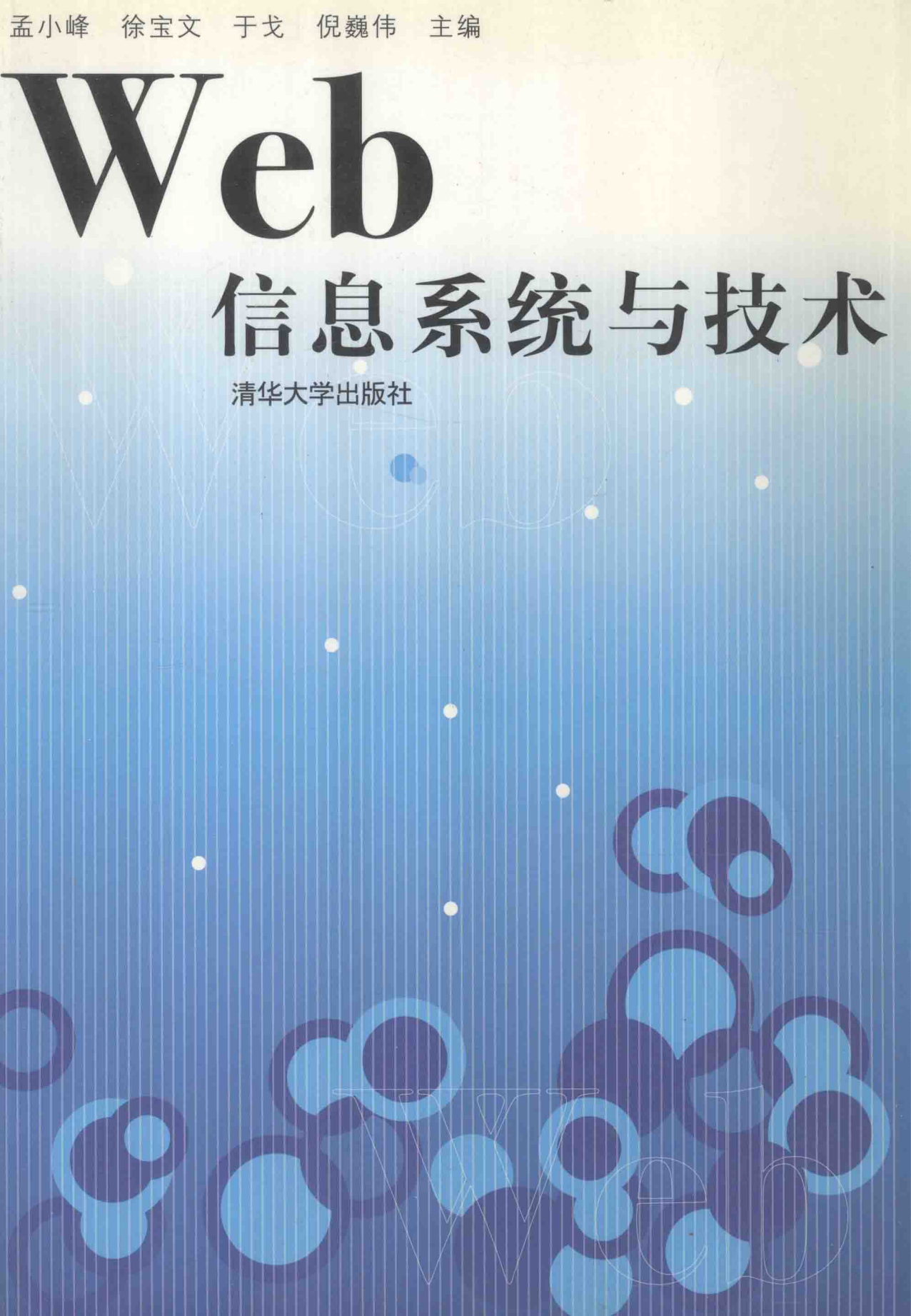
孟小峰　徐宝文　于戈　倪巍伟　主编

Web 信息系统与技术

清华大学出版社



孟小峰　徐宝文　于戈　倪巍伟　主编

Web 信息系统与技术

清华大学出版社
北京

内 容 简 介

随着运行于 Internet 平台上的各种应用和服务的快速发展，适用于网络平台交互、动态特性的各种技术层出不穷。人们迫切需要有效的手段和方法，以便更好地组织、管理和利用 Internet 和 Web 信息资源。

本书针对 Web 信息系统应用中这些急需解决的问题，结合 Web 信息系统与技术在国内的发展及应用情况进行探讨，主要内容涵盖以下几个方面：Web 信息挖掘与检索、语义 Web 与智能 Web、Web 数据管理和信息集成、Web 应用框架和体系结构、Web 与信息系统安全性、Web Service 和工作流模型、自动文本索引与分类技术、Web 和网格技术等。

图书在版编目（CIP）数据

Web 信息系统与技术 / 孟小峰，徐宝文，于戈，倪巍伟主编. —北京：清华大学出版社，2005.9
ISBN 7-302-11513-3

I. W… II. ①孟… ②徐… ③于… ④倪… III. 计算机网络-程序设计 IV. TP393.09

中国版本图书馆 CIP 数据核字（2005）第 087657 号

出 版 者：清华大学出版社　　地　　址：北京清华大学学研大厦
　　http://www.tup.com.cn　　邮　　编：100084
　　社 总 机：010-62770175　　客户服务：010-62776969
组稿编辑：索　梅
文稿编辑：张为民
印 装 者：北京鑫海金澳胶印有限公司
发 行 者：新华书店总店北京发行所
开　　本：185×260　印张：22.75　字数：565 千字
版　　次：2005 年 9 月第 1 版　2005 年 9 月第1 次印刷
书　　号：ISBN 7-302-11513-3/TP · 7557
印　　数：1 ~ 500
定　　价：75.00 元

中国计算机学会
第二届全国 Web 信息系统及其应用学术会议（WISA2005）
组织机构

主办单位： 中国计算机学会电子政务与办公自动化专业委员会

承办单位： 东北大学
中国人民大学
东南大学
辽宁省计算机学会
辽宁省信息中心

大会主席： 于　戈

副 主 席： 何炎祥　卢正鼎　沈钧毅

程序委员会

主　席： 孟小峰

副主席： 徐宝文

委　员：（以姓氏笔画为序）

于　戈	王国仁	卢正鼎	孙志挥	刘　青	沈均毅
何炎祥	李　曦	李师贤	李庆忠	李瑞轩	邢春晓
孟小峰	杨　楠	陆建江	徐宝文	徐立臻	倪巍伟
彭智勇					

组织委员会

主　席： 王国仁

副主席： 徐立臻　刘　青

委　员：（以姓氏笔画为序）

王　斌	王大玲	邓庆绪	朱仕文	许　蕾
周晓宇	陈　林	赵志滨	倪巍伟	鲍玉斌

友情赞助单位

本次会议得到以下单位的大力支持，在此表示衷心的感谢！

Neusoft东软

东软软件股份有限公司
Neusoft Co.,Ltd.

DELL™

Dell(中国)有限公司
Dell Computer(China) Co., Ltd.

hp invent

中国惠普有限公司
China Hewlett-Packard Co., Ltd.

IBM

IBM 中国有限公司
IBM (China) Co., Ltd.

Microsoft® & O

辽宁立科信息工程有限公司
辽宁微软技术中心
Liaoning LIC Information Engineering Co., Ltd.
Liaoning Microsoft Technology Center

GENETEK
开元通盛

沈阳开元通盛科技有限公司
ShenYang KaiYuanTongSheng Technologies,INC.

前　言

随着信息化时代的到来，运行于 Internet 平台上的各种应用和服务得到了蓬勃发展，适用于网络平台交互、动态特性的各种技术层出不穷，在提高办公效率、节约资源消耗和扩大信息共享等方面发挥了巨大的作用。但随着 Internet 和 Web 技术的不断发展和更新，信息资源的数量越来越多且结构和内容越来越复杂，这使得管理、查询和使用这些数据信息资源变得越来越困难，人们迫切需要新的手段和方法，以更好地组织、管理、利用 Internet 和 Web 信息资源。

中国计算机学会电子政务与办公自动化专业委员会已经举办过 5 届年会，前 4 届的名称为“全国电子政务与办公自动化学术会议”。为适应形势发展的需要，专委会在 2003 年 4 月扬州的“第四届全国电子政务与办公自动化学术会议”上决定，将其年会改名为“全国 Web 信息系统及其应用学术会议（WISA）”。第 1 届会议已于 2004 年 10 月在武汉举办。本次是“第二届全国 Web 信息系统及其应用学术会议（WISA2005）”，将于 2005 年 9 月下旬在沈阳举办，由东北大学、东南大学和中国人民大学承办。

本次会议于 2004 年末开始征文。在征文过程中得到了中国计算机学会总部以及《计算机学报》、《软件学报》、《计算机研究与发展》、《计算机科学》、《小型微型计算机系统》、《计算机工程》、《计算机工程与应用》、《计算机应用研究》、《微机发展》、《计算机工程与设计》、《系统工程与电子技术》、《计算机世界》、《中国计算机报》等数十家报刊、期刊编辑部的大力支持，他们免费为我们发表或转载了会议消息或征文通知。至征文截止日 2005 年 4 月 10 日，本次会议共收到应征论文 586 篇，经程序委员会 4 月 16 日在北京召开的审稿会议决定，录用其中的 239 篇，录用率约 41%。其中，62 篇作为英文稿在《武汉大学学报（英文版）》（EI 源刊）上发表，118 篇在《计算机科学》上发表，59 篇在清华大学出版社出版的会议论文集《Web 信息系统与技术》上发表。

东北大学承担了大量的会务工作，中国人民大学和东南大学为论文的征集、审稿、编辑、出版做了大量的工作。录用的全部论文委托东南大学徐宝文教授审阅。东南大学陈林和周晓宇分别负责《武汉大学学报（英文版）》和《计算机科学》发表的论文的通联、编辑工作，东南大学陆建江、倪巍伟具体负责清华大学出版社出版的论文集《Web 信息系统与技术》的通联、编辑工作。

《武汉大学学报（英文版）》编辑部、《计算机科学》编辑部和清华大学出版社对本次会议论文的编辑、出版给予了极大的支持，我们在此代表全体论文作者及与会代表对各有关单位和个人致以衷心的谢意！

尽管我们在本论文集的编辑、出版过程中尽了很大的努力，但因才疏学浅，一定还会存在不少错漏与谬误，祈望各位同行、论文作者及与会代表不吝赐教和批评指正。

第二届全国 Web 信息系统及其应用学术会议程序委员会

2005 年 7 月

目　录

Web 信息挖掘与检索

基于距离的划分聚簇算法……叶若芬　李春平（3）
一个内容择优的 Web Word 文档自动搜集系统……钱丽萍（8）
基于 Web Services 的分布式元搜索引擎的设计……魏振达　阳小华　刘军（13）

语义 Web 与智能 Web

一种基于语义的系统集成模型的研究与应用……连惠群　郭红（21）
基于从 OWL-S 到 BPEL 映射的 UDDI……王艳（27）
语义 Web 服务标记语言 OWL-S……姜久雷　郝克刚（33）

Web 数据管理和信息集成

XML 在 PDM 系统中的 Web 设计与应用……刘舒（37）
超负载 Web 服务器的性能研究……唐龙业　孙倩　周宗平（43）
基于 Web 的医学脑图像数据库……汤天宇　焦蕴　阮宗才（48）

Web 应用框架和体系结构

系级教务管理支撑平台的设计与实现……鲍丽薇　叶晓俊（57）

Web 与信息系统安全性

网络入侵检测中数据采集技术的研究……窦伟平　邱伟　李传林（65）
基于 Web 服务的统一身份认证服务代理的实现……都艺兵　于华　杨莉萍（70）
实现 Web 服务器安全的移动监控……刘友生　杨宇　陈一平（75）
基于 Kylin 系统的加密文件系统设计与实现……钟经伟　付松龄　廖湘科（81）
加密技术在政府专网中的应用……王浩　徐心和（86）
基于 ASP.NET 页面指纹的分组式权限管理方案……刘军　阳小华　杨星（92）
一种信任类型动态定义模型的研究……张仕斌　何大可（97）
一体化物流在电子商务中的应用……夏露（103）
基于 Web 的可扩展建模与仿真框架的安全性研究……何明　裘杭萍　刘晓明（110）

基于 Globus 的网格安全认证机制的研究与实现 …… 夏捷 蔡洪斌 杨春（114）
一种基于代理的分布式入侵检测系统设计 …… 韦大伟（120）

Web 服务和工作流模型

Web 服务在企业信息资源规划中的应用研究 …… 刘纪敏 贺国平（129）
Web 服务请求和响应模式的一种改进 …… 刘振鹏 石磊 薛玉倩 牛晓霞（135）
一种基于本体的 Chord 网服务发现方法 …… 刘彬 张世栋（140）
基于数据挖掘的电子商务系统 …… 汤宗健 梁革英（145）
一种采用 Web Service 技术基于 Ontology 的语义检索 …… 王宗纬 朱国进 苏祥（149）
基于 Web 服务的网络教学软件开发模型 …… 肖化昆（154）
LF-Grid 服务描述的语义扩展 …… 李霞 李庆忠（160）
在面向服务的架构 SOA 中集成 Web 服务与网格服务 …… 王克敏 王永滨 刘曙元（167）
基于 Web 服务的电子商务信息共享系统的研究 …… 刘琼 赵韩 董玉德（173）
基于工作流的动态 CAD 图纸管理模型 …… 程维 朱志良 杨广明（177）
基于工作流的临床诊治系统的设计 …… 徐大华 赵阳（182）
基于 Web Service 技术构筑企业应用集成 …… 于华（187）

自动文本索引与分类技术

并行算法的同构化表示法 …… 周启海（195）

Web 和网格技术

计算网格中任务分配的设计与实现 …… 向俊凌 蔡洪斌 吴跃（201）
计算网格互认证响应时间的一种模型及其模拟验证
…… 林欣 刘晖 李明禄 孟彦文（207）
电力应用网格的研究 …… 杨薛明 张宪 毛慧娜 袁江野（215）
基于 API 调用的异常检测的缓冲区溢出防范方法 …… 林国敏 谭毓安 曹元大（220）
UML 建模过程中一致性和完整性的保证机制 …… 王祯 赵合计（224）
一种分布协作的 Web 代理缓存系统 DCPCS …… 韩向春 郭婷婷 王璿 林星宇（231）
基于 Web 的个性化智能教学系统 …… 刘敏昆 李志平（237）
基于 Web 方式的行业电子政务系统应用分类和技术实现方式 …… 杜守国 李文（242）
基于 Web 远程计算机审计平台的研究 …… 黄作明 丛秋实（247）
一个基于 Web 的网上商品质量保证平台的解决方案 …… 张文东 刘胜全（251）
电子政务安全中无线局域网认证的实现 …… 刘斌 李仁发 喻飞 徐成（255）
网络拓扑自动发现算法研究与优化 …… 吕爱丽 魏海平 叶小涛 赵林（263）

其他相关技术

“一站式”电子政府数据模型研究……李莉 林硕 刘辉林 王国仁（271）
电子政务安全体系模型研究与实践……沈瑞鑫 胡华平（278）
图像特效处理设计与实现……杨慧娟 张森（286）
另一种在基于 Lotus Domino 的 OAS 中实现对 Word 文档在线编辑的方法
……郑炎雄 陈传波 金寿吉（291）
电子政务中的信息安全性设计……赵永斌 范通让 胡予濮（298）
电子政务工程民间组织网络信息管理系统的设计与实现……应红燕 杨伟兵（303）
分布式企业信息管理系统的负载测试……刘靖 叶新铭（309）
基于分布式应用技术的医院信息系统……马德新 路新春（315）
基于 MS Word 2000 平台的计算机辅助教学系统的开发
……严明 陈欣 张丹 郝春炜（320）
知识管理理论框架研究……尚凌卉 刘鹏（327）
专题资源信息仓库的技术研究与实现……杨聚祥 尹健 刘永昌 张静（333）
设计模式及其中 AMCCS 中的应用……王正俊 顾宏斌（339）
基于构件技术实现 ERP 系统的开发过程以及构件检索方法……鲁佩云（346）

Contents

Web Information Mining and Retrieval

Distance-based Partition Clustering Algorithm ····· Ye Ruofen Li Chunping (3)
An Automatic Collector for Web Word Documents with User Preferred Content ····· Qian Liping (8)
The Design of the Meta-search Engine Based on Web Services ····· Wei Zhenda Yang Xiaohua Liu Jun (13)

Semantic Web and Intelligent Web

The Research and Application of Semantic Based System Integration Model ····· Lian Huiqun Guo Hong (21)
UDDI Based on the Mapping from OWL-S to BPEL ····· Wang Yan (27)
The Semantic Web Services Markup Language OWL-S ····· Jiang Jiulei Hao Kegang (33)

Web Data Management Information Integration

Web Design and Application of XML in PDM ····· Liu Shu (37)
Research on the Performance of Overloaded Web Server ····· Tang Longye Sun Qian Zhou Zongping (43)
A Brain Atlas on Web ····· Tang Tianyu Jiao Yun Ruan Zhongcai (48)

Web Application Framework and Architecture

Design and Implementation of Supporting Platform of Department Teaching Management ····· Bao Liwei Ye Xiaojun (57)

Web Information Security

The Research of Data Collection Technology in Intrusion Detection System ····· Dou Weiping Qiu Wei Li Chuanlin (65)
The Implementation of a Single Logon Authentication Service Agent Based on Web Services ····· Du Yibing Yu Hua Yang Liping (70)

Mobile Monitoring for Web Server's Safe Realization
..... Liu Yousheng Yang Yu Chen Yiping (75)
Design and Implementation of the Cryptographic File System for Kylin
..... Zhong Jingwei Fu Songling Liao Xiangke (81)
The Application of Encryption Technology to Government Special Net
..... Wang Hao Xu Xinhe (86)
An Authority by Group Management Scheme Based on the Fingerprint of Web Page in ASP.NET ..... Liu Jun Yang Xiaohua Yang Xing (92)
Research of a Dynamic Definition Model of Trust Class ..... Zhang Shibin He Dake (97)
Electronic-commerce Based Cooperative Logistics and Development Strategy ..... Xia Lu (103)
Study on the Security Based on XMSF ..... He Ming Qiu Hangping Liu Xiaoming (110)
Research and Implement of Globus-Base Grid Security Authentication Infrastructrue ..... Xia Jie Cai Hongbin Yang Chun (114)
Design of a Proxy-Based Distributed Intrusion Detection System ..... Wei Dawei (120)

Web Service and Workflow Models

The Application and Research of Web Services in Enterprise Information Resource
..... Liu Jimin He Guoping (129)
An Improved Web Services Request and Response Pattern
..... Liu Zhenpeng Shi Lei Xue Yuqian Niu Xiaoxia (135)
An Ontology-Based Approach to Web Services Discovery in Chord
..... Liu Bin Zhang Shidong (140)
An Electronic Commerce System Based on Data Mining
..... Tang Zongjian Liang Geying (145)
An Ontology-Based Semantic Retrieval Using Web Service Technology
..... Wang Zongwei Zhu Guojin Su Xiang (149)
Development Model of Networked Teaching Software Based on Web Services
..... Xiao Huakun (154)
A Semantic Extension of LF-Grid Service Description ..... Li Xia Li Qingzhong (160)
Using Web Service and Grid Service in the Service-Oriented Architectures
..... Wang Kemin Wang Yongbin Liu Shuyuan (167)
Web Service-Based System for Information Sharing between E-Commerce Systems
..... Liu Qiong Zhao Han Dong Yude (173)
Workflow-based Dynamic CAD Document Management System
..... Cheng Wei Zhu Zhiliang Yang Guangming (177)
Designing on Treatment Information System Based on Workflow
..... Xu Dahua Zhao Yang (182)
Building Enterprise Application Integration Based on Web Service ..... Yu Hua (187)

Automatic Text Indexing and Classification

An Isomorphized Representation for Parallel Algorithm Design ····· Zhou Qihai (195)

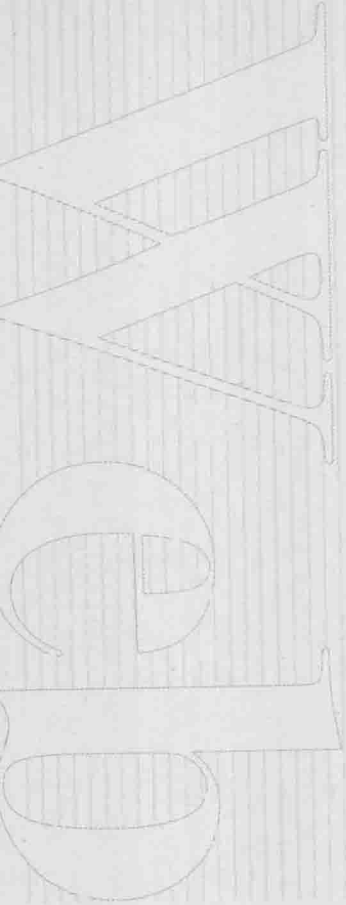
Web and Grid Technology

Design and Implementation of Computional Grid for Task Distribution
····· Xiang Junling Cai Hongbing Wu Yue (201)
A Model to Compute Response Time of Mutual Authentication in Grids and
Its NS Simulation ····· Lin Xin Liu Hui Li Minglu Meng Yanwen (207)
The Study on the Framework of Power Applied Grid
····· Yang Xueming Zhang Xian Mao Huina Yuan Jiangye (215)
Buffer Overflow Protection Based on Anomaly Detection of DLL Functions
····· Lin Guomin Tan Yuan Cao Yuanda (220)
Mechanism for Ensuring Consistency and Completeness in UML Design
····· Wang Zhen Zhao Heji (224)
A Distributed and Cooperative Web Proxy Caching System
····· Han Xiangchun Guo Tingting Wang Xuan Lin Xingyu (231)
Web-based Personalized Intelligent Tutoring System ····· Liu Minkun Li Zhiping (237)
A Classification and Implementation of E-government Based on Web-Server
····· Du Shouguo Li Wen (242)
Web-based Remote Computer Auditing Platforms ····· Huang Zuoming Cong Qiushi (247)
A Solution for Guarantee Platform of Goods Quality Based on Web
····· Zhang Wendong Liu Shengquan (251)
Research on WLAN Authentication by Security E-government
····· Liu Bin Li Renfa Yu Fei Xu Cheng (255)
Research and Optimization of Automatic Network Topology Discovery Algorithm
····· Lv Aili Wei Haiping Ye Xiaotao Zhao Lin (263)

Other Related Technologies

Studies of one-stop E-government Data Model
····· Li Li Lin Shuo Liu Huilin Wang Guoren (271)
The Study and Practice for E-government Security Framework Architecture Model
····· Shen Ruixin Hu Huaping (278)
Designing and Realization of Image Specially Good Effect Processing
····· Yang Huijuan Zhang Sen (286)

Another Way to Edit Word Documents Online in OAS Based on Lotus Domino
..... Zheng Yanxiong Chen Chuanbo Jin Shouji (291)
The Security Modal of E-government System
.....Zhao Yongbin Fan Tongran Hu Yupu (298)
Design and Implementation of Electronic Government Affair, MIS for Non-Governmental Organization..... Ying Hongyan Yang Weibing (303)
Workload Testing of Distributed Enterprise Management Information System
..... Liu Jing Ye Xinming (309)
Hospital Information Systems Based on Distributed Application Technology
.....Ma Dexin Lu Xinchun (315)
A Research into Computer Assistant Teaching System Inside MS Word 2000
..... Yan Ming Chen Xin Zhang Dan Hao Chunwei (320)
Research on Framework of Knowledge Management .....Shang Linghui Liu Peng (327)
The Research and Implementation of a Special Subject Information Storehouse
..... Yang Juxiang Yin Jian Liu Yongchang Zhang Jing (333)
Design Pattern and Its Application in AMCCS ..... Wang Zhengjun Gu Hongbin (339)
ERP System Development Based on Commonent and Commonent Retrieval Method ..... Lu Peiyun (346)



8 Web信息挖掘与检索

Web信息检索与挖掘

基于距离的划分聚簇算法

叶若芬 李春平

（清华大学软件学院 北京 100084）

摘 要：k-means 算法在聚簇大的数据集时是公认比较有效的算法之一，然而它只能应用在具有数值属性描述的数据对象集合上，这种数据对象叫做数值数据；却无法应用于真实世界中具有其他形形色色属性的数据对象集合上，比如颜色、纹理、形状等特征描述的数据对象集合，这种数据叫做分类数据。为了能对分类数据进行聚簇，对 k-means 算法进行了扩展，出现两种新的算法：一种是 k-modes 算法，另一种是 k-prototypes 算法。但这两种算法都需要用户事先确定聚簇数 k、阈值 t 和聚簇中心 Q，在不明白数据分布状况的情况下能较准确地确定这3个参数值是很不容易的，改进的k-modes算法有效解决了这一问题。

关键词：聚簇，k-means，k-modes，k-prototypes，相异度

Distance-based Partition Clustering Algorithm

Ye Ruofen Li Chunping

（*School of Software, Tsinghua University, Beijing 100084, China*）

Abstract: The k-means algorithm is well known for its efficiency in clustering large data sets. However, working only on numeric values prohibits it from being used to cluster real world data containing categorical values, such as those data whose attributes is color, texture and shape etc. To cluster categorical values,the k-modes algorithm and k-prototypes algorithm were presented. Yet it is necessary for users to predefine the number of clusters, the center of a cluster and the initial threshold for these algorithms. It is difficult to judge the number of clusters and the initial threshold while not understanding the distribution of the original data. The issue is addressed in this paper for an improved k-modes algorithm.

Key words: Cluster，k-means，k-modes，k-prototypes，Dissimilarity

1 引言

数据挖掘是数据库研究、开发和应用最活跃的分支科学之一，从大量数据中用非平凡的方法发现有用的知识和人们感兴趣的数据模式成了人们的一种自然需求[1]。随着数据挖掘研究的蓬勃发展，出现很多数据挖掘的方法，其中聚簇是最基本的方法，它既可以独立

作者简介 叶若芬（1968—），女，硕士生，主要研究方向为数据挖掘，E-mail：yeruofen@eyou.com，电话：13503178621。李春平（1965—），男，副教授，主要研究方向为数据挖掘、智能信息处理等，E-mail：cli@tsinghua.edu.cn，电话：010-62795437。

地应用，也可以作为其他数据挖掘方法的前期工作。在聚簇方法中，k-means 算法是最著名和最常用的划分法之一，根据实际需要相继出现了许多 k-means 算法的变种。k-means 算法能有效地处理规模较大和高维的数据集合，但却只能聚簇数值数据，因为数值数据能用欧几里德距离测量不同数据对象之间的相异度；不能处理非数值数据，即分类数据。不过现实生活中碰到的数据类型是各种各样的，要发挥数据挖掘工具应有的作用，设计混合型数据的数据挖掘工具已经成为必然趋势。为了能对分类数据进行聚簇，对 k-means 算法进行了扩展，出现两种新的算法：一种是 k-modes 算法，另一种是 k-prototypes 算法。k-modes 算法用了一种简单的相异度测量处理分类数据，用新的相异度值进行聚簇的过程和 k-means 算法是一样的；k-prototypes 算法结合了 k-means 算法和 k-modes 算法的相异度测量方法聚簇数值型和分类型的混合数据[2, 3]。这两种扩展的算法对聚簇大的数据集以及高维的数据集也都是很有效的，不足之处在于要预先确定把原始数据集分成几个簇，聚簇数 k 的值对聚簇结果能产生很大影响，关于这一点，本文提出了一种比较有效的解决方法。

2 介绍 k-means 算法

把数据分成几组，按照定义的测量标准，同组内数据与其他组数据相比具有较强的相似性，这就叫聚簇[4]。聚簇是数据挖掘最基础的操作，但现在存在的一些传统聚簇方法已不能满足处理复杂类型的、高维的、任意分布形状的数据集合的需要。

k-means 算法就是用得最多的一种传统的聚簇方法，是一种划分法，相似度的计算是求数据对象与簇中心的距离，与簇中心距离近的就划为一个簇。其工作流程如下：首先，随机地选择 k 个对象，每个对象初始地代表了一个簇的平均值或中心。对剩余的每个对象，根据其与各个簇中心的距离，将其赋给最近的簇。然后重新计算每个簇的平均值，求出新的簇中心，再重新聚簇。这个过程不断重复，直到准则函数收敛。该算法的时间复杂度是 O（nkt），其中 n 是所有对象数目，k 是簇的数目，t 是迭代次数。它的效率比较高；缺点是只能处理数值型数据，不能处理分类数据，对例外数据非常敏感，不能处理非凸面形状的聚簇[1]。

3 介绍 k-modes 算法

k-modes 算法改变了 k-means 算法的相异度测量方法，用一个简单的匹配相异度测量对分类数据进行聚簇处理。

3.1 简单的匹配相异度测量

设 $\boldsymbol{X}$、$\boldsymbol{Y}$ 是分类数据集中的两个对象，该对象是 $m(x_1, x_2, \cdots, x_m)$ 维的，则这两个对象之间的相异度为：

$$d(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

当 $x_j = y_j$ 时，$\delta(x_j, y_j) = 0$；当 $x_j \neq y_j$ 时，$\delta(x_j, y_j) = 1$。