基于 OLAP 和聚类分析结合的美国专利挖掘系统

吕 琳,朱东华,刘玉琴
LV lin,ZHU Dong-hua,LIU Yu-qin

北京理工大学 管理与经济学院,北京 100081
School of Management and Economics,Beijing Institute of Technology,Beijing 100081,China
E-mail:llxpx@bit.edu.cn

LV lin,ZHU Dong-hua,LIU Yu-qin.Design and realization of US patent mining system based on combination of OLAP and clustering analysis.Computer Engineering and Applications,2007,43(25):186-187.

Abstract: For US issued patent databases as example,OLAP and clustering technologies are in-depth discussed.In view of the similarity and difference of OLAP and clustering,the design and realization scheme of US patent mining system based on the combination of them is presented and their visualization outcomes are displayed.Based on that,a general frame of data mining system is constructed.The results show that the combinative method of OLAP and deep-seated mining technologies will be a trend of data mining development.

Key words: OLAP;clustering;data mining

摘 要:以美国授权专利数据库为实例,对 OLAP 及聚类分析技术进行了深入而细致的探讨。针对它们的共通性和差异性,提出了两者结合的美国专利挖掘系统的设计与实现方案,并给出了可视化结果。在此基础上,构建了数据挖掘系统的通用框架。结果表明,将 OLAP 和数据深层挖掘技术紧密配合、协调使用将是数据挖掘发展的一个方向和趋势。

关键词:OLAP;聚类;数据挖掘
文章编号:1002-8331(2007)25-0186-02 **文献标识码**:A **中图分类号**:TP311

1 OLAP 与数据挖掘技术概述

OLAP(Online Analytical Processing)与数据挖掘技术是近年来数据库领域的研究重点和热点。建立在数据仓库基础上的 OLAP 以多维分析为基础,提供数据多层面、多角度的逻辑视图[1];数据挖掘则通过对大量的历史存储数据的分析和分类,从中得到有意义的模式和关系[2]。OLAP 与数据挖掘都是决策支持工具,但两者却具有明显的区别。OLAP 主要允许客户端设计汇总表来存储数据,便于数据的修复和导航。它可以用来尝试发现新的数据,但因为数据发现工作更多地依靠用户输入的问题和假设,所以在 OLAP 协助下所做的数据发现由于用户先入为主的局限性限制了问题和假设的范围,从而使最终的结论变得局限、偶然和不完全。而数据挖掘则是自动地发现可以应用到预测未来结果的新的模式和规则,用户不必提出确切的问题,只需用工具去挖掘隐藏的模式并预测未来的趋势,这样有利于发现未知的事实。从数据分析的深度看,OLAP 位于较浅的层次,数据挖掘则是更深层的分析,它可以发现 OLAP 所不能发现的更为复杂而细致的信息。

2 美国专利数据挖掘系统框架

综上所述,OLAP 与数据挖掘无论在所用技术、适用范围还是在用途上都存在着较大的差异。但数据挖掘不仅可以由 OLAP 来分析或展现,且其结果又可以指导 OLAP 多维模型,所以两者密切相关。因此在决策分析中必须协调使用它们才能发挥出最佳的作用。针对这一特点,以美国授权专利数据库为实例,本文设计与实现了基于 OLAP 和数据挖掘的重要方法之一——聚类分析结合的美国专利挖掘系统(US Patent Mining System),其总体框架如图 1 所示。该框架也同样适用于其它的挖掘系统,是一个可扩展的通用框架。

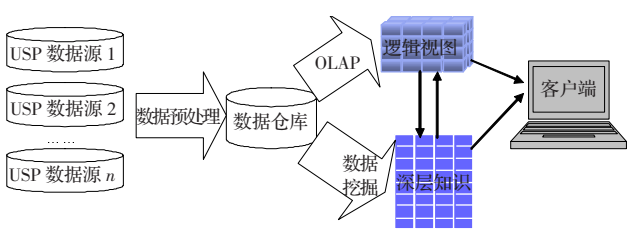

图 1 美国专利挖掘系统整体框架

3 美国专利数据的 OLAP 分析及其可视化结果

3.1 数据预处理

数据预处理是成功进行 OLAP 和数据挖掘的前提和基础。

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.70031010);985 哲学社会科学创新基地建设研究论文之一。

作者简介:吕琳(1973-),女,博士后,主要研究方向:数据挖掘,自然语言理解,人工智能;朱东华(1963-),男,博士生导师,教授,主要研究方向:数据挖掘,技术监测,知识发现。

据统计，它约占整个数据挖掘过程的 60%左右，所以数据预处理的好坏将直接影响到数据挖掘的质量和效果，它是数据挖掘过程中不可或缺的关键环节。我们的 US Patent Data Mining System 通过对美国专利数据网页的所有 30 个域的抽取以及不同地域数据的集成、转换（包括字段名转换、日期格式转换等等）、清洗、规约、离散、概念层次化和装载等预处理，最终构建成以美国专利基本信息表 “basicinfo”为核心的由一个事实表和五个维表组成的美国专利数据仓库。

3.2　构建 OLAP 立方体

OLAP 分析以数据仓库为数据源构建立方体。它提供了三种不同的存储选项以最优化数据检索：(1)MOLAP（基于多维数据库的 OLAP），它用于在数据仓库上建立多维立方体；(2)ROLAP(基于关系型数据的 OLAP)，它使用原来的 RDBMS 作为存储机制；(3)HOLAP（混合 OLAP），它整合了 MOLAP 和 ROLAP 的功能，在 MOLAP 立方体中存储高级别的聚集，并在关系型数据库中保持低级别的聚集和单行明细项[3]。

上述三种 OLAP 技术各有所长。由于本系统处理的是数据量巨大的美国专利数据(从 1985 年到 2004 年共 20 年约 300 G 的美国专利数据)，而 ROLAP 在处理大规模数据方面具有独特的优势，所以 ROLAP 成为首选。ROLAP 采用星型模式、雪花模式或父子模式组织多维数据[1]。此处采用星型模式对数据仓库层中的美国专利信息进行组织，由于篇幅有限，以 2003 年全年的数据为例，它由一个事实数据表和五个维度表构成一个星型构架，如图 2 所示。构架的核心是事实数据表，即美国专利基本信息表，它是数据立方体中度量值的源，其度量值为授权专利的数量；维度表由发明人维，分类维，专利主要信息维，范畴分类维和时间维组成，它是数据立方体中维度的源。

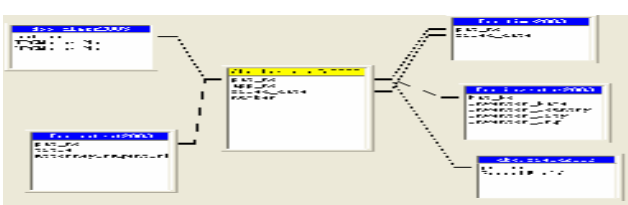

图 2　美国专利 2003 年的数据仓库星型结构

3.3　OLAP 可视化结果

根据构建的美国专利 2003 年全年的数据仓库和 OLAP 立方体，已成功实现了任意行维、列维及条件维的 OLAP 分析及图表显示。以 MS Analysis Services 2000 为开发环境，图 3~5 分别展示了由一个事实表，两个行维（时间维、IPC 维）和两个列维（发明人维、代理机构维）组成的 OLAP 原始数据表及其不同类型的可视化结果。

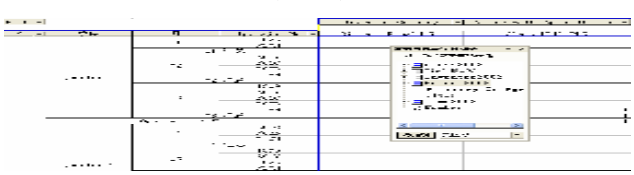

图 3　OLAP 原始数据表

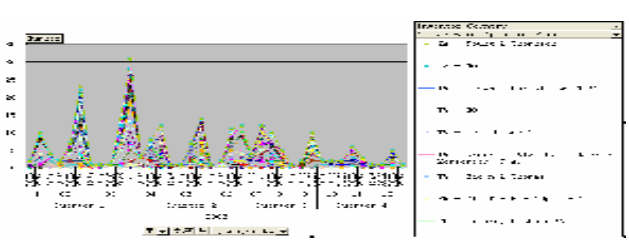

图 4　OLAP 图展示 1——彩色折线图

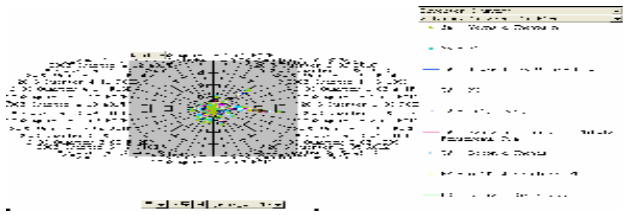

图 5　OLAP 图展示 2——雷达图

4　美国专利数据的聚类分析及其可视化结果

深层的数据挖掘技术大体可以分为六类：关联分析、分类、聚类、有序模式判别、时序相似性分析及偏差分析。此处重点讨论聚类分析。

4.1　聚类算法概述

聚类分析是一种间接的数据挖掘方法，它用来寻找数据库内固有的基本分类，是对群体及其成员进行分类的递归过程[4]。聚类算法将数据划分为一系列有意义的子集，使得子集间的差别尽可能大，而子集内的差别尽可能小。简言之，聚类的所有算法就是找到记录并将这些记录指定给它定义的群体，它通过不断迭代和调整将记录分组到距离其最近的子集中，从而获得记录组群间存在的固有的相似性。优序排列、间隔值、度量值和分类值是影响聚类分析的四大要素。

4.2　K-Means 算法

K-Means 算法是生成一组聚类的常用方法之一，也是本挖掘系统采用的聚类算法。其主要特征是可以根据需要预先确定 K 个聚类，即 K 个变量。对于 K 值的选取没有一个固定的快速规则可循，通常一个比较理想的方法是对各种值进行试验。该算法主要包括查找聚类、寻找聚类中心和调整边界三个步骤。

4.3　聚类挖掘模型

聚类模型采用基于概率的聚类方法，算法假设属性之间的概率分布是相互独立的，根据概率密度对数据进行分组。在创建聚类挖掘模型时，需要选取源类型，选择挖掘技术，设置事例、培训数据和聚类个数等。其中，选取源类型是基础和关键，传统的算法一般选取一个关系型数据库，而我们的系统则以 OLAP 立方体作为数据源，这是本系统的最大特色之一。只有这样才能充分结合 OLAP 和挖掘算法的优势，从而使聚类效果达到最佳。挖掘技术则选择了上述的 K-Means 算法。

4.4　可视化结果

以美国专利 OLAP 立方体为数据源，在 MS Analysis Services 2000 的开发环境下，分别用 K-means 算法建立了聚类，成功实现了 UPC，IPC，Inventor，Patent，FiledTime，Field 等字段的聚类分析及其可视化。图 6 和图 7 分别是基于数据仓库中的两个表 basicinfo 和 inventor，分析专利发明人国家的聚类原始表和效果图(K—means 算法：K=20)。

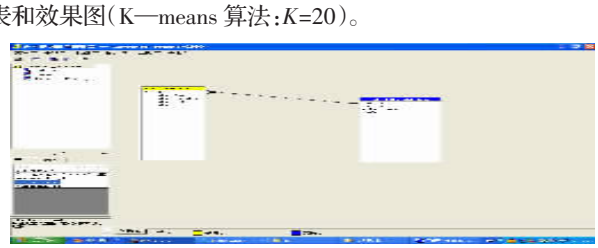

图 6　聚类分析原始数据表

（下转 217 页）