数据挖掘专利综述

刘晓东[1,2],刘大有[1,2]
(1. 吉林大学计算机科学与技术学院,吉林长春 130012; 2. 吉林大学符号计算与知识工程教育部重点实验室,吉林长春 130012)

摘 要: 尽管科学研究专利是反映科学研究成果的一个重要方面,专利申请本身是一项重要的科学研究工作,但是长期以来,专利所包含的科学研究成果在文献中却没有得到充分的反映. 由此,对著名的美国专利和商标委员会数据库(US PATENT & TRADEMARK OFFICE DATABASE)中数据挖掘专利的授权情况进行了分析. 对于专利授权比较集中的领域—关联规则、互联网挖掘、聚类算法和并行数据挖掘等方面中的代表性专利进行了总结和分析. 最后,指出了当前数据挖掘专利的一些空白领域.

关键词: 数据挖掘; 专利; 关联规则; 互联网挖掘; 聚类算法; 并行数据挖掘
中图分类号: TP311, TP18 **文献标识码**: A **文章编号**: 0372-2112 (2003) 12A-1989-05

Data Mining Patent Summarization

LIU Xiao-dong[1,2], LIU Da-you[1,2]
(1. *School of Computer Science and Technology, Jilin University, Changchun, Jilin* 130012, *China*;
2. *Ministry of Education Key Laboratory of Symbolic Computation and Knowledge Engineering, Jilin University, Changchun, Jilin* 130012, *China*)

Abstract: Patent of scientific research is one important aspect which reflects the fruits of scientific researches, and patent application itself is one of the most important scientific researches as well, whereas it has not been paid enough attention for a long time in scientific literature. We analyze data mining patent grant in us patent & trademark office database, which is authoritative in the world. We introduce some typical patents in the fields which are concerned with most patents, including association rules, web mining, clustering algorithms, and parallel data mining etc. In the end, we present data mining topics which are currently blank in the field of data mining patent.

Key words: data mining; patent; association rules; web mining; clustering algorithms; parallel data mining

1 引言

1989 年 8 月,在美国底特律召开的第 11 届国际人工智能联合会议的专题讨论会上,"数据库中的知识发现"这一概念首次被提出. 在 1991、1993 和 1994 年都举行了数据挖掘专题讨论会. 随着参加会议人数的增多,从 1995 年开始,每年都举办一次有关数据挖掘技术研究的国际会议. 1997 年,介绍数据挖掘研究的杂志《知识发现与数据挖掘》创刊. 随着数据挖掘研究的不断深入,其研究成果越来越受到业界的关注[1].

众多的数据挖掘与知识发现系统和工具不断投入市场. 较有代表性的数据挖掘工具主要有:美国 George Mason 大学 Ryszard S Michalski 领导开发的 INLEN 系统[2],它结合了数据库、知识库和一个适用范围较宽的机器学习技术来辅助数据分析专家从数据库或知识库中抽取知识并发现有价值的规律;美国 Kansas 大学开发的 LERS 系统[3],是基于粗集理论的数据挖掘工具;美国 SPSS 公司研制了著名的数据挖掘工具箱 Clementine[4],Clementine 主要使用了神经网络、决策树和规则推导等技术,在实际应用中取得了很好的效果;美国德士古公司的 GeoProbe 数据挖掘系统对地震数据进行评估,在发现尼日利亚近海的阿哥巴米大油田的过程中起了重大作用,该油田储量大约为 14.5 亿桶[5];中国科学院计算技术研究所智能信息处理开放实验室研制成功的多策略数据挖掘平台,提供决策树、支持向量机、粗糙集、模糊聚类、基于范例推理、统计方法、神经计算、可视化等多种数据挖掘算法,支持特征抽取、分类、聚类、预测、关联规则发现、统计分析等数据挖掘功能,并支持高层次的决策分析功能[6].

近年来,数据挖掘研究取得了重要进展,数据挖掘专利的数目与日俱增. 考虑到专利是反映科学研究成果的一个重要方面,本文以著名的美国专利和商标委员会数据库中数据挖掘专利授权情况为背景[7],介绍数据挖掘研究取得的进展和发展趋势. 通过检索,我们得到 1997 年至 2002 年已授权的数据挖掘专利 101 项,表 1 列出了已授权的数据挖掘专利按年份的分布情况:

从表 1 中可以观察到:从 1997 到 2002 年,已授权的数据

收稿日期:2003-04-02;修回日期:2003-10-20
基金项目:国家自然科学基金(No. 60173006)

挖掘专利数目逐年递增,这反映了数据挖掘领域近年来越来越活跃的研究趋势.下面我们按专利授权比较集中的领域——关联规则、互联网挖掘、聚类算法和并行数据挖掘等,分别介绍了一些有代表性的专利.这些专利反映了相对应领域的研究成果和研究趋势.最后,本文讨论了当前数据挖掘研究的一些热点问题.

表 1 数据挖掘专利按年份分布

年 份	1997	1998	1999	2000	2001	2002
专利数	2	10	12	19	24	34

2 关联规则

在已授权的 101 项数据挖掘专利中,与关联规则(association rules)挖掘技术有关的专利 19 项.关联分析(association analysis)发现关联规则,这些规则展示了属性—值频繁地在给定数据集中一起出现的条件.关联分析广泛用于购物篮和事务数据分析.其形式化描述为:关联规则是形如 $X \Rightarrow Y$,即"$A_1 \wedge A_2 \wedge \cdots \wedge A_m \Rightarrow B_1 \wedge B_2 \wedge \cdots \wedge B_n$"的规则,其中,$A_i(i \in \{1, \cdots, m\})$,$B_j(j \in \{1, \cdots, n\})$是属性—值对.关联规则 $X \Rightarrow Y$ 解释为"满足 X 中条件的数据库元组多半也满足 Y 中条件"[8].

关联规则一直是热点研究课题.研究内容主要包括一般关联规则挖掘方法、定量关联规则挖掘方法、在挖掘中去除垃圾规则的方法(既减少无用关联规则,又提高挖掘效率)、提高挖掘效率的并行挖掘方法、减少 CPU 和 I/O 成本的关联规则挖掘方法、在线挖掘方法以及基于大型数据集和集成数据库的关联规则挖掘方法等等.从与关联规则有关的专利中,我们能看到近几年来关联规则的研究成果和研究重点,下面对这方面有代表性的专利进行介绍.

在 19 项与关联规则有关的专利中,与一般性的关联规则挖掘方法有关的专利共有 10 项.IBM 公司 Agrawal 等人最早获得与关联规则挖掘有关的专利.这里主要介绍 Agrawal 等人的 2 项专利,《数据库中一般关联规则的挖掘方法和系统》(5,615,341)和《在大型关系表中挖掘定量关联规则的方法和系统》(5,724,573).括号内是按时间排序的与专利一一对应的专利号,凭专利号可直接到数据库中查询到该专利.专刊《数据库中一般关联规则的挖掘方法和系统》提供了发现消费者购买趋势的方法和系统.数据库存放的是有关消费者购买各种物品的历史数据,消费者的一次购买活动对应一个数据记录(即一行数据),一种物品对应一列数据.通过计算每两个物品或更多物品间共同发生的频率大小,来决定哪些物品之间有关联,从而生成关联规则.专利《在大型关系表中挖掘定量关联规则的方法和系统》提供了解决连续属性离散化的一个方法,它首先将选定的定量属性的取值划分为区间(intervals),合并相邻的属性值和区间为一个范围(ranges),在产生备选项目集的基础上决定频繁项目集,输出满足预先设定条件的关联规则.其余 8 项专利是《数据挖掘中的维数约简方法》(6,032,146)、《用关联规则进行维数约简的数据挖掘方法》(6,134,555)、《生成加权关联规则的方法和工具》(6,173,280)、《在数据中挖掘关联规则的方法》(6,185,549)、《发现属性间关联规则的数据挖掘工具》(6,272,478)、《使用循环关联规则的数据挖掘方法》(6,278,998)、《使用对象关系扩展的关联规则挖掘方法》(6,301,575)和《挖掘具有权值的关联规则挖掘方法》(6,415,287).

与提高关联规则挖掘效率方法有关的专利共有 4 项,其中有 3 项是 IBM 公司 Agrawal 等人申请的.我们主要介绍一下 Agrawal 的专利《基于属性约束的关联规则挖掘方法》(6,061,682),该专利主要解决关联规则垃圾问题.该专利实现了在挖掘过程中根据约束条件筛选关联规则,而不是产生所有关联规则后再进行筛选,该方法明显提高了挖掘效率.其余 3 项专利是《数据库中快速挖掘关联规则的方法和系统》(5,794,209)、《数据库中关联规则的并行挖掘方法和系统》(5,842,200)和《在关联规则挖掘中减少 CPU 和 I/O 成本的改进方法和系统》(5,813,003).

与特殊环境下关联规则挖掘方法有关的专利共有 5 项.IBM 公司 Aggarwal 等人申请的.下面我们主要介绍一下 Aggarwal 的《在线挖掘定量关联规则的方法》专利(6,092,064),该专利主要解决在线环境下关联规则的挖掘问题.该专利分为预处理和规则生成 2 个阶段.在预处理阶段首先针对前件属性(antecedent attributes)构建多维检索结构,规则生成阶段在此基础上合并符合条件的区域给出有层次的规则集.其余 4 项专利是《在线挖掘关联规则的方法》(5,920,855)、《消除在线挖掘关联规则产生冗余的方法》(5,943,667)、《基于集成数据库的数据挖掘系统》(6,324,533)和《在大型数据库中计算关联规则的方法》(5,983,222).

3 互联网挖掘

在已授权的数据挖掘专利中,与互联网挖掘有关的专利 16 项.目前互联网是一个巨大的、分布广泛和全球性的信息服务中心,它涉及新闻、广告、消费信息、金融管理、教育、科技、政府、电子商务和许多其他信息服务.互联网还包含了丰富的动态超链接信息和互联网页面信息,这为数据挖掘提供了丰富的资源.互联网挖掘可分为三类:互联网内容挖掘(Web content mining),互联网结构挖掘(Web structure mining),和互联网使用记录的挖掘(Web usage mining).

与互联网内容挖掘有关的专利共有 9 项,其中有 4 项是 Appleman 等人申请的.1999 年和 2000 年,他们就同一题目获得 2 项专利,是《协作式互联网数据挖掘系统》(5,918,010)和《协作式互联网数据挖掘系统》(6,081,788).协作式互联网数据挖掘系统将自动处理每一个网页,并将特征代码插入到相应的网页中,在网页搜索服务和广告代理搜索中都有广泛应用.其余 7 项专利是《电子商务应用中挖掘因果关系规则的方法》(5,832,482)、《训练和管理网络向导的系统和方法》(6,157,926)、《基于向导的互联网目录系统和方法》(6,195,681)、《提供与特定主题相关的互联网资源定位系统》(6,336,132)、《网络中目标标记方法》(6,377,936)、《网络中传递和重建个性化数据的方法》(6,477,565)和《自动无线网络趋势检测方法》(6,487,404).

与互联网结构挖掘和互联网使用记录挖掘有关的专利共有 7 项. 2001 年 7 月,Dryken Technologies 公司 Vanderveldt 等人获得《动态数据挖掘和在线定制信息传递系统和方法》专利,专利号是 6,266,668. 该专利主要完成互联网动态查询功能. 该方法包括如下步骤:首先针对用户的查询命令产生精确查询简要表(Search-specific profile),然后将精确查询简要表输入到数据挖掘搜索引擎(Data-mining search engine)中. 数据挖掘搜索引擎挖掘精确查询简要表,产生兴趣主题(Topic of interests). 寻找与兴趣主题相关的目标数据站点(Destination data sites),最后至少用一种查询工具和一个目标数据站点与兴趣主题相匹配,将相关信息提供给用户. 其余 6 项专利是《分布式网络环境基于支持向量机的知识发现》(6,157,921)、《从互联网和内部网大型数据库挖掘共同模式和推理规则》(6,094,645)、《自动网页生成系统中更改所包含文件的系统和方法》(6,226,648)、《从互联网挖掘共同模式和推理规则》(6,263,327)、《访问间断连接的移动客户端页面的方法》(6,505,242)和《建造、下载和访问间断连接的移动客户端页面的方法》(6,507,867).

互联网挖掘研究越来越热. 现在互联网挖掘在互联网向导、自动网页生成和在线搜索等课题取得了一定的进展,在相当长的一段时间里它们仍是互联网挖掘的研究难点和热点.

4　聚类算法

在已授权数据挖掘专利中,与聚类算法(Clustering algorithm)有关的专利 9 项. 聚类(clustering)就是将数据对象分组成为多个类或簇(cluster),同一个簇中的对象具有较高的相似度,而不同簇中的对象差别较大. 聚类分析有很广泛的应用,包括市场或客户分割、模式识别、生物学研究、空间数据分析、互联网文档分类及许多其它方面.

2000 年 1 月,微软公司 Fayyad 等人获得《大型数据库 K-means 增量式聚类系统》专利,专利号是 6,012,058. 该专利实现了一个在数据库中分析数据的数据挖掘系统,在分析开始之前,给 K 一个初值,就是说,想要将数据分成 K 个类,然后再给各个类一个初始平均值或中心点. 从数据库中读取一部分数据,更新平均值或中心点,如果满足设定的聚类条件,则停止,以当前的平均值或中心点即可生成 K 个类;如果不满足设定的聚类条件,则继续读取数据. 2002 年 4 月,Fayyad 等人获得《大型数据库增量式聚类系统》专利,专利号是 6,374,251. 与上面的专利相比,该专利不只局限于 K-means 聚类方法,有更广泛的应用范围. 另外与聚类算法有关的 7 项专利是《减少 K-means 数据聚类计算要求的方法和仪器》(5,983,224)、《数据挖掘中高维数据下层空间自动聚类方法》(6,003,029)、《针对小型和大型数据库聚类的初始条件精炼方法》(6,115,708)、《大型数据库期望值最大化增量式聚类系统》(6,263,337)、《在多维数据库结构中问题自动隔离系统和方法》(6,330,564)、《在大型数据库增量式聚类系统中变化聚类数目》(6,449,612)和《电子税收信息的聚类和转换方法》(6,473,741).

近年来,各种文献提到数十种聚类算法. 例如,层次聚类(hierarchical)[9,10]、K-means 聚类[11]、自组织映射聚类、K-中心点、使用样本的聚类(Clustering using representatives)和使用层次的补偿叠代缩减聚类(Balanced iterative reducing clustering using hierarchies)等等. 但从 1999 年到 2002 年,每年只通过与聚类算法有关的专利 2 项. 这反映出聚类算法申请专利具有较高的难度,同时也反映出专利要有原创性的特点.

5　并行数据挖掘

在已授权的数据挖掘专利 101 项中,与并行数据挖掘有关的专利 8 项.

2001 年 7 月,IBM 公司 Dhillon 等人获得《分布式多处理器系统并行数据聚类方法和系统》专利,专利号是 6,269,376. 该专利将数据集划分成很多数据块(Data blocks),对每一数据块都初始化 K 个中心点. 同时对数据块进行异步操作,每一个异步操作都将数据块中的点划分到与之距离最近的中心点中. 计算所有数据块中数据点到各自中心点的距离累加值. 再根据距离累加值重新计算 K 个中心点. 当数据块的距离累加值在有限的次数内不再减少时,该数据块计算过程结束. 当所有数据块计算结束后,聚类操作完成. 1998 年 5 月,IBM 公司 Chen 等人获得《并行数据挖掘高效信息搜集方法》专利,专利号是 5,758,147. 该专利提出多节点数据的搜集方法,解决了分布式数据库环境的数据挖掘问题.

与并行数据挖掘有关的其它 6 项专利是《在并行多处理系统中定位和抽样数据》(6,049,861)、《动态计算大型项目集的方法和仪器》(6,185,559)、《内存共享多处理器系统并行分类方法》(6,230,151)、《稀疏数据挖掘自组织映射增量式并行算法》(6,260,03)、《数据分析方法》(6,510,457)和《用户反馈和处理系统》(6,510,427).

1998 年到 2000 年与并行数据挖掘有关的专利只有 2 项,而从 2001 年到 2002 年通过 6 项. 这与数据量的迅速增加有着密切的关系,并行数据挖掘研究在处理海量数据和提高挖掘效率等方面越来越引起有关研究人员的关注.

6　其他典型数据挖掘专利

6.1　与面向对象数据挖掘有关的专利

1998 年 7 月,IBM 公司 Bigus 等人获得《面向对象数据挖掘》专利,专利号是 5,787,425. 该面向对象的数据挖掘框架包含了核心功能和可扩展功能,能够灵活地完成多种数据挖掘任务. 1999 年 2 月,Chang 获得《面向对象数据挖掘与决策系统》专利,专利号是 5,875,285. 该专利采用面向对象技术,实现了面向对象的数据挖掘子系统和面向对象的决策制定子系统.

6.2　与决策树生成有关的专利

1998 年 7 月,IBM 公司 Agrawal 等人获得《基于 MDL(最小描述长度)和预先分类决策树的数据挖掘方法和系统》专利,专利号是 5,787,274. 该专利首先对数值属性进行排序,然后基于宽度优先生成决策树,再基于 MDL 进行剪枝. 2000 年 8 月,IBM 公司 Coppersmith 等人获得《寻找离散属性二元决策树最佳阈值方法》专利,专利号是 6,101,275. 2001 年 9 月,纽约

大学 Tuzhilin 等人获得《从数据库决定行为模式方法》专利,该专利主要给出了生成层次行为树的方法,专利号是 6,292,797.

6.3 与数据挖掘界面和结果输出方法有关的专利

2000 年 8 月, IBM 公司 Medl 等人获得《数据挖掘的 GUI 向导》(6,108,004),该专利提供了图形化用户界面,为数据挖掘结果的结构化表示提供了方便. 2001 年 7 月,Unica 公司的 Lee 等人获得《数据挖掘软件的可视化显示技术》专利,专利号是 6,269,325. 2002 年 11 月,日本富士公司 Yaginuma 等人获得《数据挖掘结果中的多维数据显示方法》专利,专利号是 6,477,538.

6.4 与数据挖掘系统建造有关的专利

1997 年 11 月 25 日,Lockheed 导弹和空间公司的 Simoudis 等人获得《在计算机系统中生成预测模型的方法》专利,专利号是 5,692,107. 该专利实现了具有自顶向下和自底向上两种数据分析方法的数据挖掘系统. 1998 年 6 月,日本日立公司 Taniguchi 等人获得《以公共记录比率为相似标准的数据挖掘方法和仪器》专利,专利号是 5,764,975. 该专利提供了发现属性关联的一种数据分析方法和系统. 1998 年 8 月,NYNEX 科学与技术公司的 Fawcett 等人获得《自动设计欺骗发现系统》专利,专利号是 5,790,645. 该专利使用机器学习和数据挖掘技术,生成客户的规则描述. 用户提供数据后,就能自动生成特定领域的欺骗发现系统. 其余 11 项专利是《关系数据多次关系增强过滤数据挖掘系统和方法》(5,884,305)、《用于保险收益分析的数据挖掘系统》(5,970,464)、《神经元代理数据挖掘系统》(5,970,482)、《内存驻有数据挖掘发动机的计算机系统》(6,029,176)、《用消费者和市场数据选择潜在消费者系统》(6,061,658)、《提供实体间关系模式的系统、方法和计算机程序产品》(6,073,138)、《交互式监视挖掘质量的数据挖掘方法、设备和计算机程序产品》(6,112,194)、《为数据挖掘提供数据库的方法和仪器》(6,185,561)、《临时模式数据挖掘系统和方法》(6,189,005)、《数据挖掘与商业活动管理集成系统》(6,240,411)和《实体间关系模式发现系统、方法和计算机程序产品》(6,324,541).

7 热点问题

7.1 主动挖掘(Active mining)

主动挖掘主要指如何自动选取挖掘对象[13]. 在数据挖掘中常有"垃圾进,垃圾出(Garbage in,garbage out)"的情况. 数据挖掘对象的选择直接影响着挖掘结果. 例如,一个金融机构可能对"客户为什么中断与我们的业务关系?"这样的问题感兴趣. 我们期望在数据挖掘过程中,程序会主动的要求增加某些数据或要求补充某些数据的细节. 在上例中,我们期望程序会主动要求输入某些已经中断业务的客户的补充资料,这样我们就不用盲目的收集所有客户的数据. 甚至因为成本的关系我们根本就不可能收集到所有客户的数据.

7.2 累积挖掘(Cumulative mining)

一些实际问题与数据的持续增加有关[14]. 例如,客户的交易数据与医院的数据每天都在增加. 因为数据的清理和算法复杂度的关系,我们不能每天将所有的数据重新挖掘. 这样我们必须进行累积挖掘分析. 累积挖掘算法的成功将会大大推广数据挖掘应用的领域.

7.3 超大数据集挖掘(Extremely large datasets mining)

有些数据集太大不能被计算机读取几遍甚至一遍也不可能. 例如百货商场的交易数据每天都以 Gigabyte 数量级增长. 互联网上的数据更是增长迅速,很难实现全部数据的访问. 现有的算法处理大型数据集的能力较差. 设计超大数据集挖掘算法成为一个挑战.

7.4 超小数据集挖掘(Extremely small datasets mining)

另一个极端,有些数据集太小,现有算法不能有效处理. 例如在脸部识别问题中,一个人通常只有一张照片,在其他情况下识别这个人就比较困难. 一些科学领域,如无机化学沸石分子筛合成反应数据库的合成数据只有 400～500 条,而作为输出端的结构类型却至少有 4 种,分析起来比较困难. 如何分析稀少数据?是否一定要用先验知识?

7.5 使用先验知识的挖掘(Mining with prior knowledge)

有些时候,某些应用领域有充足的先验知识[15]. 例如,一个人可能拥有有关政治事件、法规、个人偏好或经济上汇率的知识. 一些科学领域可能有公式或规则等比较成熟的知识,我们能否将这些多变的、抽象的或者不确定的知识加入到我们的算法中?

7.6 混合数据类型挖掘(Mixed media data mining)

许多数据集包含的数据类型多于一种类型[16]. 例如,医院的数据库通常包含数值类型(例如测试结果)、图像类型(例如 *X*-rays)、二值属性数据类型(例如抽烟/不抽烟)和声音数据类型(例如医生的声音)等等. 现有的算法通常只能处理一种类型数据. 如何进行多数据类型的数据挖掘?是分别处理各数据类型再将结果集成到一起,还是直接处理多类型数据?

7.7 可视化与交互式(Visualization and interactive)

在许多应用中,数据挖掘是一个包括自动数据分析和专家咨询的交互过程. 例如,消费数据库和医院数据库挖掘经常是一个交互的过程,包括专家审查数据、重新安排数据和发现特殊模式等等. 当数据是高维的时候,数据可视化比较困难,尤其当数据属性里面包含非数值属性时,例如文本类型,数据可视化更加困难. 有没有好的方法解决高维属性和非数值属性的可视化问题?如何解决交互式数据挖掘问题?

8 结论

本文分析了美国专利和商标委员会数据库数据挖掘专利的授权情况,重点对关联规则、互联网挖掘、聚类算法和并行数据挖掘等有关专利进行了介绍和讨论. 最后本文阐明了数据挖掘研究所面临的挑战、任务和机遇. 一方面,本文着重从专利方面反映了数据挖掘的研究现状,以期弥补专业文献对数据挖掘研究现状介绍的不足;另一方面,本文也试图对申请数据挖掘专利的研究人员在专利申请方面有一定的帮助.

参考文献:

[1] U M Fayyad, G Piatetsky-Shapiro, P Smyth. From Data Mining to

Knowledge Discovery:an overview,Advances in Knowledge Discovery and Data Mining[M]. AAAI/MIT Press,1996.

[2] The internet address of INLEN syetem[DB/OL]. http://www.mli.gmu.edu/projects/inlen.html.

[3] J W Grzymala-Busse. Rough sets in knowledge discovery[J]. Physica-Verlag,1998:562 - 565.

[4] The internet address of Clementine[DB/OL]. http://www.spss.com/clementine/.

[5] Otis Port. Virtual prospecting[N]. Business Week,New York,MARCH 23. 2001.

[6] SHI Zhong-zhi. AI prospecting[C]. AI Prospecting in China,Beijin: Beijing University of Posts and telecommunications Press,2001.

[7] The internet address of US PATENT&TRADEMARK OFFICE DATABASE[DB/OL]. http://www.uspto.gov/patft/index.html.

[8] HAN Jia-wei,Micheline Kamber. Data Mining:Concepts and Techniques[M]. Higher Education Press,2001.

[9] M B Eisen,P T Spellman,P O Brown,D Botstein. Cluster analysis and display of genome-wide expression patterns[J]. Proc. Natl. Acad. Sci. USA,1998 - 95:14863 - 14868.

[10] X Wen,S Fuhrman,G S Michaels,D B Carr,S Smith,J L Barker,R Somogyi. Large-scale temporal gene expression mapping of central nervous system development[J]. Proc. Natl. Acad. Sci. USA,1998 - 95:334 - 339.

[11] R Herwig,A J Poustka,C Mller,H Lehrach,J O Brien. Large-scale clustering of cDNA-fingerprinting data[J]. Genome Res,1999 - 9:1093 - 1105.

[12] P Tamayo,D Slonim,J Mesirov,Q Zhu,S Kitareewan,E Dmitrovsky,E S Lander,T R Golub. Interpreting patterns of gene expression with self-organizing maps:methods and application to hematopoietic differentiation[J]. Proc. Natl. Acad. Sci. USA,1999 - 96:2907 - 2912.

[13] M Klemettinen,H Mannila,P Ronkainen,H Toivonen,A I Verkamo. Finding interesting rules from large sets of discovered association rules[A]. Proceedings of the Third International Conference on Information and Knowledge Management (CIKM'94)[C]. Gaithersburg,MD,ACM Nov,1994. 401 - 407.

[14] W Kloesgen. Efficient discovery of interesting statements in databases[J]. Journal of Intelligent Information Systems,1995,4(1):53 - 69.

[15] H Mannila,H Toivonen. Multiple uses of frequent sets and condensed representatioms[A]. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)[C]. Portland,Oregon,AAAI Press,1996. 189 - 194.

[16] U M Fayyad,G Piatetsky-shapiro,P Smyth,R Uthurusamy. Advances in Knowledge Discovery and Data Mining[M]. AAAI Press,Menlo Park,CA,1996.

作者简介:

刘晓东 男,1973 年生于辽宁省宽甸县,博士研究生,主要研究方向包括知识发现,机器学习和数据挖掘等. Email:xiaodong@jlu.edu.cn.

刘大有 男,1942 年生于吉林省长春市,教授,博士生导师,主要研究方向包括人工智能,空间推理,智能软件,数据结构,计算机算法,粗集和数据挖掘. Email:liudy@jlu.edu.cn.