摘 要

随着企业信息化的不断深入，以往数据处理已经不能满足企业信息化发展的需求，企业对数据进行整合与分析的需求更加强烈，如何从这些海量数据信息中提取出对企业有用的信息，构建统一的数据分析平台已经成为了企业信息化进程中亟待解决的问题。企业级的数据分析系统优劣决定着企业信息化的成败。

本文从系统工程的角度全面系统地分析了构建企业智能决策支持系统的理论与技术，针对在烧结生产控制过程中，信息孤岛现象严重，数据分析能力欠缺等问题，采用系统工程的思维方式，结合烧结生产管理的具体需求，设计并实现了一个烧结生产信息数据分析子系统，用以指导企业生产。详细阐述了烧结生产信息数据分析子系统中数据仓库(Data Warehouse)、联机分析处理(OLAP)、数据挖掘(Data Mining)相关技术的理论及具体使用。将智能决策支持的设计思想引入到烧结生产控制过程中，包括数据仓库模型构建，数据预处理的策略设计，以及基于数据仓库的 OLAP 和数据挖掘算法应用研究。

本文从理论、系统模式和实施技术等方面将智能决策支持引入到烧结生产控制过程中。本文设计实现的烧结生产信息数据分析子系统也只是企业数据分析系统过程中的一次尝试，为烧结厂未来的系统开发提供了一个可行的、有实际参考依据的、体现了先进的数据分析和决策支持思想的系统方案。

关键词 智能决策支持系统；数据仓库；数据挖掘；OLAP；数据预处理

Abstract

Along with the developing of informationization of enterprise, traditional data processing technology cannot satisfy the demand of informationization development of modern enterprises. Now enterprises more cry for data's conformity and analysis.How to pick up information available to enterprises from massive data and construct unified data analysis platform has already become a problem which is urgent to be solved during the course of enterprises informationization. The enterprise level data analysis system determines the success or failure of enterprise informationization.

This article has comprehensively and systematically analyzed the construction enterprise intelligence decision support system both in theory and in technology at systems engineering's view, which aiming at the questions that the phenomenon of information isolated-to-island is serious and data analysis ability is not enough during the agglutination production control process, adopting the systems engineering methods, together with specific management needs in agglutination production, designed and realized a subsystem of agglutination production message data analysis to instruct the enterprises' production. Elaborated in detail the theory of correlation technique and the concrete use of the subsystem of agglutination production message data analysis in the data warehouse (Data Warehouse)、on-line analysis processing (OLAP) and the data mining (Data Mining). Introduced the design concept of intelligent decision-making support technology to the agglutination production control process, including the construction of data warehouse model, the policy design of data pretreatment, as well as the applied research of OLAP and data mining algorithm which based on data warehouse.

This article introduces the intelligent decision-making support technology to the production control system of agglutination plant from the theory, system model and implementation technology aspects and so on. The subsystem of agglutination production message data analysis designed and realized in this article also is an attempt in business data analysis system processing, which provides a systematic scheme to agglutination plants in the future which is feasible、has actual reference、also embodies the advanced data analysis and the policy-making support thought.

Key words IDSS; DW; DM; OLAP; Data Preprocessing

河北科技大学学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品或成果。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：[signature]　　指导教师签名：[signature]

2009 年 6 月 8 日　　2009 年 6 月 8 日

河北科技大学学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权河北科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

□保密，在___年解密后适用本授权书。

本学位论文属于

☑不保密。

（请在以上方框内打“√”）

学位论文作者签名：[signature]　　指导教师签名：[signature]

2009 年 6 月 8 日　　2009 年 6 月 8 日

第1章 绪 论

1.1 课题研究的背景

当前，在全国和冶金行业各企业都面临的问题是：如何提高企业的经济效益，更加合理地共享信息资源，科学而有效地进行资源配置，提高企业的管理水平和生产指挥的水平、尤其是针对冶金行业提出的管控网分开等情况，提出用信息技术去改造传统产业，改进生产方式和管理模式，促进冶金行业中企业的发展，实施企业信息化建设[1]。

信息化已经成为一种趋势，是参与国际竞争与合作的重要条件，也是应对加入世界贸易组织挑战的迫切需要。现在国外大的汽车生产商，零部件都实行全球网上采购，我国汽车零部件生产企业如果不能进入全球采购系统，就可能被淘汰。据哈尔滨飞机制造公司介绍，他们与法国合作设计的直升飞机中，对方要求中方要能读懂他们的程序软件，具备远程网上合作设计的能力，否则，就没合作的基础。由此可见，我们的企业要参与国际合作与竞争，即使能够生产出与跨国公司同样质量的产品，在信息化方面的差距也将使我们在竞争中处于被动。

总的来说，企业信息化建设是一场革命，在提高企业管理水平，促进管理现代化，转换经营机制，建立现代企业制度，有效降低成本，加快技术进步，增强市场竞争力，提高经济效益等方面都有着现实和深远的意义，是带动企业各项工作创新和升级的重要突破口。我们广大企业特别是各个行业的骨干企业，应当在企业信息化建设方面先行一步，跟上时代发展的步伐，增强市场竞争的能力。

河北某钢铁集团烧结厂建立了先进的公司企业计算机网络系统，研制开发出一些应用系统，能够提供配料等生产过程监测数据，确实提高了生产过程的自动化程度，解决了一些问题。但是，因为这些系统的相对复杂性和封闭性，使不同系统之间的配合关系，以及与管理信息系统之间的配合关系，并没有达到信息及时、顺畅的同步，导致每个自动化单元之间缺乏必要的整合，对整个生产厂来说它们都是分散的信息孤岛，只能为己所用，上、下游信息沟通闭塞，不能有机的将前后工序衔接在一起。同时，由于没有合适数据分析平台，使信息的反馈滞后，导致信息在纵向上缺乏集成，使很多应用都无从实现，这是影响企业信息化整体水平的“鸿沟”。

为了充分地发挥公司企业网的作用和更好地利用现有的数据及信息资源，还有许多工作要做：

1) 充分地利用烧结厂企业网的设备和网络资源，大力开发信息资源，加大在企业网上传送的信息量。信息资源包括：生产动态信息、企业对外宣传信息和企业员工工作方面的信息，使得信息资源在企业网得到利用与共享。

2)针对生产管理实际工作情况，将检测数据加工、整合为管理方面的信息，提高生产经营的管理水平与现代化水平。

3)将生产方面的数据加工处理成信息，便于领导生产指挥决策。

1.2 本课题的研究对象和目的

本人通过对河北某钢铁集团烧结厂的实地考察，以烧结厂的生产管理为依据，对建立烧结生产信息数据分析子系统进行了研究。针对“信息孤岛”、数据分析能力欠缺等问题，在本系统中引入了智能决策支持系统的概念，以期利用智能决策支持系统技术对生产控制过程中长期积累的大量生产数据进行管理与分析，从海量数据发掘生产规律，为各级调度和厂领导能够实时掌握本厂生产数据、设备状态等生产经营情况提供信息分析平台。使企业的生产效率大大提高，具有重要的理论意义和实用价值。

1.3 智能决策支持系统的研究现状

智能决策支持系统(Intelligent Decision Support Systems，简称 IDSS)起源于八十年代初期。首先，Bonczek 等人提出 DSS 与专家系统(Expert System，简称 ES)相结合，分别发挥 DSS 数值分析与 ES 的符号处理的特点，用于有效地解决定量与定性的问题以及半结构化、非结构化的问题。这种 DSS 与 ES 结合的思想即构成的 IDSS 的初期模型。IDSS 的这种模型扩大了 DSS 处理问题的范围,提高了决策能力因此它具有很强的生命力，并且在应用中发挥了巨大的作用，因而成为目前 DSS 发展中的重要方向[2]。

目前的 IDSS 的构造模型与应用范围在逐步扩大，所开发的系统也日益成熟，如 Inforym 公司推出的 REVEAL 系统与 Caregie-Mellon 大学开发的 JMS 系统均在实际应用中发挥出很大的作用。

我国开发 DSS 与 IDSS 应用系统的工作始于八十年代初，在八十年代中期已研制出 DSS 生成器，并开发出 DSS 应用系统，如煤炭部计算中心与哈尔滨工业大学 1986 年所研制的 DSS 生成器、华中工学院开发的资源分配 IDSS 应用系统等均属此类产品，此后陆续研制开发的 IDSS 的系统很多，如上海科技大学的 IDSS 生成器 IDSSG-KD. 1，南京大学的生成器 NCIDSSG，复旦大学的生成器 DISSET-L。此外，还有一些 IDSS 应用系统如上海科大、复旦大学利用其生成器在上海石洞口电厂开发的“发电成本目标管理系统的知识处理子系统”，南京大学用其生成器开发的“产品质量 IDSS”系统，最近，利用面向对象技术将 IDSS 作为一种综合集成处理也在我国开始研制。

目前，按照智能决策方法，可将智能决策支持系统的研究现状，大致可以把分为 5 类[3]：

(1) 基于ES(Expert System)的IDSS ES是目前AI中应用较成熟的一个领域，一般由知识库、推理机及数据库组成，它使用非数量化的逻辑语句来表达知识，用自动推理的方式进行问题求解，而DSS主要使用数量化方法将问题模型化后，利用对数值模型的计算结果来进行决策支持。

(2) 基于机器学习的IDSS 机器学习是通过计算机模拟人类的学习来获得人类解决问题的知识。机器学习由于能自动获取知识，在一定程度上能解决专家系统中知识获取“瓶颈”问题。

机器学习通过在数据中搜索统计模式和关系，把记录聚集到特定的分类中，产生规则和规则树。这种方法的优势在于不仅能提供关于预测和分类模型，而且能从数据中产生明确的规则。递归分类算法、神经网络、模糊逻辑、遗传算法、粗糙集理论等被广泛应用于机器学习。

(3) 基于Agent的IDSS Agent是目前AI领域的研究热点，主要有智能型Agent研究、Multi-Agent系统研究和Agent-oriented的程序设计研究三个方面。Agent自身应该具有知识、目标和能力。知识是Agent对其周围环境和要求解的问题的某种描述。目标是Agent解决问题能所能达到的程度。能力就是Agent自身具有的解决问题的技能。

针对不同的具体任务，人们构造不同种类的Agent来满足需要。界面Agent：是由人和计算机通过人机界面组成的一个有机的整体。信息Agent：是对系统中的信息进行各种操作的一种智能体。移动Agent：是指能在复杂的网络系统中自由移动，并通过与服务设施和其它Agent相互协作来完成全局性目标。协作Agent：协作Agent是定义Agent之间的协作关系的Agent，包括各种协作协议、策略、对协作的处理和评估。

(4) 基于数据仓库的IDSS 数据仓库通过多数据源信息的概括、聚集和集成，建立面向主题、集成、时变、持久的数据集合，从而为决策提供可用信息。与数据仓库同时发展起来的OLAP（联机分析处理）技术通过对数据仓库的即时、多维、复杂查询和综合分析，得出隐藏在数据中的总体特征和发展趋势。OLAP进行的多维数据分析有切片和切块、旋转、钻取等方式。

(5) 基于范例推理的IDSS 基于范例推理是从过去的经验中发现解决当前问题线索的方法。过去事件的集合构成一个范例库(casebase)，即问题处理的模型。当前处理的问题成为目标范例，记忆的问题或情景成为源范例，CBR处理问题时，先在范例库中搜索与目标范例具有相同属性的源范例，再通过范例的匹配情况进行调整。基于范例推理简化了知识获取的过程，对过去的求解过程的复用，提高了问题求解的效率，对有些难以通过计算推导来求解的问题，可以发挥很好的作用。

当前智能决策支持已广泛应用于金融、保险、电信、零售业等多个领域。随着

计算机系统应用范围的不断扩大和数据量的迅速增加，这项技术的发展正方兴未艾，其前景无限光明。

1.4 论文工作和章节安排

本文借助现代管理科学及智能决策支持系统理论知识，把智能决策支持系统应用到企业的生产管理中，参照国内外最新研究成果，开发了本企业的生产信息数据分析子系统。本文的主要工作：

1) 搜集国内外已有的决策支持理论，追踪最新决策技术的前沿理论，作为本课题实现的理论基础。

2) 对烧结生产管理进行分析，了解项目对生产信息需要进行数据分析的需求。

3) 根据需求，对系统的结构进行设计研究。

4) 通过编制软件，为企业开发出一个比较实用的生产信息数据分析子系统。

本文的内容安排如下：

第 1 章绪论。总体论述了课题的背景、本课题的研究对象及目的和智能决策支持系统的研究现状。

第 2 章烧结生产管理过程。主要对烧结厂的生产管理进行介绍，为系统的开发奠定基础。

第 3 章相关理论概念综述。对 IDSS 相关技术进行了简单介绍，包括 DW 及 OLAP、DM 的知识和理论。

第 4 章烧结生产决策支持系统的设计与研究。主要有系统的设计；数据仓库模型的设计等；基于数据仓库的 OLAP 应用研究；基于数据仓库的数据挖掘算法研究等。

第 5 章烧结生产信息数据分析子系统的实现。针对上述技术的研究，本章介绍了系统的平台及功能实现。

最后在结论中，论文最后对课题的成果和不足作了总结，并对烧结生产信息数据分析子系统在生产的应用进行了展望。

第 2 章　烧结生产管理过程

所谓“烧结”就是将粉状物料进行高温加热，在不完全熔化的条件下冷结成块的一种冶金造块方法。采用该方法的目的是将在理化性能上不能满足高炉要求的粉状物料加工成为在物理、化学、冶金性能上能满足下道工序的人造富矿。烧结的任务就是满足高炉对入炉矿石日益提高的要求。因此烧结生产管理包括以下几个方面：

2.1　原料管理

烧结生产使用的主要原料有含铁原料、熔剂、燃料。原料数量大，品种繁多，质量不均一。为保证获得高产、优质的烧结矿，做到均衡进料、均衡供料、确保原料质量稳定是关键。因此对原料进行精心管理是烧结生产中一个十分重要的生产环节。原料管理的主要内容包括原料质量管理，进料、用料计划的编制，验收入库，贮存与加工，供应输送，原料入库及消耗的统计，原料成本的控制等。

2.2　烧结调度管理

调度管理是为实现并强化调度系统功能而开展的一种管理实践活动，是企业管理范围内的一项高度综合性的管理工作。搞好烧结调度工作，从而确保整个烧结生产活动的正常有序进行，完成预定的生产经营目标任务[3]。

2.2.1　调度室管理

烧结厂生产技术专业部门是烧结生产活动的指挥中心和信息中心，调度室是其指挥调度功能的具体执行部门，是烧结厂与上至公司、下至各车间以及相关厂矿之间联系的桥梁和纽带。在中夜班和节假日调度室担负着全厂生产组织的重任，在生产组织中行使厂长授予的职权，因而烧结调度室在烧结生产中具有举足轻重的地位。厂调度室的工作职责和内容如下：

在生产技术专业部门有关责任人的领导下，厂调度室当班要为完成日产量、质量计划，进一步实现厂月目标负责。具体的工作职责和内容如下：

1)厂调度室根据厂调度会的安排来组织烧结机的生产运行或检修，并适时地掌握各烧结机的运行情况及检修进度，落实厂调度会布置的解决生产、设备问题及检修计划的兑现情况，未完成的应追查其原因，并向生产技术专业负责人汇报。中班负责搞好当班生产，信息管理系统收集各车间当班生产中反映的生产状况及第二天要检修的设备情况。夜班主要根据白班的计划检修组织生产，抓好烧结矿或原料的备料，为白班安排检修创造条件。

2)负责班中的各类生产、设备、安全问题或事故的处理。对白班出现的生产、设备问题，根据高炉烧结矿槽存情况和问题的严重程度决定是否安排调修及调修的

次序，若决定临时的设备调修，应通知相关责任单位的负责人组织检修人员到现场处理，并随时掌握进度。对中夜班和节假日发生的设备问题，根据设备问题的严重程度酌情处理；当发生较大或重大生产、设备、安全事故时，除按正常程序组织外，还应及时向有关领导和公司总调度室汇报。

3)及时准确地贯彻执行公司总调度室的指令与相关单位的生产协调。应经常与高炉调度室联系，了解高炉的生产情况和烧结矿槽存情况。做好烧结矿对各高炉供应，高炉返矿对各烧结车间的分配工作；与能源总调度室联系解决生产中供电、供水、煤气等能源介质问题；与运输部原料站联系解决入厂原料翻、卸工作中出现的问题。根据用料计划及所掌握的各种烧结原料的库存情况，负责入厂原燃料的接收和调度。负责入厂原、燃料的质量把关，接到原料车间有关入厂原、燃料的质量异议，应立即下令停止翻(卸)车。抓好烧结矿实物质量和节能降耗工作。

4)负责各类调度信息的收集、传递和反馈工作，并将所需的各项数据真实可靠和准确无误地输入计算机内。负责夜班打印《调度日报》，为厂调度会提供生产信息。

2.2.2 调度信息管理

调度信息是企业生产经营活动中各种发展变化和特征的真实反应，包括一切与生产有关的人员、设备、能源、原料及产品本身等各方面的信息，它是生产组织及其调整的直接依据，调度信息的传递和反馈是否准确及时，直接关系到生产能否顺利进行，因而调度信息管理是调度管理中的重要一环。调度信息的传递和反馈层次较为复杂，一旦某一环节出差错，后面得到的就都是错误信息，这将给生产带来极大的波动。为了确保调度信息的真实性、实效性、系统性和目的性，就必须对工段负责人和车间主控室岗位工作人员从严要求。对影响生产的重大信息，厂调度室负责人要到现场了解，核实后再将信息准确、及时地向厂领导和公司总调度室传递，若发生较大的、重大的生产或设备事故，厂调度室负责人可以在 30min 之内向厂领导和公司总调度室汇报。各车间主控室岗位人员在交接班前必须将本班的生产、设备、安全情况及时反馈录入生产信息数据分析子系统，厂调度室和各车间主控室均应按要求做完善、矫正生产信息数据分析子系统《调度日报》记录。厂调度室应按规定时间将本厂生产、设备、安全及有关技术经济指标等情况通过网络反馈给公司总调度室。

2.3 烧结统计管理

统计工作的基本任务是运用一切科学的方法收集、整理、研究生产经营活动中的数据及其相关文字资料，建立统计台帐，编制统计报表，进行统计分析，实行统计监督，开展统计信息咨询服务。其目的是实现统计指标体系完整化、统计分类标准化、统计调查工作科学化、统计基础工作规范化、统计计算和统计数据传输技术

现代化、统计服务优质化[4]。

2.3.1　统计原始数据管理

烧结生产统计原始数据主要来自基层生产岗位、公司质检中心驻烧结厂检查站和化验室，这些单位将每天生产的原始数据、检验数据首先报告给各车间主控室，将这些原始数据全部输入计算机，作为烧结信息数据分析系统的数据库数据。这些原始数据作为烧结生产、经营活动过程的第一手数据资料或文字记载，是未经任何加工整理的原始材料。

烧结各项统计记录较烦琐，全厂的统计记录多达 385 项，统计记录作为记载生产经营活动的原始资料凭证，为了确保其准确、及时、齐全和整洁，厂制定了严格的程序管理文件，对全厂的质量记录结合岗位特点和生产经营的需要重新制定，并在实际工作中切实执行。

2.3.2　统计台帐的管理

统计台帐是基层单位按照日报统计报表和经营管理的需要，将原始记录依时间顺序，经登记、汇总、积累的资料。建立统计台帐，是为了把原始记录的结果，分门别类，定期录入。有了全面、系统的统计台帐，便于本单位领导及时掌握生产经营的动态，也便于填报统计报表。统计台帐分厂部和车间两级台帐，原始记录由车间统计员汇总登入车间统计台帐，并据以编制厂内报表。厂内报表由厂专业部门统计汇总，登入厂级统计台帐，再据以编制统计报表。统计台帐可以由烧结生产信息数据分析子系统来实现。

2.3.3　统计生产指标体系

2.3.3.1　烧结矿产量统计指标

烧结矿产量的计算方法为：

1)烧结矿的产量=当日烧结车间出厂计量原始-记录所列数量(生产量)-高炉的返矿量(*t*)

2)入炉烧结矿的产量=当日进入高炉计量原始记录所列数量(*t*)

3)入炉烧结矿的产量=高炉上料的烧结矿量士中间矿槽及堆场本月和上月贮存差额(*t*)+运输途耗量(*t*)(约 1%)

2.3.3.2　烧结矿生产主要技术经济指标

烧结矿生产主要技术经济指标是直接反映烧结的生产、技术、管理水平和产品质量高低的综合指标体系。技术经济指标的原始资料通过取样、理化检验、技术测定等方法获得。由于烧结矿不进行总体检验，因此有关烧结矿产品质量的指标以其被检品代表总体作为统计依据，所有数据都应进入烧结生产决策支持系统[3]。

2.3.4 统计报表管理

统计报表的编制是统计工作的中心环节。烧结统计报表严格按统计指标体系规定的计算方法编制，所有的计算过程都由计算机完成，同时根据公司报表制度规定的表种、表式、填报单位、填报范围、报送时间、受表单位、报送份数进行填报。

2.4 烧结生产质量管理

随着炼铁技术的发展，高炉对烧结矿产品质量的要求越来越高，因此，烧结矿质量标准也随之提高，目前烧结生产根据公司的要求，烧结厂实施了此标准，以充分满足高炉冶炼的需要。如表(2-1)所示：

表 2-1 某钢铁集团烧结厂烧结矿质量指标规定

Tab. 2-1 a sintering plant in Iron and Steel Group sinter quality requirements

级别	执行标准	化学成分(质量分数%)				物理性能指标(%)		
		TTe	RO 倍	TeO	S	转鼓指数 +6.3mm	筛分指数 -5mm	抗磨指数 -0.5mm
优质	内控	A±0.3	B±0.05	≤8	≤0.05	≥76	≤5	≤7
一级	内控	A±0.5	B±0.08	≤10	≤0.05	≥73	≤7	≤8
合格	内控	A±0.1	B±0.12	≤14	≤0.08	≥70	≤9	≤9

2.5 过程控制管理

过程是产品在生产过程中质量特性发生变化的加工单元，是人员、原料、设备、环境、方法对产品质量起综合作用的环节。过程控制是全面质量管理的重要内容，是烧结生产烧结矿质量保证体系的重要环节。搞好过程控制，不仅可以提高烧结矿质量，还可以增加烧结矿产量，降低能源消耗，有利于生产成本的降低[5]。

2.5.1 过程控制点

控制点所要控制的特性或对象，要尽可能地用数据表示。对生产现场来说，针对过程的问题点，把关键过程和存在的过程中的某些特征控制起来，就是过程控制点。一个过程控制点可以是产品的一个关键质量特征，如烧结矿的碱度、FeO 等；也可以是实际生产中一道关键过程的特性，与温度、透气性等。建立过程控制点就是要把管“结果”(质量特性)转换成管“原因”(人员、设备、环境、材料、方法、环境)。具体来说，就是对控制质量特征，利用因果分析图法进行过程分析，找出支

配性要素，并进行一次、两次或多次展开，直到便于管理为止。然后制定管理标准，规定这些过程要素的管理项目、检测方法、允许界限值以及责任者等，通过控制这些过程要素来达到预报和预控产品质量的目的。

2.5.2　烧结矿过程分析

过程分析是过程控制的基础，不通过过程分析找出过程支配因素及制定标准，过程控制将无据可依，过程质量也将无法保证。在生产过程中，为保证产品质量，通常是预防为主，即以工作质量保过程质量，以过程质量保产品质量。

在实际烧结过程中，烧结厂一般确定原料验收、加工与准备、配料以及烧结为关键过程，控制点也一般建立在这些环节上。过程控制信息量大，只有建立强大的系统作数据分析，才能使生产管理水平迈上一个新台阶。

2.6　本章小结

在本章中主要介绍了烧结生产管理的相关知识，为后面的烧结生产决策支持系统的设计开发奠定了需求分析基础。

第 3 章　相关理论概念综述

3.1　智能决策支持系统

3.1.1　智能决策支持系统的定义

智能决策支持系统(Intelligent Decision Support Systems，简称 IDSS)是将人工智能技术引入决策支持系统而形成的一种新型信息系统。它是以信息技术为手段，应用管理科学、计算机科学及有关学科的理论和方法，针对半结构化和非结构化的决策问题，通过提供背景材料、协助明确问题、修改完善模型、列举可能方案、进行分析比较等方式，为管理者做出正确决策提供帮助的智能型人工交互式信息系统[6]。

高层管理领域中的管理决策者常常遇到一些结构不良问题，由于这些问题无法准确描述处理原则且极其复杂，因而不能应用标准程序性过程进行求解。为了解决这种情况决策支持系统应运而生。从 DSS 产生至今的 20 多年里，DSS 在概念内涵，结构设计和应用研究诸方面取得较快发展。尽管如此，由于传统 DSS 的设计强调对数据、模型和两者集成的支持，其实现起先主要局限在单独和特定的问题领域，因而存在领域依赖和用户接口友好性较差等不足，因此，传统的 DSS 以数据和数学模型分析技术为特征，具有阶段性和局限性。

进入 80 年代后，人工智能(artificial intelligence，简称 AI)技术尤其是专家系统(expert systems，简称 ES)的蓬勃发展为 DSS 的发展注入了新的活力。如何将 DSS 同 AI 尤其是同 ES 相结合以形成一个充分利用人类专家在解决不良结构问题时的知识、方法和经验以达到更有效决策的和用户接口友好的决策支持环境，亦即构成智能决策支持系统(Intelligent Decision Support Systems，简称 IDSS)或称之为基于知识的决策支持系统(knowledge based decision support system，简称 KB-DSS)，便成为 DSS 的重要发展趋势。

智能决策支持系统是在决策支持系统的基础上集成人工智能的专家系统而成的。IDSS 的核心思想是将人工智能技术和其它相关学科的成果及其技术相结合，使 DSS 具有人工智能的行为能够充分利用人类的知识。

随着 IDSS 的发展，人们不断将 IDSS 的智能部件进行扩展，使 IDSS 的智能并不仅仅局限于对知识库的使用上，对其他子系统也加入了智能部件。对模型库而言，它可以实现模型自动选择和生成；对于人机界面部分，它可以使其更容易使用和可以理解决策者的思维，具有学习功能；对于数据库部分，数据仓库、联机分析处理和数据挖掘技术的应用，可以对数据进行复杂的分析处理，同时可从数据(仓库)库中挖掘出隐含的知识，增强原来的知识库，以达到增强系统智能决策的目的[7]。

(1) IDSS 的结构　综合分析已有的 IDSS 研究，可以得出，IDSS 智能的实现从总体可以分为三种：利用 AI 实现系统的智能；利用数据库领域的新工具数据仓库、联机分析处理及知识挖掘技术来帮助实现智能；利用其他技术来实现 IDSS 各部件统一，使 IDSS 在整体上统一表示、相互协调以实现系统的整体智能行为。其中利用 AI 实现系统的智能，又可以分为基于传统 AI 的 IDSS、基于机器学习的 IDSS 和基于 AI 新技术 Agent 的 IDSS。不同的 AI 技术与 DSS 结合，形成不同形式的 IDSS。

DSS 主要是由对话子系统、数据库子系统和模型库子系统组成。ES 是目前 AI 中应用较成熟的一个领域，一般主要由知识库、推理机和知识库管理系统三者组成，它使用非量化的逻辑语句来表达知识，用自动推理的方式进行问题求解，而 DSS 主要使用数量化方法将问题模型化，利用对数值模型的计算结果来进行决策支持。DSS 和 ES 相结合即成为 IDSS。在 IDSS 中，将 DSS 和 ES 结合，主要有三种结合方式。

1) ES 和 DSS 中某个部件相集成

可以将 ES 和 DSS 的数据库、模型库、人机界面部件分别相结合。与数据库相结合，ES 可指导用户的数据操作和模型库相结合，ES 帮助选择模型参数，进行模型计算结果分析，以使用户深入了解和识别决策问题，在今后类似问题求解中顺利选择模型；与人机交互界面相结合，生成友好的人机界面，使用户与计算机交互自然流畅，并对模型计算结果进行合理的解释。

这类集成方式实现起来相对比较简单，但其适用面较窄，系统的智能性有限，当需要表示复杂系统时，就不能满足需要。

2) ES 作为 DSS 的独立部件

这种形式在传统的 DSS 中增加一个知识库子系统或知识处理系统，利用知识库中已有的知识，结合 DSS 模型库的计算结果，经过推理机的推理及分析，得到经定性分析与定量计算的结果，用以指导决策。这种结果形式中数据库是核心部分，起着信息传递和交互作用，其信息传递量较大。转换速度快，已构建的 IDSS 中很多采用这种集成形式，这种集成形式也叫以 DSS 为主体的 IDSS。其逻辑结构如图 3-1 所示。

3) DSS 作为 ES 的一个扩充部件

ES 辅助决策者进行问题求解，在需要具体数据时，从 DSS 的数据库中调用，或运行 DSS 模型间接得到数据。这种集成方式适合于问题求解需要大量知识、经验和判断的决策支持。

将 DSS 和 ES 集成，把 ES 的知识处理融入 DSS，使 DSS 具有一定的智能性。DSS 作为一种推理形式出现，受 ES 中的推理机所控制。这种集成形式也叫以 ES 为主体的 IDSS。

(2) IDSS 的特点　①基于成熟的技术，容易构造出实用系统；②充分利用了

各层次的信息资源；③基于规则的表达方式，使用户易于掌握使用；④具有很强的模块化特性，并且模块重用性好，系统的开发成本低；⑤系统的各部分组合灵活，可实现强大功能，并且易于维护；⑥系统可迅速采用先进的支撑技术，如 AI 技术等。

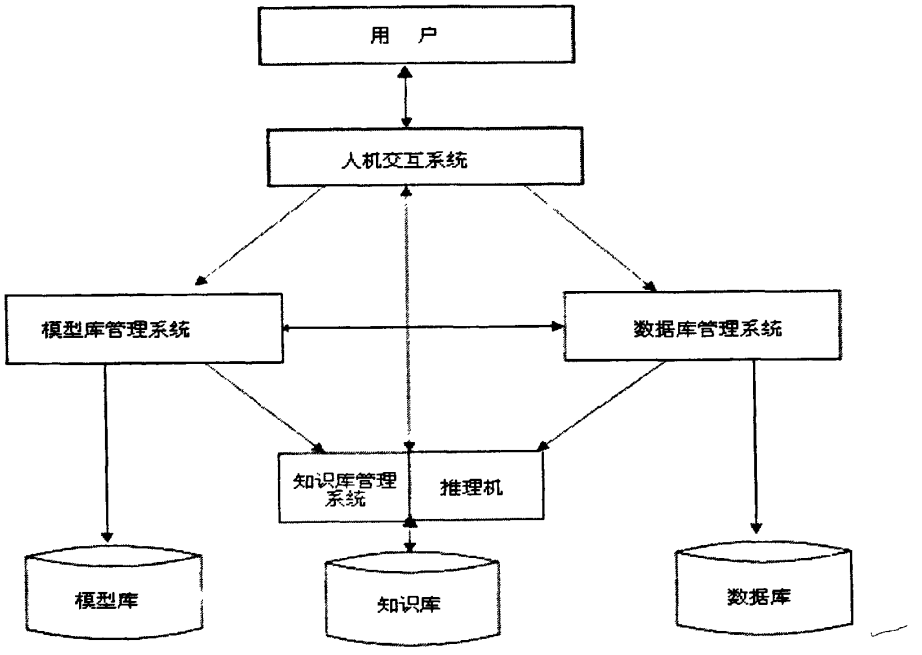

图 3-1 IDSS 结构图

Fig.3-1 IDSS structure drawing

3.1.2 智能决策支持系统的新技术

(1) 联机分析处理 联机分析处理(online Analysis process，简称 OLAP)是一种验证型的分析软件，它具有归纳的作用，它将数据仓库中的数据作为分析对象，通过多种复杂操作，可以为高层管理人员的决策提供有力支持[8]。

OLAP 技术是近年来对 IDSS 研究和应用的一个新进展。OLAP 是在某个假设的前提下通过数据查询和分析来验证或否定这个假设，属于验证型分析。OLAP 能够提供数值，统计等分析信息，缺乏分类，预测等分析能力，特别是从原有信息中发现知识是 OLAP 力不能及的。

(2) 数据仓库 数据仓库(Data Warehouse，简称 DW)的概念是 Prism Solutions 公司副总裁 W. H. Inmon 在 1992 年出版的《建立数据仓库》(Building the DataWarehouse)中提出的。数据仓库的提出是以关系数据库、并行处理和分布式技术的飞速发展为基础，它是解决信息技术(Information Technology，简称 IT)在发展中一方面拥有大量数据，另一方面有用信息却很贫乏(Data rich-Information poor)这种不正常现象的综合解决方案。W. H. Inmon 在《建立数据仓库》一书中，对数据仓库

的定义为：数据仓库是面向主题的、集成的、稳定的、不同时间的数据集合，用于支持经营管理中决策制定过程[9]。

数据仓库是在原有关系型数据库基础上发展形成的，但不同于数据库系统的组织结构形式，它从原有的业务数据库中获得的基本数据和综合数据被分成一些不同的层次(levers)。一般数据仓库的结构组成如图 3-2 所示。包括当前基本数据(current detail data)、历史基本数据(older detail data)、轻度综合数据(lightly stunmarized data)，高度综合数据(highly summarized data)、元数据(meta data)。

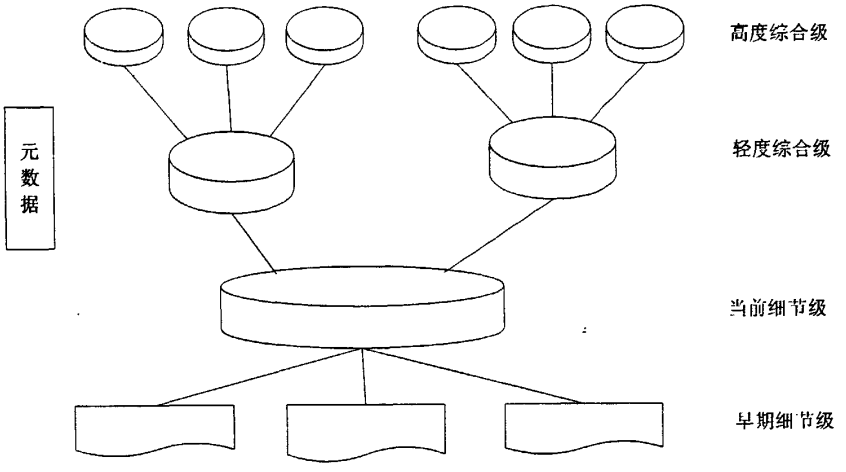

图 3-2　数据仓库的结构组成图

Fig.3-2　The structure constitutional diagram of Data warehouse

数据仓库与关系数据库不同，它没有严格的数学理论基础，而更偏向于工程。由于数据仓库的这种工程性，因而在技术上可以根据它的工作过程分为：数据的抽取、数据的存储和管理、数据的表现以及数据仓库的设计的技术咨询四个方面。

(3)　数据挖掘　1995 年在加拿大召开了第一届知识发现(Knowledge Discovery in Database，简称 KDD)和数据挖掘(Data Mining，简称 DM)国际学术会议以后，“数据挖掘”开始流行，它是“知识发现”概念的深化，知识发现与数据挖掘是人工智能、机器学习与数据库技术相结合的产物[10]。

知识发现被认为是从数据中发现有用知识的整个过程。数据挖掘被认为是 KDD 过程中的一个特定步骤，它用专门算法从数据中抽取模式(patterns)。KDD 过程是多个步骤相互连接起来，反复进行人机交互的过程。具体说明如图 3-3 所示：

第一步，建立某个应用领域：包括应用中的预先知识和目标。

第二步，建立一个目标数据集：选择一个数据集或在多个数据集的子集上聚焦。

第三步，数据清洗和预处理：去除噪声或无关数据，去除空白数据域，考虑时间顺序和数据的变化等。

第四步，数据转换：找到数据的特征进行编码，减少有改变量的数目。如年龄，

10年为一个，定为10级。

第二步，选定某个数据开采算法：决定数据开采的目的，用KDD过程中的准则选择某一特定数据开采算法(如汇总、聚类、分类、回归等)用于搜索数据中的模式，它可以是类似的。

第三步，数据开采：搜索或产生一个特定的感兴趣的模式或一个特定的数据集。

第四步，解释：解释某个发现的模式，去除多余的不切题意的模式，转换某个有用的模式为知识。

第五步，评价知识：将这些知识放到运行系统中，考查这些知识的作用，或者证明这些知识，用预先可信的知识检查和解决知识中可能的矛盾。

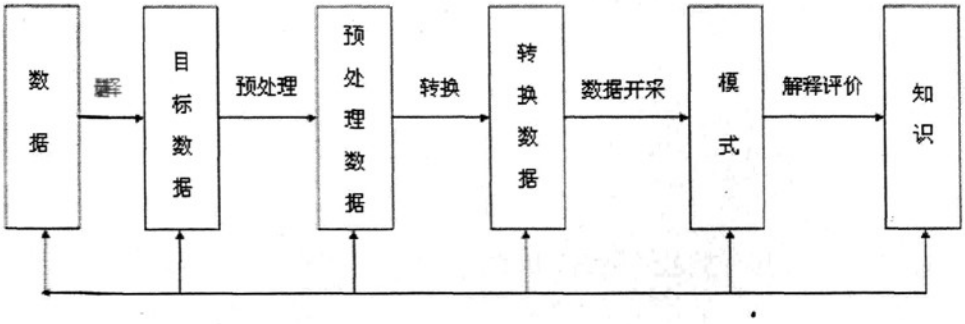

图3-3 KDD过程图

Fig.3-3 The process map of KDD

3.1.3 智能决策支持系统的新体系结构

所谓IDSS的新体系结构就是将IDSS与“DW和DM”结合起来，即将两种不同的辅助决策方式结合起来，起到相辅相成的作用，进一步提高辅助决策的手段。IDSS的新结构是在IDSS的结构上增加DW和DM的结构，统一由“问题综合与交互系统”进行综合集成[11]。

新结构的共同基础是数据库。新结构包括两个主体，一个主体是知识库系统和模型系统，它为决策问题提供定性分析(知识推理)和定量分析(模型计算)相结合的辅助决策信息。另一个主体是DW和DM，它从数据库、数据仓库中提取有用的信息和知识，这些信息和知识反映了大量数据内在的规律和知识。如图3-4所示。

3.2 数据仓库技术

3.2.1 数据仓库的定义

市场需求是技术发展的源动力。数据库技术在企业应用的早期，计算机系统所处理的是系统手工业务自动化的问题。例如银行的储蓄系统、电信的计费系统，它们都属于这类型的联机事务处理系统。联机事务处理系统只涉及当前数据，系统积累

下的历史业务数据往往被转储到脱机的环境中。

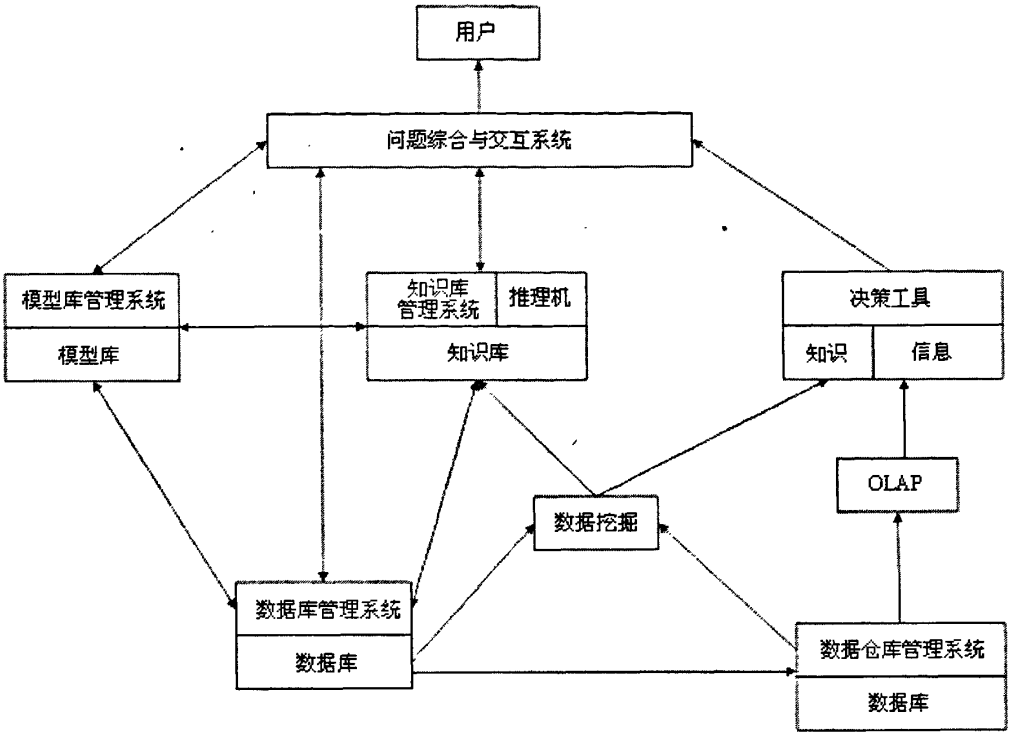

图 3-4　智能决策支持系统的新体系结构

Fig.3-4　The new architecture of Intelligent Decision Support System

应用在不断地进步，企业管理者们发现单靠拥有联机事务处理系统已经不足以获得市场竞争的优势；他们需要对其自身业务的运作以及整个市场相关行业的态势进行分析，从而做出有利的决策。这些决策需要对大量的业务数据包括历史业务数据进行分析才能得到，而这种基于业务数据的决策分析，我们把它称之为联机分析处理。如果说传统联机事务处理强调的是更新数据库—向数据库中添加信息，那么联机分析处理就是要从数据库中获取信息、利用信息。

事实上，将大量的业务数据应用于分析和统计原本是一个非常简单和自然的想法。但在实际的操作中，人们却发现要获得有用的信息、并非想象的那么容易：第一，业务数据往往被存放于分散的异构环境中，不易统一查询访问，而且还有大量的历史数据处于脱机状态，形同虚设；第二，业务数据的模式是针对事务处理系统设计的，数据的格式和描述方式并不适合非计算机专业人员进行业务上的分析和统计。针对这些问题，人们专门为业务的统计分析建立一个数据中心，它的数据可以从联机的事务处理系统、异构的外部数据源、脱机的历史业务数据中得到。通过它可以满足决策支持和联机分析应用所要求的一切数据，这个数据中心就叫做数据仓库。

数据仓库并非是一个仅仅存储数据的简单信息库，因为这实际上与传统数据库没有两样。数据仓库实际上是一个“面向主题的、集成的、不可更新的、随时间不断变化的数据集合”。如果说传统数据库系统的重点与要求是快速、准确、安全、可靠地将数据存进数据库中的话，那么数据仓库系统的重点与要求就是能够准确、安全、可靠地从数据库中取出数据，经过加工转换成有规律信息之后，再供管理人员进行分析使用。

如果需要给数据仓库一个定义的话，那么可以把它看作一个作为决策支持系统和联机分析应用数据源的结构化数据环境。数据仓库所要研究和解决的问题就是从海量数据中获取信息。W. H. Inmon 在其著作《Building the Data warehouse》一书中对数据仓库给予了如下描述：数据仓库(Data Warehouse)是一个面向主题的(Subject-oriented)、集成的(Integrate)、相对稳定的(Nonvolatile)、反映历史变化(Time-variant)的数据集合，用于支持管理决策。对于这个描述我们可以从两个层次予以理解，首先，数据仓库用于支持决策，面向分析型数据处理，它不同于企业的操作型数据库；其次，数据仓库是对多个异构的数据源有效集成，集成后按照主题进行了重组，并包含历史数据，而且存放在数据仓库中的数据一般不再修改。

从广义的实践的角度来理解，数据仓库是一个体系结构，一个以所定义的数据集合为中心的以决策支持为主导的支持企业运作的 IT 体系结构。数据仓库的关键技术分为数据的抽取、存储与管理以及数据的表现等三个基本方面[12]：

(1) **数据的抽取** 数据的抽取是数据进入仓库的入口。由于数据仓库是一个独立的数据环境，它需要通过抽取过程将数据从联机事务处理系统、外部数据源、脱机的数据存储介质中导入到数据仓库。数据抽取在技术上主要涉及互连、复制、增量、转换、调度和监控等方面。数据仓库中的数据并不要求与联机事务处理系统保持实时同步，因此数据抽取可以定时进行。

(2) **存储和管理** 数据仓库的核心是数据的存储和管理。数据仓库的组织管理方式决定了它有别于传统数据库，同时也决定了其对外部数据的表现形式。要决定采用什么产品和技术来建立数据仓库，需要从数据仓库的技术特点着手分析。

(3) **数据的表现** 数据表现实际上相当于数据仓库的门面，其性能主要集中在多维分析、数理统计和数据挖掘方面。而多维分析又是数据仓库的重要表现形式。

3.2.2 数据仓库的系统结构

数据仓库系统由数据仓库、仓库管理和分析工具三部分组成[13]。其系统结构形式如图 3-5 所示：

(1) **仓库管理** 在确定数据仓库信息需求之后，首先进行数据建模，确定从源数据到数据仓库的数据抽取、清理和转换过程。元数据是数据仓库的核心，它用于存储数据模型、定义数据结构、转换规划、仓库结构、控制信息等。仓库的管理工

作需通过数据仓库管理系统来完成。

(2)　数据仓库工具集　由于数据仓库的数据量大，必须有一套功能很强的分析工具集来实现从数据仓库中提供辅助决策的信息，完成决策支持系统的各种要求。分析工具集分两类工具：

1) 查询工具

数据仓库的查询不是查询记录级数据，而是查询分析要求。查询工具一般含可视化工具和多维分析工具。

2) 数据挖掘工具

从大量数据中挖掘出有规律性的知识需要利用数据挖掘工具。

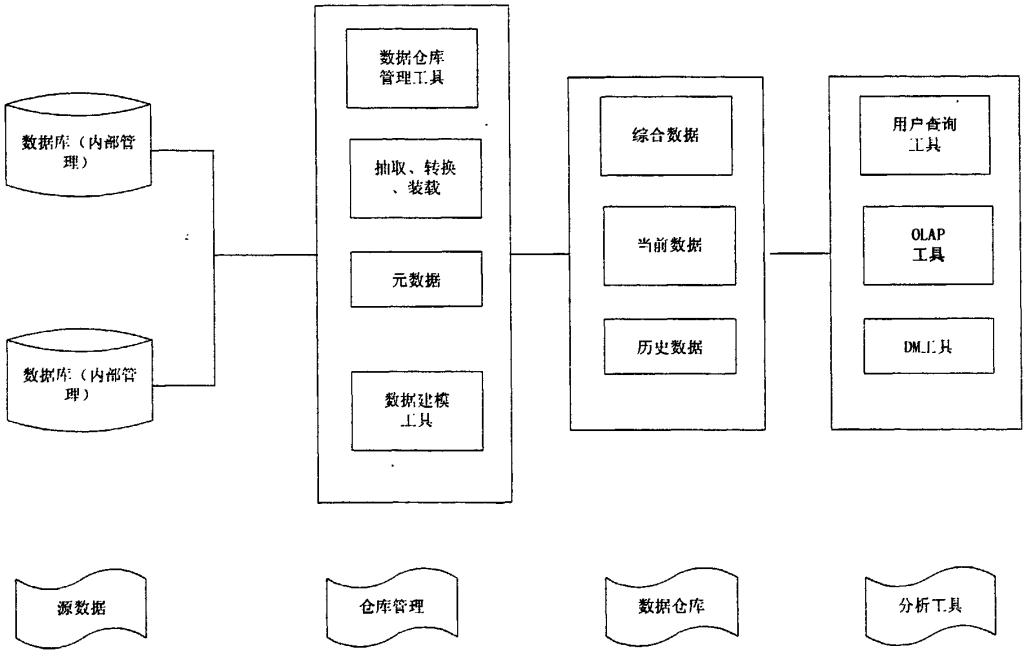

图 3-5　数据仓库结构图

Fig.3-5　The structure drawing of Data Warehouse

3.2.3　数据仓库的数据获取

数据在进入数据仓库前，必须按照数据质量标准进行净化和清洁[14]。

(1)　数据质量的标准　①数据的准确性；②数据符合它的类型要求和取值要求；③数据具有完整性和不冗余；④数据是集成的和一致的；⑤数据是及时的，遵循业务规则，满足业务要求。

(2)　数据变换　在业务数据加载到数据仓库之前，完成对数据的内容和结构变化的整个过程称为数据变换。数据变换的目的是改进数据仓库中数据的质量，提高数据仓库中数据的可用性。

1) 简单变换

每变换一次改变一个数据属性，而不考虑该属性的背景或与它相关的信息。如数据类型转换、日期/时间格式转换等。

2) 清洁和刷洗

这是比简单变换更复杂的一种变换。清洁和刷洗主要有两种方式：一种是检验有效值，如要求数据值落在预定的范围内或在指定的值中，否则就予以剔除；另一种是重新格式化，它是将以不同方式存储在不同数据来源中的同类信息转换成统一的表示方法存入到数据仓库。

3) 集成

集成就是把不同数据源中得到的业务数据结合在一起，将它们集成为一个紧密结合的数据模型。

4) 聚集和概括

大多数数据仓库都要用到数据的某种聚集和概括。聚集是指将不同业务元素加在一起成为一个概括，概括是指按一个或几个业务维最近的数据加在一起。例如，按帐户号码和按日期概括各次事物，并在数据仓库中存入概括信息。在许多情况下，一定时期内需要很细节的数据，到了某一时期对所有这些细节数据的需要就减弱了，只需保留一些概括数据。对于各种细节数据，有许多概括的方法，每一种概括或聚集都可以在某一个或几个维中进行。

3.2.4 联机分析处理与数据挖掘

数据仓库是进行决策分析的基础，但只有通过高效的工具，数据仓库才能真正发挥出数据宝库的作用。联机分析处理和数据挖掘就是与数据仓库密切相关的分析工具[15]。

3.2.4.1 联机分析处理

OLAP 是使用户能够从多种角度对从原始数据中转化出来的信息进行快速、一致、交互地存取，从而获得对数据的更深入了解的一类软件技术。它通过快速、一致、交互地访问各种可能的信息视图，帮助数据分析人员、管理人员、决策人员洞察数据深处的奥秘，掌握隐于其中的规律。

OLAP 是数据仓库主要的前端支持工具，侧重于把数据仓库中的数据进行分析，转换成辅助决策信息。OLAP 的重要特点就是多维数据分析，这与数据仓库中的多维数据组织方式正好形成相互结合、相互补充的两个方面。OLAP 的目标是满足多维环境特定的查询和报表需求，它的技术核心是“维”这个概念，OLAP 可以说是多维数据分析工具的集合。

OLAP 的多维数据分析，主要通过以下三种方式进行：

(1) **切片和切块** 在多维数据结构中，按二维进行切片，按三维进行切块，可得到所需要的数据。如在“城市、产品、时间”三维立方体中进行切块和切片，可

以得到各城市、各产品、各期间的销售情况。

(2)　**钻取**　钻取包括向下钻取和向上钻取操作，钻取的深度与维所划分的层次相对应。

(3)　**旋转**　通过旋转可以得到不同视角的数据。

3.2.4.2　数据挖掘

数据挖掘，从技术角度来定义，就是从大量数据中提取人们感兴趣的知识，这些知识是隐含的、事先未知的潜在有用信息。提取的知识表示为概念(Concepts)、规则(Rule)、规律(Regularities)、模式(Patterns)等形式。从商业角度来定义，数据挖掘是一种商业信息处理技术，其主要特点是对商业业务数据进行抽取、转换、分析和其他模型化处理，在大量数据中找出有价值的隐藏信息，加以分析，并将这些信息归纳成结构模式，作为企业在进行决策时的参考[16]。

数据挖掘的功能是确定数据挖掘任务中要找的模式类型，模式主要有以下五种：

(1)　**分类**　分类是数据挖掘中应用得最多的功能，它是根据数据的属性将数据分派到不同的组中。在实际应用过程中，分类模式可以分析分组中数据的各种属性，并找出数据的属性模型。这样就可以利用该模式来分析已有数据，并预测新数据将属于哪一个组。

(2)　**聚类**　当要分析的数据缺乏描述信息，或者是无法组织成任何分类模式时，可以采用聚类模式。聚类是将数据中比较接近的划归为一类，合理的聚类后，每一类内就可以找出有关的特征，有利于发现真正有用的信息。聚类和分类的区别在于：分类是基于已有的分类体系表的，而聚类则没有分类表，只是基于数据之间的相似度。

(3)　**估计与预测**　估计是根据已有的长期积累的数据来推测某一属性未知的值，预测是根据数据某一属性的过去值来推测该属性未来的值。

(4)　**关联分析**　若两个或多个数据项的取值重复出现且概率很高时，它们就存在某种关联，可以据此建立起这些数据项的关联规则。如在超市销售中买面包的顾客 90%的人还同时购买牛奶。

(5)　**序列发现**　序列发现与关联分析相仿，但它需把数据之间的关联性与时间联系起来，主要用于分析数据仓库中某类同时间相关的数据，并发现某一时间段内数据的相关处理模型。为了发现序列模式，不仅需要知道某事件是否发生，而且需要确定事件发生的时间。如购买彩电的人当中的 50%会在 3 个月内购买影蝶机.

以上五种模式可以产生五种基本类型的信息：第一种是分类信息，它是最常用的一种信息，主要在于找出描述一组数据共同特性的模式；第二种是聚类信息，它把那些没有类别的数据聚集成多个类别，给用户提供“物以类聚”的宏观观念；第三种是估计和预测信息，这是最容易理解的信息，它可以通过使用隐藏在数据中的

模型来估计一些连续变量的未来值；第四种是关联信息，它可以显示与单个事件相关的信息；第五种是序列信息，它可以显示所有时间内互相链接的一些事件。

总的来说，数据挖掘的任务是找出特征、规律、联系而不是验证；数据挖掘必须是多种技术的结合，而不只是统计分析。

3.3 本章小结

在本章中介绍智能决策支持系统的含义，并追踪智能决策支持系统的新技术，介绍了数据仓库等技术的相关理论和知识，为烧结生产决策支持系统的建立奠定了技术基础。

第 4 章　烧结生产决策支持系统的设计

4.1　系统需求概述

4.1.1　烧结厂各管理层次的需求分析

(1)　集团公司需求　在已有的网络平台(广域网、企业网、工控网)、生产业务系统、生产控制系统等信息基础设施基础上，进一步整合数据资源，最大限度地发挥出信息资源在配置企业资源、辅助中高层管理决策、优化生产过程、提高企业运作效率、降低成本的重要作用，技术水平上达到行业领先(达到行业/省级科技进步奖的标准)。

(2)　烧结厂厂领导主要需求　系统能实时监测企业的生产经营状态，为厂领导生产经营决策提供依据，通过系统的运行真正起到提高产量，降低成本的目标。其中厂领导特别关注的指标是每天的品位(烧结矿)、产量(总产量，生产单元的产量)、生产完成情况、成本(总成本、作业段成本、成本构成项目的变动情况)。

(3)　生产车间主要需求　①一烧、二烧和三烧车间：系统能实时提供车间总体的生产的实时状态及与定额的比较；各种生产成分构成项目的变动情况；机旁备件的领用和使用状况；②原料车间、成品车间：成本的实时状态；各种成本构成项目的变动情况。

(4)　生产辅助部门　①动力：提供生产过程中所消耗的电、水实时信息以及关键大型设备(烧结机)、变压器的能耗情况，尤其是电力的消耗情况；各生产车间能耗情况及所占的比例；②协力：设备维修计划、维修事务管理、设备点检管理等。

4.1.2　系统的基本目标

1) 对烧结矿生产过程自动化系统采集的数据按要求进行存贮、统计查询，保证生产连续、稳定、高质和高效；通过应用生产执行系统，最终达到减员增效的目的；通过设备安全管理，达到保障设备安全、减少故障停机时间，提高设备作业率；通过对能耗、物耗的管理、生产成本的监控和物资库存优化管理与控制，建立企业内部的信息传输平台，实现“横向到部门，纵向到车间”，达到强化管理、节能降耗、降低成本目标。

2) 通过对系统研究，建立静态和动态成本数学模型、生产量预测数学模型、能耗、物耗实时监控管理等项内容，从而，达到为厂、车间领导提供决策依据，及时准确指挥生产，强化管理、提高效率、节能降耗、降低成本、提高企业综合经济效益，实现国内一流烧结矿企业的目标。

4.2 烧结生产决策支持系统架构

4.2.1 系统设计方案

在认真研究了烧结生产控制过程后，提出了基于数据仓库、联机分析处理、数据挖掘三种技术相结合的烧结生产决策支持系统的解决方案，即 DW+OLAP+DM～IDSS 的可行方案。

数据仓库用于数据的存储和组织，OLAP 集中于数据的分析，数据挖掘则致力于知识的自动发现。

新型 IDSS 架构是以数据库中大量数据为基础，特点如下[17]：

1) 在底层的数据库中保存了大量的事务级细节数据。这些数据是整个 IDSS 系统的数据来源。

2) 数据仓库对底层数据库中的事务级数据进行集成、转换、综合，重新组织成面向全局的数据视图，为 IDSS 提供数据存储和组织的基础。

3) OLAP 从数据仓库中的集成数据出发，构建面向分析的多维数据模型，再使用多维分析方法从多个不同的视角对多维数据进行分析、比较，分析活动从以前的方法驱动转向数据驱动，分析方法和数据结构实现了分离。

4) 数据挖掘以数据仓库和多维数据集中的大量数据为基础，自动发现数据中的潜在模式，并以这些模式为基础自动地做出预测。

数据挖掘表明知识就隐藏在日常积累下来的大量数据中，仅靠复杂的算法和推理并不能发现知识，数据才是知识的真正源泉。

4.2.2 系统体系结构

烧结生产决策支持系统的体系结构如图 4-1 所示。

烧结生产决策支持系统在总体上分为三层。首先是数据获取层，将原有的烧结机等生产控制机器中的相关数据进行清洗、加工、整理、抽取到数据仓库中。然后根据自身管理、业务的需要在数据仓库的基础上建立适合自身应用的多维数据集，形成第二层数据存储层。数据仓库、多维数据集中蕴含的信息可以在数据访问层通过报表、OLAP、即时查询、数据挖掘形式向 IDSS 系统使用人员展现[18]。

最终用户通过客户界面访问系统数据信息，在登录系统后，系统将按照用户的授权、角色向最终用户展现集成的信息。

4.3 烧结生产决策支持系统中的数据仓库模型设计

当前，数据仓库的模型组织方式通常主要有星型结构、雪花结构和混合结构等三种模式。

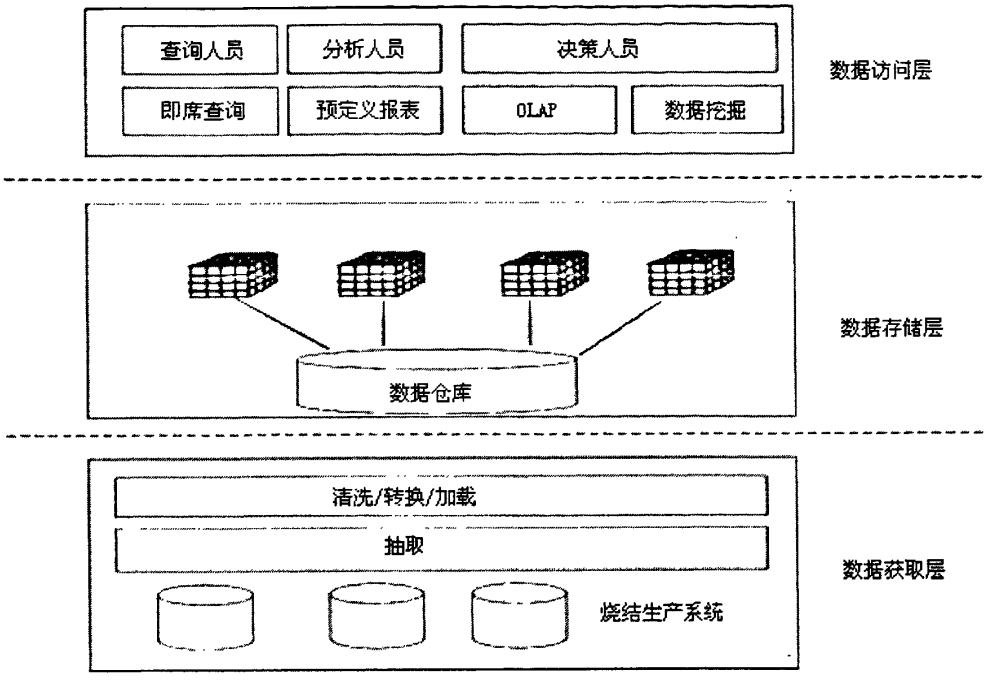

图 4-1　系统体系结构图

Fig.4-1　System Structure

星型结构是构建数据仓库最常用的一种结构，由两类基本表组成：一个事实表和多个维表；事实表包含实际事务或需要分析的值，维表包含有关这些事务或值的描述信息，每维一个表。

雪花结构是星型结构的变种，主要通过规范化星型结构的某些维表形成。然而，由于这种规范化，使得雪花结构可能降低查询的性能，其使用受到局限。

混合结构主要针对多个事实表的情形设计，在这些事实表之间存在有共享的维表。

在烧结生产决策支持系统中，数据结构比较简单，主要包括实时监控信息、效应信息、库存、工艺参数、生产等。根据现场的实际需求，在设计数据仓库时，将上述五类信息形成五个主题，每个主题都选择星型结构进行构建。

下面，以实时监控信息主题为例，说明烧结生产决策支持系统数据仓库模型结构的具体设计。

在实时监控信息主题中，包含一个事实表和两个维表。事实表是一个包含多属性值的关系表，其信息主要是原料、风速、真空度、电压、机速和烧结终点等若干属性值，构成相应的度量；两个维表分别为时间维和部门维。在实际的应用中，用户感兴趣的往往是类似于“上个月一工区的风速偏小，出矿量少”等形式的问题，这种数据分析与挖掘的方式，需要对维引入概念分层，形成相应的聚合数据。为此，

对时间维设计了年、季度、月、日、班、小时、一分钟等若干层次；部门维则包括系列、车间、工区、班组、烧结机机号等若干层次。整个主题最后构成一个分层多维的数据立方体(CUBE)。其结构如图 4-2 所示。

其它主题的设计与实时监控信息主题相类似，最终都形成相应的多层数据立方体。

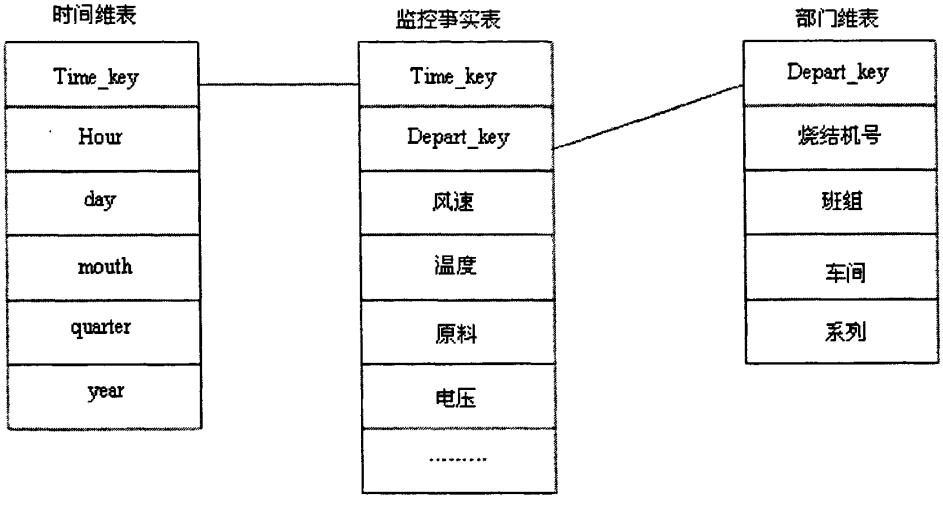

图 4-2　监控信息主题的星型模式

Fig.4-2　The star pattern of theme of Monitoring

4.3.1　烧结生产决策支持系统中完整的数据仓库模型

在烧结生产决策支持系统中，完整的数据仓库模型包括核心组成部分与外围基础部分。核心的组成部分由数据预处理、数据立方体存储和 OLAP 访问操作等构成；其外围组成则包括软件基础和硬件基础。其结构如图 4-3 所示。

数据预处理过程是数据分析和挖掘的数据基础，具体的实现步骤包括：

1) 根据需要，通过人工或程序对底层控制系统数据进行清理，如空穴填充、去除噪音等；

2) 对同构或异构的数据源建立统一名称映射表，组织数据转换策略，实现事务数据向仓库数据的转换和过滤；

3) 为数据分析和挖掘操作进行数据标准化，如对数据的初值化、均值化等。

数据立方体存储是数据仓库的数据中心，由事实表和维表构成；为了提高数据访问的效率，通常对某些列构建索引或实现物化视图。

OLAP 访问操作包括上卷、下钻、旋转等；作为其功能扩展，则包括数据挖掘、决策支持等。访问操作向用户返回最终的结果，其性能的好坏，直接影响到数据仓库构建的效果。

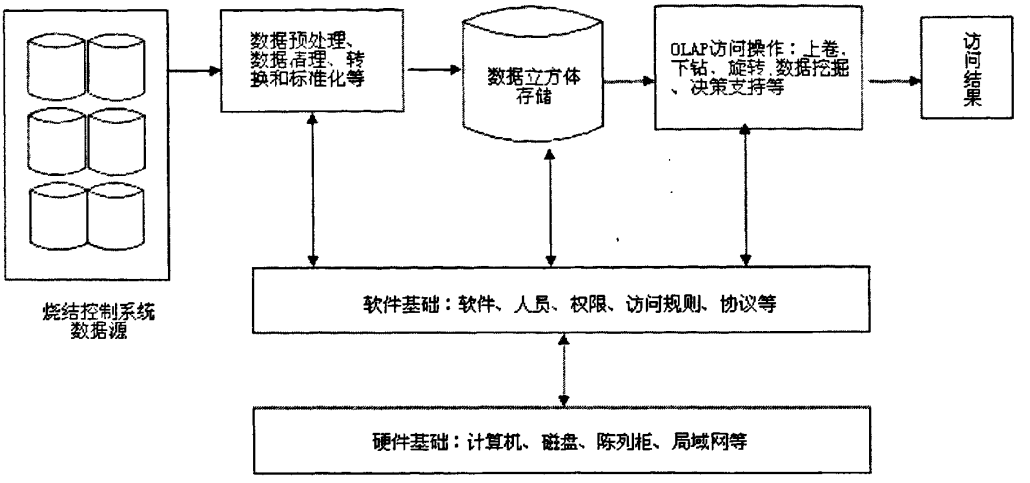

图 4-3　完整的烧结控制系统数据仓库模型

Fig.4-3　The model of data warehouse of complete Sintering control system

软件基础除了系统运行所需要的一组软件以外，还包括数据仓库的操作人员(维护人员、分析人员、开发设计人员等)，访问的权限与角色设置，OLAP 操作规则，网络访问的相关协议、数据备份策略等；硬件基础则包括数据仓库所在的计算机硬件组成，磁盘、磁带机、阵列柜等存储设备、网络环境等。

4.3.2　数据粒度的选择与数据存储

所谓粒度，是指数据的详细程度，数据越详细，粒度越小；数据的抽象程度越高，粒度越大。在数据仓库中，数据粒度的选择非常重要，直接影响到物理存储空间、系统的性能等。在选择粒度的大小之前，需要计算烧结生产决策支持系统中的数据量和记录行数，并根据服务器的空间大小和处理性能进行调整。

以实时监控系统为例，在监控信息主题中，事实表的属性个数为 30 个左右，每列的长度平均为 8 个字节。下面是其相关的计算结果：

4.3.2.1　以一分钟存储一个点考虑

1) 记录条数

单台烧结机一分钟数据：

一天：24×60=1440(条)

一个月：30×1440=43200(条)

一年：12×43200=518400(条)

2) 字节数

一条记录所占的字节数：30×8=240(B)

单台烧结机一天的数据量：1440×240=345600(B)=337.5(KB)

单台烧结机一个月的数据量：337.5×30=10125(KB)=9.89(MB)

单台烧结机一年的数据量：9.89×2=118.68(MB)

4.3.2.2 以一个小时存储一个点考虑

1)记录条数

单台小时数据：

一天：24(条)

一个月：24×30=720(条)

一年：720×12=8640(条)

2)字节数

单台烧结机一天的数据量：24×240=5760(B)=5.625(KB)

单台烧结机一个月的数据量：5.625×30=168.75(KB)

单台烧结机一年的数据量：168.75×12=2025(KB)=1.98(MB)

4.3.2.3 以一天存储一个点考虑

1)记录条数

单台日数据：

一天：1(条)

一个月：1×30=30(条)

一年：30×12=360(条)

2)字节数

单台烧结机一天的数据量：1×240=240(B)

单台烧结机一个月的数据量：240×30=7200(B)=7.03(KB)

单台烧结机一年的数据量：7.03×12=8100(KB)=84.375(KB)

上述计算结果显示了监控信息的数据量；此外，在数据仓库中使用索引(例如，建立三个主键的索引)，则索引的记录数同上，索引空间是相应数据存储空间的十分之一左右。多家烧结厂的实际应用基本与上述计算的结果相符合。

以国内3000天左右的烧结机寿命计算，则需要存储五年的数据。如此庞大的数据量和记录数，对数据仓库的查询性能和效率是一个挑战。为此，作者提出使用三重粒度级别进行数据存储：一分钟粒度、小时粒度和日粒度。其存储策略为前一种细节数据保存两年，后两种轻度综合的数据则保持五年(根据实际需求可以保存更长时间)。

4.4 数据预处理

在烧结生产运行过程中，由于其产生的数据具有特殊性，不能直接导入到数据仓库中，否则会导致数据的冗余、不完整等。所以需要对其进行预处理，

4.4.1　空穴填充

在数据预处理过程中，需要对数据进行清理、过滤等操作。空穴现象是其中的一个重要问题。产生空穴的原因主要是生产控制系统在进行采样时，由于某些无法克服的困难(如网络拥塞时产生的丢帧现象)导致数据序列出现空穴，或者是工作中的实际困难导致数据序列出现空穴(如工艺参数中分子比的测定，工厂一般一周只测 3 次，一周便会有 4 个空数据)。如何有效地处理这些空穴，将会影响到数据的分析结果。

处理空穴办法主要有三种，一种是采用过滤的方法，将带有空穴数据的记录过滤掉，留下的数据形成完整的序列；另一种方法是使用均值对空穴进行替换；第三种方法是插值，将空缺的数据近似地补回来[19]。

(1)　过滤　过滤的算法比较简单，通过简单的比较判断记录的所有属性是否为空即可完成，具体的算法描述如下：

表 4-1　烧结厂 1#机的工艺参数部分数据

Tab.4-1　Sintering plant part of the process parameters of the 1# machine

ID	日期	温度(℃)	FeO 含量(%)	出矿量(t)
1	2007-7-1	1050	20	1500
2	2007-7-2	1020		1406
3	2007-7-3	1080	19	1402
4	2007-7-4	1075		1410
5	2007-7-5	1035	19.5	1460
6	2007-7-6	1056		1400
7	2007-7-7	1045	21	1450
8	2007-7-8	1095		1500

```
Filter_algorithms(time_series_data_set:DB)
Begin
  Let   result_dbset = Φ
  For each   record   REC ∈ DB   do
    Begin
      If   REC. parameters is not null then set result_dbset=result_dbset ∪ REC
      End
  Return result_dbset
End
```

但是，过滤算法在烧结生产决策支持系统中存在有很大的缺陷。仍以工艺参数

为例，分子比每周测量三次，但是，烧结混合料温度和烧结混合料水平等参数却是每两天测量一次，而温度、出矿量等则基本上每天都有数据，表 4-1 列出了某厂 1# 机 10 天的部分工艺参数数据。

采用过滤算法计算后，虽然形成了完整的时间序列，但却只保留了 3 天的数据。表 4-2 则列出了过滤算法计算后的结果。

通过上述示例可以看出，采用这种过滤的算法，取交集后将会丢失大部分有用的数据，特别是蕴藏于其中的未知规律。所以，在进行分析或数据挖掘时，或者得不到满意的结果，或分析的结果严重偏离其实际值，算法的性能非常差，不利于管理和指导烧结厂的实际生产。

表 4-2 烧结厂 1#机的工艺参数部分数据

Tab.4-2 Sintering plant part of the process parameters of the 1 # machine

ID	日期	温度(℃)	FeO 含量(%)	出矿量(t)
1	2007-7-1	1050	20	1500
3	2007-7-3	1080	19	1402
5	2007-7-5	1035	19.5	1460
7	2007-7-7	1045	21	1450

(2) **替换** 替换算法实现空穴填充的思想是，使用序列均值或将序列分类后使用该记录所属分类的均值对空穴进行替换。序列均值具体的算法描述如下：

```
Replacewe_algorithms(time_series_data_set: DB)
Begin
  For each parameter ∈ REC. parameters do
    Let averagees_dbset(count of parameters)=average(parameters. value)
  For each record REC ∈ DB do
    Begin
      If  rec. parameters. value is null then
        Set rec. parameters. value=average_dbset. parameters. value
      End
  Return DB
End
```

序列分类均值需要先进行数据的分类，对每个分类计算相应属性的均值，并利用该均值替换相应分类中空穴数据的值。具体的序列分类均值算法本文不再详细讨论。

以均值替换算法为例，考察表 4-1 的数据，经过替换算法计算后，其结果见表

4-3。

这种替换算法的性能比过滤算法有所改进，不会丢失信息。但是，由于用同一个值进行替换，没有考虑系统运行的趋势，且在进行分析或挖掘时会获得这种特殊的规则，算法的性能仍有待改进。

表 4-3　替换后 1#机的工艺参数数据

Tab.4-3　After the replacement of 1# machine data processing parameters

ID	日期	温度(℃)	FeO 含量(%)	出矿量(t)
1	2007-7-1	1050	20	1500
2	2007-7-2	1020	19.625	1406
3	2007-7-3	1080	19	1402
4	2007-7-4	1075	19.625	1410
5	2007-7-5	1035	19.5	1460
6	2007-7-6	1056	19.625	1400
7	2007-7-7	1045	21	1450
8	2007-7-8	1095	19.625	1500

(3)　插值　利用插值方法进行空穴填充的思想是，根据系统运行的惯性，参考空穴前后的数据进行分析，获得近似的预测结果作为空穴的值。插值方法的优点是保留系统运行时产生的所有信息，只要选择合适的算法，总可以近似地找到系统运行的规律；其不足之处是，由于增加了信息，虽然这些信息可能比较接近其实际的运行值，但误差的存在是不可避免的，分析或挖掘的结果有时候会失真。但是，考虑到烧结生产固有的特点，即烧结过程是一个缓慢的变化过程，反应的持续时间长，其各种工作参数的变化基本符合其总体发展变化的趋势。考虑系统运行的惯性，插值后获得的结果基本上不会偏离实际很远，失真的概率比较小。

4.4.2　重复记录处理

在烧结控制过程中，由于一些不可避免的原因，如底层 CAN 通讯网的堵塞或工控机故障等，可能会产生重复记录。例如，在表 4-4 所示的小时数据中，在 24 点钟的时候产生了一条有时间却没有其它信息的重复记录，其记录号为 2。

这种重复记录进入数据仓库后参与数据分析，将会影响该天的统计值，得到虚假的分析结果。所以，在进行数据预处理时，必需找出所有这种重复的记录，经过比较，将产生噪音的重复记录删除。

在烧结控制过程中，类似于小时数据这种重复记录的处理，其过程主要包括三个步骤：重复记录辩识、噪音识别及记录过滤。记录过滤的方法比较简单，主要是根据获得的噪音记录号，对该条记录进行合并、删除等操作。下面，主要描述前两

个步骤。

(1) 重复记录辩识 在烧结控制过程中，小时数据每个小时产生一条记录，根据这种特点，可以对时间值进行处理，即构造一个新的整型序列，其取值为对应记录中时间属性的整点值。在处理时，可以直接对时间取小时数，结果如表 4-5 所示。

表 4-4 有重复记录的小时数据

Tab.4-4 There are duplicate records of the hours of data

记录号	烧结机号	日期	时间(h)	设定电压(V)	工作电压(V)	电流(A)	电阻(O)
1	001	2007-7-5	23:03:00	4.15	4.19	203.4	4.154
2	001	2007-7-6	00:03:00	0	0	0	0
3	001	2007-7-6	00:10:00	4.15	4.184	202.4	4.16
4	001	2007-7-6	1:03:00	4.15	4.241	206.3	4.168

表 4-5 对时间值处理后的中间数据

Tab.4-5 On the value of time after the middle of the data

记录号	烧结机号	日期	时间(h)	设定电压(V)	工作电压(V)	电流(A)	电阻(O)
1	001	2007-7-5	23	4.15	4.19	203.4	4.154
2	001	2007-7-6	0	0	0	0	0
3	001	2007-7-6	0	4.15	4.184	202.4	4.16
4	001	2007-7-6	1	4.15	4.241	206.3	4.168

对时间值处理后的中间数据，其重复记录辩识的算法如下：

```
Identify_repeat_algorithm( recordsets: RECS)
Begin
  Get main parameters including jihao，date，time，id
    /*选取辩识记录的主要属性，包括机号、日期、时间和记录号*/
  Ordered RECS by jihao,date,time
    /*依次对机号、日期、时间进行排序*/
  Let Storages= }
  Let count=0
  Let flag=false
  For i=1 to RECS.counts-1 do
    Begin
    If RECS(i).jihao.value=RECS(i+1).jihao.value and_
      RECS(i).date.value=RECS(i+1).date.value and_
      RECS(i).time.value=RECS(i+1).time.value then
```

```
/*比较第 i 条和第 i+1 条记录的选定属性值*/
Begin
  If flag is true then
      Storages(count)=Storages(count) ∪ RESC(i+1).id
  Else
   Begin
      Set flag=true
         count=count+1
      Storages(count)=Storages(count) ∪ RESC(i).id ∪ RESC(i+1).id
        End
  End
Else
Set flag=true
End
  Return count and Storages
End
```

例如，算法对表 4-5 的数据进行处理后，返回重复的记录组数为 1，其重复的记录号为{2，3}。

在算法 Identify_repeat_algorithm(IRA)中，只针对三个主要的属性进行辩识，不失一般性，将其推广到所有的属性集，同时，引入属性的权重，以区分辩识过程中不同属性所起的作用。因此，作为算法的一种扩展，考虑一般情形中的重复记录辩识问题。为了讨论的方便，先给出相关的条件描述：

设数据集有 n 个属性，Pam_i 是其中的一个属性，i=1，2，…，n；每个属性给定一个权重，设为 w_1，w_2，w_n，且满足条件：$w_1+w_2+\cdots+w_n=1$；RecA，RecB 是数据集中的两条记录；若 Pam_i 是字符串类型，则 $LenA_i$ 表示记录 RecA 的 Pam_i 属性的字符长度，$LenB_i$ 表示记录 RecB 的 Pam_i 属性的字符长度；若 Pam_i 是数值类型，则 Max_i 表示数据集中该属性序列的最大值，Min_i 表示数据集中该属性序列的最小值；记 $Len_i=Max(LenA_i, LenB_i)$；$Reduce_i=Max_i-Min_i$；对于二进制数据和逻辑型数据，其处理方式和字符型数据相类似，而日期型数据则可以转化为数字型数据，不再说明。

(2) 噪音识别　在烧结生产控制过程中，对重复记录辩识后，如何识别重复记录中的噪音是重复记录处理的另一个重要过程。通常，有两种识别方法：手工识别和计算机自动识别[20]。

手工识别的处理过程比较简单，根据重复记录辩识返回的记录号，获得所有重复记录的详细信息，通过观察，识别噪音记录，并对噪音标记，然后通过手工进行

清理；或者返回系统，进行下一步操作，如合并、删除等处理。对重复记录的分组数目不多时，可以采取这种手工处理的方式；但是，当分组数目比较大时，这种手工处理的效率比较低下，需要寻求一种自动识别模式。

在烧结生产控制过程中，对重复记录中的噪音进行自动识别，其应用的背景主要是：烧结的反应过程是一个缓慢发展变化的过程，系统的运行具有一定的惯性和稳定性。据此，算法的主要思想是将数据集构造成空间，每一条记录代表空间中的一个点；其中，空间的维数由判别噪音所需要的数值属性集确定。计算每一条重复记录与其相邻的记录之间的欧氏距离，其中，距离最小的点为需要保留的点，其余的点则标记为噪音。

例如，在表 4-5 中的数据，主要根据设定电压、工作电压、电流和电阻来判断重复记录 2、3 中的噪音，故可以构造一个包含上述属性的四维空间。

由于篇幅的关系，下面仅给出算法的简单描述：

算法：Identify_Noiseee_Algorithm（INA）

输入：数据集 RECS 重复记录号{i_1，i_2，…，i_j}；2≤j≤RECS. count}

输出：保留的记录号 Storages

Begin

对数据集 RECS 进行适当的排序；

/*排序方式同 GIRA 算法相同*/

获得重复记录号对应的顺序号，记为 k_1，k_2，…，k_j；

Select case

Case k_1=l and k_j < RECS.count

将 k_1 送入 Storages

/*可随机选择一个，此处选第一个记录号*/

Case k_1=1 and k_j < RECS.count

计算所有重复记录与第 k_j+1 条记录之间的距离 d_1，d_2，…，d_j

其中，d_j=distance(RECS(k_j)，RECS(k_j+1))

获得 Min(d_1，d_2，…，d_j)对应的记录号 k_{Min} 并送入 Storages

Case k_1>1 and k_j =RECS. count

计算所有重复记录与第 k_1-1 条记录之间的距离 d_1，d_2，…，d_j

其中，d_j=distance(RECS(k_1)，RECS(k_1-1))

获得 Min(d_1，d_2，…，d_j)对应的记录号 k_{Min} 并送入 Storages

Case k_1>1 and k_j<RECS. count

计算所有重复记录与第 k_1-1 条记录之间的距离 d_{11}，d_{12}，…，d_{1j}

计算所有重复记录与第 k_j+1 条记录之间的距离 d_{21}，d_{22}，…，d_{2j}

其中，d_{1i}=distance(RECS(k_i),RECS(k_i-1))

D_{2i}=distance(RECS(k_i),RECS(k_j+1))

计算 $d_i=\frac{d_{1i}+d_{1i}}{2}$，$i$=1，2，…，$j$；

获得 Min(d_1，d_2，…，d_j)对应的记录号 k_{Min} 并送入 Storages

End Select

Return Storages

其中，过程 distance(A,B)计算记录 A 与 B 之间对应属性的欧氏距离。仍以表 4-5 中的数据为例，根据 INA 算法，有：

$$d(2,1)=\sqrt{(4.15-0)^2+(4.19-0)^2+(203.4-0)^2+(4.154-0)^2}$$

=203.528

d(2,4)=206.427

d(3,1)=1

d(3,4)=3.9

从而，Min((d(2,1)+d(2,4))/2,(d(3,1)+d(3,4))/2)=2.45，故记录 3 是保留数据，而记录 2 是噪音数据，可以清除。

为了避免属性非规范化的影响，通常可对参与欧氏距离计算的属性序列进行标准化，再进行距离计算。

4.4.3　数据抽取、转换与加载策略

在数据仓库建设中，数据抽取、转换与加载(Extract、Transformation、Loading，ETL)的处理过程，是系统实施成功与否的最关健和最困难的部分。在考虑 ETL 策略时，必需首先确定下列问题：

1)源数据库，是指数据仓库中数据的来源，即烧结控制系统中的事务数据，需要判断是本地数据还是远程数据？是同构或是异构？

2)目的数据库，即数据仓库的数据库，通常通过建模工具进行设置；

3)数据存储粒度，即粒度的选择与存储；

4)数据加载的频度，根据源数据库的变化，确定仓库中数据的更新周期；

5)数据访问的方法、权限，主要从安全性的角度考虑，通常根据角色和用户权限来确定，避免 ETL 过程中数据的泄密；

6)元数据管理，包括 ETL 过程的描述信息等。

数据抽取的方法比较多，主要有基于存储过程的数据抽取、基于编程工具和调用接口的数据抽取、基于 ODBC 的数据抽取、基于脚本的数据抽取以及基于商业工具的数据抽取等。无论采用哪种方法，其最基本的功能主要包括：数据格式的一致性；抽取数据的高质量；数据源数据的异构获取能力；抽取、转换和加载数据的自

动化，等等[21]。

作为一种实用的 ETL 策略，利用 Microsoft SQL Server 2005 Integration Services(SSIS)进行数据转换，其具体的实施步骤包括：设置数据源、设置数据目的地、设置转换细节、执行数据转换。

4.5 联机分析处理的应用研究

基于数据仓库的应用可分为三种类型：查询报表型、验证型、挖掘型。查询报表型应用并不是象生产管理系统中的查询那样仅仅对记录级数据进行查询和制作报表(决策支持系统中也存在这类查询，但是为数不多)，这里的查询是对分析结果的查询，并且要求采用各种图形和报表工具以便于决策分析用户更方便、更清晰地了解复杂的查询结果。验证型应用是指深入了解事务，作出结论性、总结性分析，验证用户的假设和问题，也可以验证数据挖掘得出的预测性结论，防止偏差，例如 OLAP 应用；挖掘型应用是指自动发现隐藏在数据间的模式，作预测性分析，例如数据挖掘。从图 4-4 可以看出，查询报表型应用容易成功，而数据挖掘应用由于在一定程度上取决于系统对算法的选择和用户对系统的理解以及熟练程度，所以成功率较低。这一节主要分析 OLAP 的应用[22]。

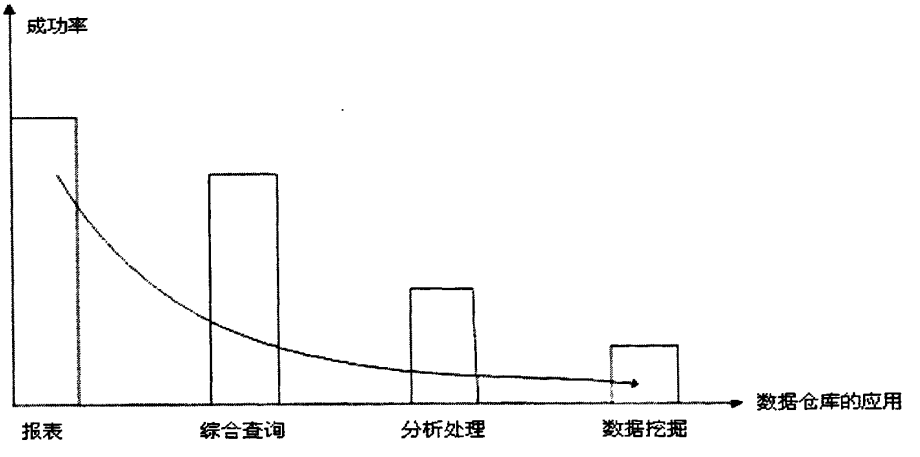

图 4-4 基于数据仓库的各种应用的成功率

Fig. 4-4 Based on the data warehouse for a variety of applications success rate

4.5.1 基于 Web 的 OLAP 体系结构

随着网络技术的发展，决策支持系统设计模式向着浏览器/服务器方向发展，OLAP 的发展趋势也向着网络化发展。如图 4-5 所示，基于 Web 的 OLAP 体系结构可以划分为表示层、逻辑层、数据层三个层次。

(1) 表示层 表示层即浏览器。在页面中，嵌入一个显示控件，它的功能是接受从服务器传来的数据，根据用户的要求组织成相应的形式，同时接受用户的输入，

分析用户的动作，生成新的多维查询要求，发送给后台的逻辑层进行处理。

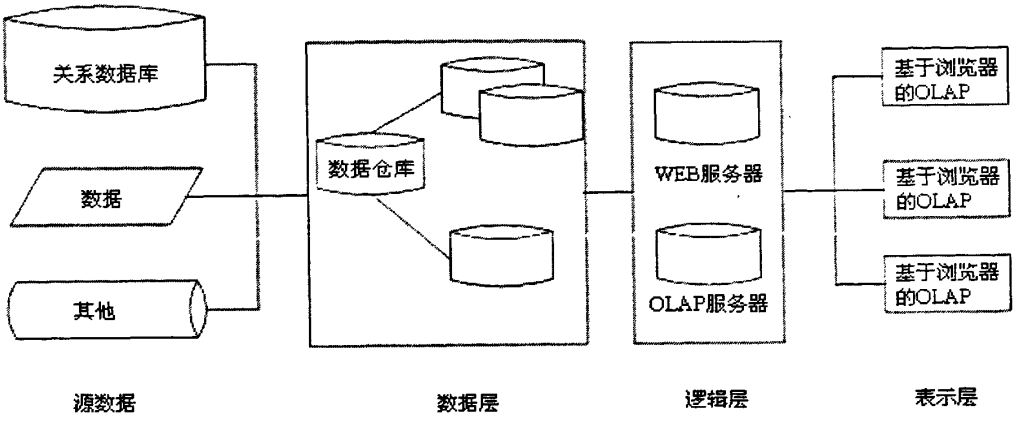

图 4-5　基于 Web 的 OLAP 体系结构

Fig. 4-5　The OLAP Architecture of Web-based

(2)　逻辑层　逻辑层是一个和 Web 服务器平行的应用服务器。它的功能是接受从客户端发送的多维查询要求，处理后对后台数据库进行操作。而后将处理结果返回给客户端。

(3)　数据层　数据层即后台数据库，它用于存储多维数据库中的数据。这种体系结构中，客户端通过浏览器对数据进行查询和处理。在浏览器和数据库服务器之间，除了通常的 Web 服务器以外，在增加一个应用服务器，存储整个系统的应用逻辑。

采用基于基于 Web 的 OLAP 体系结构开发系统有着如下的优点：

1) 对于用户而言，所要使用的仅仅是浏览器。对于绝大部分用户而言，学会使用浏览器比进行专门的培训从而掌握某种查询软件的使用是要简单的多。而且，用户可以不受地理的限制进行查询，同样，您的数据也可以提供给世界上任何一处的用户。

2) 系统的应用逻辑是存放在系统中的应用服务器上的，对系统的升级改造工作也就仅仅是在该服务器上了，这样企业的 IT 人员就可以集中精力处理整个系统的关键部分，而不必陷于琐碎的安装、修改工作中了。同样，通过浏览器工作，可以省去对每一台客户端进行配置的繁琐工作，提高效率。

3) 系统的关键部分是在服务器上，我们可以比较容易的对系统的使用人员进行控制，保证系统的安全性。

4.5.2　OLAP 应用的数据组织

根据 OLAP 服务器端的不同的数据组织方法，OLAP 分为三种结构：多维 OLAP(Multidimensional OLAP)、关系型 OLAP(Relational OLAP)以及混合型

OLAP(Hybrid OLAP)[23]。

多维 OLAP 利用一个专用的多维数据库来存储 OLAP 分析所需的数据，数据以多维方式存储，并以多维视图的方式显示。其结构如图 4-6。在 MOLAP 的结构中，OLTP 数据源中的数据经过提取、清洁、转换等提交到数据仓库服务器中。OLAP 服务器再把数据仓库服务器中的数据进行一系列的预处理，例如计算、合并，并且把结果按一定的层次结构组织成多维立方体(Cube)的形式，存入 OLAP 服务器的多维数据库中。用户通过客户端的应用软件界面递交分析需求给 OLAP 服务器，再由 OLAP 服务器检索多维数据库，将得到的结果返回给用户。

MOLAP 结构的主要优点是它能迅速地响应决策分析人员的分析请求并快速地将分析结果返回给用户，这得益于它独特的多维数据库结果以及存储在其中的预处理程度很高的数据；而且这种结构多维概念表达清晰，占用存储少。但是，在 MOLAP 结构中，OLAP 服务器主要是通过读取预处理的数据来完成分析操作，而这些预处理操作必须预先定义好，这就限制了 MOLAP 结构的灵活性，例如当数据或者计算频繁变化时，其重复计算量相当大，有时还需要重新构建多维数据库。

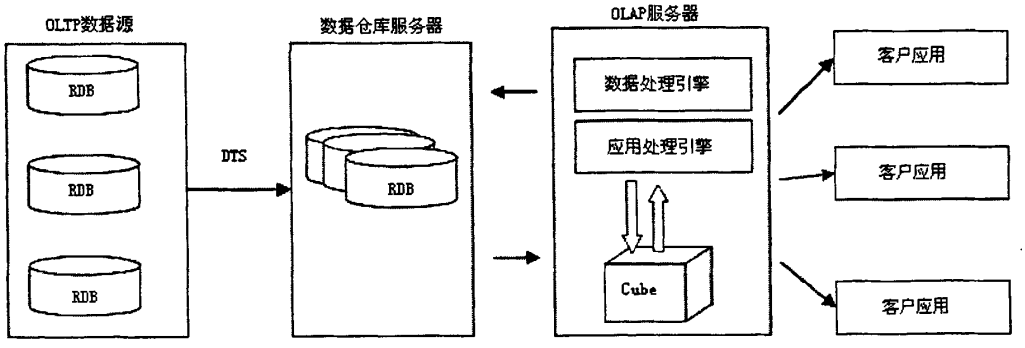

图 4-6 MOLAP 的结构

Fig. 4-6 Structure of MOLAP

关系 OLAP 的底层数据库是关系型数据库，由于数据仓库系统一般采用关系数据库，所以 ROLAP 的数据一般存储在数据仓库中。ROLAP 的结构如图 4-7 所示。在烧结生产决策支持系统中，就可以采用的这种结构。当数据仓库的数据模型确定之后，分散在企业中各 OLTP 数据库中的数据被加载到数据仓库中。OLAP 服务器对数据仓库中的数据进行预处理，并以关系型结构存回数据仓库中。当客户端用户提出 OLAP 需求时，ROLAP 服务器将这些多维分析要求转换成 SQL 语句，从数据仓库中提取数据，并将结果经多维处理转换成多维视图返回给用户。烧结生产决策支持系统中，这部分功能是利用 MS SQL Sever 的 OLAP Service 服务实现的。在图 4-7 中利用 OLAP Service 服务，组织了时间、车间、成品名称、成品等级等维度，以及销售量、销售额、销售利润等度量指标，根据二八法则运算聚合(aggregation)，并存

储到数据仓库中。客户进行 OLAP 应用时，可以利用 Pivot Table Service 服务进行多维分析，包括前面提到的切片、切块、上翻、下钻、旋转等动作。

采用这种类型的 OLAP 结构之后，数据仓库中既含有细节数据，又含有各种层次的概括数据，如果所需要的概括数据不存在，ROLAP 服务器会自动生成，比 MOLAP 增加了灵活性。ROLAP 的优点还在于关系型数据库有文件管理工具和开放式的 SQL 接口，便于前台应用的开发。由于数据仓库的数据采用的星型设计方案，减少了 RDBMS 处理时表连接的系统开销，使 ROLAP 体系结构更适用于概括数据和细节数据的动态访问。但是，ROLAP 也有缺点，它对用户的分析请求处理的时间比 MOLAP 长。

针对 ROLAP 和 MOLAP 各自的优缺点，又出现了 HOLAP 结构。迄今为止，对 HOLAP 还没有一个正式的定义。这种结构是 ROLAP 和 MOLAR 的有机组合，要求维数能够被动态更新，可以快速存取各种级别的汇总数据，可以方便地对计算和汇总算法进行维护和修改，适应大数据量的数据分析。目前 HOLAP 在技术上还不够成熟，还没有在企业中得到广泛的应用。

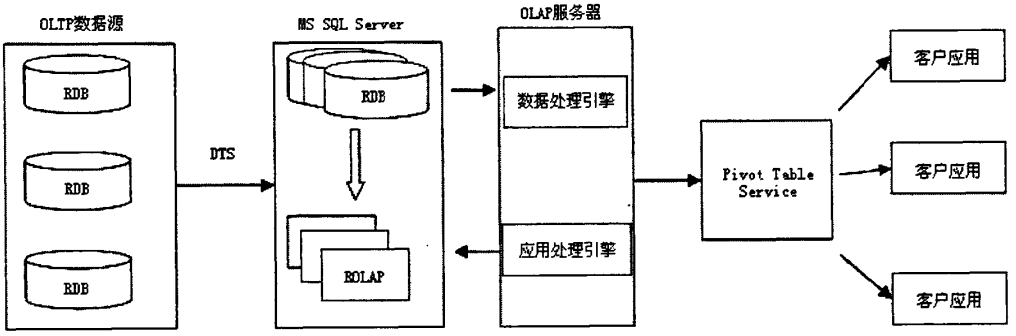

图 4-7　烧结生产决策支持系统的 ROLAP 结构

Fig. 4-7　The ROLAP structure of Sintering Production Decision Support System

因此，综合以上比较，考虑到数据访问的灵活性、可伸缩性以及技术的成熟性，烧结生产决策支持系统的 OLAP 部分采用了 ROLAP 结构。

4.5.3　多维数据分析

多维数据分析是指对以多维形式组织起来的数据采取切片、切块、旋转等各种分析动作，以求剖析数据，使最终用户能从多角度、多侧面地观察数据库中的数据，从而深入地了解包含在数据中的信息、内涵。多维分析方式迎合了人的思维模式，因此减少了混淆并且降低了出现错误解释的可能性[24]。

4.5.3.1　切片和切块

在多维数据结构中，按二维进行切片，按三维进行切块，可得到所需要的数据。

切片就是在某个或某些维上选定一个维成员，而在某两个维上取一定区间的维成员或全部维成员。如在“车间、产量、时间”三维立方体中进行切块和切片(如下图 4-8 所示)。

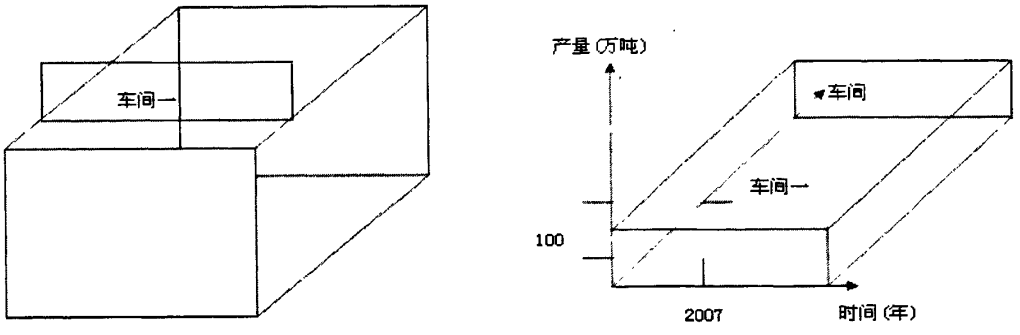

图 4-8 切片、切块示意图

Fig.4-8 Diagram of Slice and Dice

4.5.3.2 钻取

钻取包含向下钻取(Drill down)和向上钻取(Drill up)操作。从高级别数据到明细数据视图称为下钻；从明细级向上到综合级(或高级)来观察数据，称为上钻。数据库的设计以及数据的粒度级别将决定下钻或上钻的能力。以 2007 年烧结厂烧结矿产量情况为例(见表 4-6)

表 4-6 烧结矿产量情况

Tab. 4-6 The amount of mineral sintering conditions

车间	产量（万吨）
车间一	90
车间二	65
车间三	80

在时间维进行下钻操作，获得新表如下(见表 4-7)

表 4-7 烧结矿产量情况

Tab. 4-7 The amount of mineral sintering conditions

	2007 年			
车间	1 季度	2 季度	3 季度	4 季度
车间一	20	20	35	15
车间二	25	5	15	15
车间三	20	15	18	27

4.5.3.3　旋转

旋转即是改变一个报告或页面显示的维方向。例如，旋转可能包含了交换行和列：或是把某一个行维移到列维中去，或是把页面显示中的一个维和页面外的维进行交换(令其成为新的行或列中的一个)。如表 4-8 所示的例子。通过旋转可以得到不同视角的数据。如将表 4-8 旋转后得到表 4-9。

表 4-8　烧结矿产量情况（单位：万吨）

Tab. 4-8　The amount of mineral sintering conditions

	2006 年				2007 年			
车间	1 季度	2 季度	3 季度	4 季度	1 季度	2 季度	3 季度	4 季度
车间一	12	20	25	14	12	20	35	15
车间二	25	5	15	15	25	5	15	15
车间三	20	15	18	27	20	15	18	27

表 4-9　烧结矿产量情况（单位：万吨）

Tab. 4-9　The amount of mineral sintering conditions

	1 季度		2 季度		3 季度		4 季度	
车间	2006 年	2007 年	2006 年	2007 年	2006 年	2007 年	2006 年	2007 年
车间一	12	12	20	20	25	35	14	15
车间二	25	25	5	5	15	15	15	15
车间三	20	20	15	15	18	18	27	27

4.6　数据挖掘技术的研究

数据挖掘是一个交叉学科，它的主要技术由以下两种技术发展而来：传统的数理统计技术和各种智能算法。

4.6.1　数理统计技术研究

统计学方法：旨在从抽样分析中提取未知的数学模型，在数据挖掘中常会遇到大量的统计数据，通过模型分析来获得普遍运行的模式规律。它的模型主要包括回归分析(线性回归、多元回归、非线性回归等)、列连表、试验设计、探索性分析(主元分析法、相关分析法、因子分析、典型相关分析等)、判别分析(贝叶斯判别、费歇尔判别、非参数判别)、聚类分析(系统聚类、动态聚类等)、罗蒂斯蒂回归、时序分析等。

在烧结生产决策支持系统中，采用回归分析法来分析生产指标和工人人数等因素之间的因果关系。这里以分析生产量和工人人数之间的因果关系为例，介绍一元

回归分析法的使用[25]。

一元线性回归法是确定两个变量之间直线关系的一种方法。这种方法虽然简单，却具有相当的普遍性，因为任何曲线在一个小范围内，都可以近似地看作是直线。其原理是假定分析指标(产量、工人人数等)和有关的影响因素之间存在线性因果关系，并且用直线回归方程式(4-1)来代表其趋势，其中 Y 为因变量，表示分析指标的预测值；X 为自变量，是引起 Y 变化的某个因素；a，b 为回归系数。

$$y = a + bx \tag{4-1}$$

这种方法的算法是：根据已知的(x_i，y_i)(i=1，2.....，n)来确所选模型中的回归系数 a，b，再计算出不同 X 所对应的 Y 值。即求解方程式(4-2)。

$$\begin{cases} na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \tag{4-2}$$

其中 n 代表期数。根据最小二乘法，并且对 X 的取值进行简化，该方程的解是式(4-3)。

$$\begin{cases} a = \dfrac{\sum_{i=1}^{n} y_i}{n} \\ b = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x^2} \end{cases} \tag{4-3}$$

分析指标和时间的相关系数 r 可以根据式(4-4)计算：

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 * \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \tag{4-4}$$

相关系数的取值在-1 和+1 之间。如果相关系数越接近+1 或者-1，说明分析指标和时间因素之间的相关性越好，用回归直线法进行分析的结果越可靠。如果相关系数的绝对值小于一定值，说明不适宜用这种方法进行预测分析。出现后一种情况的原因有：第一，分析的这两个变量之间本来就不存在因果关系；第二，分析的这两个变量之间存在线性因果关系，但是还存在其他的起着更主要作用的变量没有被列入此分析模型，此时需要进一步用二元回归法等方法进行分析；第三，分析的这两个变量之间的因果关系是非线性关系，此时需要用非线性回归分析法进行分析。

这里可能会出现的极端情况有：r=+1 或者 r=−1 时，表示 x 和 y 完全线性相关，

完全可以用式(4-1)来描述和预测两者的关系；r=0 时，表示 x 和 y 为零相关，即两者之间没有线性关系，建立的回归方程没有实际意义，不能用于预测。

例：根据下列某烧结厂生产量与新增工人人数统计表预测未来烧结矿的生产量。如表 4-10，4-11，图 4-9 所示：

表 4-10　烧结矿的产量

Tab. 4-10　Sinter production

	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
产量（万吨）	28	31	50	53	61	70	60	66	63	65
新增工人人数	25	28	34	38	47	62	45	56	54	55

表 4-11　相关系数计算表

Tab. 4-11　The correlation coefficient calculated table

	y_i	x_i	$x_i y_i$	x_i^2	y_i^2
1989 年	28	25	700	625	784
1990 年	31	28	868	784	961
1991 年	50	34	1700	1156	2500
1992 年	53	38	2014	1444	2809
1993 年	61	47	2867	2209	3721
1994 年	70	62	4340	3844	4900
1995 年	60	45	2700	2025	3600
1996 年	66	56	3696	3136	4356
1997 年	63	54	3402	2916	3969
1998 年	65	55	3575	3025	4225
Σ	547	444	25862	21164	31825

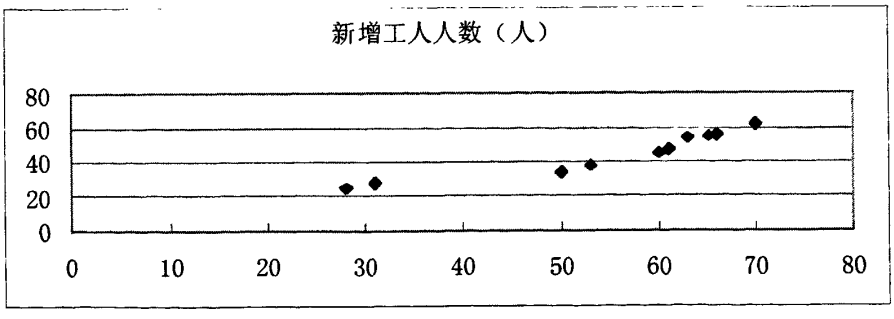

图 4-9　新增工人散点图

Fig.4-9　New workers scatter

根据回归方程进行预测，假定1999年的新增工人人数为57人，2000年为59人，则：

1999年的预测值为：5.89+1.1×57=68.6万吨

2000年的预测值为：5.89+1.1×59=70.79万吨

4.6.2 智能算法研究

智能算法也可称为机器学习方法：让计算机从经验中学习从而获得知识，它包括人工神经网络(ANN)、决策树方法、遗传算法(GA)、逻辑归纳算法、支持向量机(SVM)等。

(1) 人工神经网络(ANN) 人工神经利用大量的简单计算单元(即神经元)，连成网络来实现大规模并行计算，通过学习改变神经元之间的连接强度。可细分为：前向神经网络(BP算法等)、自组织神经网络（自组织特征映射、竞争学习等）等。神经网络近来越来越受到人们的关注，因为它为解决大复杂度问题提供了一种相对来说比较有效的简单方法。神经网络可以很容易的解决具有上百个参数的问题(当然实际生物体中存在的神经网络要比我们这里所说的程序模拟的神经网络要复杂的多)。神经网络常用于两类问题：分类和回归[26]。

在结构上，可以把一个神经网络划分为输入层、输出层和隐含层。输入层的每个节点对应一个个的预测变量。输出层的节点对应目标变量，可有多个。在输入层和输出层之间是隐含层(对神经网络使用者来说不可见)，隐含层的层数和每层节点的个数决定了神经网络的复杂度。

在使用神经网络时有几点需要注意：

1)神经网络很难解释，目前还没有能对神经网络做出显而易见解释的方法学。

2)神经网络会学习过度，在训练神经网络时一定要恰当的使用一些能严格衡量神经网络的方法，如前面提到的测试集方法和交叉验证法等。这主要是由于神经网络太灵活、可变参数太多，如果给予足够的时间，神经网络几乎可以“记住”任何事情。

3)训练时间比较长。除非问题非常简单，训练一个神经网络可能需要相当可观的时间才能完成。当然，一旦神经网络建立好了，在用它做预测时运行速度还是很快的。

4)建立神经网络需要做的数据准备工作量很大。一个很有误导性的神话就是不管用什么数据神经网络都能很好的工作并做出准确的预测。这是不确切的，要想得到准确度高的模型必须认真的进行数据清洗、整理、转换、选择等工作，对任何数据挖掘技术都是这样，神经网络尤其注重这一点。比如神经网络要求所有的输入变量都必须是0-1(或-1～+1)之间的实数，因此像“地区”之类文本数据必须先做必要的处理之后才能用作神经网络的输入。

(2)　支持向量机(SVM)　SVM 是在统计学习理论基础上发展起来的一种新的机器学习方法。统计学习理论对有限样本情况下模式识别中的一些根本性问题进行了系统的理论研究，很大程度上解决了模型选择与过学习问题、非线性和维数灾难问题、局部极小点等问题，因此成为目前研究的热点。支持向量机的研究范围和成果不断扩大。近年来，对 SVM 的研究主要集中在对 SVM 本身性质的研究和完善以及加大 SVM 应用研究的深度和广度两方面。到目前为止，支持向量机已成功应用于模式分类、回归分析、函数估计等领域。在孤立手写字符识别、网页或文本自动分类、说话人识别、人脸检测、计算机入侵检测、基因分类、遥感图象分析、目标识别、函数回归、估计、函数逼近、密度估计、时间序列预测及数据压缩、文本过滤、数据挖掘、非线性系统控制等问题中，都有支持向量机的成功应用[27]。

在烧结生产决策支持系统中，采用时间序列预测法来分析和预测产量随时间的变化情况。时间序列预测法是借助统计方法来分析预测目标的时间序列的发展变化过程的规律性，建立数学模型，据此推测预测目标的发展趋势与水平的一种定量预测方法。下面以预测产量为例，介绍本系统使用的算法[28]。

我们认为影响产量变化的因素有四类：长期趋势因子(T)、季节变动因子(S)、周期波动因子(C)和不规则变动因子(I)。长期趋势因子是时间序列变量(产量)在较长的持续时间内的某种发展总趋势，是分析预测的重点。季节变动因子是由于季节更换的固定规律作用而发生的周期性变动。周期波动因子也称循环变动因子，是呈涨落相间的波浪式的起伏变动，属于环境影响。不规则变动因子也称随机变动因子，是偶发事件导致时间序列中出现数值忽高忽低的无规则可循的变动，从长期观察，可以期望它们相互抵消。因此，一般情况下主要是考虑长期趋势因子和季节变动因子，利用乘法模式进行分析和预测：

$$Y=T*S \tag{4-5}$$

本系统在计算式(4-5)中 T 的时候，采用的以下算法：如果时间序列观察值变化幅度不大，即随机因素的干扰比较小，那么用移动平均法计算 T；否则，用二次曲线趋势预测法计算 T。具体算法是：

1)移动平均法

它是在算术平均法的基础上发展起来的一种预测法。当实际数据含有不明显的趋势变化、周期变化和随机变化时，可以用移动平均法消除或者减少这些变动因素的影响，分析时间序列的趋势，进行预测。所以移动平均法适用于分析短期内略有波动的产量等指标。其原理是假定预测事物的未来状况只与最近若干期状况有关，而与较远期状况无关，根据时间序列，逐项移动、依次计算包括一定项数的序列平均数，形成一个序列平均数的时间序列。

移动平均法的算法是每次取一定数量周期的数据平均，按照周期次序逐次推进，

每推进一个周期，舍去前面的一个数据，增加下一个周期的数据，再进行平均。依次不断向前推进，以这种平均值作为下一个周期的预测值。基本公式为：

$$Y_t = \frac{y_t + y_{t-1} + y_{t-2} + \cdots + y_{t-n+1}}{N} \tag{4-6}$$

其中：Y_t 为 t 期的移动平均数，作为 t+1 期的趋势值；N 为每次移动平均包含的数据个数；Y_t，Y_{t-1}，……Y_{t-n} 为时间序列观察值。

将公式(4-5)略作变化，可推出一个简便的计算公式：

$$\begin{aligned} Y_t &= \frac{y_t + y_{t-1} + y_{t-2} + \cdots + y_{t-n+1}}{N} \\ &= \frac{y_t + y_{t-1} + y_{t-2} + \cdots + y_{t-n+1} + y_{t-n} - y_{t-n}}{N} \\ &= \frac{y_{t-1} + y_{t-2} + \cdots + y_{t-n+1} + y_{t-n}}{N} + \frac{y_t - y_{t-n}}{N} \\ &= Y_{t-1} + \frac{y_t - y_{t-n}}{N} \end{aligned} \tag{4-7}$$

公式(4-7)实际上是一个递推公式，只要计算 $y_t - y_{t-n}/N$，就可以很方便地得到新的移动平均值，即新的预测值。新的移动平均值是对以前的移动平均值所进行的调整，因此大大简化计算。例烧结厂 1 月到 12 月的产量，如表 4-12、图 4-10 所示：

表 4-12　移动平均预测结果

Tab. 4-12　Results of Moving average forecast

月份	产量	三期移动平均值（T=3）	六期移动平均值（T=6）
1	650		
2	678		
3	720		
4	785	682.67	
5	859	727.67	
6	920	788.00	
7	850	854.67	768.67
8	758	876.33	802.00
9	892	842.67	815.33
10	920	833.33	844.00
11	789	856.67	866.50
12	844	867.00	854.83

2）二次曲线趋势预测法

二次曲线趋势预测法是把分析指标和时间之间的关系曲线看作抛物线形状，两者的数值符合式(4-4)，其中 Y 为因变量，表示分析指标的预测值；t 为自变量，在此表示时间因素：a，b，c 为待定系数。

$$y = a + bt + ct^2 \tag{4-8}$$

在实际的系统中，一般认为 t 代表等间距的时间序列在程序设计时，可以令：$\sum_{i=1}^{n} t_i = 0$，并且对代码进行处理：当 n 为奇数时，取间隔期为 1，将 t=0 置于预测基础期的中间；当 n 为偶数时，取间隔期为 2，将 t=-1，t=1 分别置于预测基础期中间的上下两期。对 t 的取值简化之后，可以求得式(4-9)的解：

$$\begin{cases} na + b\sum_{i=1}^{n} t_i + c\sum_{i=1}^{n} t^2 = \sum_{i=1}^{n} y_i \\ a\sum_{i}^{n} t_i + b\sum_{i=1}^{n} t_i^2 + c\sum_{i=1}^{n} t_i^3 = \sum_{i=1}^{n} t_i y_i \\ a\sum_{i=1}^{n} t_i^2 + b\sum_{i=1}^{n} t_i^3 + c\sum_{i=1}^{n} t_i^3 = \sum_{i=1}^{n} t_i^2 y_i \end{cases} \tag{4-9}$$

$$\begin{cases} a = \dfrac{\sum_{i=1}^{n} t_i^4 \sum_{i=1}^{n} t_i - \sum_{i=1}^{n} t_i^2 \sum_{i=1}^{n} t_i^2 y_i}{n\sum_{i=1}^{n} t_i^4 - (\sum_{i=1}^{n} t_i^2)^2} \\ b = \dfrac{\sum_{i=1}^{n} t_i y_i}{\sum_{i=1}^{n} t_i^2} \\ c = \dfrac{n\sum_{i=1}^{n} t_i^2 y_t - \sum_{i=1}^{n} t_i^2 \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} t_i^4 - \left(\sum_{i=1}^{n} t_i^2\right)^2} \end{cases} \tag{4-10}$$

这样就可以将 a，b，c 代入式(4-8)进行分析预测。这种方法认为分析指标(产量)是受多因素影响的，趋势变动线是曲线。

例：根据下表中烧结厂历年产量，预测 2005 和 2006 年的烧结矿产量。如表 4-13，图 4-11 所示。

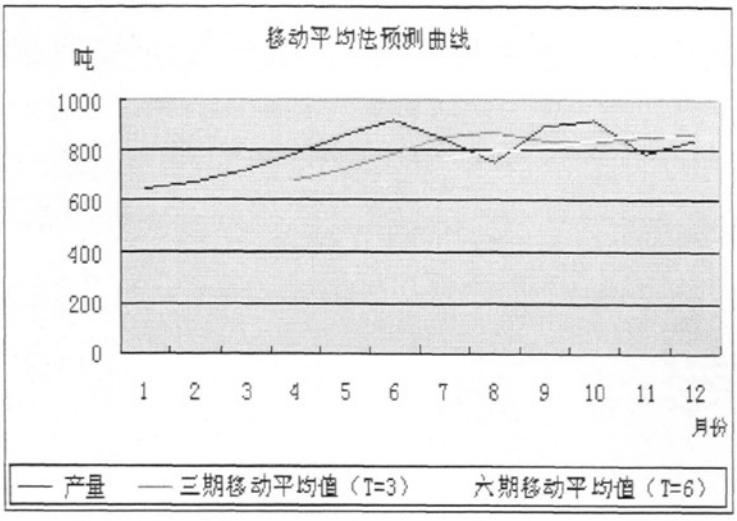

图 4-10　移动平均法预测曲线

Fig.4-10　Curve of Moving-average forecast

表 4-13　烧结厂历年烧结矿产量

Tab. 4-13　Sales of a calendar year of Supermarket

	Y（万元）	t	t^2	t^4	tY	t^2Y	$\hat{Y}_t = a + bt + ct^2$
1998	350	-3	9	81	-1050	3150	334.52
1999	300	-2	4	16	-600	1200	303.57
2000	250	-1	1	1	-250	250	300.00
2001	350	0	0	0	0	0	320.81
2002	400	1	1	1	400	40	375.00
2003	450	2	4	16	900	1800	453.57
2004	550	3	9	81	1650	4950	559.52
n=7	ΣY=2650		Σt^2=28	Σt^4=196	ΣtY=1050	Σt^2Y=11750	

应运用二次曲线进行预测。求得趋势曲线：Y_t=323.81+37.5t=13.69t^2，将 2005 年和 20 年的时间序列变量值 t 和 t^2 代入，求出：

$\hat{Y}$ 2005=692.85　　$\hat{Y}$ 2006=853.56

本系统在计算式(4-5)中 S 的时候，采用的是直接平均法测定季节指数。具体算法是：计算各年同月或者同季度观察值的平均数(用 A 表示)；再计算历年所有月份或者季度的总平均值(用 B 表示)；然后计算各月或者各季度的季节指数，即 S=A/B。

用以上方法计算出 T 和 S 值之后，代入式(4-5)，就可以计算出预测值。

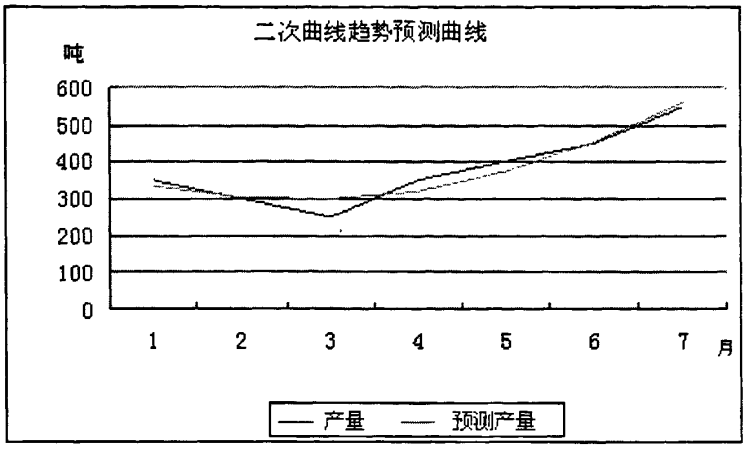

图 4-11　二次曲线趋势预测曲线

Fig. 4-11　The Curve of Quadratic trend forecast

4.7　本章小结

本章主要介绍了烧结生产信息分析系统架构的研究。重点介绍数据仓库的设计过程，数据预处理的方法和策略、联机分析处理的应用和数据挖掘算法等设计工作，为系统的实现提供数据的依据，使数据分析的实现成为可能。

第 5 章　烧结生产信息数据分析子系统的实现

5.1　系统的开发环境

5.1.1　SQL Server 2005 技术

SQL Server 2005 是一个全面的数据库平台，使用集成的商业智能(BI)工具提供了企业级的数据管理。SQL Server 2005 基于 SQL Server 2000 的强大功能之上，提供了一个完整的数据管理和分析解决方案，它给不同规模的组织带来帮助。SQL Server 2005 数据库引擎为关系型数据和结构化数据提供了更安全可靠的存储功能，可以构建和管理用于业务的高可用和高性能的数据应用程序。

SQL Server 2005 结合了分析、报表、集成和通知功能。可以构建和部署经济有效的 BI 解决方案，并通过记分卡、Dashboard，Web services 和移动设备将数据应用推向业务的各个领域。

SQL Server 2005 具备许多优势，SQL Server 2005 的数据挖掘与 SQL Server、SQL Server Integration Services、Analysis Services 都进行了很好的集成。SQL Server 2005 具有易用性，可伸缩性和可扩展性等特点，同时它包含简单而丰富的 API。

与 Microsoft Visual Studio，Microsoft Office System 以及新的开发工具包(包括 Business Intelligence Development Studio)的紧密集成使 SQL Server 2005 与众不同。无论是开发人员、数据库管理员、信息工作者还是决策者，SQL Server 2005 都可以提供创新的解决方案，以从数据中更多地获益。

5.1.2　Analysis Services 和 Reporting Services

Microsoft SQL Server 2005 Analysis Services(SSAS)为商业智能应用程序提供联机分析处理(OLAP)和数据挖掘功能。Analysis Services 允许您设计、创建和管理包含从其他数据源(如关系数据库)聚合的数据的多维结构，以实现对 OLAP 的支持。对于数据挖掘应用程序，Analysis Services 允许您设计、创建和可视化处理那些通过使用各种行业标准数据挖掘算法，并根据其他数据源构造出来的数据挖掘模型。

Microsoft SQL Server 2005 Reporting Services(SSRS)提供企业级的支持 Web 的报表功能，从而使您可以创建从多个数据源提取内容的报表，以各种形式发布报表，并可以集中管理安全性和订阅。

5.1.3　Visual Studio 2005 工具

Visual Studio 2005 和.NET Framework 2.0 在应用程序开发的所有方面取得了大幅的进展。首先，Visual Studio 2005 根据开发人员个人的需要调整软件开发体验，

设置新的开发人员工作效率标准。这一“个性化工作效率”将在开发环境和.NET Framework 类库中提供相应的功能，以帮助开发人员在最少的时间内克服其最为紧迫的困难。其次，Visual Studio 2005 使开发人员能够通过与 Microsoft Office System 和 SQL Server 2005 的更好集成，在更广泛的应用程序开发方案中应用现有的技能。最后，Visual Studio 2005 提供一组新的工具和功能，以满足目前大规模企业的应用程序开发需要。

5.2　系统的设计原则

在企业生产过程中，车间生产计划和生产进度控制是相互交织进行的。在生产系统中，上工序的工作就是为下工序做生产准备，如果上一道工序的进度没有控制好，就可能造成下道工序和所有的后续工序出现停工待料的局面，生产系统运行的秩序将被打乱，出现停产或半停产情况。所以，每一个工序的进度控制都必须做到准确无误。但是，由于受到外部(外购品交货是否准时，工装设备的保证是否完好)、内部(次品的出现、设备故障的发生等)因素的影响，每道工序都能按照计划时间100%准时地完成也是不可能的，所以适当的半成品存货是解决这类问题的主要措施。在生产准备中半成品的存货多少为合适，要根据具体的生产条件(如产品的复杂程度、工序的复杂程度等)来确定。如果半成品的存货太多，会给公司的资金运营、物流系统造成很大压力；存货太少，又无法保证生产系统正常运行的要求[30]。

在传统生产管理系统中，特别是一些多品种、大批量的生产系统中，要处理好上述问题，需要大量人工来处理很多的信息和数据，而且仍然可能出现疏漏或错误。在现代生产管理体系中，通过使用先进的生产管理软件，就可以很好地处理以上问题。由此，对于企业生产信息数据分析系统的基本要求包含下面几方面：

5.2.1　系统功能要求

1) 系统功能的易用性

企业生产现场管理，事务繁多，数据复杂，因此，友好的交互界面和各种功能的易用性具有更大的现实意义。

2) 系统功能的灵活性

灵活处理并满足中高层领导的决策需要，功能模块之间的互相兼容和调用，信息查询、数据修改、决策传达、问题反馈、参数设置等方面都要求能灵活运行。

3) 系统功能的可扩充性

系统首先根据当前企业生产管理层提出的经验需求以及从科学生产管理的学术理论方面的需求，实现对生产管理辅助决策的基本功能，同时企业生产管理层还可能随时遇到新的管理问题，并希望得到系统的决策辅助，因此系统功能的可扩充性对处理临时性问题具有很大的意义。

5.2.2 系统性能要求

为了真正对生产车间管理层起到提供实时信息支持，进而辅助决策的作用，企业生产信息数据分析系统需要具备以下的基本要素：

1)系统设计的前瞻性：对应系统功能的可扩展性，随着企业信息化的日益深入，管理层对系统的认识也会不断地发展，在彼此的交互过程中新的功能需求会不断地涌现，这要求系统的设计应具备良好的扩展性，对后台数据仓库、联机事务处理、数据挖掘、查询功能、数据接口等部分的设计应该考虑到管理中可能发生的变化，以适应现代制造企业管理层的长期应用。

2)运行环境的网络化：企业生产信息数据分析系统必须在网络环境下运行，采用 B/S(浏览器/服务器)模式，使得车间管理层领导能够很方便地在任何网络可及的地方进行数据分析和生产监控，同时也有利于系统的实施和维护。

3)系统运行的高效性：系统需采用先进的技术和处理算法流程，保证系统在大数据量的情况下高效运行。

4)系统数据的准确性：企业生产信息数据分析系统是应用在车间主任、分厂厂长办公室、各个业务管理办公室及各工段生产控制处的一个对决策起重要作用的系统，因此系统功能模块处理所得的数据必须保证非常高的准确度，从而有效起到辅助决策的作用。

5.3 系统的功能实现

烧结生产信息数据分析软件的统一平台采用整体考虑、分模块设计的方法，覆盖进厂、配料、混合、烧结、成品、外发等多个生产区域和环节，体现流程性特点，实现流程跟踪、生产过程监测，对生产组织各环节数据需求进行分析、归纳、总结，确定数据模块，分析数据来源、去向。

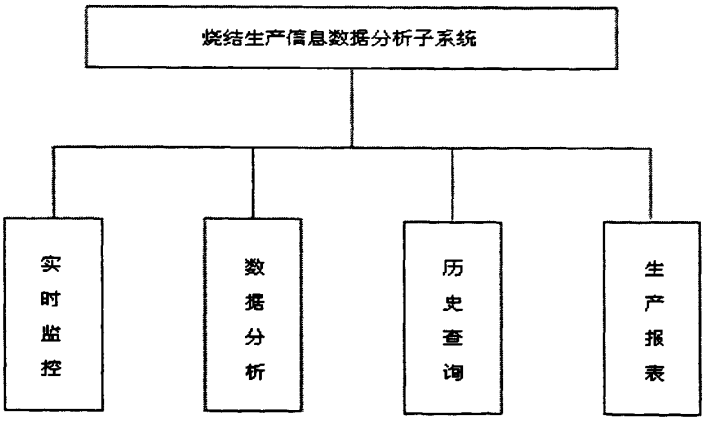

图 5-1 烧结生产信息数据分析功能模块图

Fig.5-1 The function map of Sinter production analysis system

该系统采用数据集中存储、分布处理的方式，数据以 SQL SERVER 的形式存储在机房工控服务器上，所有生产数据设计成一个数据库系统，数据表分系统、分类型单独设计。开发工具运用微软公司基于 WEB 的 ASP. NET 系统。主要包括实时监控、数据分析、历史查询、生产报表等四个大的模块，如图 5-1 所示。

5.3.1　实时监控模块

该模块主要实现以下几个功能，如图 5-2 所示：

1) 生产调度集中监测：以全厂区域划分，按照流程设计，把烧结厂调度关注的设备状态、仓存、生产数据，以图形化的方式直观反映。

2) 配料区域集中监测：按照流程设计，把原料集控室关注的烧结厂各流程环节的设备状态、仓存、生产数据，以图形化的方式直观反映。

3) 烧结区域集中监测：集中反映烧结集控室关注的各流程环节的设备状态、仓存、生产数据。

4) 成品区域集中监测：集中反映成品集控室关注的各流程环节的设备状态、仓存、生产数据。

5) 分系统的实时监测：包括各个烧结机生产量，烧结机参数，配料和成品、原燃料的各种仓存，皮带秤数据等数据。

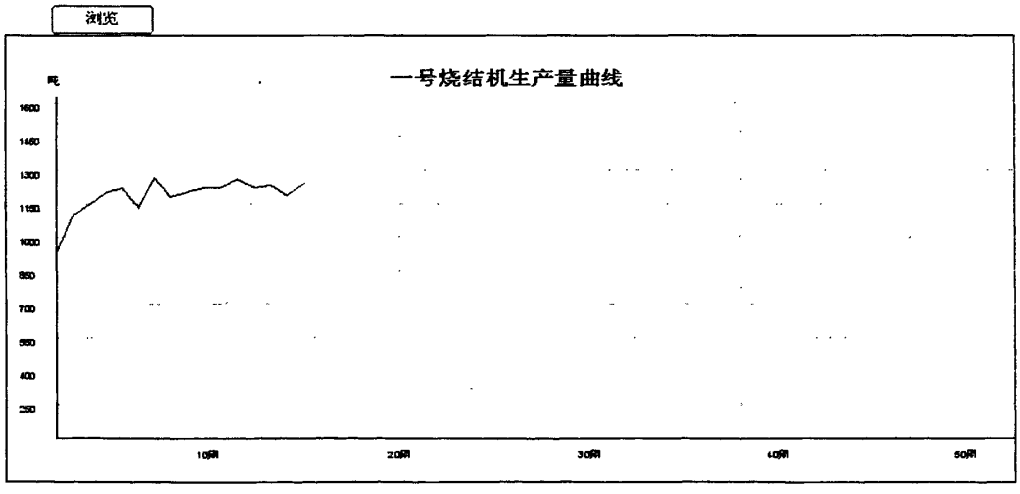

图 5-2　烧结机生产量实时数据监控

Fig.5-2　Sintering machine output real-time data monitoring

5.3.2　数据分析模块

该模块主要实现以下几个功能，如图 5-3 所示：

1) 曲线(柱形图)对比分析功能：根据专业要求，对重要数据进行对比分析，分析趋势，寻找规律，为指导生产提供依据，积累经验。比如烧结机生产能力分析功

能，可以提供考核依据。

2）数据分析功能：通过计算，得到相应数据分析结果，提供管理的依据，比如预测产量，评估生产质量等。

预测 一号烧结机 下 6 个星期的生产量 预 测

星期	生产量（吨）
14	176.27
15	195.79
16	273.17
17	131.01
18	113.34
19	51.85

图 5-3(a) 预测产量

Fig.5-3(a) Output prediction

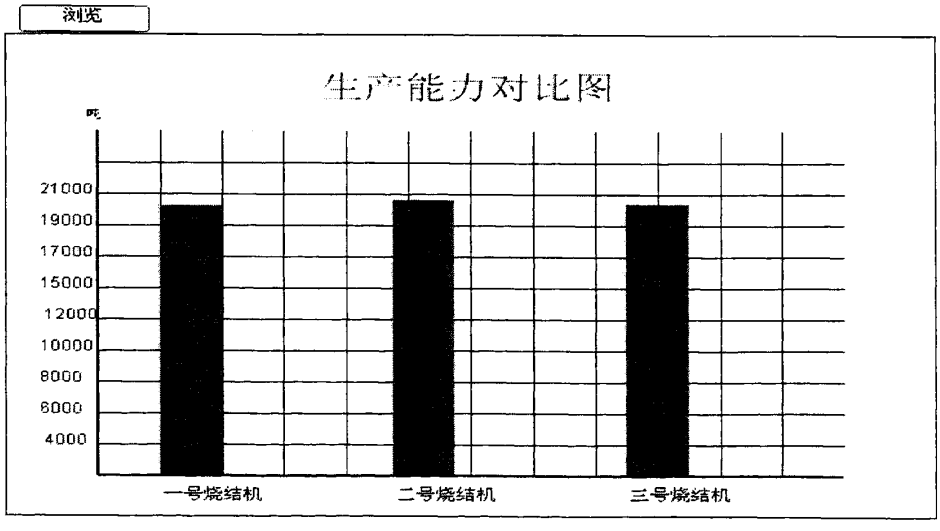

图 5-3(b) 生产能力对比柱形图

Fig.5-3(b) The Column chart of Capacity comparison

5.3.3 历史查询模块

该模块主要根据生产、设备、物资、质量等专业需求，进行管理层数据查询。还包括烧结机、混合机、配料仓存、成品仓存、原燃料各种仓存、皮带秤计量等数据的历史查询，可以实现按日期、分类等综合查询和累计计算，如图 5-6 所示。

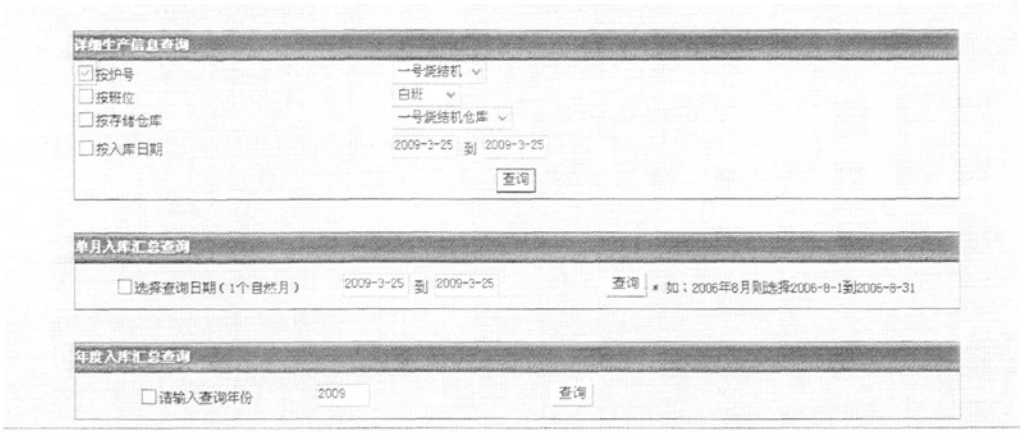

图 5-4　生产信息查询图

Fig.5-4　The map of production information inquiries

5.3.4　生产报表模块

该模块主要实现的功能：主要包括厂调度报表、原料集控报表、烧结主控室报表、成品主控室报表，解决手工记录、电话报数的问题，对网上已有的数据直接利用，对需要手工录入的数据，由数据来源部门直接录入，减少数据传递层次，保证数据的准确性，自动形成生产报表，并保存历史报表，实现查询功能，如图 5-7 所示。

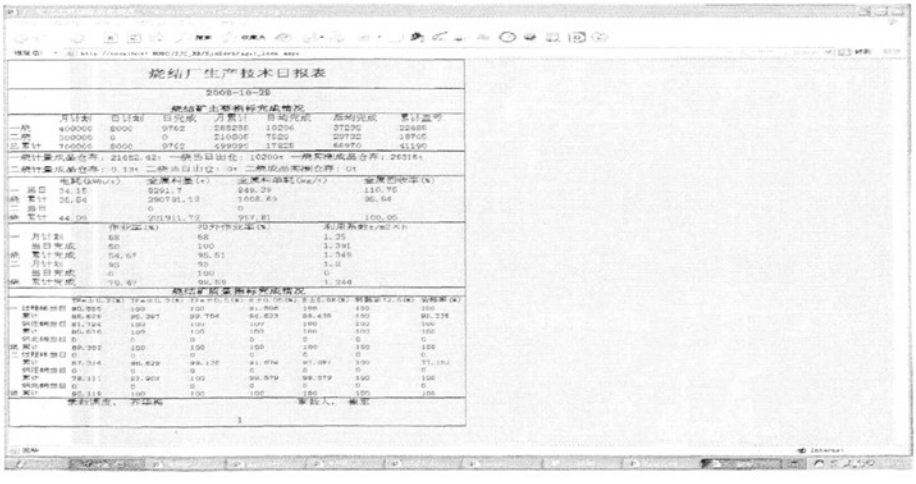

图 5-5　生产技术报表

Fig.5-5　The statements of Production technology

5.4　本章小结

本章将主要介绍系统主要功能模块的实现，通过对烧结厂生产控制过程数据建立统一分析研究的软件平台，可为各级专业从整体上分析烧结生产过程提供强有力

的工具，进而提升专业管理水平。实现了烧结信息化建设水平整体提升，解决“信息孤岛”、生产数据分析不足等的问题。

结 论

本文的所有研究工作主要是集中在如何在烧结生产控制过程中实现智能决策支持，论文系统分析了烧结生产控制过程中决策支持系统的构建问题及相关技术，提出了以.NET为架构，以数据仓库，数据挖掘、联机分析处理为典型应用的烧结生产控制过程信息系统的设计方案。并以某烧结厂信息化建设为背景，设计并实现了烧结生产信息数据分析子系统，证实烧结生产控制过程决策支持系统的设计思想是可行的。本文的研究工作主要体现在以下几个方面：

1)针对烧结厂“信息孤岛”、数据分析能力不足，探讨了建设烧结生产控制决策支持系统的必要性，并进行了系统需求分析的研究。

2)给出了烧结生产智能决策支持系统的架构，并对其数据仓库进行了设计，针对其数据来源不同，对其数据预处理策略进行了研究。

3)基于数据仓库分别对其进行了联机分析处理和数据挖掘的应用研究，可以进行深层次的数据分析。

4)选择了系统开发环境，并根据设计原则，开发了烧结生产信息数据分析子系统。实现了对生产过程的实时监控，并且可以对生产能力等进行对比、预测等功能。

由于时间和其他主客观条件的限制，本文建立的烧结生产信息数据分析子系统，还存在着不少需要完善的地方：

(1) 对生产数据挖掘的深入研究 烧结生产数据的智能分析只是从预测和序列发现这两个方面进行了知识挖掘，在今后的工作，要建立多种挖掘模型，从不同的角度对烧结生产数据进行分析，综合考虑挖掘结果，择优选择，从而得到有用的知识，为烧结生产提供决策支持。

(2) 提高系统交互性 强化用户与系统的交互能力，让用户根据经验来选择挖掘模型。对某种挖掘模型的参数能进行调整，便于实现用户的个性化查询。

(3) 提高系统的友好人机交互界面 系统的开发时间仓促，在计算机技术方面的知识不够丰富，从而使得系统在程序开发工具利用、界面的美观程度等方面存在不足。

由于本人能力所限和时间紧迫，对于本系统的研究和实现可能还存在不足之处，请各位老师予以批评指正。

参考文献

[1] 冯仁德, 孙在国. 现代企业. 成都：西南财经大学版社, 2003: 40-85

[2] 陈文伟, 决策支持系统及其开发.第 2 版. 北京: 清华大学出版社, 2000: 252-356

[3] 何晓义. 包钢炼铁厂管理信息系统_烧结部分_开发与应用. [内蒙古大学硕士论文]. 2006

[4] 贾艳. 铁矿粉烧结生产. 北京: 冶金工业出版社， 2006

[5] 范晓慧.王海东, 烧结过程数学模型与人工智能. 长沙: 中南工业大学出版社, 2002

[6] 陈文伟, 决策支持系统及其开发.第 2 版. 北京: 清华大学出版社, 2000: 252-356

[7] 孟波. 计算机决策支持系统. 武汉: 武汉大学出版社, 2001: 31-44

[8] 林宏谕. 决策分析——OLAP 建置与应用. 北京: 中国铁道出版社, 2001

[9] 林杰斌, 刘明德, 陈湘. 数据挖掘与 OLAP 理论与实务. 北京: 清华大学出版社, 2003: 156-160

[10] 高洪深. 决策支持系统(DSS)理论方法案例.第 3 版. 北京: 清华大学出版社, 2005: 178-180

[11] 王珊. 数据仓库技术与联机分析处理.科学出版社, 1998: 178-182

[12] 陈京民. 数据仓库原理、设计与应用. 北京: 中国水利水电出版社, 2004: 50-80

[13] 邵玉样. 连锁销售决策支持系统的研究与开发. [武汉理工大学硕士论文]. 2003: 45-80

[14] Harry sign. Data Warehousing Concept: Technology Implementation managing. MIT Press, 2001: 163-164

[15] 成明. 数据挖掘技术在运营分析与决策支持系统中的应用. [吉林大学硕士论文]. 2006

[16] H.INMON.SULDING The data warehouse third edition.John Wiley & sons, 2003

[17] 李江峰. 基于联机分析和数据挖掘的决策支持系统的研究与应用. [浙江工业大学硕士论文]. 2003: 50-55

[18] 钟金柱. 数据挖掘在库存决策支持系统中的应用研究. [西安理工大学硕士论文]. 2004

[19] U.Fayyad, Shapiro.Piatetsky, Smyth.Uthurusamy. Advances in Knowledge Discovery and Data Mining.MIT Press, 1996

[20] Efrem, G.Mallach, 决策支持与数据仓库系统. 李昭勇译. 北京: 电子工业出版社, 2003: 156-160

[21] 沈兆阳. SQL Seever200O OLAP 解决方案——数据仓库与 Analysis Services. 北京: 清华大学出版社, 2001: 80-100

[22] 刘代飞. 烧结过程工艺参数优化模型的研究. [中南大学硕士论文]. 2004

[23] 管襄华. 烧结过程控制技术研究. [山东科技大大学硕士论文]. 2002

[24] Efrem, G.Mallach, 决策支持与数据仓库系统. 李昭勇译. 北京: 电子工业出版社， 2003: 156-160

[25] H.INMON. DATA STORES, Data Warehousing and the Zach man Framework.1sted. 北京: 世界图书出版公司北京公司, 1999

[26] JOYCE BISCHOFF, TED ALEXANDER. 数据仓库技术. 成栋等译. 北京: 中国电子出版社, 1998

[27] H.Gill, P.Rao. 数据仓库——客户/服务器计算指南. 王仲谋, 刘书舟译. 北京: 清华大学出版社, 1997

[28] TOM HAMMERGREN. 数据仓库技术. 曹增强, 王备战译. 北京: 中国水利水电出版社, 1998

[29] 樊重俊, 韩崇昭. 企业经营计划决策支持方法与决策支持系统研究. 系统工程理论与实践, 1998, 1: 60-64

[30] 鲁敏. 大型决策支持系统的数据仓库设计与实施中关键技术. 计算机工程与应用, 1999, 9: 94-97

[31] MICHAEL COREY 著, SQL Server7 数据仓库. 希望图书创作室译. 北京: 北京希望电子出版社, 2000

[32] ERIK THOMSEN, GEORGE SPOFFORD. MicroSoft OLAP 解决方案. 北京: 人民邮电出版社, 2000

[33] 卢玉民. 计算机管理决策支持系统. 北京: 中国铁道出版社, 1993

[34] 张根保. 企业信息化. 北京: 机械工业出版社, 1999

[35] 吕俊亚. 微型计算机在企业管理中的应用. 成都: 成都科技大学出版社, 1995

[36] 马芸生, 杜俊俐. 决策支持系统与智能决策支持系统. 北京: 中国纺织出版社, 1995

[37] 陈景艳. 决策支持系统. 成都: 西南交通大学出版社, 1995

[38] A.BERSON , S.J.SMITH. Data Warehouse, Data Mantilla, and OLAP. McGraw.Hill, 1997

[40] 李敏强, 潘振江. 基于数据仓库技术的决策支持系统的研究与应用.系统工程理论与实践. 1998, 3: 14-19

[41] P.SELLIS T K.A survey of logical model for OLAP databases. ACM SIGMOD RECORD.1999, 28(4)

[42] PANES JACOBSON, SQL Seever2000 Analysis Services. 江帆文化艺术中心译. 北京: 机械工业出版社, 2001

[43] 张新香. 综合数据仓库的新型决策支持系统. 现代计算机.2004, 2

[44] 袁长河, 吴永明. 基于数据仓库的决策支持系统研究与建设. 计算机工程与应用.2001, 16

[46] 康晓东. 基于数据仓库的数据挖掘技术. 北京: 机械工业出版社, 2004

[47] 石丽, 李坚. 数据仓库与决策支持. 北京: 国防工业出版社, 2003

[48] MANCUSO, AI MORENO. The Role of OLAP in the Corporate Information Factory. DM Review, November, 2002

[49] CHAUDHURI Dayai U. Data warehousing and OLAP for decision support. ACM SIGMOD RECORD.1997, 26(2)

[50] S.GUMUSOGLU, H.TUTEK. An Analysis Method in Project Management using Primal-dual Relationships. International Journal of Project Management, 1998, 16(5): 321-327

[51] L. J. KIM, K. Y. Dae. Search Heuristics for Resource Constrained Project Scheduling. Journal of the Operational Research Society, 1996, 47(5): 678-689

[52] T. AHN, S. S. Erengue. The Resource Constrained Project Scheduling Problem with Multiple Crushable Modes: A Heuristic Procedure. European Journal of Operational Research, 1998, 107: 250-259

[53] S. D. PERES, W. ROUX, J. B. LLASSERRE. Multi-resource Shop Scheduling with Resource Flesibility. European Journal of Operational Research, 1998, 107: 289-305

[54] M. HAPKE, A. JASZKIEWICZ, R. SLOWINSKI. Interactive Analysis of Multiple-criteria Project Scheduling Problem. European Journal of Operational Research, 1998, 107: 315-324 [50] S. M. YACOUB, H. H. AMMAR. Pattern-Oriented Analysis and Design: Composing Patterns to Design Software Systems. Boston: Addison-Wesley Professional, 2003

[55] D. ALUR, J. CRUPI, D. MALKS. J2EE 核心模式. 牛志奇, 丁天, 田缊哲译. 北京: 机械工业出版社, 2002

[56] R. C. MARTIN, D. RIEHLE. Pattern Languages of Program Design. Boston: Addison-Wesley Professional, 1999

[57] D. BROEMMER. J2EE Best Practices: Java Design Patterns, Automation and Perfomance. Hoboken: Wiley, 2002

攻读硕士学位期间所发表的论文

[1] 段勋，高鸿斌，李明. 烧结生产信息数据分析系统中数据仓库的应用研究. 中国科技博览，2008,17:61-62

[2] 李明，刁彦华，段勋. 基于小波分析的电缆故障行波测距. 中国科技博览，2008,17:30-31

致 谢

本论文是在高鸿斌老师的悉心指导下完成的。不论在论文选题方面，还是在文字的修改、润色方面，高老师都给我提出了许多宝贵的指导和建议。在我的 3 年研究生学习中，无论是在学习上，还是在生活中，高老师都给了我最大的帮助和鼓励。可以说，我的每一点成绩的取得，无不浸润着高老师的巨大心血。感谢高老师一直以来对我学习上的谆谆教导，生活上无微不至的关怀！

在这篇论文完成的过程中，还得到了首钢矿业公司烧结厂李涛同志的帮助，使我能顺利完成论文。

最后，我要感谢计算机教研室的全体老师，感谢诸位老师的关心与指导，还有全体同班同学对我的无私帮助。使我在课题期间少走了弯路。还有，我要深深的感谢我的家人，他们在我攻读硕士学位期间所给予我的支持和无微不至的关怀，使我的论文能够顺利完成。