湖北经济学院学术文库

电子商务环境中分布式数据挖掘的研究

DianZiShangWu HuanJingZhong FenBuShi ShuJu WaJue De YanJiu

余小高◎著

湖北长江出版集团
湖北人民出版社

湖北经济学院学术文库

电子商务环境中分布式数据挖掘的研究

DianZiShangWu HuanJingZhong FenBuShi ShuJu WaJue De YanJiu

余小高◎著

湖北长江出版集团
湖北人民出版社

鄂新登字01号
图书在版编目(CIP)数据

电子商务环境中分布式数据挖掘的研究/余小高著.
武汉:湖北人民出版社,2008.4

ISBN 978-7-216-05506-2

Ⅰ.电…
Ⅱ.余…
Ⅲ.数据采集—计算机应用—电子商务
Ⅳ.F713.36-39

中国版本图书馆CIP数据核字(2008)第013070号

电子商务环境中分布式数据挖掘的研究　　余小高 著

出版发行:湖北长江出版集团 湖北人民出版社
地址:武汉市雄楚大街268号
邮编:430070

印刷:鄂州市立龙印刷服务有限责任公司
经销:湖北省新华书店
开本:787毫米×1092毫米 1/16
印张:16.125
字数:315千字
插页:1
版次:2008年4月第1版
印次:2008年4月第1次印刷
书号:ISBN 978-7-216-05506-2
定价:38.00元

本社网址:http://www.hbpp.com.cn

摘　要

现代新兴商业模式的电子商务的蓬勃发展，使得 Internet 上的资源和服务更加丰富多彩，这些丰富的资源和服务每天都会产生许多新的、蕴涵着大量重要信息的海量数据，这些数据往往是异构的、不确定性的和非结构化的，其复杂程度已远远超出了人类目前已有的分析和理解能力。因此，研究有效利用这些复杂资源的新技术，数据挖掘具有重要的现实意义。就信息处理而言，数据挖掘是致力于数据分析和理解数据内部蕴藏知识的技术，它成为未来信息技术应用的重要目标之一。

本书在对分布式数据挖掘、Web 服务及 Agent 相关技术进行了分析的基础上，为解决电子商务环境中分布式数据挖掘的核心问题，从关键算法和架构两个方面进行了深入研究。

针对 k 最近邻搜索算法存在的问题；提出了电子商务环境中一种自适应的基于 P2P 的 k 最近邻搜索算法 P2PAKNNS。探讨了度量空间、相似性查询和GHT * 规则，自定义了高维数据的相似度函数 HDSF（X，Y），论述了 GHT * 中插入算法及范围查找算法和搜索算法。在此基础上，具体给出了 P2PAKNNS 算法的实现方法，并通过实验，验证了其正确性。

同时本书对 DENCLUE 算法进行了研究，为使其适合电子商务环境、解决其存在问题，结合 P2PAKNNS 算法的优点，提出了电子商务环境中基于距离和密度的无监督聚类算法 KNDC。论述了模糊簇的划分及参数 k，讨论了参数 σ 和 ξ 的设置，给出了 KNDC 算法的具体实现方法，并予以验证。

本书还针对电子商务环境中分布式数据挖掘的关联规则，在研究 Apriori 关联规则算法、多重最小支持度 Apriori 算法、相关支持度 Apriori 算法 RSAA、平均项目集合分割法的基础上；改进了阈值的制定方法；为提高挖掘有价值的稀有数据的效率和精确度，根据 RSAA 和基于无向项集图算法 BOUIGA 提出了 RSAA-BOUIGA 算法，分析了其正确性。

在此基础上，结合业界和学术界对 Web 服务和移动 Agent 的研究成果，将 Web 服务和 Agent 最新技术引入了电子商务环境中分布式数据挖掘，提出了电子商务环境下基于 Web 服务和移动 Agent 技术的数据挖掘架构 BWADM，并论述了组合服务规范与组合服务的执行，具体阈述了此架构的数据预处理组件、算法管理组件、控制中心组件、算法库组件和模型表示组件。

然后建立了 BWADM 原型，结合 Web 服务技术，给出了基于服务的数据挖掘系统逻辑结构，设计并实现了该系统，验证了 BWADM 的合理性和上述算法在效率、精确度等方面的优越性。

最后分析并指出了目前电子商务推荐系统存在的问题，并将电子商务环境中分布式数据挖掘技术应用到推荐系统中。为提高协同过滤推荐效率和精确度，根据 P2PAKNNS 提出了基于 P2PAKNNS 的协同过滤推荐算法，并根据 KNDC 提出了基于 KNDC 的协同过滤推荐算法，分别给出了这两个算法的具体实现方法，并予以验证。在此基础上，为解决目前电子商务推荐系统实时性和可扩展性不足、推荐工具种类繁多却单一、推荐结果解释性差等问题，将 BWADM 应用到电子商务推荐系统中，研究了基于隐式评分的推荐系统，并设计和实现了基于 BWADM 的电子商务推荐系统原型系统 BDBRS，验证了其正确性和上述推荐算法在效率、精确度等方面的优越性。

Abstract

The explosive growth of Electronic Commerce makes the resources and services more plentiful with current fire-new business on Internet, everyday these rich resources and services generate volumes of heterogeneous, uncertain, and unstructured data which so complexity that far beyond human's current capacity to interpreting and digesting them. Hence, it is practical important to develop some new techniques for making use of these valuable complex resources sufficiency. Data mining is a technique that aims to analyze and understand large source data and reveal knowledge hidden in the data. It has been viewed as an important evolution in information processing.

Key algorithms and architecture are studied to solve the core problems of distributed data mining in Electronic commerce environment, after distributed data mining, Web services and Agent technology are analyzed in the book dissertation.

Firstly, an adaptive distributed algorithm, called P2PAKNNS for P2P k-nearest neighbor search, in high dimensions is proposed to solve the shortcomings of KNNs in Electronic Commerce environment. Metric Space, Similarity Queries and Principles of GHT* are discussed. Similarity measure functionis HDSF (X, Y) given. Insert, Range find and Search Algorithms in GHT* are discussed. The detailed P2PAKNNS algorithm is given and discussed with experiment.

Secondly, DENCLUE is analyzed. In order to solve its shortcomings and be applied in Electronic Commerce environment, a new clustering algorithm of distributed data mining based on P2PAKNNS and DENCLUE called KNDC is proposed for Electronic Commerce environment. Fuzzy Clusters division, parameter k, parameter σ and ξ are discussed, the detailed algorithm is given and discussed with experiment.

Thirdly, these association rules algorithms of data mining are introduced detailedly; they are Apriori algorithm, relative support Apriori algorithm, mean itemset dividing method. Based on those analyises the method of threshold's settings is improved. Then RSAA-BOUIGA algorithm is proposed to improve the precision and efficiency of valuable rare data mining according to BOUIGA and RSAA algorithm.

Fourthly, after reviewing former research and combincing the solution of industry and academy, a new architecture called BWADM based on the researches above is pro-

posed, it is a distributed data mining system based on Web services and Agent in Electronic Commerce environment. Web service composition rules and execution of Web service composition are discussed. These modules are introduced detailedly in Electronic Commerce environment; there are algorithm management module, control center module, algorithm database module and model representation module.

Then, a prototype system of distributed data mining is given for Electronic Commerce. The logic structure of data mining system based on Web services is proposed according to Web services technology; the project of data mining based on Web services is designed and implemented. These prove the reasonableness of the BWADM and the algorithms discussed above are more efficacious and precise than current algorithms.

Lastly, the problems are analyzed and indicated in current Electronic Commerce system, the technology of distributed data mining is applied to Recommendation System. In order to improve the efficacious and precise of Recommendation Algorithms, two Collaborative Filtering Recommendation Algorithms are proposed for Recommendation System, which are based on P2PAKNNS and KNDC, and the detailed algorithms are given and discussed with experiment. In order to solve the problems that the real-time and the scalability for further development may not be enough, the number of the recommendation tools is large but they lack varieties, and the explanations for recommendations are not reliable enough, BWADM is applied for Recommendation System and studied the Recommendation System based on the implicit rating; the project of Recommendation System based on BWADM called BDBRS is designed and implemented. These prove the reasonableness of BDBRS and above these recommendation algorithms are more efficacious and precise than current algorithms.

目　录

第一章　绪论

第一节　研究背景及意义

当今时代是网络化、信息化的时代，Internet 在全球迅速扩张，并且逐渐渗透到社会的各个领域。人们利用它可以方便地收发邮件、浏览网页、聊天、下载软件、购买商品等等。针对 Internet 上涌现出的这些商机，许多企业纷纷把自己的业务转移到 Internet 上，建立自己的网站，实现各种商业活动，这便是电子商务。

电子商务让企业和客户通过网络交互，取代了传统的面对面的交互方式。由于 Internet 的普及与便利，电子商务企业要面对的客户群是巨大而复杂的，市场也变得更大、更复杂。这个时候，一方面企业比以往更加期望能够对用户和市场进行把握，而电子商务企业对市场和客户的知识往往是非常缺乏的；另一方面，电子商务在运转的过程中又积累了大量有关用户本身和用户商务行为的数据，电子商务企业面临数据丰富、知识贫乏的尴尬。

数据挖掘，也称知识发现，被描述为从数据中抽取出隐含的、具有潜在用途的、人类可理解的模式。数据挖掘通过发现有用的新规律和新概念，提高了数据拥有者对大量原始数据的深层次理解、认识和应用。数据挖掘技术是目前国际上数据库和信息决策领域最前沿的研究方向之一，引起了学术界和工业界的广泛关注[1]。数据挖掘是一个多学科交叉的新兴研究领域，在这个新兴领域中，汇集了来自机器学习、模式识别、数据库、统计学、人工智能以及管理信息系统等各学科的成果[2]。多元化的投入，使得这一学科得以蓬勃发展，而且已初具规模。除学术上的需要外，极大的商业应用前景更是推动数据挖掘研究不断深入的关键因素[3]。

将数据挖掘技术应用于电子商务，对这些数据进行挖掘，就可以找出这些有价值的“知识”。企业用户可以根据这些“知识”，把握客户动态，追踪市场变化，做出正确的针对性的决策，比如改进网站、向各类用户推出个性化的页面，或者向高流失客户群提供优惠政策进行挽留等等。因此，发掘电子商务系统积累的大量数据，将使企业能够及时应对市场变化、占领市场的制高点，在电子商务平台上进行数据挖掘就成为一个研究热点[4]。

电子商务是基于 Internet/Intranet 或局域网、广域网，包括了从销售、市场到商业信息管理的全过程。在电子环境中，许多大型数据库都是以分布的形式存在的。随着 WWW 应用的日趋普及，Internet 已成为当今世界最大的分布数据源，Internet 中的数据正以几何级数增长，而且 Internet 本身就是一个巨大的分布式系统。如何应用计算机网络中的庞大数据资源，发现和获取其中有价值的知识，已经成为人们必须正视的问题。而分布式数据挖掘是在电子商务环境中发现和获取有用知识的最佳方法之一。分布式数据挖掘为从“数据海洋”中开采有用的知识提供了有效途径，它将在金融投资、电信、市场营销、气象和灾难预报、科学决策、Internet 信息浏览等方面发挥巨大作用，具有广阔的应用前景[65]。

电子商务推荐系统模拟商店销售人员向客户提供商品推荐，帮助客户找到所需商品，从而顺利完成购买过程，因此可以有效保留客户，提高电子商务系统的销售水平；商家也可以通过推荐系统保持与客户的联系，以建立良好客户关系，期待更高的客户回店率。推荐系统实现了客商的“双赢”。电子商务系统需要推荐系统的大力支持帮助客户找到所需商品，同时电子商务系统环境也很适合推荐系统的实施，因为在电子商务中能够方便地收集丰富的电子化数据，能够很好地检验推荐的效果。成功的电子商务推荐系统将会产生巨大的经济效益。电子商务推荐系统具有良好的发展和应用前景。研究表明，电子商务的销售行业使用个性化推荐系统后，销售额能提高 2% ~8%[5]，尤其在书籍、电影、CD 音像、日用百货等产品价格相对较为低廉且商品种类繁多、用户使用个性化推荐系统程度高的行业，推荐系统能大大提高企业的销售额。目前，几乎所有大型的电子商务系统，如 Amazon、CDNOW、eBay 和 DangDang 等，都不同程度地使用了各种形式的推荐系统。各种提供个性化服务的 Web 站点也需要推荐系统的大力支持。电子商务推荐系统理论也越来越成熟，应用也越来越广泛。但是，随着客户和商品数量的增加，电子商务的规模已经越来越大，推荐系统必须面对更大规模的数据，必须及时地、有效地处理这些大规模的数据，实时地作出推荐，并就推荐结果进行有效地解释、说服客户。面对如此大规模的数据，目前的推荐系统的效率和准确度已经不能适应实际的需要。

第二节　国内外研究现状与分析

一、研究现状简述

随着通信和计算机网络技术的飞速发展以及 Internet 的普及，商业空间的规模发展到全球成为可能，电子商务就是为了适应这种以全球为市场的变化而发展起来的一种商业模型。但是在电子商务系统中，通常面临着以下问题[6]：（1）

随着互联网的普及和电子商务的发展，电子商务系统在为用户提供越来越多选择的同时，其结构也变得更加复杂，用户如何才能不会迷失在大量的商品信息空间中，顺利找到自己需要的商品？（2）商家如何投其所好，为用户实现主动推荐，提供个性化服务？（3）企业如何根据用户的访问规律，快速调整企业的经营管理策略，留住老客户，挖掘潜在客户，寻找新的商机？（4）企业如何适应市场的变化，并进一步优化网站组织结构和服务方式，以提高网站的效率？电子商务网站的后台数据库中丰富的数据资源包含了对市场分析及预测非常有益的潜在信息，如何进行充分地挖掘和利用意义重大。由此，电子商务数据挖掘应运而生，前景广阔[1]。

将数据挖掘技术与电子商务结合起来，进行电子商务方面的数据挖掘，可以帮助人们更有效地从电子商务网站数据中获取有用的信息。然而电子商务环境下的数据挖掘与传统的数据挖掘相比有很多不同之处，电子商务下的数据挖掘的对象是大量异构的、分布的、半结构化的数据，其自身的特殊性决定了在数据挖掘之前必须进行数据预处理[6]。研究电子商务环境下的数据挖掘技术，并将它用于商业站点的开发，对发现电子商务智能性、提高站点实施的促销效果等决策具有实际的意义[3]。分析互联网背后的用户行为，获取用户的行为模式，进而调整页面的结构设计，推荐用户最可能感兴趣的商品，预测用户的行为，为用户提供更好的服务，给商业站点带来利润。对于一个电子商务公司每天搜集和处理的大量数据，利用数据挖掘技术可以帮助他们高度自动化地分析数据，做出归纳性推理，从中挖掘出潜在的模式，并预测未来，帮助企业决策者调整市场策略，减少风险，做出正确决策，从而给公司带来巨大的利润。

在电子商务中使用数据挖掘技术可以发现隐藏知识，提高企业竞争力。越来越多的企业使用数据挖掘来加强电子商务系统的智能，许多企业都在其具体的应用中使用数据挖掘技术。目前，以电子商务应用为背景的数据挖掘和知识发现的研究，主要是根据商业中对条码机数据的分析，发现顾客购物规律，采用的数据主要是 Web 日志。通过对大量事务数据的挖掘以产生有益于买卖双方的概要信息、客户分析和事务规则，这在电子商务中变得日益重要[7]。例如，客户的购物行为模式源自客户分析而被用作引导个性化市场、发展新客户、发掘商机和探测商业诈骗行为；事务规则被用来解释销售陡然升降的内因、分析商业走势等等。利用数据挖掘技术建立电子商务推荐系统，公司通过分析大量的交易记录，可以预测用户未来的购买需要，向用户推荐其可能感兴趣的商品。实际上，电子商务推荐系统能够直接与用户交互，模拟商品销售人员向用户提供商品推荐，帮助用户找到所需商品，从而顺利完成购买过程。从用户角度来看，通过对收集到的用户的访问行为、访问频度、访问内容等浏览信息进行挖掘，提取用户的特征，获取用户访问 Web 的模式，为用户实现主动推荐，提供个性化服务；从企

业角度来看，企业希望能够获取用户的访问规律，以帮助企业确定顾客消费的生命周期，针对不同的产品制订相应的营销策略，进一步优化网站的组织结构和服务方式，以提高网站的效率。在日趋激烈的竞争环境下，电子商务推荐系统能够实现个性化服务，从而有效保留用户，防止用户流失，提高系统的销售。推荐系统在电子商务系统中具有良好的发展和应用前景，逐步成为电子商务信息技术的一个重要研究内容，并受到越来越多研究者的关注，为电子商务建立以数据挖掘为核心的客户关系管理系统，挖掘电子商务系统积累的大量数据，将使企业能够及时应对市场变化，占领市场的制高点[8]。

在电子商务中，数据挖掘的方法虽然多种，但主要是三种，即协同过滤、聚类分析、关联规则[9]。协同过滤技术主要采用最近邻技术，利用客户的历史喜好信息计算客户之间的距离，目标客户对特定商品的喜好程度由其最近邻居对商品评价的加权平均值来计算。k 最近邻法（KNN）是一个理论上比较成熟的方法，在电子商务数据挖掘中有广泛的用途。KNN 查询是多媒体数据库管理系统中最具代表性的查询方式之一，它将 k 个与查询点最接近的对象作为查询结果返回[10]，它是向量空间模型中最好的文本分类算法之一[11]。数据分类技术是一种强有力的分析手段，它旨在生成一个分类函数或分类模型，由该模型把数据库中的数据项映射到某一给定类别中。现有的数据分类算法大体可以划分为两大类：积极学习方法与消极学习方法。消极方法使用很多不同的局部线性函数来形成对目标函数隐含的全局逼近，具有比积极方法更丰富的假设空间。其中消极学习型中应用最广泛的是 k 最近邻法[12]。KNN 的优点是易于快速实现，分类效果好。同时它也存在一些限制，比如存储训练例子的数量问题、相似性度量问题等[13]。

聚类是电子商务数据挖掘中一门非常有用的技术，可以用于从大量数据中寻找隐含的数据分布和模式。通过聚类，将数据划分为若干类，然后在每一类中寻找模式和各种潜在的有用信息[14]。聚类分析可以帮助市场人员发现顾客群中所存在的不同特征的组群，并可以利用购买模式来描述这些具有不同特征的顾客组群。最近邻（KNN）算法可以快速进行聚类并且能有效处理噪声点，但当数据密度和聚类间的距离不均匀时聚类质量较差[15]。在电子商务数据挖掘的常用聚类算法中，DENCLUE 是一种优良的算法。它能够识别各种复杂形状的聚类，能有效排除噪声干扰，并且聚类结果不受输入顺序的影响。但是其参数需要人工来确定，需要大量的内存，同时影响精度和效率[16]。

关联规则挖掘是电子商务数据挖掘中最活跃的研究方法之一。最早是由 Agrawal 等人提出的。最初提出的动机是针对购物篮分析问题提出的，其目的是为了发现交易数据库中不同商品之间的联系规则，这些规则刻画了顾客购买行为模式，可以用来指导商家科学地安排进货、库存以及货架设计等。之后诸多的研究人员对关联规则的挖掘问题进行了大量的研究，他们的工作涉及到关联规则的

挖掘理论的探索、原有的算法的改进和新算法的设计、并行关联规则挖掘以及数量关联规则挖掘等问题。在提高挖掘规则算法的效率、适应性、可用性以及应用推广等方面，许多学者进行了不懈的努力[17]。

目前，国外对于数据挖掘的研究重点逐渐从发现方法转向系统应用，注重多种发现策略和技术的集成，以及多种学科之间的相互渗透。当前，世界上比较有影响的典型数据挖掘系统有 Cover Story，EXPLORA，Knowledge Discovery Workbench，DB Miner，Quest 等[18]。在应用方面包括：KDD 商业软件工具不断产生和完善，注重建立解决问题的整体系统，而不是孤立的过程。用户主要集中在大型银行、保险公司、电信公司和销售业。国外很多计算机公司非常注重数据挖掘的开发应用，比如 IBM 和微软成立了相应的研究中心进行这方面的工作[1]。此外，一些公司的相关软件也开始在国内销售，如 IBM 公司的 Intelligent Miner，SAS 公司的 Enterprise Miner。

国内从事数据挖掘研究的主体主要是大学，也有研究所和公司在从事这方面的研究，这些工作一般集中于学习算法和有关数据挖掘理论方面的研究[1]。例如，复旦大学、南京大学、西安交通大学、东南大学、国防科大等单位已经在挖掘算法效率改进等方面做了不少的工作。在具体应用方面，中科院计算所智能处理开放实验室的史忠植等人设计了一个数据挖掘工具 MSMiner，使用决策树算法为广东地税提供纳税人异常情况检测，复旦德门公司开发的“天眼”数据挖掘工具集 Dminer 集成了多种数据挖掘算法，取得了较好的挖掘效果[4]。目前，国内外对电子商务环境下的分布式数据挖掘研究集中在理论研究和应用研究两个方面，国外在发现用户兴趣模式的理论体系研究和个性化服务方面都取得了较大的进展。与国外相比，国内的分布式数据挖掘起步较晚，这一领域正处在研究开发阶段，应用和产品还相对滞后，但是在理论和应用研究上有很多成果。这些研究的重点在局部挖掘算法的设计、分析和改进，较少对数据挖掘系统自身的构建、开发模式进行系统论述[19]。电子商务环境下的分布式数据挖掘系统是一个有机的整体，各个部分有着密切的联系，这一新兴领域不但有很好的研究和应用前景，而且有很好的商业机会[20]。因而，有必要根据当前数据存储、应用环境的特点，构建一个实用的分布式数据挖掘原型系统，以指导今后设计实际的分布式数据挖掘工具，以及支持在此基础上的渐进开发。

目前，推荐系统已广泛运用到各行业中，推荐对象包括书籍、音像、网页、文章和新闻等。推荐系统主要包括基于内容的推荐（CB）和基于协同过滤（CF）两种[21]。基于内容的推荐技术是信息检索领域的重要研究内容[22][23]。基于内容的推荐系统需要分析资源的内容信息[24][25]，根据用户兴趣建立用户档案（Profile），然后根据资源内容与用户档案之间的相似性向用户提供推荐服务[26]。内容分析技术[27][28][29]挖掘资源的特征，分析 Web 日志捕捉潜在规则。用户档