doi:10.3969/j.issn.1000-436x.2012.z1.003

基于情境和主体特征融入性的多维度个性化推荐模型研究

琚春华，鲍福光
(浙江工商大学 计算机与信息工程学院，浙江 杭州 310018)

摘 要：个性化推荐准确率的高低是互联网应用成功与否的关键因素，针对传统推荐模型的不足，提出一种基于情境和主体特征融入性的多维度个性化推荐模型，该模型能够充分利用地域文化背景、领域主题情景、主体特征等信息，避免了传统算法把用户整体作为单个向量的弊端，克服了数据稀疏性等问题。实验结果表明，该模型的推荐质量比传统的协同推荐模型高，更有针对性地向用户推荐他们感兴趣的项目。

关键词：推荐模型；个性化；多维度；情境；特征选取
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2012)Z1-0017-11

Research on a multidimensional personalized recommendation model based on a situation and characteristics of the users

JU Chun-hua, BAO Fu-guang
(School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China)

Abstract: The accuracy of personalized recommendation was the key factor of Internet application to success. Because of the deficiency of the traditional recommend model, a multidimensional personalized recommendation model based on a special situation and main characteristics of the users was proposed. This model could make full use of regional culture background, field scene, characteristics of the users and so on, avoided the disadvantages of traditional algorithm, put the user's overall characteristics as a single vector, and overcomed the problem of sparse data. The experimental results show that the quality of this recommendation model is better than traditional collaborative recommend model with the more specific items match user interests.

Key words: recommend model; personalized; multidimensional; situation; feature selection

1 引言

随着互联网技术的发展和网络应用在全球的不断升级，网络正影响着人们的工作和生活方式。然而，现有的网络信息服务还存在着明显的不足，比如资源量大且分散，对于不同用户提供的信息几乎是相同的，多数服务还处于有求则应状态。对于大量普通用户而言，目前互联网上的“信息过载”、“信息迷航”等问题日益严重。信息时代在给用户带来了丰富信息资源的同时，也给用户在迅速准确寻找所需信息方面造成了巨大的困难。推荐系统正是为了解决互联网信息过载等问题而提出的一种

收稿日期：2012-08-06
基金项目：国家自然科学基金资助项目（71071141）；国家教育部博士点基金资助项目（2103326110001）；浙江省自然科学基金（重点）资助项目（Z1091224）；浙江省新苗人才计划基金资助项目（2012R408063）；浙江工商大学研究生科研创新基金资助项目
Foundation Items: The National Natural Science Foundation of China (71071141); Doctoral Program Fund of National Ministry of Education (2103326110001); The Natural Science Foundation of Zhejiang Province (Z1091224); Zhejiang Provincial University Students Innovative Project (2012R408063); Zhejiang Gongshang University Graduated Students Scientific Innovative Project

智能优化系统，从大量的互联网及其应用的信息资源中推荐符合用户兴趣偏好的资源。但前期的推荐系统多数是非个性化推荐系统，如电子商务网站的销量排行、价格排行等。随着智能信息挖掘技术的不断发展，现代互联网及其应用迫切需要一种能够根据用户的主体特征及其行为组织来调整信息或推荐用户感兴趣项目的服务模式，将互联网及其应用从被动接受用户的请求转变为主动感知并分析用户的信息需求，实现对用户的个性化主动推送信息的服务，即个性化推荐服务。所谓个性化推荐服务是根据用户的兴趣特征及偏好和行为，向用户推荐用户感兴趣的信息或商品的服务模式[1]。例如在电子商务领域，随着电子商务规模的迅速扩大，商品数量和品项的急剧增长，用户花大量的时间和精力也不一定能找到自己感兴趣的商品。这种需要浏览大量无用信息的过程无疑会使消费者产生抵触情绪，从而造成客户流失。为了解决这些问题，个性化推荐系统的各方面内容已经成为了国内外学者的研究热点和实践家们的商业模式试探。个性化推荐服务系统是一种建立在海量数据挖掘基础上的高级商务智能平台，以帮助互联网及其应用平台为其用户提供个性化的推荐服务和决策支持。

2 推荐模型及问题分析

2.1 典型的推荐模型

自从1995年，卡耐基梅隆大学（Carnegie Mellon university）的Robert Armstrong等和斯坦福大学的Marko Balabanovic等在美国人工智能协会上分别推出个性化导航系统Web Watcher和个性化推荐系统LIRA以来，国内外研究和实践人员纷纷对不同的个性化推荐服务进行了深入广泛的研究。根据系统获取用户特征信息的方式、系统推荐所采用的模式以及推荐信息的个性化程度，可以将目前的信息推荐系统分为非个性化商务推荐系统、基于项目属性的推荐系统、基于历史行为的推荐系统、用户相关性协同过滤推荐系统和基于项目相关性的协同推荐系统等。

1) 非个性化商务推荐系统。向当前用户推荐的信息或项目一般是基于系统的点击率排行，销量排行，或是评价均值排行等。这类推荐技术是基于用户的单个会话，独立于用户。每位用户得到的推荐结果都是一致的。如Amazon提供的“客户评价”推荐，淘宝网提供的“价格排序、好评排序和热卖推荐”等。

2) 基于项目属性的推荐系统。根据用户浏览或搜索项目的属性特征产生相近属性项目的列表推荐给用户。这种方式类似于搜索引擎。通过分析项目的内容和属性来确定用户的兴趣[2]基于项目属性的推荐需要用户显式提供相应的属性特征，属于半自动推荐。如Amazon 提供“更多关于XX的商品”等。

3) 基于历史行为的推荐系统。这类系统会根据用户的历史购买行为和浏览行为，对用户当前情况产生推荐列表。这类推荐通常是基于用户多次会话实现的。

4) 用户相关性协同过滤推荐系统[3]。首先寻求当前用户的最邻近用户，再依据最邻近用户的历史购买行为和历史评分情况向当前用户推荐相应的项目。这种个性化推荐服务系统通常不需要用户的显式输入。所进行的推荐通常是基于用户多次会话实现的。

5) 基于项目相关性的协同过滤推荐系统。一种是Sarwar[4](2001)提出的，寻找与目标项目最相似的项目或信息来估计某用户对目标项目或信息的评分，将其中评分最高的项目或信息推荐给该用户。另一种是分析项目间存在的相关性，再向用户推荐其可能同样感兴趣的项目。例如Amazon 提供的“经常一起购买的商品”和“购买此商品的顾客也同时购买（customers who bought this X also bought）”推荐等。

Bettman(1998)[5]和Prahalad(2004)[6]比较早地研究复杂变化场景对用户行为和潜在需求的影响机制，研究发现场景的变化可能会影响用户的行为或决策。Gorgoglione和Palmisano(2008)[7]通过用户购物行为的实验证实在用户的行为模式中如果能考虑这些变化的场景因素可以提高对用户行为的预测能力。Panniello(2009)[8]研究认为考虑情景因素可以在用户购物历史记录的数据中识别出更具有相似性的行为模式，可以更好地预测用户潜在的需求，激发购买欲望。随着移动商务的发展，更多的信息推荐服务发生在复杂多变的场景中，因此研究者越来越关注情景对用户潜在需求、推荐准确性、推荐效率等的影响，产生了一些新的研究发展和成果。

2.2 问题分析

用户具有诸多的个性特征，不同类别的特征属

性对不同领域的推荐会有不同影响权重；不同的情境与文化社会环境同样会影响用户的个性化选择。如上海的“海派文化”、北京的“京派文化”、广州的“羊城文化”等都会影响处于这种文化背景中的个体进行消费行为，所处的时代情境或实时事件也对个体的消费行为起到了很大的作为。同时，处于相近社会背景或实时事件下的个体可以利用长尾理论创造更多的用户兴趣、提升销售效益。如宁波与上海之间的文化社会影响是长期、紧密、互相的，那么一件在上海兴起并流行的商品可以推荐给在宁波的相似用户，实现长尾理论的效益。然而，传统的个性化推荐系统难以实现上述融入文化背景和实时事件等情境的多维度个性化推荐。文献[13]和文献[14]提出了集成情境的多维度推荐系统, 将情境和基于用户协同过滤相结合的方法进行评分推荐。虽然它们考虑了情境, 但都是针对整个用户群体得到整体泛化的情境。

传统的协同推荐算法，通常把用户所有属性或项目属性作为一个向量，笼统地进行相似度的计算，以选择最邻近用户或项目。这类算法存在以下不足。1) 用户相似性只能反映不同用户对所有项目偏好的相似性。但事实上几乎不可能存在几个不同用户对所有项目都具备共同兴趣。2) 视所有的兴趣和偏好具有同等重要性，而事实上，不同的主体属性和兴趣偏好对不同类型的项目或领域的影响程度往往是不同的。因此传统的相似性也不具备实用性和代表性。3) 相似性中没有涉及情境信息，一是用户针对性，处于不同地域环境下的用户会有不同的兴趣偏好，这就需要引入宏观的情境；二是推荐领域或对象的针对性，针对不同的推荐领域或对象，影响用户偏好的属性会有所不同，这就需要融入微观的情境。情境信息往往对不同用户推荐具有很高的参考价值。此外，传统推荐还存在数据稀疏性、扩展性、冷启动等问题。

本文在分析传统推荐模型的基础上，引入城市社会背景情境包括待推荐的领域、文化社会环境以及微观情境（场景）。提出了基于情境和主体特征融入性的多维度个性化推荐模型。

3　融入情境和主体特征的协同推荐模型

本文提出构建一种融入情境的多维度个性化推荐模型的基本思想是：把传统的最邻近用户扩展到各个情境下的邻居，情境包括待推荐的领域、文化环境以及微观情境（场景）等。为了实现优质准确的个性化推荐服务，首先需要跟踪和学习用户的兴趣和行为特征及其所处的文化环境和实时事件等情境。根据当前用户和待推荐项目所处的情境，寻找当前用户的最邻近用户，再根据最邻近用户对某项目感兴趣程度进行合理推荐。

1) 城市（或地域）社会背景，属于宏观的情境。不同的城市在经济条件、产业结构、历史文化和开放程度等方面的差异性使得消费者的消费行为各有不同。2) 领域主题要求，属于微观的情境信息。在不同的推荐领域下，影响消费者购物行为的因素是不同的。3) 主体特征，即消费者的属性和行为。属性往往是静态的或者说是在一定时期内比较稳定，例如性别、学历、职业、出生年月等。行为特征主要是指消费者的消费行为和评价行为。

3.1　基本定义

定义 1　主体　用户 u 是在某网站上具有唯一访问账号的注册用户。主体信息特征包括：主体的相关注册信息，登录后进行点击、浏览、购买等行为信息。将网站上注册用户的集合定义为用户集 $U=\{u_1,u_2,\cdots,u_N\}$。

定义 2　主体属性　用户背景属性集 UBE 是用户 u 已知存在多种背景因素的集合，主要有地域(Region)、性别(Gender)、年龄(Age)、婚姻(Marriage)、教育程度(Education)、专业（Major）和收入(Income)等，则定义用户背景集为

UBE = {*Region, Gender, Age, Marriage, Education, Major, Income,*⋯}

定义 3　主体行为　用户行为集 UIB 是 u 在页面上访问时所有行为信息的集合。主要包括用户访问网页时的几类行为数据：标记行为，例如添加到收藏夹(Save)、增加书签(Book)、放入购物车等；操作行为，例如拖动滚动条(Scroll)、点击某个超链接(Click)、浏览页面的时间(Times)等；购买与评价行为，这是用户最关键且最有价值的行为。则定义兴趣行为集为

UIB = {*Book, Save, Scroll, Times, Click, Buy, evaluate,*⋯}

定义 4　情境　领域主题情景，属于微观的情景信息。情境对象是信息推荐过程中任何关联的对象，存在一组描述该对象特征的非空属性集 S_i = （$S_{i1}, S_{i2},\cdots, S_{im}$），每个属性 S_{ij} (j=1,2,⋯,m)都有一组可选的属性值 S_{ij} ={ $S_{ij1},S_{ij2},\cdots,S_{ijr}$ }；对于推

荐过程中的时刻 t， S_i 都唯一具有一个属性值 s'_{ij} ($s'_{ij} \in S_{ij}$)。相应地在时刻 t，情境对象 S_i 都具有特定状态 $s'_i = \{s'_{i1}, s'_{i2}, \cdots, s'_{im}\}$。在不同的推荐领域或主题下，影响消费者或用户行为的因素是不同的。

定义 5 用户兴趣内容集 UIC 表示在网站中所有用户可访问资源分类后的兴趣内容集合：

$$UIC=\{P_1,\cdots,P_l\} \cup \{L_1,\cdots,L_m\} \cup \{T_1,\cdots,T_n\}$$
$$= \{UIC_1, UIC_2,\cdots,UIC_M\}$$

其中，P 是一个网站组件频道；L 是一条超链接内容；T 是一个标签页面；UIC 是采用概念分层方法分类生成的兴趣内容，则有对应的兴趣概念集：$\Sigma = \{\sigma_x \mid 1 \leqslant x \leqslant Z\}$，$\exists UIC \mid\!\rightarrow \sigma_x$，$\sigma_x$ 为兴趣内容特征概念，$\mid\!\rightarrow$ 表示兴趣内容到特征概念的映射关系。

定义 6 令用户 u 在同一个会话时间段 T 中的访问过程可顺序记录为一条访问事务 tr，定义为多元组：

$\{tr.u,\ (tr.content_1,\ tr.time_1,\ tr.background_1,\ tr.behavior_1), \cdots, (tr.content_p,\ tr.time_p,\ tr.Background_p,\ tr.behavior_p)\}$。其中，$tr.u \in U$ 表示访问用户；四元组 $(tr.content, tr.time,\ tr.background,\ tr.behavior_1)$ 表示用户的每一次访问操作，$tr.content \in UIC$ 表示具体的兴趣内容对象，$tr.time(tr.time_p - tr.time_1 \leqslant T)$ 表示访问时间戳；$tr.background \in UBE$ 表示用户的具体背景因素；$tr.behavior \in UIB$ 表示用户的具体兴趣行为。因此,将所有访问事务 tr 按照会话时间顺序组成该用户在浏览网站过程中的访问事务集：$TR_u = \{tr_i \mid 1 \leqslant i \leqslant |TR_u|\}$，$|TR_u|$ 为用户的会话总数。

3.2 模型的构建

从系统学角度考虑，推荐模型由输入、推荐过程、输出等 3 部分组成，其中，输入部分主要包括 2 类数据，显性数据和隐性数据。显性数据是指用户自行注册输入的特征信息，比如年龄、性别、学历、收入等。隐性数据是来自于数据的提取过程，比如从用户历史购买记录中推测用户偏好等。网络用户的兴趣特征主要由外部和内部因素所决定。外部因素主要包括家庭、文化和社会经济因素；内部因素主要有职业、年龄、性别、收入、自我观念等。这些因素综合在一起对用户行为产生影响。不同用户之间存在着各方面差异，对商品的兴趣程度和所关注的重点也有所不同，且往往只是关注某一两个特定领域的资源子集。

用户兴趣模型(UIM, user interest model) 是个性化兴趣推荐服务的重要部分和依据，可以将得到的用户兴趣偏好用结构化形式动态地保存为个体用户兴趣模型[8]。领域本体(domain ontology) 是在特定领域内可以重用的，用来提供该领域地概念定义和概念之间的关系，提供该领域中的活动以及该领域的主要理论和基本原理等[9]。本文使用领域本体的一个子集, 即一个小型的领域本体来构建用户的初始个性化用户兴趣本体，实现领域本体的构建。个性化用户兴趣本体 Personal IO(personal interest ontology) 是基于用户研究领域构建的初始个性化用户兴趣本体，是领域本体的一个子集。个性化兴趣本体 Personal IO 是领域本体在不同用户需求描述的基础上通过本体投影获取的。

基于情境和主体特征融入性的用户模型构建框架如图 1 所示, 由个性化用户兴趣本体即用户模型的获取、更新和用户群的构建 3 部分组成。其中，个性化用户兴趣本体的获取包括获得用户的个人信息、构建领域本体等；用户模型的更新是根据用户浏览或检索信息的行为构建参考本体，把它归并到个性化用户兴趣本体中，实现用户模型的学习更新；用户群是每个个性化用户兴趣本体通过相似度计算得到的。推荐流程如图 2 所示。

3.3 模型的算法描述

定义 7 切片操作 O1（slicing） 按照城市进行切片操作，提取不同城市或区域的分数据集。

定义 8 城市相似性操作 O2 根据城市的经济状况、产业结构、历史文化、开放程度、文化形式和地理位置等因素，对所分析的城市或区域进行针对性地聚类。

定义 9 特征选取操作 O3 主体情境相关性分析，在具体主题情境下进行主体特征的属性与行为的关联分析，提取关键属性和行为，进而更新领域本体。

算法的描述主要有前期准备、算法的输入、输出和计算流程等。关键处理部分流程如图 3 所示。

1) 前期准备：城市文化背景分析。根据城市的经济条件、产业结构、历史文化和开放程度等不同或相似度进行分析和聚类。一般按城市划分区域，再按城市的文化相似度进行交叉推荐，例如上海和宁波的消费者。

2) 输入：①按城市或地域进行切片的用户行为数据 UIB；②用户注册集和属性数据表 U，UBE；③项目信息数据表 IT；④用户兴趣内容集。（UID 表示用户 ID，IID 表示待推荐项目的 ID）。

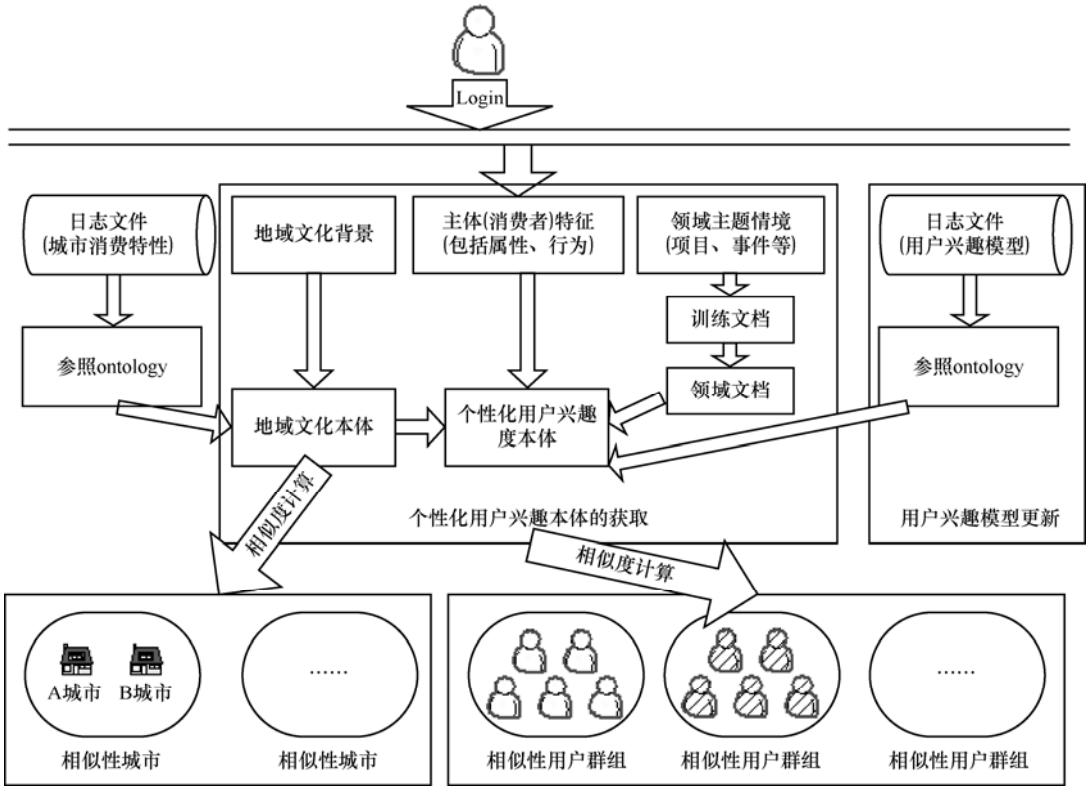

图 1　基于情境和主体特征融入性的用户兴趣模型

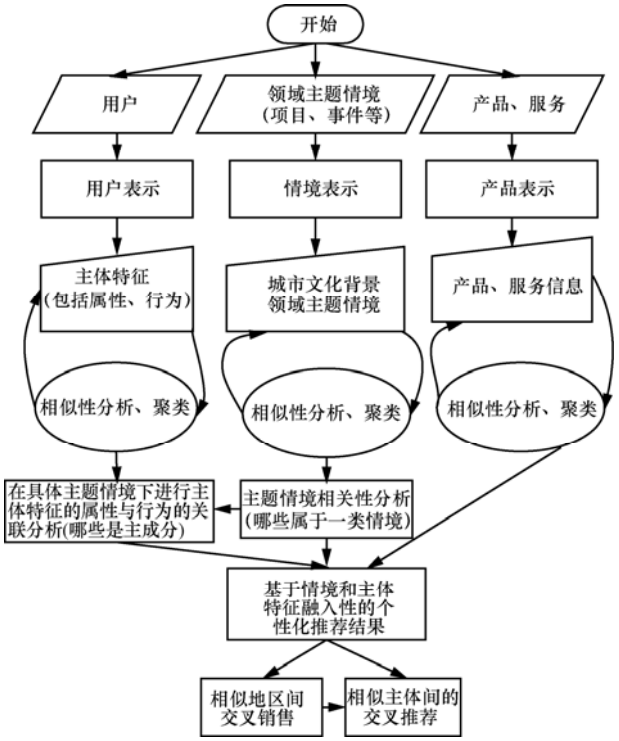

图 2　推荐流程

3) 输出：目标用户 *UID* 对待推荐项目 *IID* 的感兴趣程度。

4) 计算流程

① 针对推荐项目类别，对城市或地区的背景因素主成分选取，并对相似的城市聚类。

② 按城市或地区的聚类结果对用户行为数据和用户注册属性数据进行切片。

③ 在不同的推荐领域或主题下，根据项目信息数据(*IT*)和用户 *UIB* 数据，计算出情境内容集合

$$I=\bigcup_{i=1}^{k}C_i \tag{1}$$

其中，k 为情境个数，S_i 表示其中的第 i 个情境所包括的项目集合：$S_1=\{i_{11},i_{12},\cdots,i_{1_{j_1}}\}$，$S_2=\{i_{21},i_{22},\cdots,i_{2_{j_2}}\}$，…，$S_k=\{i_{k1},i_{k2},\cdots,i_{kj_k}\}$ 由于一个项目可以属于多个情境，故 $Num(I)\leqslant sum\{j_1,j_2,\cdots,j_k\}$，其中，$Num(I)$ 为 I 所含项目的总数。

④ 根据情境集合 I，寻找项目的邻近项目。即计算项目 j 与项目 q 之间的相似性(品牌、价格、评分、销量)，记为 $Sim(j,q)$；再根据 $Sim(j,q)$ 大小顺序排列，确定邻近项目。

⑤ 根据情境集合 I，寻找该情境下的主要影响因素。即计算用户不同属性与用户行为以及用户不同行为之间的关联性，进行基于因子分析的特征选取或主成分分析。

⑥ a 根据情境集合 I，结合步骤④用户特征选

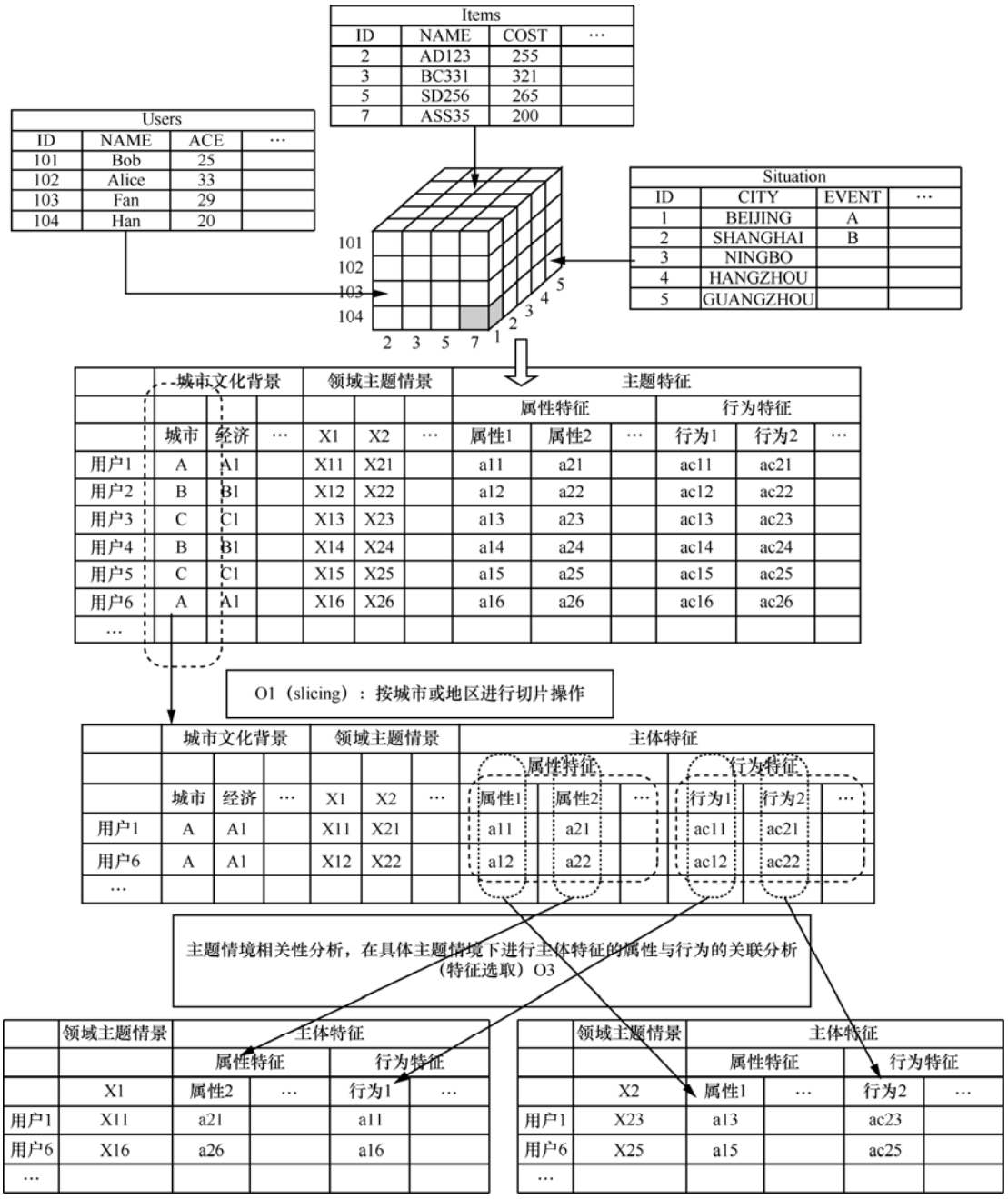

图 3　关键处理部分流程

取的关键属性、用户行为和用户评分数据表，计算用户在各个情境下的相似度，即主体与主体的相似性（评分、行为、属性、…），形成用户相似矩阵。相似矩阵 $\boldsymbol{Sim}(s,i,j)$ 为三维矩阵，每一个元素 $\boldsymbol{Sim}(SID,UID_i,UID_j)$ 表示在情境 SID 下用户 i 与用户 j 的余弦相似度。

$$\boldsymbol{Sim}(s,i,j)=\cos(\vec{\boldsymbol{V}}_{s,i},\vec{\boldsymbol{V}}_{s,j})=\frac{\vec{\boldsymbol{V}}_{s,i}\times\vec{\boldsymbol{V}}_{s,j}}{\left\|\vec{\boldsymbol{V}}_{s,i}\right\|\left\|\vec{\boldsymbol{V}}_{s,j}\right\|} \tag{2}$$

其中，$\vec{\boldsymbol{V}}_{s,i}$，$\vec{\boldsymbol{V}}_{s,j}$ 分别为用户 i，j 在同一情境 s 的属性特征和行为特征向量。

在面向情境的协同过滤推荐算法中，一个用户在每一个情境中均存在一个最近邻居集，用户 j 的最近邻居集为：$F_j=\{F_{j,s_1},F_{j,s_2},\cdots,F_{j,s_k}\}1\leqslant j\leqslant Num(U)$，其中，$Num(U)$ 为用户总数，$s_1,s_2,\cdots,s_k$ 为 k 个情境。F_{j,s_i} 为用户 j 在第 s_i 个情境中的最近邻居集合，其中的元素按照相似度降序排列，每个情境下的最近邻居集合中邻居用户个数可以相同，也可以不同，比如最近邻居个数与情境规模相关联等。

b 根据情境集合 I，结合步骤④用户特征选取

的关键属性、用户行为和用户评分数据表，运用 C5.0 算法进行决策分析。

⑦ 根据步骤③和⑤得出的在各个情境下的主体相似性和项目相似性进行推荐。根据得到的目标用户的最近邻居集合 $F_{UID,IID}$ 产生推荐：用户 *UID* 对项目 *IID* 的预测评分 $S_{UID,IID}$ 可以通过 $F_{UID,IID}$ 中各个用户对项目 *IID* 评分的加权平均得到，具体公式如下

$$S_{UID,IID}=\frac{\sum_{n\in F_{UID,IID}}[R_{n,IID}\cdot Sim(UID,n)]}{\sum_{n\in F_{UID,IID}}|Sim(UID,n)|} \tag{3}$$

其中，$R_{n,IID}$ 为用户 *n* 对项目 *IID* 的评分，$Sim(UID,n)$ 为用户 *UID* 与用户 *n* 的相似度。

算法中各个步骤可划分为 2 个阶段：步骤①到步骤⑤为机器学习阶段，步骤⑥和步骤⑦为项目推荐阶段。项目推荐步骤的时间复杂度为 $O(KN)$，其中，*K* 为待推荐项目所属场景个数，*N* 为用户个数 $Num(U)$，且 *M* 往往比较小，故算法的推荐速度较快，能够很好地满足在线推荐的要求。

4　实证研究

本文选取了全国 4 个直辖市和 15 个副省级城市 2010 年度的社会经济指标和消费性支出与结构的相关数据（2011 年中国统计年鉴）作为区域协同的实证研究，对本文模型进行了分析，如图 4 所示，其相关数据如表 1 所示。

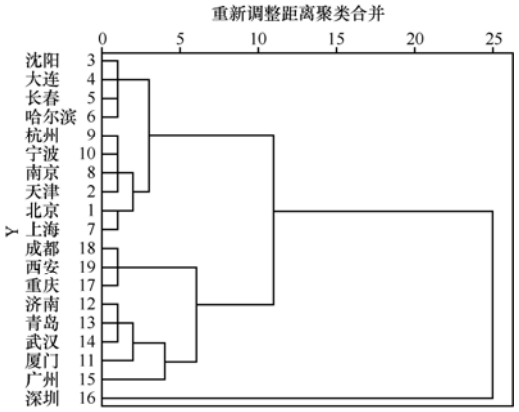

图 4　2010 年度基于地域划分的 19 市消费特征聚类树状示意

表 1　　2010 年度的社会经济指标和消费性支出与结构的相关数据

城市	总人口/万人	第三产业生产总值	城乡居民人均储蓄年末余额/元	岗职工平均工资/元	社会商品零售总额/万元	社会商品人均零售额/万元	普通高校在校生数/人	人均 GDP/元	城镇居民人均消费支出/元	城市居民人均可支配收入	剧场影剧院/个	区域
北京	1 257.8	10 600	134 157	65 682	62 292 986	49 525	577 828	112 208	19 934	29 073	182	1
天津	984.85	4 238	57 210	52 963	29 025 506	29 472	429 224	93 663	16 562	24 293	27	1
沈阳	719.6	2 254	46 390	41 899	20 658 671	28 708	348 583	69 726	16 961	20 541	46	3
大连	586.44	2 188	57 547	44 615	16 397 554	27 961	245 756	87 957	16 580	21 293	6	3
长春	758.89	1 356	27 180	35 721	12 867 475	16 955	365 734	43 867	16 580	17 922	23	3
哈尔滨	992.02	1 867	26 008	32 411	17 701 558	17 843	481 589	36 943	13 939	17 557	70	3
上海	1 412.32	9 833	115 053	71 875	60 705 033	42 982	515 661	121 544	23 200	31 838	86	2
南京	632.42	2 660	55 530	48 781	22 887 409	36 190	793 405	81 127	18 156	28 312	28	2
杭州	689.12	2 896	72 425	48 771	21 460 790	31 142	434 811	86 329	20 219	30 035	42	2
宁波	574.08	2 073	57 695	43 476	17 044 539	29 690	140 818	89 935	19 420	30 166	24	2
厦门	180.21	1 012	76 836	40 282	6 850 248	38 012	128 470	11 315	19 961	29 253	5	4
济南	604.08	2 057	36 215	37 853	18 024 610	29 838	642 541	64 735	15 973	25 321	12	5
青岛	763.64	2 630	38 137	37 803	19 611 331	25 681	284 788	74 199	17 531	24 998	40	5
武汉	836.73	2 863	42 911	39 302	25 704 037	30 719	881 433	66 520	14 490	20 806	64	6
广州	806.14	6 557	115 393	54 494	44 763 780	55 528	843 934	133 330	25 011	30 658	23	7
深圳	259.87	5 051	258 477	50 455	30 007 629	115 471	67 324	368 703	22 806	32 381	16	7
重庆	3 303.45	2 881	17 677	35 367	28 780 433	8 712	565 868	23 991	13 335	17 532	33	8
成都	1 149.07	2 785	44 134	38 604	24 176 000	21 039	617 482	48 311	15 511	20 835	15	8
西安	782.73	1 691	46 986	37 871	16 370 367	20 914	734 350	41 412	16 543	22 244	31	8

表 2　　**用户的背景因子及初始主题兴趣表（部分）**

序号	性别	学历	职业	地域	婚否	月收入/元	年龄	专业	兴趣 1	兴趣 2	兴趣 3
1	女	本科	公司职员	长春	是	14 000	46	建筑	服装	美容	食品
2	男	博士	教师	成都	否	13 000	31	计算机	服装	文体	娱乐
3	男	高中及以下	—	上海	是	8 000	42	—	数码	文体	娱乐
4	男	博士	公务员	广州	是	15 000	35	法律	服装	家居	
5	男	硕士	国企职员	杭州	是	12 000	39	电子	服装	数码	家居
6	男	博士	教师	宁波	否	23 000	30	食品	数码	食品	文体
7	女	本科	教师	宁波	否	8 000	32	学前教育	服装	数码	娱乐
8	男	硕士	公务员	哈尔滨	否	5 400	28	工商管理	数码	文体	
9	女	本科	外企职员	上海	否	14 000	27	营销	配饰	美容	食品
10	女	本科	公司职员	杭州	是	5 000	33	信息管理	母婴	家居	服装
11	女	博士	外企职员	深圳	否	14 000	31	企业管理	美容	娱乐	
12	男	本科	教师	广州	是	4 500	38	体育	数码	家居	文体
13	女	本科	—	南京	是	5 000	40	—	家居	母婴	文体
14	女	本科	私营业者	沈阳	否	9 000	28	信息管理	配饰	服装	美容
15	女	硕士	学生	北京	否	2 000	25	营销	数码	配饰	文体
16	女	高中及以下	民营企业	宁波	是	6 500	46	—	服装	美容	
17	男	高中及以下	民营企业	天津	是	6 000	42	—	服装	美容	娱乐
18	女	本科	公司职员	大连	是	5 000	27	应用物理	服装	配饰	文体
19	女	博士	心理师	成都	否	12 000	26	心理学	娱乐	服装	食品
…	…	…	…	…	…	…	…	…	…	…	…

本实验的数据来自某购物网站，根据用户的注册信息以及历史购买行为数据得到用户的背景因素和最感兴趣的 3 个主题，其中，背景因素主要是地域、性别、学历、职业、年龄、婚否、收入和教育等方面，主题主要分为服装、配饰、美容、数码、家居、母婴、食品、文体、娱乐等类别。用户的背景因子及初始主题兴趣如表 2 所示，其中，兴趣主题按强弱排列，即兴趣主题 1>兴趣主题 2>兴趣主题 3，空表示兴趣没有出现。

针对“服装”主题，进行主体地域聚类和主体特征抽取，结果如图 5 和图 6 所示。

根据针对“服装”消费的文化背景相似性聚类分析，可以得到上海、杭州、南京、宁波等长三角地区的城市具有高度相似的“服装”消费特点。本文以上海、杭州、南京、宁波等长三角地区为例，进行基于因子分析的特征抽取。结果发现，年龄（age）、职业(occupation)、性别（gender）收入（inc.）婚姻（marriage）等属性是影响消费者对“服装”选择的最重要因素。教育、专业、组织等对“服装”选择影响较小。长三角地区的 4 个城市的服装消费数据分析结果显示：最注重“品牌”的人群是高收入的中青年男性，但是网购比例不高；最喜欢网上购买服装是中等收入的高学历女性和学生；最关注“性价比”的是中等收入的高学历人士，其中，已婚女性尤为突出。具体推荐规则如表 3 和表 4 所示，其中，部分推荐规则已经合并。

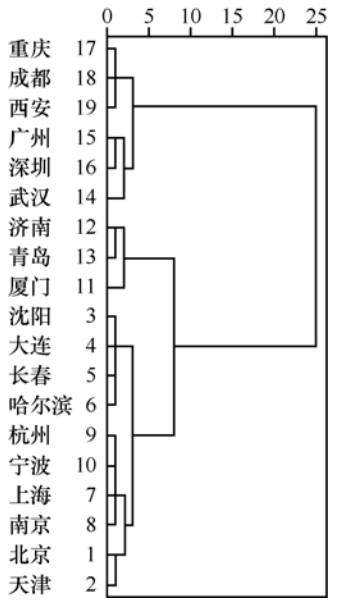

图 5　19 个城市“服装”消费特征聚类

图 6　影响消费者对“服装”选择的因素

表 3　　**长三角城市“服装”网上消费的推荐规则（部分）**

序号	性别★	学历	职业★	地域	婚否★	月收入/元★	年龄★	关注 1	关注 2	推荐类型	置信度
1	男	—	—	长三角	是	>10 000	>30&<50	优品牌	质量	成熟型、正装	89.5%
2	女	本科以上	—	长三角	—	>3 000 & <8 000	>20&<40	流行	性价比高	休闲装、正装	76.2%
3	女	本科以上	—	长三角	是	>3 000 & <8 000	>30&<40	性价比高	质量	休闲装、配饰	81.4%
4	男	本科以上	学生	长三角	—	<1 000	>20&<30	价格低	质量	青春型、休闲装	59.6%
5	男	—	公司职员	长三角	是	>3 000 & <5 000	>30&<40	价格低	实用	休闲装	77.3%
6	女	—	公务员	长三角	否	>3 000 & <8 000	>40&<50	质量	性价比高	休闲装、正装	86.7%

注：“★”表示影响用户对“服装”消费的重要因素；“—”表示该因素的值对该规则的影响不大。

表 4　　**长三角城市“女装”网上消费的推荐规则（部分）**

序号	职业★	婚否★	月收入（元）★	年龄★	关注 1	关注 2	推荐类型
1	—	是	>80 000	>30&<40	优品牌	做工	成熟型、旗袍、高腰裙
2	公司职员	是	>5 000 & <8 000	>30&<40	质量	性价比高	A 字裙、旗袍
3	—	—	>3 000 & <5 000	>20&<30	流行	质量	休闲装、配饰
4	学生	否	<1 000	>20&<30	质量	价格低	青春型、公主裙
5		是	>3 000 & <5 000	>30&<40	价格低	实用	低腰裙、休闲裤
6	公务员	否	>5 000 & <8 000	>40&<50	质量	性价比高	休闲装、正装

注：“★”表示影响用户对“服装”消费的重要因素；“—”表示该因素的值对该规则的影响不大。

5　实验评价

5.1　评价指标

推荐模型的优劣很大程度上取决于所推荐的项目是否符合用户的需求和兴趣，是否能够挖掘用户的隐性需求。与多数文献一样，本文采用较常见的统计精度方法，以预测推荐值和实际值的平均绝对误差(MAE, mean absolute error)作为推荐模型推荐效果的衡量标准。

MAE 法通过计算推荐模型产生的目标项目的预测值与用户的实际值之间的偏差来衡量推荐的准确性。

$$MAE=\frac{\sum_{i=1}^{n}\left|p_i-q_i\right|}{N} \tag{4}$$

式(4)中，p_i 表示对第 i 个用户推荐预测值的情况，用户推荐集合表示为$\{p_1,p_2,\cdots,p_n\}$；q_i 表示对应用户的实际值，用户实际值集合为$\{q_1,q_2,\cdots,q_n\}$。

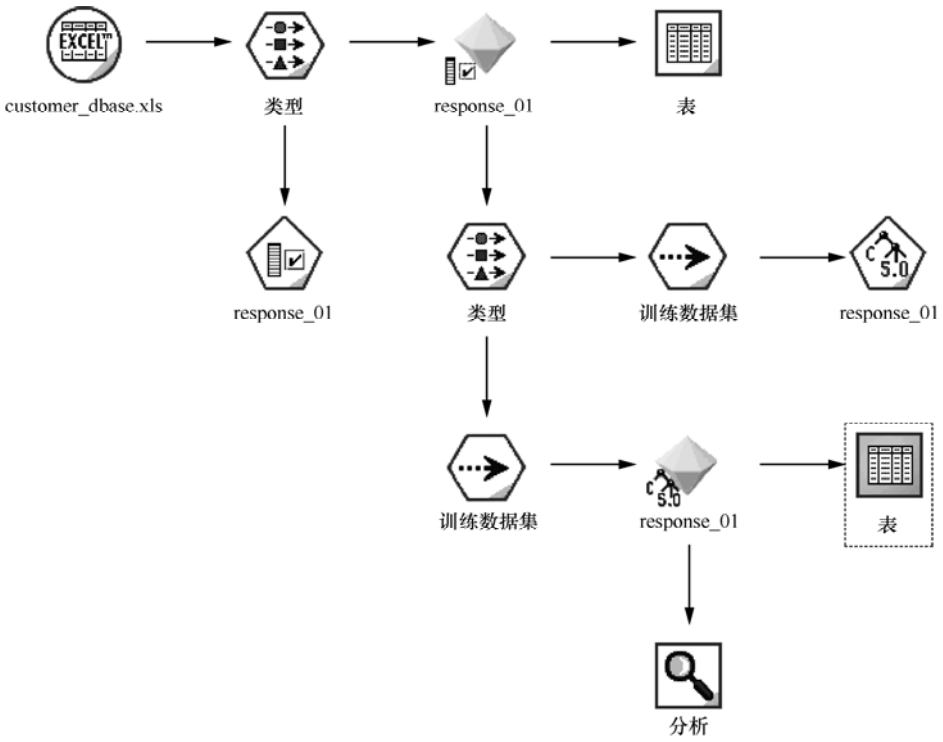

图 7　在 clementine 中的模型过程

5.2　实验结果

本文选取了 Clementine 自带的 customer_dbase 和 NewsChan 数据集作为实验数据，对本文模型进行了分析评估，如图 7 所示。customer_dbase 数据包含 132 个字段，5 000 条记录。为了模型的评估，本文选取了其中 2 500 条记录作为训练集数据项，剩余 2 500 条记录为预测推荐数据集。本文采用预测者与真实值的平均绝对误差(MAE)对算法进行精确度评价，如表 5 所示。

表 5　　customer_dbase 推荐准确率对比

推荐项目	全特征个性化推荐模型	基于情境的全特征个性化推荐模型	基于主体特征抽取个性化推荐模型	基于情境和主体特征融入性的多维度个性化推荐模型
Response-01	53.6%	71.3%	91.8%	92.3%
Response-02	48.2%	72.9%	87.6%	88.2%
Response-03	54.1%	56.5%	89.8%	91.2%

6　结束语

随着智能信息挖掘技术的不断发展和主体个性化特征的融入，现代互联网及其应用迫切需要一种能够根据用户的主体特征及其行为组织来调整信息或推荐用户感兴趣项目的服务模式。那么，个性化推荐服务的准确率高低是互联网应用能否成功的关键因素，解决这些问题的关键在于将互联网及其应用从被动接受用户的请求转变为主动感知并分析用户的信息需求，实现互联网及其不同的网络应用系统对用户的个性化主动推送信息服务。本文针对传统推荐模型的不足，提出一种基于情境和主体特征融入性的多维度个性化推荐模型，该模型能够充分利用地域文化背景、领域主题情景、主体特征等信息，避免了传统算法把用户的整体作为单个向量的弊端，克服了数据稀疏性等问题。实验结果表明，该模型的推荐质量比传统的协同推荐模型更高，更有针对性地向用户推荐他们感兴趣的项目。

参考文献：

[1] 石美红, 王婷, 陈永当等. 基于业务过程和知识需求的知识推送系统[J]. 计算机集成制造系统, 2011,17(4):882-887.
SHI M H, WANG T, CHEN Y D, *et al*. Knowledge push system based on business process and knowledge need[J]. Computer Integrated Manufacturing Systems,2011,17(4):882-887.

[2] PAZZANI M J, BILLSUS D. Content-Based Recommendation Systems[M]. Berlin: Springer-Verlag,2007.

[3] 张光卫, 李德毅, 李鹏等. 基于云模型的协同过滤推荐算法[J]. 软件学报,2007,18(10):2403-2411.
ZHANG G W, LI D Y, LI P, *et al*. A collaborative filtering recomm-

endation algorithm based on cloud model[J]. Journal of Software, 2007,18(10):2403-2411.

[4] SARWAR B, KARYPIS G, KONSTAN J, *et al.* Item-based collaborative filtering recommendation algorithms[A]. Proceedings of the 10th International World Wide Web Conference[C]. New York, N Y, USA, 2001. 285-295.

[5] BETTMAN J R, LUCE M F, PAYNE J W. Constructive consumer choice processes[J]. Journal of Consumer Research Consum, 1998, 25(3):187-217.

[6] PRAHALAD C K, BEYOND CRM C K. Prahalad Predicts Customer Context is the Next Big Thing[M]. NewYork: AMACOM, 2004.

[7] PALMISANO C, TUZHILIN A, GORGOGLIONE M. Using context to improve predictive models of customers in personalization applications[J]. IEEE T Knowl Data Engineer, 2008, 20(22): 1535- 1549.

[8] PANNIELLO U, TUZHILIN A. Comparing Pre-filtering and Post- filtering Approach in a Collaborative Contextual Recommendation System[M]. Berlin:Springer, 2009.348-359.

[9] 刘永利, 欧阳元新, 闻佳等. 基于概念聚类的用户兴趣建模方法[J]. 北京航空航天大学学报, 2010,36(2):188-192.
LIU Y L, OUYANG Y X, WEN J, *et al*. Approach to modeling user interests using conceptual clustering[J]. Journal of Beijing University of Aeronautics and Astronautics,2010,36(2):188-192.

[10] 严隽薇, 黄勋, 刘敏等. 基于本体用户兴趣模型的个性化推荐算法[J]. 计算机集成制造系统, 2010,16(12): 2757-2762.
YAN J W, HUANG X, LIU M, *et al.* Personalized recommendation algorithm for user interest model based on ontology[J]. Computer Integrated Manufacturing Systems, 2010, 16(12): 2757-2762.

[11] HUANG Z, CHEN H, ZENG D. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering[J]. ACM Transactions on Information Systems, 2004, 22(1): 116-142.

[12] BRECHEISEN S, KRIEGEL H, PFEIFLE M. Efficient density based clustering of complex objects[A]. Proceedings of the 4th IEEE International Conference on Data Mining[C]. Washington D C, USA, 2004. 43-50.

[13] 张光卫, 康建初, 李鹤松等. 面向场景的协同过滤推荐算法[J]. 系统仿真学报, 2006, 18 (z2): 595-601.
ZHANG G W, KANG J C, LI H S, *et al*. Context based collaborative filtering recommendation algorithm[J]. Journal of System Simulation, 2006, 18 (z2): 595-601.

[14] ADOMAVICIUS G, SANKARANARAYANAN R, SEN S, *et al.* Incorporating contextual information in recommender systems using a multidimensional approach[J]. ACM Transactions on Information Systems (TOIS), 2005, 23 (1):103-145.

[15] GAO M, WU Z F. Incorporating pragmatic information in personalized recommendation systems[A]. The 11th International Conference on Informatics and Semiotics in Organizations[C]. Beijing, China, 2009. 156-164.

[16] 邹博伟, 张宇, 范基礼等. 基于改进 TextTiling 方法的用户新兴趣发现的研究[J]. 计算机研究与发展, 2009,46(9):1594-1600.
ZOU B W, ZHANG Y, FAN J L, *et al.* Research on personalized information retrieval based on user's new interest detection[J]. Journal of Computer Research and Development,2009,46(9):1594-1600.

作者简介：

琚春华（1962-），男，浙江常山人，博士，浙江工商大学教授、科技处处长，主要研究方向为智能信息处理、数据挖掘、电子商务与物流优化等。

鲍福光（1986-），男，浙江余姚人，浙江工商大学硕士生，主要研究方向为智能信息处理、数据挖掘和供应链协同合作。