

·信息工作·

数据挖掘研究现状综述

王立伟 (上海社会科学院图书馆 上海 200235)

摘要: 数据挖掘作为情报学最常用的分析手段得到各个领域的广泛关注, 每年KDD、PAKDD和ECML/PKDD三大学术会议的召开也给各国和地区进行学术交流提供便利。文章基于PAKDD学术会议和KDDuggets公司的统计数据对当前数据挖掘现状进行综述分析。

关键词: 数据挖掘 PAKDD

中图分类号: 351.11

文献标识码: A

文章编号: 1003-6938(2008)05-0041-06

The Summarization of Present Situation of Data Mining Research

Wang Liwei (The Library of Shanghai Academy of Social Sciences, Shanghai, 200235)

Abstract: The data mining, as the most useful analysis means of the information studies, is highly concerned from all the fields. Annually, the top three academic conferences which are KDD, PAKDD and ECMLPKDD also offer the convenience for the different countries and religions to communicate with each other academically. This thesis is based on the PAKDD academic conference and the statistics from KDDuggets company, analyzing the present situation of data mining comprehensively.

Key words: data mining; PAKDD

CLC number: G351.11

Document code: A

Article ID: 1003-6938(2008)05-0041-06

1 引言

上世纪九十年代,随着数据库系统的广泛应用和网络技术的高速发展,数据库技术也进入一个全新的阶段,即从过去仅管理一些简单数据发展到管理由各种计算机所产生的图形、图像、音频、视频、电子档案、Web页面等多种类型的复杂数据,并且数据量也越来越大。在给我们提供丰富信息的同时,也体现出明显的海量信息特征。

信息爆炸时代,海量信息给人们带来许多负面影响,最主要的就是有效信息难以提炼。过多无用的信息必然会产生信息距离 the Distance of Information-state Transition,信息状态转移距离,是对一个事物信息状态转移所遇到障碍的测度,简称DIST或DIT^[1]和有用知识的丢失。这也就是约翰·内斯伯特 John Naisbert)称为的“信息丰富而知识贫乏”窘境。因此,人们迫切希望能对海量数据进行深入分析,发现并提取隐藏在其中的信息,以更好地利用这些数据。但仅以数据库

系统的录入、查询、统计等功能,无法发现数据中存在的关系和规则,无法根据现有的数据预测未来的发展趋势,更缺乏挖掘数据背后隐藏知识的手段。正是在这样的条件下,数据挖掘技术应运而生。

2 数据挖掘研究现状

2.1 学术研究

(1) KDD(Knowledge Discovery in Databases)国际学术大会

数据挖掘技术出现于20世纪80年代末,它促成了数据库中的知识发现(KDD)产生。在1989年美国底特律召开的第十一届国际联合人工智能学术会议上首次提到知识发现这一概念,到1993年,美国电气电子工程师学会(IEEE)的知识与数据工程 Knowledge and Data Engineering 会刊出版了KDD技术专刊,发表的论文和摘要体现了当时KDD的最新研究成果和动态。

随着来自各个领域的研究人员和应用开发者不断增多,

1995年在加拿大蒙特利尔召开了首届KDD国际学术年会,会上把数据挖掘技术分为工程领域的数据挖掘与科研领域的知识发现。^[2]此后,此类会议每年召开一次,数量和规模逐渐扩大,从专题研讨会一直发展到国际学术大会,并成为当前计算机领域的研究方向和研究热点。目前对KDD的研究主要围绕理论、技术和应用这三个方面展开。

据统计显示,从1995年至2007年召开的13次KDD国际学术大会中,9次都在美国主要城市(如纽约、芝加哥、华盛顿等)举办,其余4次均在加拿大举办(见表1),从未在北美以外地区举办过。

表1 KDD (Knowledge Discovery and Data Mining) Meetings^[3]

International Conference on KDD	Date	City
13 th	August 2007	San Jose, CA, USA
12 th	August 2006	Philadelphia, PA, USA
11 th	August 2005	Chicago, IL, USA
10 th	August 2004	Seattle, WA, USA
9 th	August 2003	Washington, DC, USA
8 th	August 2002	Edmonton, Alberta, Canada
7 th	August 2001	San Francisco, CA, USA
6 th	August 2000	Boston, MA, USA
5 th	August 1999	San Diego, CA, USA
4 th	August 1998	New York, NY, USA
3 th	August 1997	Newport Beach, CA
2 th	August 1996	Portland, OR
1 th	August 1995	Montreal, Canada

(2) PAKDD (Pacific-Asia Conference on KDD) 学术会议

1997年,也就是首届蒙特利尔KDD国际学术大会召开之后的2年,PAKDD学术会议(Pacific-Asia Conference on KDD)在亚太地区顺利召开,这标志着亚太地区数据挖掘研究进入发展时期。PAKDD会议每年召开一次,从1997年至2007年的11年中,亚洲和大洋洲的主要国家都成功举办过该项会议(见表2)。其中,新加坡第十届PAKDD会议除了进行数据挖掘学术研究外,还与新加坡统计协会(SIS)、新加坡模式识别和机器智能协会(PREMIA)共同组织了一场基于解决电信运营商问题的数据挖掘竞赛。其内容为“如何区分移动通讯

网客户中使用第二代(2G)和第三代(3G)服务的用户”,旨在明确目前2G网络用户中哪些使用者具有巨大的潜在可能性转移到使用移动运营商的3G移动网络和服务上。

表2 Pacific-Asia Conference on KDD (PAKDD)^{[4][5]}

PAKDD	Date	City
11 th	May 2007	Nanjing, China
10 th	April 2006	Singapore
9 th	May 2005	Hanoi, Vietnam
8 th	May 2004	Sydney, Australia
7 th	April 2003	Seoul, Korea
6 th	May 2002	Taipei, Taiwan
5 th	April 2001	Hong Kong, China
4 th	April 2000	Kyoto, Japan
3 th	April 1999	Beijing, China
2 th	April 1998	Melbourne, Australia
1 th	1997	Singapore

与KDD国际学术会议(ACM SIGKDD International Conference on Knowledge Discovery and Data Mining)或ECML/PAKDD学术会议(European Conference on Machine Learning & European Conference on Principles and Practice of Knowledge Discovery in Databases)定期举办竞赛模式不同,新加坡PAKDD会议是继2000年第四届京都PAKDD会议后,第二次举办类似的比赛。之前,京都PAKDD会议曾有过使用医学数据进行数据挖掘比赛的历史记录。^[6]

2001~2007共7年时间中,PAKDD会议依次由香港、台北、首尔、悉尼、河内、新加坡和南京主办。根据对主办方出版的论文集Advances in Knowledge Discovery and Data Mining统计显示,7年中共有32个国家和地区共计593篇参会论文被收录论文集。其中澳大利亚、韩国、加拿大、美国、日本、台湾、香港和中国内地每届被收录的论文总和超过论文总数的60%。2001年香港会议收录论文最多的为美国和香港,所占比例均为12.70%;2002年台北会议收录论文最多的为台湾,所占比例为21.43%;2003年首尔会议收录论文最多的为韩国,占20.00%;2004年悉尼会议收录论文最多的为澳大利亚,占19.28%;2005年河内会议收录论文最多的为中国,占19.00%;2006年新加坡会议收录论文最多的为美国,占18.81%;2007年南京会议收录论文最多的为中国,占45.38%(见表3)。可见,PAKDD会议的主办权对一个国家数据挖掘研究具有非常积极的促进作用。

统计显示,上述国家和地区中,仅美国每届被收录的参会论文比重超过10%,最高时为2006年新加坡会议,比重为

表3 2001-2007年主要国家和地区被收录论文比重表

	2001 年香港	2002 年台北	2003 年首尔	2004 年悉尼	2005 年河内	2006 年新加坡	2007 年南京
澳大利亚	11.11%	3.57%	10.00%	19.28%	8.00%	4.95%	8.46%
韩国			20.00%	3.61%	5.00%	5.94%	6.92%
加拿大	4.76%	12.50%	6.67%	1.20%	2.00%	2.97%	0.77%
美国	12.70%	12.50%	15.00%	14.46%	13.00%	18.81%	10.00%
日本	7.94%	8.93%	8.33%	6.02%	9.00%	6.93%	3.85%
台湾	9.52%	21.43%	3.33%	3.61%	3.00%	9.90%	4.62%
香港	12.70%	10.71%	6.67%	7.23%	1.00%	2.97%	2.31%
中国内地	6.35%	7.14%	6.67%	12.05%	19.00%	11.88%	45.38%
总和	65.08%	76.78%	76.67%	67.46%	60%	64.35%	82.31%

18.81%,可见美国数据挖掘研究实力和研究水平。中国对数据挖掘研究起步晚于美国,在2001-2003年中被收录论文比重较为稳定,为6%~7%。2004年比重有明显提高,较上年上升80%,并在2005年河内会议论文收录比重首次超过美国(美国为13.00%,中国为19.00%),在2007年南京会议中收录比重达到顶峰,比重接近50%。其余5个国家和地区每次收录论文比重多为10%以下,鲜有较高的收录比重。

通过对2001-2007年参会论文集进行目录词频分析,算法和最优算法^[7]研究从2001年开始一直成为PAKDD学术会议参会论文的重要组成部分,也是被选最多的论文主题。和算法相关的论文2001年有10篇,2002年有9篇,2003年有8篇,2004年有4篇,2005年有15篇,2006年有5篇,2007年有20篇。支持向量机(Support Vector Machines)和支持向量回归(Support Vector Regression)成为近年来研究的新方向,相关论文2005年收录4篇,2006年收录6篇,2007年收录10篇。

2.2 应用研究

(1) 应用领域

数据挖掘应用研究是指开发各种数据挖掘系统和工具,并在各个行业中的应用。目前的典型应用领域包括:市场分析和预测;如英国BBC广播公司进行的收视率调查、大型超市销售分析与预测、销售渠道与价格分析等;工业生产:主要用于发现最佳生产过程;金融:采用统计回归式神经网络构造预测模型,如自动投资系统(Automated Investor)、可预测最佳投资时机;科学研究;贝克(Bacon)对于天文定理的发现、地震发现者(Quake finder)用于分析地壳的构造活动等;Web数据挖掘;站点访问模式分析、网页内容自动分类、聚类;工程诊断。数据挖掘作为一种新的知识发现手段,还引起了工程诊断领域的重视,许多国家和研究机构都在监测诊断项目中加入了对数据挖掘的研究。^[7]

根据KDNuggets公司做的调查统计显示(见表4),2003~2005年期间,数据挖掘技术应用领域比重排在前3位的依次是CRM(客户关系管理)占34.90%,Banking(银行业)占34.23%和Credit Scoring(信用得分)占23.49%。2006年,数据挖掘技术应用领域比重前3位略有变化,除CRM仍然占据首位位置外(占38.74%),第二和第三依次是Fraud Detection(欺诈检测)占21.62%和Direct Marketing/Fundraising(直销/募款)占19.82%。2007年数据挖掘应用领域比重首位仍然是CRM占26.10%,第二位回归于银行业(占23.90%),第三位为直销/募款(占20.30%)。

随着数据挖掘研究的不断深入,数据挖掘应用领域的规模正在逐步扩大,其中较为显著的依次为Banking(银行业),Entertainment/Music(娱乐/音乐),Science(科学)和Health care/HR(卫生保健/人力资源),它们在2007年的应用比重较2006年增长100%以上,增长比率依次为3200%,200%,117%和100%(见表5)。

(2) 软件产业

由于数据挖掘技术在各领域被广泛应用,其软件市场需求量也变得很大。因此,包括国际知名公司在内的软件公司纷纷加入数据挖掘工具研发的行列中来。

根据National Center for Data Mining at UIC(University of Illinois at Chicago)的R.Grossman观点,数据挖掘软件的发展经历了4个时代:^[8]

第一代数据挖掘软件,支持一个或少数几个数据挖掘算法,这些算法设计用于数据向量挖掘,多用于商业系统。Salford Systems公司早期的CART系统就属于这种系统。新加坡国立大学研制的CBA,其基于关联规则的分类算法,能从关系数据或者交易数据中挖掘关联规则,利用关联规则进行分类和预测。

表4 数据挖掘应用领域比重统计表^{[9][10][11]}

Industries/fields	Proportion (%) 2003~2005	Proportion(%) June 2006	Proportion(%) June 2007
CRM	34.90	38.74	26.10
Banking	34.23	0.90	23.90
Direct Marketing/ Fundraising	22.82	19.82	20.30
Science	11.41	10.81	18.80
Fraud Detection	20.81	21.62	18.80
Telecom	15.44	12.61	15.20
Credit Scoring	23.49	18.92	13.80
Other	7.38	13.51	13.00
Biotech/Genomics	7.38	15.32	11.60
Web usage mining		10.81	10.10
Retail	16.78	9.91	10.10
Medical/Pharma	8.05	7.21	9.40
Insurance	16.11	10.81	8.70
Health care/HR	10.07	4.50	7.20
Government/Military	8.05	6.31	7.20
Financials/Leading			7.20
Web content mining/ Search		13.51	6.50
Manufacturing	12.75	6.31	6.50
E- commerce	7.38	5.41	5.80
Entertainment/Music	2.68	1.80	4.30
Social Policy/Survey analysis			3.60
Security/Anti- terrorism	3.36	4.50	3.60
Investment/Stocks	3.36	9.91	2.90
Travel/Hospitality	5.37	4.50	2.20
Junk email/Anti- spam	3.36	1.80	2.20
Web	6.04		
Gambling	1.34		
Voters	149	111	138

注：本文基于调查源数据对2003 -2005和2006年数据做了修正，修正公式：比重（Proportion）=单项选择数（Reply）/有效样本数（Voters）。

第二代数据挖掘软件系统与数据库管理系统（DBMS）集成，支持数据库和数据仓库，具有高性能的接口，具有较高的

可扩展性。能够挖掘大数据集以及更复杂的数据集和高维数据，但这一代的数据挖掘软件只注重模型的生成，典型代表有DB Miner和SAS Enterprise Miner。

表5 2007年不同领域应用数据挖掘技术较2006年的增长率^[12]

Industries/fields	Proportion of growth from 2006 to 2007
Banking	3200%
Entertainment/Music	200%
Science	117%
Health care/HR	100%
Medical/Pharma	63%
Junk email/Anti- spam	50%
Telecom	50%
Government/Military	43%
E- commerce	33%
Manufacturing	29%
Direct Marketing/Fundraising	27%
Retail	27%
Other	20%
Web usage mining	17%
Fraud Detection	8%
Insurance	0%
Security/Anti- terrorism	0%
Biotech/Genomics	- 6%
Credit Scoring	- 10%
CRM	- 16%
Travel/Hospitality	- 40%
Web content mining/Search	- 40%
Investment/Stocks	- 64%

第三代数据挖掘系统的特点是和预言模型系统之间能够实现无缝的集成，使得由数据挖掘软件产生的模型的变化能够及时反映到语言模型系统中，由数据挖掘软件产生的预言模型能够自动地被操作型系统吸收，从而与操作型系统中的语言模型相联合提供决策支持的功能。它能够挖掘网络环境下（Internet/Intranet/Extranet）的分布式和高度异质的数据，并且能够有效地和操作型系统集成。其缺点是不能支持移动环境。这一代数据挖掘系统关键的技术之一是提供对建立在异质系统上的多个预言模型以及管理这些预言模型的元数据提供第一级别的支持。SPSS Clementine 就是属于这一代的产品。

第四代软件能够挖掘嵌入式系统、移动系统和普遍存在

的计算设备产生的各种类型的数据。2001~2006年Kargupta作为马里兰巴尔的摩州立大学 University of Maryland Baltimore County) 正在研制的CAREER 数据挖掘项目的负责人, 其研究目的是开发挖掘分布式和异质数据的第四代数据挖掘系统。

目前国外已有很多技术成熟、有较强产业化能力的数据挖掘软件, 其中主要的有:

SAS Enterprise Miner: SAS系统全称为Statistics Analysis System, 是美国使用最为广泛的三大著名统计分析软件(SAS, SPSS和SYSTAT)之一, 被誉为统计分析的标准软件。1997年SAS发布了SAS Enterprise Miner, 这个工具为用户提供了用于建模的一个图形化流程处理环境, 并且它有一组常用的数据挖掘算法, 包括决策树、神经网络、回归、关联等, 还支持文本挖掘。

SPSS Clementine: SPSS是世界上最早的统计分析软件之一。1998年末SPSS收购了英国ISL公司, 通过继承获得了这家公司的Clementine数据挖掘包。Clementine是首次引入数据挖掘流概念的产品之一。它允许用户在同个工作流环境中清理数据、转换数据和构建模型。

IBM Intelligent Miner: 包括分析软件工具Intelligent Miner for Data和Intelligent Miner for Text, 不仅可以寻找包含于传统文件、数据库、数据仓库和数据中心中的隐含信息, 更允许企业从文本信息中获取有价值的客户信息。Intelligent Miner使用预测模型标记语言(Predictive Modeling Markup Language, PMML)来导出挖掘模型, 这种语言由数据挖掘协会(Data Mining Group, DMG)定义。

Insightful Mine(I-Miner): 由美国Insightful公司开发的具有高度可扩展性的数据分析和数据挖掘软件。目前在金融、生物科技、政府机构等企事业单位应用非常广泛。

此外, 还有Oracle公司从Thinking Machines公司取得的Darwin; Unica公司开发的Affinium Model; Angoss Software所开发的Knowledge SEEKER; 加拿大Simon Fraser大学开发的DB-Miner; SGI公司和美国Stanford大学联合开发的Minset; HNC公司开发的用于信用卡诈骗分析的Database Mining Workstation; IBM公司Almaden研究中心开发的Quest; Neo Vista开发的Decision Series; 以及KEFIR系统、SKICAT系统等。

国内也有不少新兴的数据挖掘软件:

DMiner: 由上海复旦德门软件公司开发的具有自主知识产权的数据挖掘平台。

iDMiner: 由海尔青大公司开发的具有自主知识产权的数据挖掘系统。其对国际通用业界标准的大胆采用, 为该软件今后的发展预留了很大的空间, 同时也为国内同类软件融

入世界及开发提供了一条新的思路。

MSMiner: 由中科院计算技术研究所智能信息处理实验室开发的多策略数据挖掘平台。

除此之外, 也有一些相关数据挖掘产品的报道, 如复旦德门公司开发的AR Miner和CIAS、东北大学开发的面向先进制造企业的综合数据挖掘系统Scope Miner、东北大学软件中心基于SAS开发的Open Miner以及长春工业大学开发的数据挖掘工具软件等。

根据Kdnuggets公司2007年5月做的调查统计显示, 商业数据挖掘软件使用比重较高的前5种数据挖掘软件均为国外软件, 其使用比重依次为: SPSS Clementine(21.72%)、Salford CART/MARS/TreeNet/RF(19.85%)、Excel(17.60%)、SPSS(17.04%)和SAS(14.98%)(见表6), 排名前10位中未有国内研发的数据挖掘软件。

3 结论

3.1 美国是全球数据挖掘研究最繁荣的地区, 并占据着研究的核心地位

作为在全球具有较大影响的KDD学术会议, 从1995年召开至今的13次会议中, 有9次在美国本土主要城市召开。此外, 作为亚洲及太平洋地区最重要的数据挖掘学术会议PAKDD会议虽说美国尚未担任过主办国, 但每届会议美国学者被收录的论文比重均超过10%, 最高时比重为18.81%。全球最著名的数据挖掘应用软件, 如SPSS、SAS和Excel均由美国公司研发, 目前应用比重排在调查的前五位。

3.2 亚太地区数据挖掘正逐步从理论研究走向应用研究

自1997年召开首届PAKDD学术会议以来, 理论学术研究一直是该会议的主要讨论内容, 但2000年第四届京都会议和2006年第十届新加坡会议都举办应用性较强的数据挖掘竞赛(京都会议举办医学方面的数据挖掘竞赛, 新加坡会议举办电信方面的数据挖掘竞赛), 从而使PAKDD会议和KDD国际学术会议与ECML/PKDD学术会议一样, 兼顾理论研究和应用案例研究。自此, 也标志着亚太地区数据挖掘逐步走向应用研究。

3.3 我国数据挖掘研究水平正在飞速发展

我国数据挖掘研究晚于美国, 21世纪方才起步。对PAKDD学术会议参会论文统计显示, 2001~2003年被收录论文比重仅为6%~7%, 2004年开始论文收录比重明显提高, 并逐年递增, 2005年河内会议论文收录比重首次超过美国(美国为13.00%, 中国为19.00%), 在2007年南京会议中收录比重达到顶峰, 比重接近50%。可见, 我国数据挖掘研究水平2004~2007年三年中正飞速发展。

表6 Data Mining/Analytic Software Tools (May 2007) ^[13]

Name	Proportion
Commercial Data Mining Software SPSS Clementine	21.72%
Salford CART/MARS/TreeNet/RF	19.85%
Excel	17.60%
SPSS	17.04%
SAS	14.98%
Angoss	14.61%
KXEN	13.11%
SQL Server	7.12%
MATLAB	5.62%
SAS E- Miner	4.68%
Other commercial tools	3.93%
Statsoft Statistica	2.81%
Insightful Miner/S- Plus	2.62%
Oracle DM	2.25%
Tiberius	2.06%
FairIsaac Model Builder	0.56%
Xelopes	0.37%
Miner3D	0.37%
Megaputer	0.19%
Your own code	11.42%
Free Data Mining Software	
Yale	19.29%
Weka	8.99%
R	7.87%
Other free tools	5.62%
C4.5/C5.0/See5	2.62%
Orange	2.25%
KNIME	0.37%

3.4 国内数据挖掘软件产业还不成熟,产品尚未被国际市场认可

国内有不少自主开发研制的数据挖掘软件,如DMiner、iDMine、MSMiner、AR Miner、CIAS、Scope Miner、Open Miner

等,但根据Kdnuggets公司2007年5月做的调查统计显示,排名前10位中未有国内研发的数据挖掘软件。可见,虽说国内各软件公司依托国内知名高校的科研实力开发数据挖掘软件,但是尚未被国际市场认可,在国际上的使用更为数甚少,国内数据挖掘软件产业还不成熟。

参考文献:

- [1] 王浣尘.信息距离与信息 [M].北京:科学出版社,2006: 26-27.
- [2] 陈文臣.Web日志挖掘技术的研究与应用 [D].北京:中国科学院研究生院,2005.
- [3] Past KDD(Knowledge Discovery and Data Mining Meetings [EB/OL] [2007-06-17].http://www.kdnuggets.com/meetings/past-meetings-kdd.html.
- [4] Meetings/Conferences in Data Mining, Knowledge Discovery, Web Mining [EB/OL] [2007-06-17].http://www.kdnuggets.com/meetings/index.html.
- [5] Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD) [EB/OL] [2007-06-17].http://www.informatik.uni-trier.de/~ley/db/conf/pakdd/index.html.
- [6] Nathaniel B.Norieh, Chew him Tan.A Look Back at the PAKDD Data Mining Competition 2006 [J].International Journal of Data Warehousing & Mining, 2007, (2) : 1- 11.
- [7] 蒋晓静,周定康.一种新的数据挖掘处理模型 [J].计算机与现代化,2003, (2) : 18- 20.
- [8] [13] Data Mining/Analytic Software Tools(May 2007) [EB/OL] [2007-06-17].http://www.kdnuggets.com/polls/2007/data_mining_software_tools.htm.
- [9] Successful Data Mining Applications\$ 2003- 2005[EB/OL] . [2007-06-17].http://www.kdnuggets.com/polls/2005/successful_data_mining_applications.htm.
- [10] Data Mining Applications- Industries(June 2006) [EB/OL] . [2007-06-17].http://www.kdnuggets.com/polls/2006/data_mining_applications_industries.htm.
- [11] Data Mining Applications by Industry(June 2007) [EB/OL] . [2007-06-17].http://www.kdnuggets.com/polls/2007/data_mining_applications.htm.
- [12] 吴婕.浅析数据挖掘软件的发展 [J].情报理论与实践, 2004, (2) : 212- 214.

作者简介:王立伟(1981-),男,硕士,上海社会科学图书馆助理工程师。