基于大数据挖掘的智能交通决策分析系统[1]

张慧哲 保丽霞

【摘要】数据仓库与数据挖掘技术的应用是目前实现智能交通系统信息资源共享与综合利用开发的重要途径。本文以开展数据仓库与数据挖掘技术研究和推动技术在交通系统中的应用为主线，提出了 ITS 数据挖掘平台的框架，重点对智能交通挖掘系统构建的关键技术、理论以及实现方法进行了较为深入的研究。通过仿真对所设计的数据挖掘系统进行了验证，试验结果说明了其模型的有效性，充分显示了此系统解决交通问题的优越性和发展的潜力，同时也为智能交通系统的建设提供了新的思路。

【关键词】数据仓库；数据挖掘；智能交通系统

1 引言

目前数据挖掘技术在交通方面的应用已经成为学术界和企业界研究与应用的热点领域，是提高交通管理的效率、改善交通服务水平的重要手段[1]。以下对国内外研究人员从事的数据挖掘技术在 ITS 系统中的应用研究做以总结。Baldi 和 Der-Horng Lee 等[2,3]全面论述了国内外数据挖掘技术在交通系统中的应用动态，指出数据挖掘技术可以作为改善交通状况的有效手段，并展望了其今后的发展趋势。Bing Wu 和 Yun S.Y 等[4,5]提出基于神经网络的数据挖掘模型来处理实时输入信息，从而对短时交通流数据进行预测，得到了较好的预测结果。李相勇和蒋葛夫等[6]将模糊理论引入了道路交通服务水平的评价中，并建立了相应的评价模型，为交通系统的评价提供了一种有效的方法。Wei-Hsun Lee 等人[7]从采集点线圈所采集的交通流信息中挖掘出交通模式信息，并将其转换为用于行程时间预测的规则，从而提出了一种基于知识的实时行程时间预测模型。Y. Chong, C. Quek[8]利用模糊神经网络可以自学习的特点来实现交通灯的智能自适应控制，取代了在不同的交通状况下需要人工对交通灯周期进行相应的调整及设定。Nale Zhao 等[9]提出了基于支持向量积(SVM)的多数据源的 ITS 融合技术，其中分为 SVM 训练、训练结果评价以及 SVM 的测试，文中还对 SVM 数据融合技术与传统方法进行试验对比，表明 SVM 方法可以实现数据质量的控制从而保证了数据的准确性。

由上可以看出，数据挖掘技术被研究人员用于解决交通中存在的各类问题，并取得了一定的成果，然而数据挖掘理论在应用中还存在一定的不足之处。由于智能交通系统本身的复杂性，各子系统的相互依赖性不断增强，孤立的子系统将很难有所作为，将智能交通信息进

基金项目：基于大数据智能分析的交通状态感知与交互发布技术研究与示范（13QB1402700）；

行整合与集成的各种共享平台也相继推出[10-13]，但平台目前还缺少对数据有效组织并进行知识规律等分析、挖掘及发现的过程，这也使得各种数据挖掘方法不能得到很好的研究和应用。本文将探索从数据仓库和数据挖掘理论的指导作用和其 ITS 系统应用两个角度，解决 ITS 中存在的实际问题，并进行了系统的设计与实现。

2 ITS 数据挖掘平台框架

根据信息处理的流程和信息平台的功能要求，以及数据和信息提取、处理、分析和服务为主体的信息平台的一般构架形式，结合考虑 ITS 共用信息平台的功能要求及其所面对的处理对象，图 1 给出了基于数据仓库和数据挖掘理论的 ITS 数据挖掘平台框架图。

数据组织策略是将来自各子系统的数据进行信息交换、信息融合和系统集成、ETL 处理后组成交通数据仓库，利用数据挖掘方法对此数据仓库按照需求进行模式、规律等的挖掘，最终将得到的结果由智能交通应用接口进行转换提供给其他管理部门的子系统共享，实现各 ITS 应用子系统功能的集成。

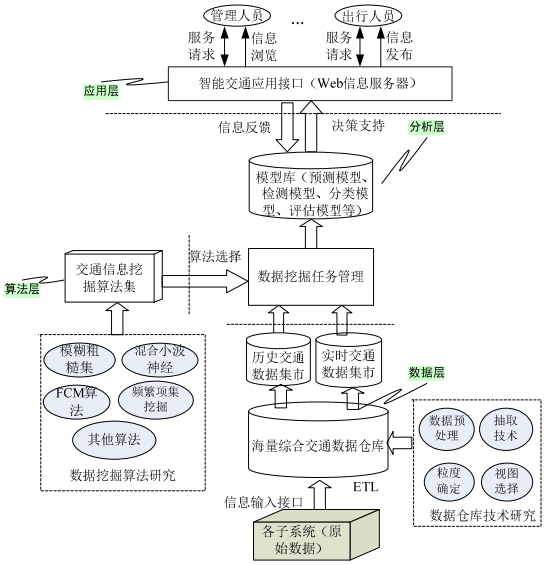

图 1 ITS 数据挖掘平台框架图

所提出的系统框架将数据仓库和数据挖掘技术融入系统模型中，提供了一个多层的应用体系结构，包括数据层、算法层、分析层和应用层。多层体系结构能够在跨平台、网络环境下应用，系统可以根据需要采用灵活的方式，如 B/S、C/S 等。

本文将数据挖掘和数据仓库结合起来，统一设计、处理，一方面可以使交通信息平台数据得到更好的组织，数据仓库系统得到更大的发展，同时也使数据挖掘真正不脱离实际而得到广泛的应用。

3 交通挖掘系统中数据仓库的设计

3.1 海量交通数据仓库的概念模型

数据仓库是面向主题的，其目的是针对主题进行数据组织，同一主题的数据往往在一个事实表中，并且只有符合主题的数据才可进入数据仓库。按照分析需求交通数据仓库系统的各个主题模型的事实属性和维度属性如表 1 所示。

表 1 概念模型中主题的描述

决策分析主题	事实表属性字段	维度属性字段	公共键码
短时交通流预测	流量，速度，占有率	时间，路段，天气情况	路段 ID
交通状态辨识	流量，速度，占有率，状态等级	时间，路段，流量等级，速度等级，占有率等级，天气情况，状态类型	路段 ID
交通事件检测	流量，速度，占有率，有无交通事件	时间，路段，流量等级，速度等级，占有率等级，天气情况	事件 ID
其他主题	…	…	…

采用 E-R 图作为描述主题及主题之间关系的工具。在 E-R 图中，长方形表示实体，在数据仓库中就表示主题，椭圆形表示主题的属性，用无向边把主题与其属性连接起来，用有向边表示主题之间的联系(单向边表示一对多关系，双向边表示多对多关系)，主题之间的无向边表示一对一的关系。以交通状态辨识为主题的概念模型如图 2 所示。

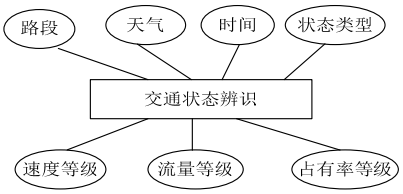

图 2 交通状态辨识概念模型图

3.2 海量交通数据仓库的逻辑模型

交通系统数据仓库的数据模型设计采用基于关系数据库的星形模型。这里以交通状态辨识为例，其星型模型实例如图 3 所示。

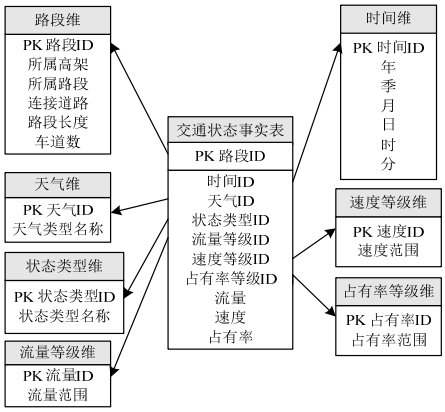

图 3 交通状态事实表与维表的星型模型

事实表包含了适当的数字型值内容，如流量、速度、占有率等。上图 3 中间的事实表表示与交通流情况相关的字段，四周表示的是与之相关的信息属性包含了主要的数据(日期，时间，地点，天气，交通流等级等)即维属性，所以只要扫描事实表就可以查询，而无须把多个庞大的表连接起来，同时维表一般比较小，与事实表连接时其速度较快，大大加快了查询速度。

3.3 海量交通数据仓库的物理模型

数据仓库的物理模型主要解决数据的索引策略、数据的存储策略以及存储分配优化等问题。物理模型建模的主要目的有两个，一是提高系统性能，二是更有效地管理存储的数据。访问的频率、数据容量、存储介质的配置都会影响物理设计的最终结果。

数据仓库的数据量很大，在多维查询的 OLAP 中，索引扮演了重要的角色，因此需要对数据的存取路径进行仔细的设计和选择，而有效的索引能够缩短读取时间并提高数据检索效率。在数据仓库中常用的索引策略有四种：B-Tree 索引、位图索引、广义索引以及连接索引[14,15]。当今主流商业数据中已普遍实现了位图索引，因此本系统采用位图索引是合适的。

4 智能交通数据挖掘系统的实现

4.1 系统的软硬件环境

系统运行环境中包括数据区、交通应用区、GIS 区、通信区、和终端区几个子系统，分别布设数据库服务器、交通应用服务器、GIS 服务器、通信服务器和一系列操作终端，各子系统实现不同的功能。

数据仓库和数据挖掘的应用系统部署在综合交通应用服务器系统中。其中综合交通应用服务器操作系统为 Red Hat Enterprise Linux，数据库服务器采用 Oracle 10g。

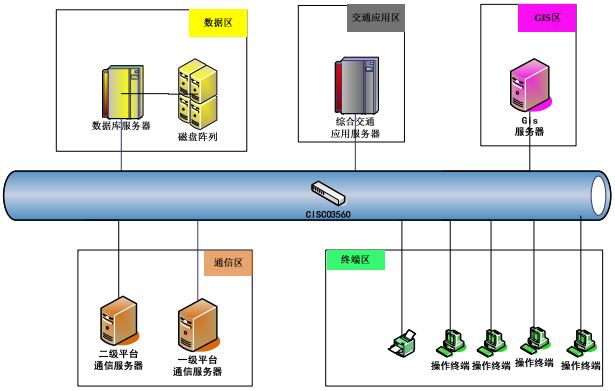

图 4 系统网络拓扑结构

4.2 系统的设计

以数据仓库的数据组织与显示方式为基础，采用 B/S 的体系结构快速开发出交通数据挖

掘原型系统，系统主要包括多维数据集定制模块、OLAP 分析方式选择模块、图表显示模块、数据挖掘模型选择模块等。各模块功能设计如下：

多维数据集定制模块：可以方便地对所要分析的多维数据集的维度进行设置，导入所用分析的维度，或过滤掉此维度。

OLAP 分析方式选择模块：OLAP 也可以简单定义成使用户能够以多维视图分析数据的工具。本系统 OLAP 分析基于 BI Beans 技术进行设计，方式选择模块分为向上钻取、向下钻取和保持升降序子模块。

图表显示模块：可以灵活配置图表类型、横轴、纵轴的名称，显示的字体，图表的大小，图例显示等并实现结果显示功能。

数据挖掘模型选择模块：可以根据所要分析数据的特点以及分析的目标，合理的确定所要应用的挖掘模型。

4.3 系统的实现

下图是数据仓库中所管理的上海市 SCATS 所采集的实时线圈数据，由于对交通流分析的一般都是将路网划分为路段，路段中又以断面为单位，数据仓库系统可按照分析的需要通过上钻操作汇总线圈数据，形成相应的断面的实时数据，如图 5。可见数据仓库系统为数据挖掘提供了良好的数据环境。

图 5 数据仓库的后台线圈及断面数据管理

将文献[15]所提出的预测模型应用于南北高架短时交通流预测，并且进行交通流状态的划分。首先以 2007 年 8 月 7 日早上 7 点为预测开始时间，预测 5 分钟以后的南北高架交通流，然后进行交通状况辨识，并以南北高架 5 分钟之后的真实情况作为预测分类结果的对比依据，这样可以对南北高架的短时交通流预测和分类结果进行很好的检验。

如图 6 中的最左边的为实时的交通状态情况，其中黄色表示交通状况拥挤，绿色表示交通状况稳定。中间的图是系统中通过数据挖掘模型预测出的 5 分钟以后的交通状况。最右边的图表示 5 分钟之后的真实交通状况。这里将南北高架分成 29 个路段，通过对比我们不难看出预测出交通状态结果和真实情况，只有 2 个路段有偏差，偏差率为 6.8%。

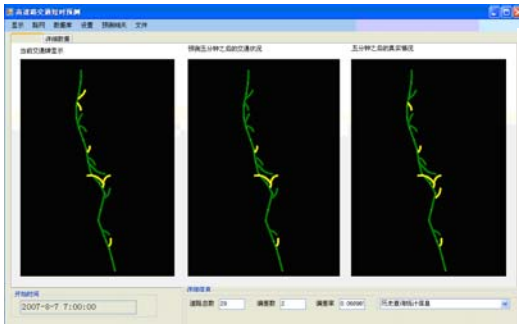

图 6 交通流预测结果

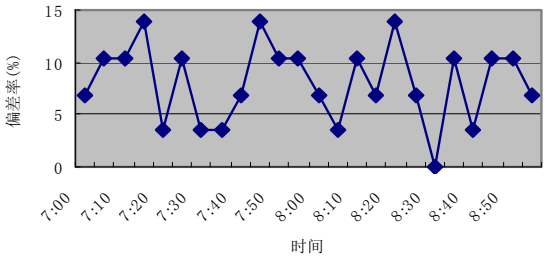

图 7 预测结果偏差统计图

我们将 2007 年 8 月 7 日 7:00~9:00 时间段内预测结果与真实情况进行了比较，得出偏差率，具体的结果如图 7 所示。

实践证明，通过数据挖掘系统能够比较真实的预测短时交通状况，证明了模型的有效性。此系统输出结果可辅助于情报板的信息发布，为交通管理人员提供较准确的路况信息，对出行者进行诱导。

5 结论

数据挖掘在智能交通领域中有广阔的发展空间，可以促进交通系统向更高层的辅助管理、辅助决策系统等信息化方向发展。本文探索从数据挖掘理论的指导作用和其 ITS 系统应用两个角度，解决 ITS 中存在的实际问题。首先对智能交通系统、数据挖掘技术在 ITS 中的应用进行介绍，在此基础上提出了 ITS 数据挖掘平台的框架。并分别对交通数据仓库概念、逻辑、物理模型的构建及相关理论进行了研究。最后对基于 DW 和 DM 的城市交通挖掘系统软硬件环境及设计、实现作以简要介绍，所建立的 ITS 数据挖掘系统已在上海市电器科学研究所进行了性能测试，运行结果良好。

【参考文献】

[1] Daimler Chrysler Corporation. Cross industry standard process for data mining [EB/OL]. http://www. Crisp-dm. org，1999

[2] Baldi，M. Baralis，E. Risso，F. Data mining techniques for effective and scalable traffic analysis[C]. Integrated Network Management，9th IFIP/IEEE International Symposium on，2005，105~118

[3] Der-Horng Lee，Shin-Ting Jeng. Applying data mining techniques for traffic incident analysis[J]. Journal of The Institution of Engineers，2004，89~94

[4] Bing Wu，Wen-Jun Zhou，Wei-Dong Zhang. The applications of data mining technologies in dynamic traffic prediction[C]. China Communication Press（2nd Edition），Beijing，2003，219~221

[5] Yun S. Y，Namkoong S. A performance evaluation of neural network models in traffic volume forecasting[J]. Journal of Trans，portation Engineering，1998，27（6）: 293~310

[6] 李相勇，蒋葛夫. `城市道路服务水平的模糊综合评判[J]. 交通运输系统工程与信息，2002，51~55

[7] Wei~Hsun Lee，Shian-Shyong Tseng，Sheng-Han Tsai. A knowledge based real-time travel time prediction system for urban network[J]. Expert Systems with Applications，2008，15(2):79~92

[8] Y. Chong，C. Quek，P. Loh. A novel neuro-cognitive approach to modeling traffic control and flow based on fuzzy neural techniques[J]. Expert Systems with Applications，2008，3(3): 16~32

[9] Nale ZHAO，Lei YU，Yanbin GENG. Support vector machine based approach to data-layer multi-source ITS data fusion[J]. Journal of Transportation Systems Engineering and Information Technology，2007，7（2）: 32~37

[10] 章威. 广州市 ITS 共用信息平台体系结构与关键算法研究[D]: [博士学位论文]. 广州: 华南理工大学，2007

[11] Florida Department of Transportation. District Surface Security and Reliability Information System Model[J]. Deployment Final Deployment Plan，2004

[12] 杨宏旭. 上海市公路网交通信息化与智能化关键技术研究[D]: [博士学位论文]. 上海: 同济大学，2006

[13] 商琳，骆斌. 一种基于数据仓库的数据挖掘系统的结构框架[J]. 计算机应用研究，2000， 17（9）: 63~65

[14] 王珊. 数据仓库技术与联机分析处理[M]. 北京: 科学出版社，1998

[15] 张慧哲，王坚. 数据挖掘在短时交通流预测模型中的应用研究[J].《计算机集成制造系统》. 2008，14（4），690~695

【作者简介】

张慧哲，女，博士，上海市城市建设设计研究院。电子信箱：zhanghuizhe168@163.com