基于分布式数据挖掘的电子商务推荐系统①

Distributed Data Mining-Based E-Commerce Recommendation System

余小高 (湖北经济学院 信息管理学院 湖北 武汉 430205)

摘 要: 为解决电子商务推荐系统开放性、效率和精确度问题，给出了基于分布式数据挖掘的电子商务推荐系统(BDBRS)功能结构，提出了该系统的体系结构，然后介绍了 BDBRS 所应用的技术和 BDBRS 的设计与实现，最后描述了 BDBRS 的部分功能模块及人机界面，验证了 BDBRS 的正确性和本文研究的推荐算法在效率、精确度等方面的优越性。

关键词: 分布式数据挖掘 推荐系统 体系结构 推荐算法 开放性 效率

1 引言

目前几乎所有大型电子商务系统不同程度地使用了推荐系统，如 Amazon、CDNOW、eBay 等。现有推荐系统体系结构主要有基于服务器端的强耦合系统 CISR 和基于客户端的强耦合系统 RCIS。对于 CISR，推荐系统在逻辑上位于 Web 服务器之后，只能依赖 Web 服务器获取用户信息，推荐系统直接使用由 Web 服务器收集的客户个性化信息或利用 Web 服务器提供的接口间接完成与用户的交互。推荐功能必然受到 Web 服务器的限制，增加了 Web 服务器开销，通用性较差。对于 RCIS，其功能和效率在很大程度上受到客户端硬件配置和运行环境等条件影响，系统收集到的仅是个别用户甚至是单一用户的访问信息，影响推荐效果。

同时，当前大部分电子商务推荐系统只利用了一部分可用信息来产生推荐，只能提供一种推荐模型[1]。由于电子商务系统本身复杂性，单一类型的推荐系统并不适用于整个电子商务系统。目前电子商务推荐系统主要存在如下问题：(1)不能灵活提供多种推荐功能。(2)与商业系统的接口多采用紧耦合方式。(3)难以动态有效地管理和维护多个推荐工具和大量数据。(4)不能灵活地调用适用的谈判协商模型和策略来支持对推荐结果的解释和讨价还价。随着推荐系统在电子商务系统中的广泛应用，带来了大量推荐工具、数据、应用接口等如何有效管理和维护的问题。随着研究的深入，如何有效集成电子商务中各种类型数据，综合使用多种推荐工具，提供多种推荐模型，产生更加有效的推荐，满足不同类型的推荐需求，得到了越来越多研究者的关注[2]。

本文在已有的研究成果“电子商务环境中分布式数据挖掘的研究[3]”基础上，提出了一个基于分布式数据挖掘的电子商务推荐系统(BDBRS)。该系统采用了一个新的开放式推荐系统构架，独立于具体的推荐应用和推荐算法，具有良好的开放性；支持完整的推荐管理功能和统一的推荐管理平台；支持多种推荐功能，实现不同的推荐模型的“即插即用”，实现多种推荐模型的组合推荐；集成多种谈判模型、协议和策略，支持对推荐结果的解释和客商之间讨价还价。

2 BDBRS功能结构图

电子商务系统推荐流程基本如下：根据应用的需要配置推荐策略，确定具体的推荐模型；再根据推荐模型的要求从多种数据源中进行数据挖掘；然后由推荐引擎接受用户的推荐请求，根据相应的推荐策略，运行相应的推荐模型，产生推荐结果。BDBRS 功能结构如图 1 所示。

① 基金项目:湖北省教育厅重点项目(D20081902,2008d095)
收稿时间:2009-03-03

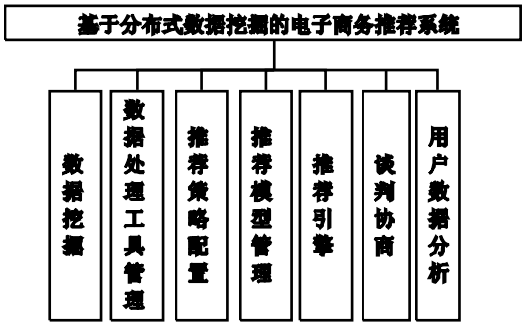

图 1　BDBRS 功能结构图

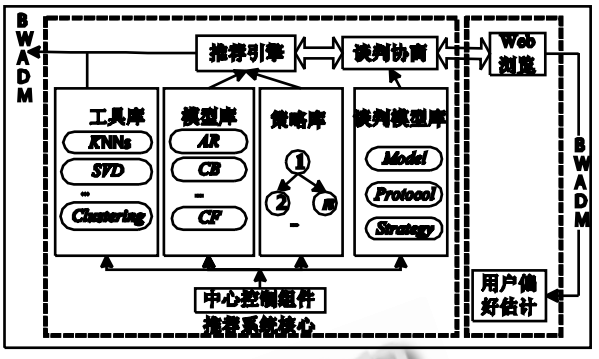

图 2　BDBRS 体系结构

数据挖掘主要是按照推荐模型的要求对数据源中数据进行处理，着重解决一些特定问题。推荐策略配置模块主要是配置具体的推荐过程，不同的推荐策略采用不同的推荐模型和推荐算法，或不同的组合推荐模型，提供不同的推荐服务。推荐模型管理模块就是根据具体的推荐应用，使用合适的推荐模型生成工具建立推荐模型，存储在推荐模型库中。

推荐引擎根据推荐策略，调用推荐模型和数据处理工具，实现具体的推荐功能。用户数据分析模块主要是根据推荐引擎的要求，收集客户数据，采用机器学习、数据挖掘等对客户数据进行分析，估计客户偏好，支持推荐引擎建立客户档案和产生实时推荐。谈判协商模块主要根据应用的特点，采用一定的谈判模型和谈判策略，实现支持客商双方展开推荐解释和讨价还价的功能。

3　BDBRS原型系统的体系结构

BDBRS 推荐系统的体系结构如图 2 所示。该系统综合了 CISR 和 RCIS 的优点，明确了推荐引擎和商务应用系统各自的角色。整个系统分为三大部分：电子商务环境中分布式数据挖掘(BWADM)[3]体系、推荐系统核心和用户数据处理系统。BWADM 体系主要是从商务系统中抽取、转换推荐系统所需要的数据，包括项档案和用户档案，建立数据仓库。推荐系统核心部分包括协同过滤推荐算法组件、推荐引擎组件、谈判协商组件、中心控制组件和推荐模型库等[4]。具体说明如下。

(1)　BWADM 体系。连接数据源，将数据转换成系统认可的数据格式和数据模型，自动监测数据源数据变化，对数据进行过滤、总结，并对各种数据进行描述。

(2) 推荐模型库。存放多种推荐模型的数据库，用于提供各种不同的推荐服务。

(3) 协同过滤推荐算法组件。存放多种数据处理工具，供数据处理时调用。该工具库由中心控制组件直接管理。

(4) 推荐策略库。根据应用的需求，存储具体的推荐策略，供推荐引擎调用，以实现具体的推荐服务。

(5) 数据处理工具库。根据推荐引擎组件的要求，从协同过滤推荐算法组件中调用一定的处理工具，对挖掘出的数据进行快速处理，辅助推荐。

(6) 推荐引擎组件。接收用户推荐请求，根据推荐策略，从推荐模型数据库中调用推荐模型，并调用协同过滤推荐算法组件，产生推荐结果。

(7) 中心控制组件。提供协同过滤推荐算法组件、模型库、推荐策略库和谈判模型库的接口，实现具体管理功能。

(8) 谈判协商组件。分析用户购买意图，根据谈判模型库中模型和策略，有效地向用户解释推荐产生的原因，说服用户采用推荐系统的推荐结果，支持买卖双方讨价还价。

(9) 用户数据处理系统。该系统是电子商务系统向用户提供推荐服务的应用程序接口，根据推荐引擎的需要，收集用户的访问日志，分析用户的浏览模式和行为模式，采用 BP 神经网络等人工智能技术，分析用户偏好，上传给推荐系统数据仓库；同时，把以 XML 等形式提交的推荐请求翻译并传送给推荐引擎；同时接收推荐引擎提供的推荐结果，将推荐结果以一定的形式返回给电子商务系统。

该体系结构也可以分为离线部分和在线部分两部分。离线部分实现数据的预处理，包括数据仓库的创

建、数据处理工具库和推荐模型库的创建以及推荐数据的预处理等；在线部分由谈判协商组件、用户数据处理系统和推荐引擎组件，以及协同过滤推荐算法组件等几部分组成。

BDBRS 构架独立于具体的推荐算法，它提供了一个开放的运行环境，其特点如下：

(1) 支持多种数据处理算法、多种推荐模型，及其它们多种形式的组合：BDBRS 系统提供多种推荐模型以适应复杂的电子商务系统的不同需要，实现具体数据处理算法和具体推荐算法的“即插即用”；同时，提供多种模型的组合推荐，互补不同推荐模型的长短，以提高推荐效率和质量。

(2) 提供谈判协商功能：只有有效地支持用户和推荐系统的谈判协商，才能有效地说服用户做出购买行动。BDBRS 系统采用谈判支持系统的理论和方法，根据应用的需求，从谈判模型库中调用谈判模型和策略，提供谈判协商功能，支持客商双方就推荐结果的谈判协商。

4 BDBRS的设计与实现

4.1 项档案的建立

项档案也即项的物理特征档案。物理特征是项在尺寸、颜色、价格等方面的物理属性。不同类别项的物理属性不尽相同，必须对每一类别的项建立一个物理特征档案。项的客观可识别特征都保存在产品数据库中，项特征档案可以根据产品特征数据库来创建，其特征可以表示为如下向量：

$$P_{(m)}=\begin{pmatrix}f_1\\f_2\\\cdots\\f_n\end{pmatrix}=\begin{pmatrix}f_{11},f_{12},\ldots,f_{1k}\\f_{21},f_{22},\ldots,f_{2j}\\\ldots\\f_{n1},f_{n2},\ldots,f_{nh}\end{pmatrix},\quad W_{(m)}=\begin{pmatrix}w_1\\w_2\\\cdots\\w_n\end{pmatrix}$$

$$\sum_{i=1}^{n} w_i=1,\ m=1,2,\ldots,M$$

在上式中，M 代表项总数，n 表示项 m 的特征数，w_i 表示第 i 个特征的权重，而 f_{ij} 代表项 m 的第 i 个特征的第 j 个属性的属性值，其中不同特征的属性数量是不一样的。通过参考项的说明信息等，项供应商可以取得其主要特征值。每个特征附有一系列的调查问卷，分别有一些不同的选项，以收集用户对项特征的满意程度反馈。

4.2 用户档案的建立

用户档案包括用户主观评价模型、偏好区间、理想特征值和用户偏好解释等，分四个阶段建立用户档案：(1)用户行为整理。(2)BP 预测模型的建立。(3)主观评价模型的建立。(4)综合统计用户各偏好信息，建立用户偏好档案。

4.2.1 用户行为的整理

该模块运行于客户端，根据用户导航模式和行为模式收集用户数据，并进行预处理。导航模式包括浏览、查找、点击、购物篮放置和购买等；行为模式包括项的点击率、观察时间、观察次数、打印、收藏等。虚拟商厦中，所有商品是根据一定分类标准存放[5]，该分类标准呈现为一种树形结构并可获取。在该模块的研究中，用户查看一个商品的数据表示为：*USER_DATA={User_ID,Item_ID, Number, Time_Length, Click_Type, Category_Click,Action, FB_of_F}*。因此，以 *User_ID* 和 *Item_ID* 为关键词，根据研究中所收集的数据，就可建立用户行为数据库。

4.2.2 BP 预测模型

该模块运行于客户端，用 BP 算法训练多层前馈网络模型，预测用户对点击过、没有购买且有放置购物篮等行为的项评分。用户的所有浏览模式和行为模式集合随机分为训练集和测试集两个部分，将每个项的浏览模式和行为模式作为输入变量，具体包括(项，次数，时间，类点击率，点击类型，行为)，行为模式的评分作为目标；如果用户有多种行为，则评分公式为 $r(i)+\sum_{j=u+1}^{4}(action(j)*w(j))$，其中 i 为 action 中二进制为 1 的最低位。根据浏览模式和行为模式，用该模型预测用户对点击过但没有显式行为的项评分，用户忽略的所有的项都有一个在 [0，1]中的预测值，这些预测值就作为用户对该项的预测评分，记为 S(i)，i 代表项。

4.2.3 主观评价模型

该模块运行于客户端，分析用户的历史和当前的行为模式与浏览模式，发现用户对项特征的偏好，建立主观评价模型，并将该模型上传给服务器端。

对用户有潜在影响的项特征偏好估计是推荐系统的基础。主观评价模型[6]是估计用户对项的主要特征的偏好，是用户档案的重要组成成分。主要思路如下：当用户浏览某项时，总是选择并综合评价该项的多个特征进行选择。在此过程中，每一个用户对同一项都有可能表现出不同的解释和评价，该过程隐藏着用户对项的深层信息。因此，可对用户对项的反馈过程进

行建模[7]，挖掘项物理特征、用户心理评价和用户解释三者之间的关系。这三者之间的关系就是用户的主观评价模型[2]。由于用户是在观察、评价项各主要特征之后再做出购买意向，该评价过程隐含着更深层次的用户信息。因此，在此阶段，系统根据项档案生成一些描述项主要特征的问卷供用户做出个人反馈。

由于综合预测评分 S(i)已预测出，因此，建立该模型的目的就是根据用户对项特征的反馈和综合预测评分，挖掘用户对项各主要特征的评价和偏好区间，以便有效进行推荐。

由此可得，若用户对项第个 i 主要特征的第 j 个属性 f_{ij} 作出了选择，则该用户对项该特征属性的心里评价为 sf(ij)=w(i)×S(i)，用户对该特征的其它属性的心里评价为：sf(ik)= fw(ik)/fw(ij) sf(ij),1≤k≤n，其中 n 为特征 f(i)的属性数量。用户对项特征 fij 的总评价 CTI(ij) 为：CTI(ij)=∑sf(ij)/m，m 为用户查看相同种类项的总数。

4.2.4 建立用户档案

该模块运行于服务器端，收集、整理用户偏好信息，建立用户详细偏好档案，以便推荐。用户的兴趣和偏好会随着时间而改变，当前的行为模式和浏览模式更能反映出该用户目前阶段的兴趣[2,8]。在该研究中时间因素必须作为一个变量来考虑，最近的兴趣点拥有更高的权重。因此，用户对项特征 f_{ij} 总的评价如下式：

$$\tilde{CTI}(ij)=\begin{cases}\dfrac{\sum sf(ij)\times\alpha}{m},\alpha>1, & \text{小于}k\text{天}\\ \dfrac{\sum sf(ij)}{m}, & \text{其他}\end{cases}$$

通过用户主观评价模型，用户对项各主要物理特征的心里评价 $C=(CT\tilde{I}(1l),\ldots,CT\tilde{I}(IK),\ldots,CT\tilde{I}(il),\ldots,CT\tilde{I}(ih))$，对其进行统计汇总就可确定用户对主要物理特征的偏好区间和理想特征值，并确定了用户档案。用户档案以树的形式存储，节点为用户偏好类别，树叶为用户档案描述。

协同过滤推荐技术需要用户评分表，记录着用户对项的评分信息，共有 3 个字段，即 User _ID(用户 ID 号)、Item_ID(商品号)、Rating (评分)。

4.3 BDBRS 系统部分模块设计介绍

BDBRS 在运行中首先要通过中心控制组件对工具库、模型库进行设置等，然后根据推荐需要定义推荐策略，最后由推荐引擎根据推荐策略调用各模块实现推荐。

4.3.1 推荐引擎

推荐引擎组件主要是接收用户推荐请求，根据一定的推荐策略，调用相关数据处理工具，对数据进行处理，并根据相关的单个推荐模型或混合推荐模型，执行具体推荐过程，产生推荐结果。基于内容的推荐策略和基于协同过滤的推荐策略组成的混合推荐策略描述如下：(1)该推荐策略是一种混合推荐技术，综合采用基于内容的推荐技术和协同过滤推荐技术。(2)根据用户评分表的稀疏程度，调用基于内容的推荐技术，根据项档案与用户档案的匹配程度对用户已评分的项所在类别中其它没有评分的项进行隐性评分。(3)最后采用协同过滤推荐技术，调用 KNDC 对用户进行聚类，产生虚拟用户，再调用 P2PAKNNS 搜索出用户的 k 个最近邻居，然后产生推荐结果，选择评分较高前 N 个项组成推荐列表。该策略在推荐引擎组件内部相应的推荐算法如下：

输入：用户评分表、用户档案、项档案、用户指定的邻居数 k，最大推荐数量 N

输出：推荐商品列表 Recommend_Item_List

(1) 设用户已经评分的项的同类项组成的集合为 item_set

(2) FOR each 当前用户 this_user 未评分的项 item

(3) 在项档案库中查找 item 的特征信息

(4) 在用户偏好表中查找与 item 档案最匹配的偏好信息

(5) 根据 item 特征信息与用户偏好信息的匹配程度对 item 进行隐性评分

(6) ENDFOR

(7) 调用 KNDC 对用户进行聚类，结果为 C={c_1,c_2,c_3,⋯,c_n}

(8) FOR each c in C

(9) 计算 c 中所有用户对项的平均评分，生成新的聚类质心。聚类质心与聚类中其他用户的距离之和最小，代表该聚类中用户对商品的典型评分。将该聚类质心作为该聚类相应的虚拟用户 Fictive_User，并将其加入虚拟用户集合 Fictive_User_Set.

(10) ENDFOR

(11) 调用 P2PAKNNS 在 Fictive_User_Set 中查

询当前用户的 k 个最近邻居 k_Nearest_Neighbors_Set

(12) FOR each 当前用户 this_user 未评分的项 j

(13) 当前用户对项 j 的评分 Rthis,j =0

(14) FOR each 邻居 i in k_Nearest_Neighbors_Set

(15) Rthis,j=Rthis,j+ //为邻居 i 的平均评分

(16) ENDFOR

(17) Rthis,j=Rthis,j/(当前用户与最近邻居相似性之和)

(18) Rthis,j=Rthis,j+(当前用户的平均评分)

(19) ENDFOR

(20) 选择当前用户评分表中未购买的并且评分最高的前 N 个项组成推荐列表 Recommend_Item_List 作为输出。

4.3.2 部分模块及人机界面

因篇幅所限，在此仅列举部分模块。策略定义界面如图 3 所示。谈判协商组件根据具体应用和被推荐对象特点调用谈判模型和谈判策略来支持对推荐结果的解释和讨价还价。

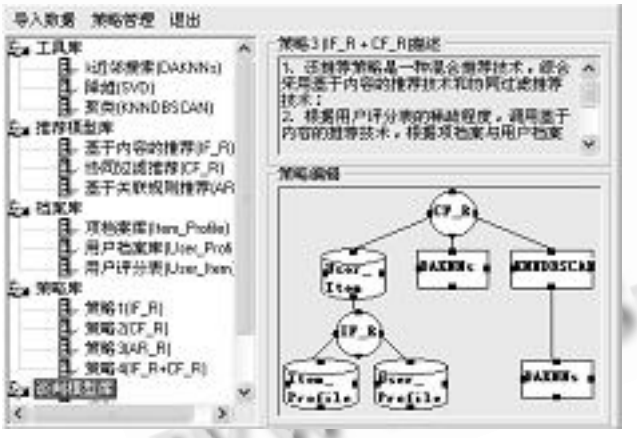

图 3 策略定义界面

图 4 是采用基于项特征的自动谈判策略时，谈判协商组件选择买方预测评分高的特征的推荐说明，向买方人员报价，以及买方人员还价的第一轮回界面。

5 结论

本文提出了一种解决电子商务环境数据复杂性的一种方法，利用分布式数据挖掘技术，对电子商务环境中用户行为和用户属性进行学习，从中获取有价值的知识，根据得到的知识产生推荐。尽管推荐系统在数据挖掘技术出现之前就已存在，但数据挖掘技术的出现，不仅给电子商务领域的海量数据处理提供了一种有效的手段，而且给电子商务推荐系统提供了自动化、智能化和更高质量的推荐结果。

图 4 议价界面

参考文献

1 邓爱林.电子商务推荐系统关键技术研究[博士学位论文].上海:复旦大学, 2003.

2 Yano E, Sueyoshi E, Shinohara I, Kato T. Development of a recommendation system with multiple subjective evaluation process models. In: Proc. of the 2003 International Conference on CyberWorlds. Washington: IEEE Computer Society Press, 2003.344－351.

3 余小高.电子商务环境中分布式数据挖掘的研究[博士学位论文].武汉:武汉理工大学,2007.

4 孟波.计算机决策支持系统.武汉:武汉大学出版社, 2001.

5 Hinneburg A, Keim DA. A general approach to clustering in large databases with noise. Knowledge and Information Systems, 2003,5:387－415.

6 Duda RO, Hart PE, Stork DG. Pattern Classification. China Machine Press, 2003:60－75.

7 Lee Y, et al. A collaborative recommendation based on neural networks. Proc. of DASFAA 2004, LNCS 2973. Berlin: Springer, 2004.425－430.

8 Weng SS, Liu MJ. Feature Based recommendations for one-to-one marketing. Expert Systems with Applications, 2004,26:493－508.