基于关联规则挖掘的分类随机游走算法

施海鹰
(上海大学 计算机工程与科学学院,上海 200444)

摘 要:随着互联网技术的不断进步和互联网的飞速发展,人们可以很方便地在互联网上寻找各种各样的信息。用户在寻找他们真正感兴趣的信息时会花费大量的时间,从而导致效率不高,这种现象被称作“信息过载”。推荐系统是解决信息过载问题的一种行之有效的方法。目前,推荐系统中应用最广泛的两种推荐技术是基于内容的推荐算法和协同过滤推荐算法,但其不能很好地处理冷启动和稀疏性问题。为了更好地解决这两个问题,在对传统分类随机游走算法进行改进的基础上,提出了基于关联规则挖掘的分类随机游走算法。该算法利用关联规则挖掘的特性,挖掘用户属性与项目之间的关联,为新用户构造初始的评分向量,弥补了经典算法的不足,较好地处理了冷启动问题。验证实验结果表明,该算法具有较好的有效性和精确性。

关键词:推荐系统;关联规则;分类随机游走算法;信息过载
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2017)09-0007-05
doi:10.3969/j.issn.1673-629X.2017.09.002

Random-walk Classification Algorithm with Association Rules Mining

SHI Hai-ying
(Schoolof Computer Engineering and Science, Shanghai University, Shanghai 200444, China)

Abstract: Along with the continuous progress of Internet technology and the rapid development of the Internet, people can easily find all kinds of information on the Internet. Users would spend a lot of time to search for information what they are really interested in, which is inefficient. The phenomenon is called information overload which is solved effectively by recommendation system as an effective method. However, the two most popular recommendation technologies in the current recommendation system are content-based recommendation and collaborative filtering recommendation, which cannot handle the problems of cold start and sparsity well. In order to better solve them, categorical random-walk algorithm based on association rules is proposed, which uses association rules to mine the association between user attributes and items and constructs the initial score vectors for new users. It has made up for the shortage of the classic algorithm and better handles the cold start problem. The results of experiments prove its effectiveness and accuracy.

Key words: recommendation system; association rules; categorical random-walk; information overload

1 概 述

随着互联网的不断发展,用户在互联网上查询感兴趣的信息的效率越来越低,用户将大量时间花在了浏览不相关的信息上。为了解决上述问题,推荐系统应运而生。推荐系统的任务主要是将信息和用户有效地联系起来,一方面让用户发现自己需要的、感兴趣的、有价值的信息;另一方面让这些信息出现在用户面前。推荐系统的出现就是为了解决“信息过载”问题[1],让用户在查询有价值的信息时具有更高的效率。推荐系统已经广泛应用于各领域,最典型的就是类似于‘亚马逊’的商业领域[2]。

推荐系统通过分析用户的历史兴趣和偏好信息,可以在项目空间中确定用户现在和将来可能会喜欢的项目,进而主动向用户提供相应的项目推荐服务。推荐系统的优劣很大程度上依靠其采取的推荐算法,不同的推荐技术具有不同的推荐质量,产生的推荐结果也不同。推荐技术可以分为基于内容的推荐[3]、协同过滤[4]、混合式推荐[5]。基于内容的推荐是最早应用的推荐算法,不需要征求用户的评价意见,而是根据用户喜欢的商品信息,分析信息的特征,根据这些信息来分析相似性。最早的基于内容的推荐应用于信息检索中,所以很多与信息检索相关的技术都可以用于基于

收稿日期:2016-07-08 **修回日期**:2016-11-02 **网络出版时间**:2017-07-11
基金项目:上海市科委重点资助项目(91330116)
作者简介:施海鹰(1979-),男,硕士研究生,研究方向为数据挖掘、人工智能;导师:杨洪斌,副教授,研究方向为数据挖掘、人工智能。

内容的推荐中。

协同过滤是推荐系统中最常用也是最成功的推荐算法。协同过滤技术的基本思想是,相似的用户喜欢的东西也可能是相似的。在现实生活中,如果一个用户想要读一本书,往往会选择和自己品味相近的朋友的推荐。基于内容的推荐算法和协同过滤算法各有各的优缺点,可以通过这两种推荐技术的组合来实现推荐算法的互补,这就是混合推荐算法。

当一个新的系统上线时,这个系统中并没有任何用户对项目的评分信息可以利用,那么协同过滤技术也就无法使用,这就是冷启动问题[6]。除此之外,当一个新用户或者新的项目加入系统时,这些项目既没有得到其他用户的评分,新用户也没有对系统中的其他项目进行过评分,这样,新项目既不会被推荐出去,同时新用户也不会得到推荐。新用户和新项目都面临着冷启动的问题。由于基于内容的推荐算法和协同过滤算法无法很好地解决冷启动问题,研究者们提出了新的算法,例如基于图的推荐算法[7-8]。

文中研究了 Zhang Liyan 等提出的分类随机游走算法[9],并在此基础上进行改进,提出了一种新的分类随机游走算法。首先建立用户—项目的相关图,在相关图上利用基于项目分类的随机游走不断迭代计算推荐结果。算法避免了传统推荐算法的缺点,同时具有较好的推荐效果。

关联规则挖掘问题是 R. Agrawal 等于 1993 年提出的[10]。关联规则是描述数据库中一组数据项之间的某种潜在关系的规则,即从数据集中识别出频繁出现的属性值集,也称为频繁项集[11],然后再利用这些频繁集的创建描述关联规则。关联规则可以找到数据项中的关联关系[12],应用到推荐算法中可以找到新用户的相关性。

为了克服分类随机游走算法不能为新用户进行推荐的缺点[13-14],将关联规则算法引入到推荐系统,提出了基于关联规则挖掘的分类随机游走算法,较好地解决了冷启动问题。

2 基于关联规则挖掘的分类随机游走算法

2.1 相关图模型

对于给定的数据集 D,该推荐算法涉及的相关数据包括用户集 $U=\{u_1,u_2,\cdots,u_{|U|}\}$、项目集 $M=\{m_1,m_2,\cdots,m_{|M|}\}$ 和用户对项目的评分 $r_{i,j}$,算法的输入可以看成是一个用户—项目矩阵 $\boldsymbol{T}$,矩阵中的元素 $T_{i,j}$ 的值为 $r_{i,j}$。

算法第一步是建立一个项目间的相关图,以此表明各个项目之间的相关性。算法用同时选择项目 m_i 和项目 m_j 的用户数量来表示两个项目间的相关性,即在矩阵 $\boldsymbol{T}$ 中第 i 列和第 j 列的值均不等于 0 的行的数量。

定义 $u_{i,j}$ 表示矩阵 $\boldsymbol{T}$ 中第 i 列和第 j 列的值均不等于 0 的行的数量,当 $i=j$ 时 $u_{i,j}=0$。由于 $\boldsymbol{T}$ 中大部分数据都为 0,为了减轻计算量,计算时对 $\boldsymbol{T}$ 中的每一行,将其不为 0 的提取出来,将其中可能的两两组合的 $u_{i,j}$ 置为 1,之后所有行将对应位置的值相加,求得最后所有的 $u_{i,j}$ 的值。

定义 $|\boldsymbol{M}|\times|\boldsymbol{M}|$ 阶的矩阵 $\tilde{\boldsymbol{M}}$,其中元素 $\tilde{M}_{i,j}$ 的值为 $u_{i,j}$。

定义 $|\boldsymbol{M}|\times|M|$ 阶的相关矩阵 $\boldsymbol{M}$,其中元素的值计算如下:

$$M_{i,j}=\tilde{M}_{i,j}/\omega_j$$

其中,ω_j 表示矩阵 $\tilde{\boldsymbol{M}}$ 第 j 列所有数值的和,如果 $\tilde{M}_{i,j}=0$,则 $M_{i,j}=0$。

基于矩阵 $\boldsymbol{M}$ 构建相关图 G,在 G 中,项目 m_i 和项目 m_j 之间存在一条边(当且仅当 $M_{i,j}>0$),边的权重等于 $M_{i,j}$。这样就构建了算法的基本模型—相关图。相关图表明了各项目间的关联度,项目 m_i 和项目 m_j 具有高的关联度,潜在的原因是它们具有某些相似的特征。

假设用户–项目矩阵 $\boldsymbol{T}$ 如下:

$$\boldsymbol{T}=\begin{bmatrix}3&4&5&0&0\\4&0&0&0&3\\0&3&0&0&4\\2&0&0&2&0\\4&4&0&0&0\\3&2&5&0&0\\0&0&4&5&0\\4&2&0&4&0\\0&3&0&0&0\\0&5&0&2&0\end{bmatrix}$$

根据定义,矩阵 $\tilde{\boldsymbol{M}}$ 为:

$$\tilde{\boldsymbol{M}}=\begin{bmatrix}0&4&2&2&1\\4&0&2&2&1\\2&2&0&1&0\\2&2&1&0&0\\1&1&0&0&0\end{bmatrix}$$

则矩阵 $\boldsymbol{M}$ 为:

$$\boldsymbol{M}=\begin{bmatrix}0&4/9&0.4&0.4&0.5\\4/9&0&0.4&0.4&0.5\\2/9&2/9&0&0.2&0\\2/9&2/9&0.2&0&0\\1/9&1/9&0&0&0\end{bmatrix}$$

根据 $\boldsymbol{M}$ 构建相关图 G,如图 1 所示。

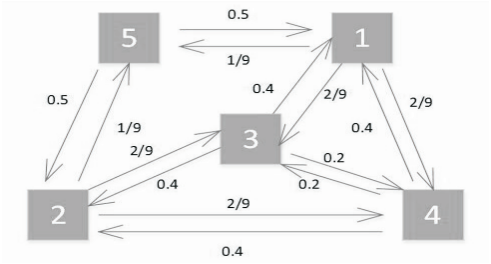

图 1　生成的相关图

2.2　算法描述

算法的基本思想是通过相关图来预测用户的喜好,相关图揭示了项目与项目之间的关联程度。训练数据集时,通过给定的用户-项目评分在相关图中依靠项目间的连接来传递。例如,如果一个项目 m_i 和用户 u_j 感兴趣的多个项目之间都存在关联,则 m_i 可以作为一个好的推荐结果推荐给用户 u_j。如果一个项目是用户所喜好的,则其一定还与该用户其他的喜好物品有较高的关联度。基于这样的结论,可以用随机游走的方法去计算用户对其他物品的评分。

用户对某一项目的评分可以从一个项目向该项目邻近的项目传递,不仅仅因为这两个项目同时出现在一个用户的喜好项目里的次数多,而且因为两个项目之间具有某些可见和不可见的相似性。由此定义一个新的概念—分类评分(Categorical Rank,CR),表示在特定分类上的项目评分,而分类随机游走算法就是迭代计算每个用户的 CR 值。

为了计算 CR 值,先迭代计算矩阵 $\boldsymbol{R}^{u_k}$,具体的迭代公式如下:

$$\begin{cases} R_{i_g}^{u_k}(0) = \dfrac{1}{|M| \times n}(1 \leqslant i \leqslant |M|, 1 \leqslant g \leqslant n) \\ F_{i_g(t)} = \left(\sum_{g=1}^{n} R_{i_g}^{u_k}(t-1)\right) \times P_{i_g} \\ R^{u_k}(t) = d\alpha M R^{u_k}(t-1) + d(1-\alpha)MF(t) + \\ \qquad (1-d)R^{u_k} \end{cases} \tag{1}$$

其中,$\boldsymbol{R}$ 为 $|\boldsymbol{M}| \times n$ 阶矩阵,$|\boldsymbol{M}|$ 是项目的总数,n 是项目类别的总数,元素 R_{i_g} 表示对于某一用户,项目 m_i 在类别 g 上的评分;$\boldsymbol{M}$ 为项目相关矩阵;$\boldsymbol{F}$ 为 $|\boldsymbol{M}| \times n$ 阶的辅助矩阵;$\boldsymbol{I}$ 为 $|\boldsymbol{M}| \times n$ 阶的矩阵,其元素的值由原始的用户—项目评分生成;d 和 α 分别为链接相关性参数和主题相关性参数,通过实验分别取 0.15 和 0.1。

式(1)表明,对于某一用户,项目 m_i 在类别 g 上的评分由三部分组成:与项目 m_i 所属类别相同的近邻项目的评分;与项目 m_i 所属类别不同的近邻项目的评分;起始的用户—项目评分。第一个方程的作用是在开始迭代之前初始化 $\boldsymbol{R}^{u_k}$;第二个方程用来计算矩阵 $\boldsymbol{F}$,P_{i_g} 表示项目 m_i 属于类别 g 的概率,由数据集中给出的项目分类信息可以计算得出;第三个方程中,I^{u_k} 中每一个元素的计算公式如下:

$$I_{i_g} = r_i^{u_k} \times P_{i_g} \tag{2}$$

其中,$r_i^{u_k}$ 表示用户 u_k 对项目 m_i 的评分。

式(2)可以使矩阵 $\boldsymbol{R}$ 在迭代过程中更快地达到收敛。当对用户 u_k 迭代完成获得矩阵 $\boldsymbol{R}^{u_k}$ 后,开始计算用户 u_k 对每个项目的 CR 评分,为此需要用到 $|\boldsymbol{M}| \times k$ 阶矩阵 CR(k 是项目的类别总数),$\mathrm{CR}_i^{u_k}$ 表示用户 u_k 对项目 m_i 的最终评估分,计算公式为:

$$\mathrm{CR}^{u_k} = \boldsymbol{R}^{u_k}(\mathrm{Prof}^{u_k})^{\mathrm{T}} \tag{3}$$

其中,$\mathrm{Prof}^{u_k} = \boldsymbol{R}^{u_k}P$,揭示了用户 u_k 对不同类别项目的兴趣。根据式(3)可以计算出用户 u_k 对所有项目的预测评分,如果一个项目最后的预测评分高,则表示用户对该项目比其他项目更感兴趣,将项目按照最后的预测评分逆序进行排序,把排在前面的项目作为最后的推荐结果推荐给用户。

2.3　改进的 FP-Growth 算法

分类随机游走算法的缺点是不能为新用户进行推荐,为了解决这个问题,需要为新用户构建一个初始的评分向量,故提出了基于关联规则的分类随机游走算法。通过关联规则找出用户属性与项目之间的关联关系,根据新用户的属性给新用户构建初始的评分向量,对经典关联规则 FP-Growth 算法进行了改进。

FP-Growth 算法需要扫描事务数据库,数据集 D 生成的用户-项目矩阵 $\boldsymbol{T}$ 就作为事务数据库。设定一个最小支持度,对 $\boldsymbol{T}$ 进行第一次扫描,抽取出那些支持度大于最小支持度的项目和用户属性,并记录其支持度计数(即其出现的次数),生成候选 1-项集,记为 L,将其按照支持度大小逆序排序。

第二次扫描 $\boldsymbol{T}$,对它的每一行,选择该行的频繁项(项目和用户属性),按照第一步中生成的 L 的顺序进行排序。之后构造 FP-tree,设定排序后的频繁项表为 $[p \mid P]$,其中 p 为第一个元素,P 为剩余元素组成的表。过程如下:如果节点 T 有子女 N 与 p 是同一个项目,则 N 的计数加 1,否则创建新的节点 N,将其计数置为 1,链接到它的父节点 T,并且通过节点链结构将其链接到具有相同项目的节点,如果 P 非空,递归调用 insert_tree(P,Tree)。

建立 FP-tree 对应的项头表(item header table),逆序遍历项头表,找出 FP-tree 中由该节点到根节点的路径。根据每个频繁元素对应的条件模式基,生成其对应的条件 FP-tree,并删除树中节点记数不满足给定的最小支持度的节点。对于每一棵条件 FP-tree,生成所有从根节点到叶子节点的路径,由路径中的集合

生成其所有非空子集,所有非空子集和每一个候选 1-项集中的元素共同构成了原始数据集中的频繁集,最后生成所有的用户属性与项目之间的关联规则。

运用改进的 FP-Growth 算法,根据已有的数据集 D 产生用户属性和项目之间的关联规则之后,对规则 $R: X \Rightarrow Y$ 中后项 Y 的每一项计算初始评分,对于 Y 中没有的项目,其初始评分设为 0,这样就构成了一个评分向量 $\boldsymbol{V}$;对于 Y 中有的项目,假设该项目在所有项目中是第 i 个项目,其初始评分的计算公式如下:

$$\text{Score} = \frac{\sum_{k=1}^{n} \text{Score}_i^k}{n} \times [1 + \text{Support}(X \cup Y) \times \text{Confidence}(R)] \tag{4}$$

其中,n 为规则前项 X 在所有用户集中出现的次数;$\sum_{k=1}^{n} \text{Score}_i^k$ 表示用户集中所有符合前项 X 的用户对后项 Y 中在项目集中排第 i 个项目的评分总和,乘式的右边表示对于支持度和置信度都高的规则,其中项目的初始评分也高。公式最后的值最高取 5。

现有一新用户 $u_k = u_{k1}, u_{k2}, \cdots, u_{kL}$,$L$ 为用户的属性总数,对于生成的规则集合 $R = \{R_1, R_2, \cdots, R_{|R|}\}$,其中的任一规则 $R_i: X_i \Rightarrow Y_i$,如果 X_i 包含于用户 u_k 的属性集,则将 R_i 加入到 R 的子集 R' 中,这样 R' 所有的规则都是前项中的用户属性是目标用户属性集的子集,即全部是与目标用户相关的关联规则,之后依据式(5)计算 u_k 的初始评分向量:

$$r^{u_k} = \frac{\sum_{i=1}^{|R'|} V_i \times \text{Support}(X_i \cup Y_i) \times \text{Confidence}(R_i')}{\sum_{i=1}^{|R'|} \text{Support}(X_i \cup Y_i) \times \text{Confidence}(R_i')} \tag{5}$$

其中,V_i 为 R' 中第 i 个规则所生成的评分向量。

式(5)根据 R' 中所有规则的评分向量的加权平均来生成 u_k 的初始评分向量。有了初始评分向量后,改进的分类随机游走算法就可以开始为 u_k 进行推荐。

3 算法流程

基于关联规则挖掘的分类随机游走算法在第一次运行时,需要进行用户属性与项目之间的关联规则挖掘,之后只需要在一个新用户对某个项目评分之后,再次进行关联规则挖掘。

算法首先建立相关图模型,之后对新用户进行推荐,需要根据挖掘的所有关联规则,选出与该用户相关的关联规则,对每个关联规则中的所有项目进行初始评分,最后对所有与用户属性相关的关联规则根据支持度和置信度计算加权平均值,得出为新用户构建的初始评分向量。将所有项目分类,计算项目类别概率矩阵 $\boldsymbol{P}$,根据式(1)初始化矩阵 $\boldsymbol{R}$,并迭代计算,之后再根据式(3)计算该新用户的 CR 值,将 CR 按照逆序排序,将前几个项目作为结果推荐给该用户。至此算法结束,具体流程如图 2 所示。

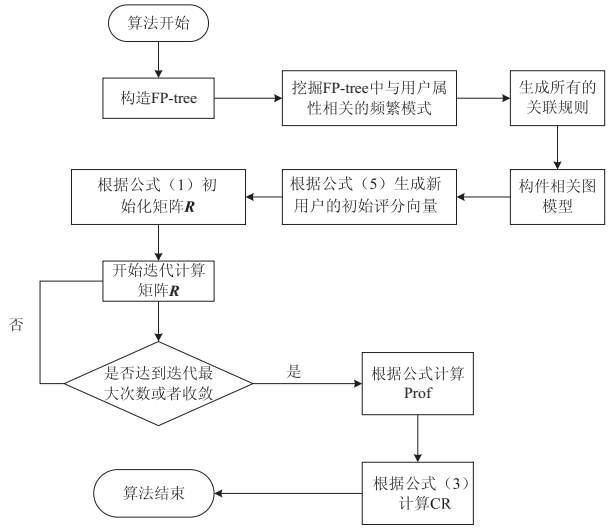

图 2 算法流程图

4 实验结果

实验环境:CPU 为 Intel(R) Core(TM) i3-2350M CPU @ 2.30 GHz;内存 4 GB;操作系统为 Windows 7(32 位);编程语言为 JAVA(JDK1.6)。

实验使用了三个数据集,分别是 MovieLens、Anonymous Microsoft Web Data 和 Entree Chicago Recommendation Data。MovieLens 由著名的 MovieLens 网站上的电影推荐组成,该网站拥有超过 50 000 名用户对3 000 多部电影进行评分,数据集中有 943 个用户对1 682部电影的 100 000 条评分,用户自己拥有三个属性,包括年龄、性别和职业。Anonymous Microsoft Web Data 记录了网站内 38 000 名匿名用户过去一周在网站上的浏览记录。Entree Chicago Recommendation Data 记录了用户对芝加哥餐馆主菜的评价。对每个数据集抽取 100 个用户的数据作为测试数据,将这 100 名用户作为新用户,在三个数据集上分别使用基于 FP-Growth 的分类随机游走算法和改进的 FP-Growth 分类随机游走算法,以均方误差(MSE)[13] 为评价指标。

在三个数据集中,分别作 10 次随机抽取 100 个用户的实验,结果如图 3~5 所示。

图 3~5 显示了三个数据集上算法的性能,算法成功地为新用户做出了推荐,且改进的 FP-Growth 分类随机游走算法的性能普遍好于基于 FP-Growth 算法的性能。将每个数据集上 10 次实验的 MSE 取平均,结果如表 1 所示。

从表1可见,改进的FP-Growth分类随机游走成

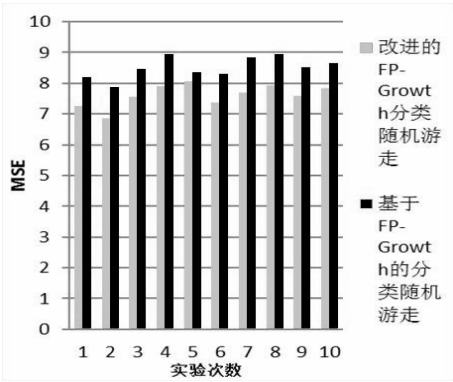

图 3　MovieLens 的实验结果

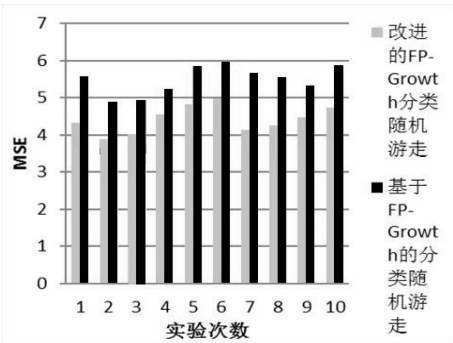

图 4　Anonymous Microsoft Web Data 的实验结果

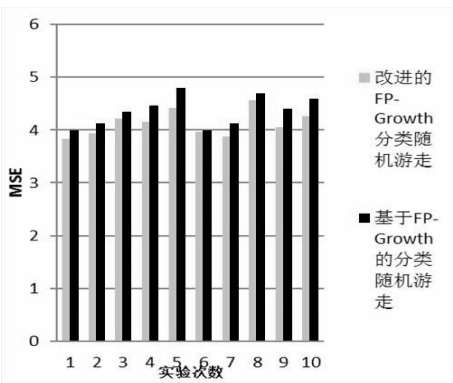

图 5　Entree Chicago Recommendation Data 的实验结果

表 1　结果对比　　s

数据集	基于 FP-Growth 的分类随机游走	改进的 FP-Growth 随机游走
MovieLens	8.50	7.60
Anonymous Microsoft Data	5.48	4.41
Entree Chicago Data	4.35	4.12

功解决了分类随机游走算法中新用户的评分向量问题,利用改进算法生成用户属性与项目之间的关联规则去构造新用户的初始评分向量,之后用分类随机游走算法为新用户进行推荐。相比于基于 FP-Growth 的分类随机游走算法可以看出,该算法对新用户具有较好的推荐结果。

5　结束语

针对随机游走分类算法的不足,提出了基于关联规则挖掘的随机游走分类算法。为了弥补分类随机游走算法不能为新用户进行推荐的缺点,即新用户没有任何评分数据的问题,该算法利用关联规则挖掘计算用户属性与项目之间的关联规则,利用这些关联规则为新用户构建一个初始的评分向量,之后为该用户计算推荐结果。实验结果表明,该算法对于新用户推荐具有较好的结果。随着信息量的扩大,算法效率也必定会受到一定的影响,为了提高算法在运算大数据量时的效率,算法的并行化和分布式计算是未来研究的重要方向。

参考文献:

[1] 邓爱林,朱扬勇,施伯乐. 基于项目评分预测的协同过滤推荐算法[J]. 软件学报,2003,14(9):1621-1628.

[2] 陈　健,印　鉴. 基于影响集的协作过滤推荐算法[J]. 软件学报,2007,18(7):1685-1694.

[3] 许海玲,吴　潇,李晓东,等. 互联网推荐系统比较研究[J]. 软件学报,2009,20(2):350-362.

[4] 曹　毅,贺卫红. 基于用户兴趣的混合推荐模型[J]. 系统工程,2009,27(6):68-72.

[5] 吴丽花,刘　鲁. 个性化推荐系统用户建模技术综述[J]. 情报学报,2006,25(1):55-62.

[6] 刘建国,周　涛,汪秉宏. 个性化推荐系统的研究进展[J]. 自然科学进展,2009,19(1):1-15.

[7] 梁昌勇,冷亚军,王勇胜,等. 电子商务推荐系统中群体用户推荐问题研究[J]. 中国管理科学,2013,21(3):153-158.

[8] Zhou T,Jiang L L,Su R Q,et al. Effect of initial configuration on network-based recommendation[J]. Physics,2008,81(5):58-62.

[9] Zhang L,Xu J,Li C. A random-walk based recommendation algorithm considering item categories[J]. Neurocomputing,2013,120(10):391-396.

[10] Agrawal R,Imieliński T,Swami A. Mining association rules between sets of items in large databases[C]//Proceedings of the 1993 ACM SIGMOD international conference on management of data. New York:ACM,1993:207-216.

[11] Minaei-Bidgoli B,Barmaki R,Nasiri M. Mining numerical association rules via multi-objective genetic algorithms[J]. Information Sciences,2013,233:15-24.

[12] Soni R K,Gupta N,Sinhal A. An FP-growth approach to mining association rules[J]. International Journal of Computer Science and Mobile Computing,2013,2(2):1-5.

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

[14] Tong Q L,Park Y,Park Y T. A time-based approach to effective recommender systems using implicit feedback[J]. Expert Systems with Applications,2008,34(4):3055-3062.