最后根据用户模型进行推荐。其统计的推荐算法是基线是 200 次单项点击。推荐过程中，为了以示区别，我们剔除了所有同类的商品，然后取非本类商品的前 10 名进行推荐。

由于缺项商品中并没有用户的任何选择，故无法直接生成用户模型，而传统的 User-User CF 推荐必须要有用户模型，因此我们无法比较 User-User CF 推荐算法。

评测中，用户只选择了影评类和音乐类，而尝试用这两类来生成影片类推荐。然后，在实验中比较 Item-based 算法与 FIP-based 算法的不同。其结果如图 3 所示。

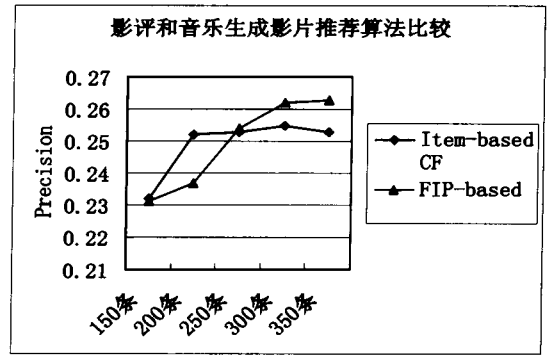

图 3　影评生成影片

仔细分析上图可以发现：

· FIP-based 与 Item-based 在推荐表现上还是比较明显的不同。随着用户兴趣数据的收集越多，FIP-based 表现越好。

· Item-based 随着数据的增加，并不会出现明显的上升，而是会逐渐平缓。

结果，新推荐模型表现出了更好的推荐效果。

小结　本文探讨了建立一个完整的多类商品模糊兴趣模型的方方面面，通过已经有模糊用户兴趣模型的挖掘，通过模糊关联规则挖掘法来生成了新的兴趣点，并将其加入到用户的兴趣模型中，从而建立了相对完整的多类商品模糊兴趣模型，并用实验证明其有效性。

参 考 文 献

1　Resnick P. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. MIT Center for Coordination Science, 1994

2　Amento R. Experiments in Social Data Mining: The TopicShop System. ACM Transactions on Computer-Human Interaction, 2003 ,10(1)

3　唐灿，朱征宇. 基于模糊兴趣模型的个性化推荐算法. 计算机工程与应用，2006(9)

4　高雅. 模糊关联规则的挖掘算法. 西南交通大学学报，2005(2)

5　www. amazon. com 采样时间：2006. 12

6　毛国君，等. 数据挖掘原理与算法. 清华大学出版社，2005