第5章

数据挖掘技术的应用

【本章要点】


观看视频

1. 数据挖掘在网络中的应用。
2. 数据挖掘在 CRM 中的应用。
3. 数据挖掘在风险评估中的应用。
4. 数据挖掘在交通领域中的应用。
5. 数据挖掘在助力疫情防控上的应用。

5.1 网络数据挖掘

Web 上有海量的数据信息，怎样对这些数据进行复杂的应用成为现今数据库技术的研究热点。数据挖掘就是从大量的数据中发现隐含的规律性的内容，解决数据的应用质量问题。充分利用有用的数据，废弃虚伪无用的数据，是数据挖掘技术最重要的应用。相对于 Web 数据而言，传统的数据库中的数据结构性很强，即其中的数据为完全结构化的数据，而 Web 数据最大的特点是半结构化。所谓半结构化，是相对于完全结构化的传统数据库的数据而言的。显然，面向 Web 的数据挖掘比面向单个数据仓库的数据挖掘要复杂得多。

1. 异构数据库环境

从数据库研究的角度出发，Web 网站上的信息也可以看作一个数据库，一个更大、更复杂的数据库。Web 上的每一个站点就是一个数据源，每个数据源都是异构的，因而每个站点之间的信息和组织都不一样，这样就构成了一个巨大的异构数据库环境。如果想

要利用这些数据进行数据挖掘，首先必须研究站点之间异构数据的集成问题，只有将这些站点的数据都集成起来，提供给用户一个统一的视图，才有可能从巨大的数据资源中获取所需的东西。其次，还要解决 Web 上的数据查询问题，因为如果所需的数据不能很有效地得到，对这些数据进行分析、集成、处理就无从谈起。

2. 半结构化的数据结构

Web 上的数据与传统的数据库中的数据不同，传统的数据库有一定的数据模型，可以根据模型来具体描述特定的数据。而 Web 上的数据非常复杂，没有特定的模型描述，每个站点的数据都各自独立设计，并且数据本身具有自述性和动态可变性。因而，Web 上的数据具有一定的结构性，但因自述层次的存在，是一种非完全结构化的数据，这也被称为半结构化数据。半结构化是 Web 数据的最大特点。

3. 解决半结构化的数据源问题

Web 数据挖掘技术首先要解决半结构化数据源模型和半结构化数据模型的查询与集成问题。解决 Web 上的异构数据的集成与查询问题，就必须要有一个模型来清晰地描述 Web 上的数据。针对 Web 上的数据半结构化的特点，寻找一个半结构化的数据模型是解决问题的关键所在。除了要定义一个半结构化的数据模型外，还需要一种半结构化模型的抽取技术，即自动从现有数据中抽取半结构化模型的技术。面向 Web 的数据挖掘必须以半结构化模型和半结构化数据模型抽取技术为前提。

4. XML 与 Web 数据挖掘技术

以 XML(eXtensible Markup Language，可扩展标记语言)为基础的新一代 WWW 环境是直接面对 Web 数据的，不仅可以很好地兼容原有的 Web 应用，而且可以更好地实现 Web 中的信息共享与交换。XML 可看作一种半结构化的数据模型，可以很容易地将 XML 的文档描述与关系数据库中的属性一一对应起来，实施精确的查询与模型抽取。

1) XML 的产生与发展

XML 是由万维网协会(W3C)设计，特别为 Web 应用服务的标准通用标记语言(Standard General Markup Language，SGML)的一个重要分支。总的来说，XML 是一种中介标示语言(Meta-Markup Language)，可提供描述结构化资料的格式，详细来说，XML 是一种类似于 HTML，被设计用来描述数据的语言。XML 提供了一种独立地运行程序的方法来共享数据，它是用来自动描述信息的一种新的标准语言，能使计算机通信把 Internet 的功能由信息传递扩大到人类其他多种多样的活动中。XML 由若干规则组成，这些规则可用于创建标记语言，并能用一种被称作分析程序的简明程序处理所有新创建的标记语言，正如 HTML 为第一个计算机用户阅读 Internet 文档提供一种显示方式一样，XML 也创建了一种任何人都能读出和写入的世界语。XML 解决了 HTML 不能解决的两个 Web 问题，即 Internet 发展速度快而接入速度慢的问题，以及可利用的

信息多，但难以找到自己需要的那部分信息的问题。XML 能增加结构和语义信息，可使计算机和服务器即时处理多种形式的信息。因此，运用 XML 的扩展功能不仅能从 Web 服务器下载大量的信息，还能大大减少网络业务量。

XML 中的标志(Tag)是没有预先定义的，使用者必须自定义需要的标志，XML 是能够进行自解释(Self Describing)的语言。XML 使用 DTD(Document Type Definition，文档类型定义)来显示这些数据，XSL(eXtensible Stylesheet Language)是一种来描述这些文档如何显示的机制，它是 XML 的样式表描述语言。XSL 的历史比 HTML 用的 CSS(Cascading Style Sheets，层叠样式表)还要悠久，XSL 包括两部分：一个用来转换 XML 文档的方法；另一个用来格式化 XML 文档的方法。XLL(eXtensible Link Language)是 XML 连接语言，它提供 XML 中的连接，与 HTML 中的类似，但功能更强大。使用 XLL 可以多方向连接，且连接可以存在于对象层级，而不仅是页面层级。由于 XML 能够标记更多的信息，因此它能使用户很轻松地找到他们需要的信息。利用 XML，Web 设计人员不仅能创建文字和图形，还能构建文档类型定义的多层次、相互依存的系统、数据树、元数据、超链接结构和样式表。

2) XML 的主要特点

正是 XML 的特点决定了其卓越的性能表现。XML 作为一种标记语言，它有以下特点：

(1) 简单。XML 经过精心设计，整个规范简单明了，它由若干规则组成，这些规则可用于创建标记语言，并能用一种常称作分析程序的简明程序处理所有新创建的标记语言。XML 能创建一种任何人都能读出和写入的世界语，这种创建世界语的功能叫作统一性功能。例如 XML 创建的标记总是成对出现的，以及依靠称作统一代码的新的编码标准。

(2) 开放。XML 是 SGML，在市场上有许多成熟的软件可用来帮助编写、管理等，开放式标准 XML 的基础是经过验证的标准技术，并针对网络进行最佳化。众多业界顶尖公司与 W3C 的工作群组并肩合作，协助确保交互作业性，支持各式系统和浏览器上的开发人员、作者和使用者，以及改进 XML 标准。XML 解释器可以使用编程的方法来载入一个 XML 文档，当这个文档被载入以后，用户就可以通过 XML 文件对象模型来获取和操纵整个文档的信息，加快了网络运行速度。

(3) 高效且可扩充。支持复用文档片段，使用者可以发明和使用自己的标签，也可与他人共享，可延伸性大，在 XML 中可以定义无限量的一组标注。XML 提供了一个标示结构化资料的架构。一个 XML 组件可以宣告与其相关的资料为零售价，及其营业税、书名、数量或其他任何数据元素。随着世界范围内的许多机构逐渐采用 XML 标准，将会有更多的相关功能出现：一旦锁定资料，便可以使用任何方式通过电缆线传递，并在浏览器中呈现，或者转交到其他应用程序进行进一步的处理。XML 提供了一个独立的运用程

序的方法来共享数据，使用 DTD，不同的组中的人就能够使用共同的 DTD 来交换数据。用户的应用程序可以使用这个标准的 DTD 来验证接收到的数据是否有效，用户也可以使用一个 DTD 来验证自己的数据。

（4）国际化。标准国际化且支持世界上大多数文字。这依靠它的统一代码的新的编码标准，这种编码标准支持世界上所有以主要语言编写的混合文本。在 HTML 中，就大多数字处理而言，一个文档一般是用一种特殊语言写成的，无论是英语还是日语、阿拉伯语，如果用户的软件不能阅读特殊语言的字符，那么用户就不能使用该文档。但是能阅读 XML 语言的软件就能顺利处理这些不同语言字符的任意组合。因此，XML 不仅能在不同的计算机系统之间交换信息，还能跨国界和超越不同文化疆界交换信息。

5. XML 在 Web 数据挖掘中的应用

XML 已经成为正式的规范，开发人员能够用 XML 的格式标记和交换数据。在 3 层架构上，XML 为数据处理提供了很好的方法。通过可升级的 3 层模型，XML 可以从现有数据中生成数据，并使用 XML 结构化的数据从商业规范和表现形式中分离信息。数据的集成、发送、处理和显示是下面过程中的每一个步骤：

1）数据集成

促进 XML 应用的是那些用标准的 HTML 无法完成的 Web 应用。这些应用从大的方面讲可以被分成 4 类：需要 Web 客户端在两个或更多异质数据库之间进行通信的应用；试图将大部分处理负载从 Web 服务器转到 Web 客户端的应用；需要 Web 客户端将同样的数据以不同的浏览形式提供给不同的用户的应用；需要智能 Web 代理根据个人用户的需要裁减信息内容的应用。显而易见，这些应用和 Web 的数据挖掘技术有着重要的联系，基于 Web 的数据挖掘必须依靠它们来实现。

2）数据发送

XML 给基于 Web 的应用软件赋予了强大的功能和灵活性，因此它给开发者和用户带来了许多好处。例如进行更有意义的搜索，并且 Web 数据可被 XML 唯一地标识。没有 XML，搜索软件必须了解每个数据库是如何构建的，但这实际上是不可能的，因为每个数据库描述数据的格式几乎都是不同的。由于不同来源数据的集成问题的存在，现在搜索多样的不兼容的数据库实际上是不可能的。XML 能够使不同来源的结构化的数据很容易地结合在一起。软件代理商可以在中间层的服务器上对从后端数据库和其他应用处来的数据进行集成。然后，数据就能被发送到客户或其他服务器进行进一步的集合、处理和分发。XML 的扩展性和灵活性允许它描述不同种类的应用软件中的数据，从描述搜集的 Web 页到数据记录，从而通过多种应用得到数据。同时，由于基于 XML 的数据是自我描述的，数据不需要有内部描述就能被交换和处理，因此，利用 XML，用户可以方便地进行本地计算和处理，XML 格式的数据发送给客户后，客户可以用应用软件解析数据并对数据进行编辑和处理。使用者可以用不同的方法处理数据，而不仅是显示

它。XML 文档对象模型(Document Object Model,DOM)允许用脚本或其他编程语言处理数据,数据计算不需要回到服务器就能进行。XML 可以被利用来分离使用者观看数据的界面,使用简单、灵活、开放的格式,可以给 Web 创建功能强大的应用软件,而原来这些软件只能建立在高端数据库上。另外,数据发到桌面后,能够用多种方式显示。

3) 数据处理

XML 还可以通过以简单、开放、扩展的方式描述结构化的数据,XML 补充了 HTML,被广泛地用来描述使用者界面。HTML 描述数据的外观,而 XML 描述数据本身。由于数据显示与内容分开,XML 定义的数据允许指定不同的显示方式,使数据更合理地表现出来。本地的数据能够以客户配置、使用者选择或其他标准决定的方式动态地表现出来。CSS 和 XSL 为数据的显示提供了公布的机制。通过 XML,数据可以粒状地更新。每当一部分数据变化后,不需要重发整个结构化的数据。变化的元素必须从服务器发送给客户,变化的数据不需要刷新整个使用者的界面就能够显示出来。但在目前,只要一条数据变化了,整一页都必须重建。这严重限制了服务器的升级性能。XML 也允许加进其他数据,例如预测的温度。加入的信息能够进入存在的页面,不需要浏览器重新发一个新的页面。XML 应用于客户需要与不同的数据源进行交互时,数据可能来自不同的数据库,它们都有各自不同的复杂格式。但客户与这些数据库间只通过一种标准语言进行交互,那就是 XML。由于 XML 的自定义性及可扩展性,它足以表达各种类型的数据。客户收到数据后可以进行处理,也可以在不同数据库间进行传递。总之,在这类应用中,XML 解决了数据的统一接口问题。但是,与其他的数据传递标准不同的是,XML 并没有定义数据文件中数据出现的具体规范,而是在数据中附加 Tag 来表达数据的逻辑结构和含义。这使 XML 成为一种程序能自动理解的规范。

4) 数据显示

XML 应用于将大量运算负荷分布在客户端,即客户可根据自己的需求选择和制作不同的应用程序以处理数据,而服务器只需发出同一个 XML 文件。例如按传统的 Client/Server 工作方式,客户向服务器发出不同的请求,服务器分别予以响应,这不仅加重了服务器本身的负荷,而且网络管理者还需事先调查各种不同的用户需求以做出相应的不同的程序,但假如用户的需求繁杂而多变,那么仍然将所有业务逻辑集中在服务器端是不合适的,因为服务器端的编程人员可能来不及满足众多的应用需求,也来不及跟上需求的变化,双方都很被动。应用 XML 则将处理数据的主动权交给了客户,服务器所做的只是尽可能完善、准确地将数据封装进 XML 文件中,正是各取所需,各司其职。XML 的自解释性使客户端在收到数据的同时也理解了数据的逻辑结构与含义,从而使广泛、通用的分布式计算成为可能。

XML 还被应用于网络代理,以便对所取得的信息进行编辑、增减以适应个人用户的需要。有些客户取得数据并不是为了直接使用,而是为了根据需要组织自己的数据库。

比如，教育部门要建立一个庞大的题库，考试时将题库中的题目取出若干组成试卷，再将试卷封装进 XML 文件，接下来在各个学校让其通过一个过滤器滤掉所有的答案，再发送到各个考生面前，未经过滤的内容可直接送到老师手中，当然考试过后还可以再传送一份答案汇编。此外，XML 文件中还可以包含诸如难度系数、往年错误率等其他相关信息，这样只需几个小程序，同一个 XML 文件便可变成多个文件传送到不同的用户手中。

面向 Web 的数据挖掘是一项复杂的技术，由于 Web 数据挖掘比单个数据仓库的挖掘要复杂得多，因此面向 Web 的数据挖掘成了一个难以解决的问题。而 XML 的出现为解决 Web 数据挖掘的难题带来了机会。由于 XML 能够使不同来源的结构化的数据很容易地结合在一起，因此使搜索多样的不兼容的数据库成为可能，从而为解决 Web 数据挖掘难题带来了希望。XML 的扩展性和灵活性允许 XML 描述不同种类的应用软件中的数据，从而能描述搜集的 Web 页中的数据记录。同时，由于基于 XML 的数据是自我描述的，因此数据不需要有内部描述就能被交换和处理。作为表示结构化数据的一个工业标准，XML 为组织、软件开发者、Web 站点和终端使用者提供了许多有利条件。相信在以后，随着 XML 作为在 Web 上交换数据的一种标准方式的出现，面向 Web 的数据挖掘将会变得非常轻松。

5.2 数据挖掘在 CRM 中的核心作用

企业发展 CRM 的目的有两方面：一是帮助营销人员管理好自己的销售过程；二是从客户数据分析中挖掘服务发展方向。其中后者是重中之重。

面临残酷的市场竞争，所有的企业都在不遗余力地争取新客户。然而，现有老客户也蕴含着巨大的商机。调查发现，大部分企业每年有 20%～50%的客户都是变动的，而这一数字在技术型公司更甚。一方面在挖空心思争取新客户，另一面却不断失去老客户。要改变这种状况，留住老客户，赢得新客户，企业必须充分挖掘现有客户的潜力。通过对客户的数据挖掘学习老客户，发掘新的目标客户，这也是很多成功企业发展 CRM 的原因。因此，一套完善的 CRM 系统在建设前期就应该认真考虑对数据挖掘的需求。

1. 需求与技术催生数据挖掘

比较常见的分类，CRM 被分为分析型、运营型、协作型，但无论哪一种，实现对客户活灵活现的了解都是最终目标，因而数据挖掘处于 CRM 系统的核心地位。

数据挖掘是提取有用信息的“数据产生”过程，是从大量数据中挖掘出隐含的、先前未知的、对决策有潜在价值的知识和规则，并能够根据已有的信息对未发生行为做出结果预测，为企业经营决策、市场策划提供依据。

数据挖掘的产生从企业需求方面讲，CRM 上线后，运营特性最先显现出来，公司日

常所有的营销业务都可以流程化和自动化地管理起来，随后客户信息的日趋复杂，客户数据的大量积累，仅限于营销流程的管理已经难以满足企业进一步的需要，企业家期待CRM扮演更重要的角色，分析大量复杂的客户数据，挖掘客户价值。因此，CRM数据应该适应多种分析需求。

2. 没有认真的客户分析，企业在市场上只能盲目探索

客户特征多维分析：挖掘客户个性需求，客户属性描述要包括地址、年龄、性别、收入、职业、教育程度等多个字段，可以进行多维的组合型分析，并快速给出符合条件的客户名单和数量。

客户行为分析：结合客户信息对某一客户群的消费行为进行分析。针对不同的消费行为及其变化，制定个性化营销策略，并从中筛选出“黄金客户”。

客户关注点分析：客户接触与客户服务的分析。

客户忠诚度分析：对客户持久性、牢固性及稳定性进行分析。

销售分析与销售预期：包括按产品、促销效果、销售渠道、销售方式等进行的分析。同时，分析不同客户对企业效益的不同影响，分析客户行为对企业收益的影响，使企业与客户的关系及企业利润得到最优化。

参数调整：为了提高分析结果的灵活度，扩大其适用范围，企业需要对有关参数进行调整。例如，价格的变化对收入会有什么样的影响，客户的消费点临近什么值开始成为“正利润”客户。企业需要通过对收集到的各种信息进行整理和分析，利用科学的方法做出各种决策。

此外，信息技术的发展对数据挖掘的产生做出了很大贡献。在IDC的调研报告中，2021年数据仓库全球市场规模达到700亿美元，数据仓库是一种面向决策主题、由多数据源集成、拥有当前及历史终结数据的数据库系统。它是一个中央存储系统，可以帮助企业员工回答来自客户的业务问题。

在CRM中，数据仓库将海量复杂的客户行为数据集中起来，建立一个整合的、结构化的数据模型，在此基础上对数据进行标准化、抽象化、规范化分类、分析，为企业管理层提供及时的决策信息，为企业业务部门提供有效的反馈数据。现在，NCR、IBM、Oracle等厂商都在数据仓库领域有所建树，一些预见性的模型和解决方案已经被建立起来，数据仓库已不仅是简单的数据存储，而成为对客户资料进行分析，挖掘客户潜力的基石。

3. 客户分析的3个阶段

客户分析过程包括3个阶段：客户行为分析、重点客户发现和效能评估。首先，将客户行为数据（反馈）和效能评估的结果集中起来进行客户行为分析，通过对重点客户的挖掘，为制定市场策略提供依据；其次，把对客户行为的分析结果以报表形式传递给市场专家，市场专家利用这些分析结果制定准确、有效的市场策略；最后，以客户所提供的市场

反馈为基础，再一次进行效能评估，为改进服务和CRM本身提供依据。

1）客户行为分析

包括行为分组、客户理解和客户组之间的交叉分析3个步骤。行为分组是关键，行为分组的分析结果使后两个步骤更加容易。

行为分组：根据不同的客户行为划分为不同的群体，各个群体有着明显的行为特征。通过分组可以更好地理解客户，发现群体客户的行为规律。分析过程中把一次市场活动后得到的客户反馈叫作“反应行为模式”，和手工销售体系中采用的“二元客户反应模式”不同，CRM采用的“分类反应行为模式”允许定义多种反应行为。定义反应行为的方法取决于企业所从事的商业领域。例如企业主营业务是服装销售，一种反应行为可以定义为“从产品目录中选购了女式服装”，也可定义为“从产品目录中选购了男式服装”。这些行为模式的定义可以根据需要非常具体（例如，购买了一件红色的男式马球牌衬衫）。

2）全面正确的客户行为分析，将使自己与客户建立“亲密”的营销关系

客户理解：其目标是将客户在行为上的共性与已知资料结合起来，对客户进行具体分析：哪些客户具有这样的购买行为？客户分布地区是哪里？此类客户给企业带来多少利润？忠诚如何？客户拥有企业的哪些产品？客户购买高峰期是什么时候？完成了这些客户理解，将为企业在确定市场活动的时间、地点、对象等方面提供确凿的依据。

组间交叉分析：客户组间交叉分析对企业来说也很重要，许多客户同属于两个不同的行为分组，且这两个分组对企业的影响相差很大。在企业中有“购买新款商品”和“购买50元以下商品”这两个行为分组。企业会认为第一个分组对企业的收益影响大，因为希望通过新款商品来扩大市场，而第二个分组对企业的收益影响小。此时，如果客户同属两个分组，我们就需要充分分析客户发生这种现象的原因。组间交叉分析为我们提供了解决方案，企业可以了解：哪些客户能够从一个行为分组跃进到另一个行为分组中；行为分组之间的主要差别；客户从一个对企业价值较小的组上升到对企业有较大价值的组的条件是什么。这些分析可以帮助企业准确地制定市场策略，以获得更多的利润。

4. 重点客户发现

CRM理论经典的2/8原则，即80%的利润来自20%的客户，重点客户发现主要应考虑以下方面：潜在客户（有价值的新客户）、交叉销售（交叉销售指企业向老客户提供新产品、新服务的营销过程）、增量销售（更多地使用同一种产品或服务）、客户保持（保持客户的忠诚度）。

假设你是一个银行的市场经理，想向现有的客户推销房屋抵押贷款和信用金卡这两个新产品以进行交叉销售。CRM进行交叉销售时，需要进行以下3个步骤。

（1）数据收集：从数据仓库中收集与客户有关的所有信息，包括客户个人信息（年龄、收入）、交易记录（最近的收支情况、消费次数和信用等级）等。

（2）进行建模：用数据挖掘的一些算法（如统计回归、逻辑回归、决策树、神经网络

等)对数据进行分析,产生一些数学公式,用来对客户将来的行为进行预测分析。

(3) 对数据进行评分:评分过程就是计算数学模型的结果。

5. 效能评估

根据客户行为分析,企业可以更准确地制定市场策略和策划市场活动。然而,这些市场活动能否达到预定的目标是改进市场策略和评价客户行为分组性能的重要指标。因此,CRM 必须对行为分析和市场策略进行评估。这些效能评估都是以客户所提供的市场反馈为基础的。针对每个市场目标设计一系列评估模板,从而使企业能够及时跟踪市场的变化。同时在这些报告中,给出一些统计指标来度量市场活动的效率,这些报告应该按月份更新,并根据市场活动而改变。在一定的时间范围内(3～6 个月)给出行为分组的报告。

5.3　数据挖掘在电信业中的应用

以杭州电信市场为例,1999 年,杭州电信开始着手数据仓库的建设,当时的主要目的是产生一些常规的统计报表。2005 年初,数据仓库的工作目标有所改变,真正开始利用数据仓库技术来进行专题分析,以帮助企业进行经营决策。经过比较,杭州电信选择了 CA 公司的数据仓库解决方案,包括 CA 的 Advantage Data Transformer 和 CleverPath OLAP。Advantage Data Transformer 具有强大的跨系统收集数据的能力,可以帮助杭州电信创建数据仓库,自动收集来自操作系统、网络管理系统和客户服务系统等不同业务系统的数据,并将其存储在数据仓库内。CleverPath OLAP 提供多种 OLAP(联机分析处理)数据分析功能,包括多维数据分析、比较分析、百分比分析等,分析结果可以转换成 Excel 形式的电子数据表格或真实图表的形式。终端用户还可直接从 OLAP 服务器端或 Web 客户机进行互动的数据分析。

杭州电信之所以选择 CA 数据仓库软件,是因为它具有两大优点:一是数据抽取、清洗、转换和展现一体化;二是在数据展现方面,报表的显示内容和形式可以动态改变,比一般报表更为灵活,可以分析,比一般报表更为深入。

杭州电信的经验表明,建立数据仓库需要注意几点:在企业级的数据共享和应用系统过程中,尤其重要的是企业数据标准的建立;将决策问题转换为分析主题;避免“一次实施,终身受益”的想法,要在实践中不断丰富和完善,不断增加新的分析主题。

1. 主题分析实例

数据仓库建成以后,杭州电信就可以根据决策支持的要求开展主题分析。目前,杭州电信开展了以下九大主题的分析。

(1) 营业受理及竣工情况分析:一是按不同业务分类统计受理及竣工情况;二是按

受理部门分类统计受理及竣工情况。根据营业受理情况调整人员配置，“九七”系统营业受理日志表中包含每一笔业务的营业员所属部门，因此可以根据各部门受理数来合理安排各营业部门的营业员配置。

（2）长话详单分析：一是分析长话话务量在时间上的分布情况；二是分析每次通话的时长分布情况；三是分析每次通话的话费分布情况。

（3）小灵通详单分析：一是分析小灵通话务量在时间上的分布情况；二是分析每次通话的时长分布情况。

（4）用户话费分析：从用户的角度，可按用户类别和话费类别在不同话费区段的用户数分布情况进行分析，也可按用户类别和话费类别的用户话费统计及时间对比进行分析；从运营商的角度，可按互联网拨号服务市场份额进行分析，也可以通过比较历史话费变化和用户类别比例进行分析，从而可以得到目前IP长话市场的运营状况。

（5）大客户情况综合分析：分析大客户每月电信消费情况及时间对比，分析大客户的行业分布，分析大客户租用电信资源情况。

（6）用户欠费情况分析：一是分析用户欠费时间分布情况；二是分析欠费用户的年龄和性别构成；三是分析欠费用户性质、种类、身份分布。

（7）201电信业务（类似于IP电话）分析：分析201通话量分校区分布情况；分析201通信量方面，电话与上网费用的比率关系；分析201通信量中国电信和其他运营商占有率情况。

（8）程控功能分析：分析电话用户选用的程控功能情况。

（9）行业分布分析：分析电信手机用户及163网易用户的行业分布情况。

2. 基础数据是关键

尽管杭州电信目前已经做了很多主题的分析，但是可以做的分析还有很多。客户的属性分析可分为两大类：一类是客户的电信消费属性；另一类是客户的社会学属性。

一般来讲，客户的电信消费属性在电信运营商的系统上是较为完整的，可以从客户打电话/上网的通信记录、客户的账务记录、客户的反馈记录中得到，运营商只要从客户的所有电信消费角度进行整理，就可以得到其电信消费属性。目前杭州电信所做的分析大多是基于电信消费属性的分析。基于客户的社会学属性的分析，对电信企业的经营决策很有价值，但很难做到，主要原因是基础数据缺乏。决策分析需要的客户社会学属性包括地理因素、人口因素、心理因素、行为因素等。

这些因素的分析对电信运营商的市场营销决策有着重要的作用，但是需要补充客户的社会属性数据。目前，电信运营商解决这个问题的办法主要有两个：一是对客户进行普查，其工作量和难度相当大；二是通过积分奖励等措施搜集部分高消费客户的社会属性资料。

3. 基础信息系统

杭州电信现有的信息系统主要包括5部分："九七"营业受理系统，它是电信"九七"工程的产物，它的功能包括营业受理、配线配号、号线维护、客户信息等；交换、传输及网管系统，负责产生通话详单、统计接通率、汇总管线资源等；计费账务系统，负责搜集计费数据、产生用户账单、统计欠费情况等；客户服务系统，负责处理114、112、180、189等特服号所提供的服务内容；财务及统计系统，这一部分与大多数单位相似。

从电信业现有系统所涵盖的业务流程来看，在市场需求分析和用户反馈两个环节方面是比较薄弱的。也就是说，一般电信运营商缺乏对客户需求的科学分析，在开展新业务时可能会冒很大的风险。

从客户关系管理的观念来看，客户信息是企业的宝贵资源。电信行业从垄断向逐步开放的进程演化时，在不断探索新的业务增长点的同时，电信公司的首要任务是争取客户并且提高客户的忠诚度。因此，信息系统必须以客户为中心，了解不同客户的不同消费模式，针对不同的用户采取不同的策略，以达到个性化服务的目标。

数据仓库的应用重点是从现有信息系统中提取有用的客户信息，辅助决策行为。

5.4　数据挖掘在风险评估中的应用

保险是一项风险业务，保险公司的一个重要工作就是进行风险评估，即对不同风险领域的鉴定和分析。风险评估对保险公司的正常运作起着至关重要的作用，保费和保单的设计都需要比较详细的风险分析。下面是一个利用KDD方法进行风险分析的实例，它从过去的保单及其索赔信息出发，利用决策树的方法寻找保单中风险较大的领域，从而得出一些实用的风险规则，对保险公司的工作起到指导作用。

评估一项保险投资组合的效果如何，既需要对该投资组合进行整体分析，又需要进行投资组合内部的分析。通过整体分析可以判断以前的投资组合是否盈利，而通过投资组合内部的详细分析可以揭示该投资组合在哪些领域盈利大，而在哪些领域损失大。投资组合内部的分析对一个保险公司来说是很重要的，因为它对于该公司是否既能保持很高的竞争力，又能保持高盈利起着很重要的作用。如果一个公司不知道其投资组合中的哪一部分存在大的风险，那么，尽管这项投资组合目前是盈利的，但要维持下去却是很难的。

投资组合的整体分析可以在总保费和总索赔的基础之上用统计的方法来实现，而对其内部的分析则需要更复杂、更精确的方法。

进行投资组合内部分析的一般方法是将该投资组合划分成一些小的风险领域，这些风险领域由一系列的风险等级来表示，风险等级则由意外事故表列出。这种分析方法将

每一个风险等级因素与索赔频率和索赔金额的关系用一个模型来表示，模型的参数用过去已有的数据（保单索赔等）来估算。参数确定后，就可以用该模型预测将来在不同的风险等级参数下的索赔频率和索赔金额。由于分析的复杂性，因此这种方法只能考虑几个参数，如索赔频率、索赔金额等。

使用这种方法，必须明确风险等级参数。如果是一个连续的参数，就需要将其划分成若干个等级。在分析详细程度和模型的可行性两方面应达到一种平衡，而且参数之间的相互作用处理起来很困难，因此也被忽略了。在参数多、多值变量多的情况下，这种相互作用是很多的。

风险分析还有其他一些方法，它们大都用在保险统计领域中。Siebes 将这一问题引入数据挖掘领域，利用概率论的方法对风险领域进行研究，将每年的保险赔偿看成是 Bernoulli 实验。这项工作导致了保险投资组合类别的相等概率及同一描述思想的发展。

在本书中，我们将保险风险分析中一些反复的、交互式的、探索性的工作看成是一种 KDD 过程，利用一些正规的分析方法，来获得这一领域中专家所具有的直觉知识。

一个保险公司投资组合数据库包含用户购买的保单集合。一个保单确保一个标定物的价值不会失去。当标定物遭到损失或丢失时，要根据保单进行索赔，以此作为补偿。一个保单在一定的时间内有效，其有效时间被称为风险期。在任一时间，投资组合数据库中的保单所对应的风险都是不同的。

保险公司成功的一个关键因素是在设置具有竞争力的保费和覆盖风险之间选择一种平衡。保险市场竞争激烈，设置过高的保费意味着失去市场，而保费过低又会影响公司的盈利。保费通常是通过对一些主要的因素（如驾驶员的年龄、车辆的类型等）进行多种分析和直觉判断来确定的。由于投资组合的数量很大，因此分析方法常常是粗略的。

一项投资组合的绩效通常用前些年的数据来评估。这种分析一般由承保人用来预测将来这项投资组合的绩效，并根据市场的变化和标定物的情况来调整保单等级结果。每年都要用这种分析来调整来年保费的设置规则。

设置保费有两种极端情况：一是所有保单都采用同一保费；二是每一保单根据具体情况单独设置保费。这两种极端情况都是不实用的。然而，一个好的保费设置应该是接近后者的。保险商比较喜欢在设置保费时考虑更多的因素。

数据挖掘提供了进行保险投资组合数据库分析的环境。ACSys 数据挖掘系统（Williams & Huang，1996）提供了风险分析框架。该系统将决策树作为知识发现算法。决策树是利用信息论中的互信息（信息增益）寻找数据库中具有最大信息量的字段，建立决策树的一个节点，再根据字段的不同取值建立树的分支，在每个分支子集中重复建立树的下层节点和分支过程，直到生成一个完整的决策树。决策树的实现需要包含以下 3 个阶段。

1. 树的生长阶段

通常在一并行体系结构中实现分类-克服策略，在一组训练例的基础上建立一个完整的决策树。

2. 树的评估阶段

利用测试例集合来评价生成的树。在这一阶段，需要对树进行适当的修剪，并选取不同的测试例集合对树的性能进行测试并进行修剪。

3. 树的应用阶段

将最后生成的树应用于未知的数据。

为进行风险分析，选取索赔金额作为目标属性，其他属性作为独立变量。所有保单被划分为两类，即有索赔的和无索赔的，将索赔金重新分类为 1 或 0，而后利用数据集合来生成一个完整的决策树。

从生成的决策树中可以建立一个规则基。一个规则基包含一组规则，每一条规则对应决策树的一条不同路径，这条路径代表它经过节点所表示的条件的一条连接。一条规则例子如下：

```
If
年龄<=20
and 性别=男性
and 保险金额>=5000
and 保险金额<10000
Then 保险声明=1,cost=0,(0,15)
```

这条规则表明在给定的条件下，一个保单被索赔。

Graham J. Williams 和 Zhexue Huang 等利用 ACSys 对 NRMA 保险公司的投资组合数据库进行了分析，得到了一些有用的规则，并在此基础上分析了一些其他公司的数据，对已有规则进行了拓宽。通过生成树中那些带索赔的叶节点，可对一些风险的重要领域进行研究。叶节点的索赔频率和索赔额提供了重要的信息，通过生成树还可得到一些其他信息，如与风险领域相关的所有保费等。

5.5　数据挖掘在通信网络警报处理中的应用

一个通信网络可以看成是由互相连接的部件组成的，如交换器、传输设备等。每个部件又包含一些子部件。分析的层次不同，部件的数目也不相同。一般来说，一个局域

电话网包含 10～1000 个部件。

在通信网络的运行过程中，网络中的每个(子)部件和软件模块都可能产生警报，这些警报描述了某些异常情况的发生，它们所指示的问题对用户来讲不一定是可见的。一个网络所发出的警报数目是相当可观的，甚至在一个小的局部通信网络中，也可能存在成千上万个不同类型的警报。警报数目可能因网络的不同、时间条件的变化而有很大差别，但通常情况下，对一个一般的通信网络，每天可能产生 200～10 000 个警报。

通信网络管理系统操作维护中心接收网络中各节点发来的警报，并将这些警报信息存储在一个警报数据库中。对这些警报的处理有多种形式，可以简单地将它们忽略，但更重要的是将这些警报提示给网络管理员，由管理员来决定怎么处理。

不同时间发生的警报组成一个警报流。处理警报流是一项十分困难的工作，主要有以下原因：

(1) 对于一个大型的通信网络来说，每天产生的警报类型和数量都是相当可观的，这表明在网络中所发生的异常情况种类繁多、数量巨大。

(2) 警报具有突发性。也就是说，在很短的时间内可能产生很多警报信息，网络管理员很难在这么短的时间内处理如此多的警报，然而，警报的突发又说明可能发生了重大的故障，网络管理员必须进行处理。

(3) 通信网络中的软硬件更新换代很快，当加入新的节点或更新旧的节点时，警报序列的特点也随之发生改变，而网络管理员要跟上这些改变是相当困难的。

人们为了解决处理警报信息的问题，采用警报过滤和关联技术，以提高提交信息的抽象级别，从而减少提交给网络管理员的警报信息的数量。

(1) 警报过滤是指在分层网络中的每一层都对下层节点发来的警报进行过滤，即一个节点只发送从子节点收到的部分警报。

(2) 警报关联是指对警报进行合并和转换，将多个警报合并成一条具有更多信息量的警报，这样可以通过发送一条警报来代替多条警报。

警报过滤和警报关联需要存储关于警报序列的知识，这些知识从原则上讲可以取自设计单个部件的工程师或有操作经验的工程师。然而，这一过程相当烦琐。警报过滤和警报关联可以用来减少提交给网络管理员的警报数量，然而，它们却不能对网络的行为做出有效的预测，从而避免重大故障的发生。网络中的故障通常出现在网络中部件间的连接上，它们的预测是一件相当困难的事，而这种预测却可带来相当可观的经济效益。

芬兰赫尔辛基大学计算机科学系的 K. Hatonen 等开发了一个基于通信网络中警报数据库的知识发现系统 TASA(Telecommunication Alarm Sequence Analyzer)。该系统是与一个通信设备生产厂商及 3 个电话经营商(两个固定城市电话网和一个国家范围的移动通信网)合作开发的，其目的是寻找有助于处理警报序列的规则，这些规则用来过

滤、转换警报，并用来预测故障。

TASA 系统将一个警报表示成一个三元组(c,a,t)，其中 c 表示发送这一警报的部件，a 是警报类型，t 是警报发生的时间。然后，利用统计方法，从一个警报序列中寻找某一情节发生的概率。

TASA 系统在警报流中计算那些经常发生的情节，根据这些情节提取有用的规律。

从一个警报序列中可以发现不同类型的知识，如神经网络、风险模型或基于规则的知识。

如果最终目的是获得好的预测性能，神经网络便是很好的选择。许多证据表明，神经网络在预测方面有很好的适用性，它将知识以连接权的形式来表示，不易理解，然而，在通信网络警报处理中，其中一个重要目的就是发现可理解的知识，通信厂商不想在他们的系统中安装任何“黑盒子”之类的东西，因此，这就排除了应用神经网络这种简单的想法。

TASA 系统知识发现中所采用的是基于规则的形式，一个一般的规则形式如下：“如果某一警报组合在一段时间内发生，那么，在给定的时间间隔内，某一类型的警报可能发生”。之所以选取这种类型的知识，是考虑到如下几条原因。

- 可理解性：这类知识易于被人们理解。当前处理警报序列的操作员喜欢用这种类型来表达他们关于警报的知识。
- 应用领域的特点：这类规则是这一领域中简单因果关系的表达，可以证明这类知识适用于通信网络。
- 存在有效算法：这种类型的规则当前有比较有效的算法来获取。

在几个 TASA 系统中发现的规则类型的例子如下：

(1) 如果 A 类型警报发生，那么，在 30 秒内，B 类型警报发生的概率为 80%。

(2) 如果 A 类型和 B 类型警报在 5 秒内发生，那么，在 60 秒内，C 类型警报发生的概率为 70%。

(3) 如果 A 类型警报在一 B 类型警报之前发生，C 类型警报发生于 D 类型警报之前，而且都在 15 秒的范围内，那么，E 类型警报在接下来的 4 分钟内发生的概率为 60%。

对于发现的规则，TASA 系统还提供了一些比较好的工具来对规则进行后处理，其中有规则的剪辑、定制、组合等，使这些规则更便于应用。

TASA 系统的第一个版本已经通过实际警报数据的测试，结果比较好，一部分通过 TASA 发现的规则正在被一些网络开发商用于产品开发中。

在市场金融方面，Integral Solution 为 BBC 开发了采用神经网络和归纳规则方法预测收视率的发现系统；在零售业，数据挖掘主要应用于销售预测、库存需求、零售点选择和价格分析，例如用自然语言和商用图表分析超市销售数据的 Spotlight 系统，及扩展到其他市场领域的 Opportunity Explorer 系统；在医疗保健方面，由 GTE 开发的 KEFIR

数据挖掘系统用于分析健康数据，确定偏差，并通过 Web 浏览器以超文本形式输出结果；在科学研究方面，SKICAT 系统能对宇宙图像数据进行分类，Quakfinder 利用卫星采集的数据监测地壳活动，HMMs 和 SAM 用于发现和构造生物模型；在司法方面，可用数据挖掘技术进行案件调查、诈骗监测、洗钱认证、犯罪组织分析，如美国财政部开发的 FAIS 系统；在制造业上，可利用数据挖掘技术进行零部件的故障诊断、资源优化、生产过程分析等。

在统计和机器学习领域中还有许多数据挖掘系统。另外，将数据仓库、OLTP、OLAP 和数据挖掘技术结合是近期数据库发展的一个趋势。数据仓库和数据挖掘都可以完成对决策技术的支持，相互间有一定的内在联系，两者集成可以有效地提高系统的决策支持能力。例如瑞典保险系统由 OLTP 系统、数据仓库、数据挖掘环境 3 部分构成。建立在 Oracle 数据库基础上的 MASY 数据仓库从多个 OLTP 信息源收集相关数据。由多种数据挖掘工具（Expla、RDT、C45 等）构成的数据挖掘环境提供动态数据分析，使用户尽可能不依赖数据挖掘专家执行多种类型的数据挖掘任务。

数据挖掘在数据库之外的其他领域也有丰硕的成果，例如统计学中已发展了许多用于数据挖掘的技术，演绎逻辑编程作为逻辑编程的一个迅速发展的分支，与数据挖掘有密切联系。

5.6 数据挖掘在交通领域的应用

大数据和数据挖掘技术的发展为解决交通中存在的问题带来了新的思路。大数据可以缓解交通堵塞，改善交通服务，促进智能交通系统更好、更快地发展。

在目前的技术条件和发展水平下，大数据在交通中的应用主要有以下几种方式：

（1）由于公共交通部门发行的一卡通大量使用，因此积累了乘客出行的海量数据，这也是大数据的一种，由此，公交部门会计算出分时段、分路段、分人群的交通出行参数，甚至可以创建公共交通模型，有针对性地采取措施，提前制定各种情况下的应对预案，科学地分配运力。

（2）交通管理部门在道路上预埋或预设物联网传感器，实时收集车流量、客流量信息，结合各种道路监控设施及交警指挥控制系统数据，由此形成智慧交通管理系统，有利于交通管理部门提高道路管理能力，制定疏散和管制措施预案，提前预警和疏导交通。

（3）通过卫星地图数据对城市道路的交通情况进行分析，得到道路交通的实时数据，这些数据可以供交通管理部门使用，也可以发布在各种数字终端供出行人员参考，来决定自己的行车路线和道路规划。

（4）出租车是城市道路的最多使用者，可以通过其车载终端或数据采集系统提供的实时数据，随时了解几乎全部主要道路的交通路况，而长期积累下的这类数据就形成了

城市区域内交通的“热力图”，进而能够分析得出什么时段的哪些地段拥堵严重，为出行提供参考。

(5) 智能手机已经很普及，多数智能手机都会使用地图应用，于是始终打开 GPS 或北斗定位系统，地图提供商将收集到的这些数据进行大数据分析，由此就可以分析出实时的道路交通拥堵状况、出行流动趋势或特定区域的人员聚集程度，这些数据公布之后会给出行提供参考。

公共交通是指城市范围内定线经营的公共汽车及轨道交通、渡轮、索道等交通方式，这些交通工具都是按照时间点发车，资源配置不合理就会导致等车时间长、乘坐拥挤、挤不上等一系列的问题。大数据技术可以实现资源的合理配置，通过站点实时客流量检测，合理分配公共资源，提高资源利用效率。此外，乘客可以通过手机 App，实时查询公交车的行驶状况、车内客流情况供乘客参考，及时更改乘坐计划，避免出现盲目等车的状况。公共交通是缓解交通拥堵的一种有效手段，完善公共交通服务质量，让市民真切地感受到公共交通带来的便利，是市民选择公共交通出行的先决条件。

随着国民经济的持续增长，交通需求越来越大，交通事故数量居高不下，道路交通安全成为全社会普遍关注的问题，减少道路交通事故的发生，提高道路交通、安全水平已经成为人们的迫切要求。

在道路交通系统中，因驾驶员的素质、车辆的安全性能、环境、道路及气候等因素的不良变化，导致这种因素组合恶化，如果这种恶化因素持续发生，就可能导致交通事故的发生。大数据的实时性及可预测性保证了交通系统对事故的主动预警，以便提前预测事故发生的可能性。例如，通过 GPS 定位技术采集车辆行驶轨迹，判断车辆是否正常行驶，若出现非正常行驶，则及时通过交通部门对车辆进行管制，通过道路环境及设施检测系统，实时采集道路环境及道路设施信息，经过云计算分析处理大数据后，及时通过交通广播发布或者通过手机短信将信息推送给附近行驶的车辆，通过大数据技术及时分析恶劣天气环境下的道路状况，减少雨天、大雾、雪天连环撞车发生的概率。

将大数据应用到应急救援系统中，可以更加准确地定位事故地点，快速通过医护及消防救援，并且可以通过大数据技术推送事故发生信息给附近行驶的车辆，让其做好让救援车队顺利通过的准备，并告知驾驶员备选路径，以便于驾驶员改变行驶路径。

大数据在交通上的应用还有一个常见的场景。随着人们生活水平的提高，道路上的机动车越来越多。套牌机动车的数量也随之增多，由于套牌机动车发现难度大，检测难度高，有许多套牌机动车并没有被发现，严重影响了道路交通安全秩序，例如随意的闯红灯、超速、跨越双实线、乱停、乱放，给人们的安全出行带来了很大的隐患，也为肇事逃逸案件的侦破增加了难度。通过大数据，可以解决套牌机动车问题，在解决交通拥挤等问题上有很大的优势。

随着车辆的增多，停车难已成为人们非常关注的问题。解决停车难问题是治理交通

拥堵工作的一部分，把大数据应用到智能交通系统中，可以通过主动式的方式向用户推送相关交通服务信息。例如利用电子车牌 GPS 定位技术获取车辆停靠位置及停靠时间信息，出现违规停靠的情况向车主手机推送相关违规信息，让其及时把车开走，这样可以缓解道路车辆乱停靠带来的交通堵塞。通过停车诱导系统获取车辆所在位置和附近一定区域内的停车场信息，预测到达停车场的时间，通过手机短信或者手机 App 的方式及时向车主推送附近停车场的信息，车主可以主动地选择停车场或者提前预订车位。

为避免乘坐高铁误点，乘客往往要提前好几个小时就往火车站赶，赶火车花费的时间甚至要比乘坐高铁的时间多出许多。把大数据技术应用到交通中，出租车公司可以联合高铁运输部门获取乘客的信息，例如手机号及乘车时间。出租车公司可以与交通信息中心联合获取出行前和出行后的交通信息，通过大数据处理技术预测从出发点到火车站的时间 t，向乘客推送路径、用时、乘车方式等信息，乘客若要乘坐出租车，则可以在合适的时间通过手机 GPS 定位技术获取出发地点及附近的出租车信息，通过实时交通信息服务，出租车司机选择最优路径，以最快的速度到达火车站，这样可以节约乘客大部分的时间。

应用大数据创建智慧城市的典型代表是杭州市。

理性的数据建模分析告诉我们：一个城市，如果把车和车、车和道路充分链接到位的话，从理论上来说，可以提升这个城市道路通行能力的 270%。在实践的层面上，在城市化快速推进的过程中，如何避免各方面“城市病”发生“共振”，从而导致系统性城市运行风险爆发，是城市管理者应当高度关注的问题。

杭州国际城市学研究中心设立的“西湖城市学金奖”奖项，面向民间领域征集破解“城市病”之道。2012 年，第二届“西湖城市学金奖”中“城市交通问题”征集成果《缓解城市交通拥堵问题 100 计》中，被杭州市交警局采纳并运用到实践中的点子比例高达 40%。杭州市交警局局长乐华说，交警局是“西湖城市学金奖”城市交通问题征集评选活动中最大的受益者。杭州的错峰限行、分区域停车费收费新政、西湖环线交通、地铁换乘优惠等交通举措都是源于“西湖城市学金奖”的金点子。

在基于大数据的智能交通应用方面，杭州国际城市学研究中心主办的“西湖城市学金奖”征集活动中也有这样的点子并已经投入使用。在第一版“杭州公共出行”应用获选西湖城市学金奖金点子后，安卓用户下载使用量达到 10 000 余次。2020 年 3 月，应用升级，在原有基础上增加了实时公交、地铁信息查询、检索功能，覆盖城市公共交通出行大范畴，并在微信平台上设服务号，通过发送关键词推送查询信息，方便除安卓系统之外的智能手机用户。

发展城市轨道交通对于解决大都市交通问题是很好的解决方案，在有效缓解城市交通的同时，也会对城市形态的发展起到积极的引导作用。在目前的形势下，发展城市轨道交通还能够在短时间内拉动固定资产投资，促进经济平稳、较快地发展。但发展城市

轨道交通投资巨大，建设一千米的地铁线路需要投资近 4 亿元人民币，因此被称为“天价工程”，其盈利模式也是世界性难题，因此对在哪些城市建设轨道交通、建设的规模有多大等重大问题，始终没有公认的判定标准。一般认为城市轨道交通建设只有与城市的发展协调同步才能取得良好的社会、经济效益，但如何界定轨道交通与城市发展的协调程度需要有科学的评价方法，基于此种考虑，城市轨道交通需与城市发展相互协调，对轨道交通与城市协调性进行定性分析，为城市轨道交通建设规模、建设时机提供决策支持。

轨道交通和城市发展协调性评价涉及社会、人口、经济、城市综合交通等各方面，包含众多因子，依照科学性、客观性、可比性和动态性原则，同时考虑各方面因素和资料占有的可能选取指标。

1. 轨道交通状况评价指标

可选取 3 个方面共 6 个原始指标评价城市轨道交通的发展状况：

(1) 表示城市轨道交通网发展规模和发展水平的指标 $A1$，包括两个子指标：轨道交通网线路长度($X1$，千米)和投入的运营车辆数量($X2$，节)。

(2) 表示城市轨道交通系统运营状况的指标 $A2$，包括两个子指标：轨道交通系统客运总量($X3$，万人)和运营车辆行驶总里程($X4$，千米节)。

(3) 表示城市轨道交通系统经营管理状况的指标 $A3$，包括两个子指标：轨道交通系统利润($X5$，万元)和轨道交通系统从业人数($X6$，人次)。

2. 城市发展状况评价指标

可选取 4 个方面共 18 个原始指标评价该城市的发展状况：

(1) 人口子系统的总量及结构($B1$)，包括 3 个指标：城市人口总量($Y1$，万人)、非农业人口总量($Y2$，万人)和从业人口总量($Y3$，万人)。

(2) 经济子系统的总量及结构($B2$)，包括 5 个指标：国民生产总值($Y4$，亿元)、第一产业生产总量($Y5$，亿元)、第二产业生产总量($Y6$，亿元)、第三产业生产总量($Y7$，亿元)和城市财政收入($Y8$，亿元)。

(3) 城市居民生活状况($B3$)，包括 5 个指标：城市消费价格指数($Y9$)、城镇居民人均住宅面积($Y10$，平方米)、城镇居民人均可支配收入($Y11$，元)、失业率($Y12$，%)和城市市政建设投入($Y13$，亿元)。

(4) 城市公共交通状况($B4$)，包括 5 个指标：城市交通投入($Y14$，亿元)、城市人均道路长度($Y15$，千米/人)、城市人均道路面积($Y16$，平方千米/人)、居民万人公交车拥有量($Y17$，辆/万人)和公交客运总量($Y18$，万人次)。

3. 具体应用

以 A 地铁运行为例，在进行设备运维管理的过程中，大数据信息挖掘技术手段在智能“轨道”交通系统中应用，服务于轨道交通设备的运维管控，动态化对智能“轨道”交通

系统中的各项设备运行情况进行数据信息管控采集，对获取的数据信息之间的因果关系进行把控，从而总结出对维保管理工作具备价值的数据信息。在实施轨道交通维保工作时，非常注重数据信息收集和数据信息挖掘，在分析各项数据信息的基础上，获取具备更高价值的维保数据信息。例如，对于车辆管理维护来说，若车辆存在既往故障数据，则可以在大数据信息挖掘的过程中，有针对性地对车辆故障历史情况进行分析，提前预测车辆各项设备的失稳潜在隐患，有序完成潜在故障设备的更换，确保列车运行的安全性和稳定性。在数据库对比分析环节，可以借助数据信息分析的形式，确定故障问题并且消除故障表现，对 A 地铁的运行安全奠定扎实的基础保障。

1）分析交通设备的用电量消耗

对于 A 地铁运行单位来说，借助大数据信息分析技术手段，对 A 地铁 2020 年 2 月至 3 月的用电消耗情况进行了数据采集和分析，采集的数据信息显示比前一年 2 月至 3 月明显少很多。

2）分析交通工具的舒适度

对于 A 地铁运营的实际情况来说，以列车稳定性数据来作为评判交通工具舒适性的重要评价因素。借助每月抽查的方式，对 A 地铁运营的舒适度进行分析，发现 2021 年开始，A 地铁运营从横纵方向上有所提速，同时整体舒适度有所降低。对 A 地铁在 2021 年 7 月 17 日和 21 日的运行情况进行分析，均存在增速稳定性问题。通过对交通工具开展调试和管理，以及大数据信息对比来看，因为 7 月正值学生们的放假季、旅游季，所以乘车人数相对较多，导致该时段的列车交通运行稳定性有所降低，这也是 2021 年 7 月 A 地铁运营舒适度降低的主要原因之一。

3）分析交通工具的维保模式

结合 A 地铁运营设备维保工作模式的实际情况来看，主要存在以下几种模式：

（1）事后维修。对于事后维修来说，便是在轨道交通运行环节出现实际故障问题之后，构建出故障报修、维修派单、维修方案校准、维修品质验收、维修成本统筹等诸多管理程序。从实施事后维修环节来说，需要首先统筹维保工作资源，实现设备故障问题检修和处理。由列车员对列车设备故障进行报修申请，对故障设备、故障问题进行有效检测，并且形成维保派单，有效执行列车检修的各项程序。在设备维修完毕之后，由大数据系统进行质量检验分析，确保列车各项设备运行稳定、运行安全之后，才能完成验收工作。

（2）故障预测。对于故障预测来说，大数据系统结合往期设备故障的表现来看，对可能引发设备故障的因素进行分析与排查，并且建立形成设备维保管理目标。结合 A 地铁运营情况来看，构建了设备阶段性功能保养、设备故障问题巡检等诸多管理机制，以期望提升设备故障预防有效性，完善故障管理体系内容。结合各项设备的功能保养需求来看，完整制定设备保障检修方案可以指导维保工作人员顺利、稳定地开展工作，此外借助

大数据信息挖掘技术手段，还能够最大限度地做好故障巡检工作，及时排查设备潜在的隐患，确保设备配件及时进行更换。对于A地铁运营来说，借助大数据信息技术手段，在2021年5月开展了2次设备保养，在2021年全年自动化故障巡检30次，指导维修工作16次。

4. 轨道交通系统与数据挖掘技术结合的应用发展策略

1）实现大数据信息内部共享交互

智能“轨道”交通系统想要充分展现出大数据信息挖掘技术的维保价值，就应该从全面的角度搜集智能“轨道”交通系统数据信息，完善信息化设备管理平台，并且大力收集数据信息，形成数据共享机制，为大数据信息挖掘工作奠定扎实基础。此外，还应能够不断提升设备故障联动有效性，减少充分劳动等诸多问题，保障设备数据同步管理的效果。因为相同部门中的设备类型具备一致性，所以能够形成模块化数据类型，并且能够从客观角度上对数据信息挖掘处理量进行处理管控。在编程数据库中，可以有效实现数据信息的导入/导出，形成数据共享处理体系，实现数据共享，为智能轨道交通系统维保管理奠定完整、全面的数据基础保障。例如，在A城市地铁运营的过程中，在智能轨道交通系统内部增设了全面化系统数据库，其中涵盖了交通设备信息管理系统数据、交通设备部件损耗管理系统数据、备用零件管理系统数据、轨道交通日常维保系统数据、交通设备档案信息数据等内容，真正实现了大数据信息内部共享交互。此外，各个数据系统之间的信息数据交互，可以在轨道交通出现故障时，开展共享式数据信息录入，形成多个系统之间的数据同步，为诸多管理工作奠定数据信息联动基础。此外，各个系统之间的数据信息共享，可以将系统入口有效地整合在相同的公用平台中，完整、精准地显示设备管理信息内容，这样可以促进轨道交通各管理部门的信息化规范性，实现轨道交通数据资源共享目标。

2）构建各个部门之间的局域用网联动机制

对于轨道交通来说，可能存在各个智能轨道交通系统设备管理和数据信息独立性，这就在一定程度上增加了大数据信息采集的片面性，对维保管理工作带来了一定难度，很难全面保障大数据信息技术分析的全面性。此外，设备之间存在相互作用的关系，所以为了实现轨道交通高质量维保管控，应该组建信息共享平台，完成各类部门的设备以及大数据信息整合，以便于强化大数据信息分析的精准性与精密性，更加清晰化地确定轨道交通各个设备和环节的运行状态。针对A城市的地铁运营实际情况来看，轨道交通工具运行很容易出现设备磨损情况，导致对电气程序、车轮性能等带来直接损害。为此，想要实现高质量的维保管控，应该强化局域用网体系联动机制开发，联系各个部门的设备故障信息，以期构建完善的大数据设备管理机制，提升数据信息的智能性水平，展现出大数据信息技术的优势，提升轨道交通运行效率。

5.7 数据挖掘技术在信用卡业务中的应用

信用卡业务具有透支笔数巨大、单笔金额小的特点，这使得数据挖掘技术在信用卡业务中的应用成为必然。国外信用卡发卡机构已经广泛应用数据挖掘技术促进信用卡业务的发展，实现全面的绩效管理。我国自 1985 年发行第一张信用卡以来，信用卡业务得到了长足的发展，积累了巨量的数据，数据挖掘在信用卡业务中的重要性日益显现。

数据挖掘技术在信用卡业务中的应用主要有分析型客户关系管理(Customer Relationship Management，CRM)、风险管理和运营管理。

1. 分析型客户关系管理

分析型客户关系管理应用包括市场细分、客户获取、交叉销售和客户流失。信用卡分析人员搜集和处理大量数据，对这些数据进行分析，发现其数据模式及特征，分析某个客户群体的特性、消费习惯、消费倾向和消费需求，进而推断出相应消费群体下一步的消费行为，然后以此为基础，对识别出来的消费群体进行特定产品的主动营销。这与传统的不区分消费者对象特征的大规模营销手段相比，大大节省了营销成本，提高了营销效果，从而能为银行带来更多的利润。对客户采用哪种营销方式是根据响应模型预测得出的客户购买概率做出的，对响应概率高的客户采用更为主动、人性化的营销方式，如电话营销、上门营销，对响应概率较低的客户可选用成本较低的电子邮件和信件营销方式。除获取新客户外，维护已有的优质客户的忠诚度也很重要，因为留住一个原有客户的成本要远远低于开发一个新客户的成本。在客户关系管理中，通过数据挖掘技术找到流失客户的特征，并发现其流失规律，就可以在那些具有相似特征的持卡人还未流失之前，对其进行有针对性的弥补，使得优质客户能为银行持续创造价值。

2. 风险管理

数据挖掘在信用卡业务中的另一个重要应用就是风险管理。在风险管理中，运用数据挖掘技术可建立各类信用评分模型。模型类型主要有 3 种：申请评分模型、行为评分模型和催收评分模型，分别为信用卡业务提供事前、事中和事后的信用风险控制。

(1) 申请评分模型专门用于对新申请客户的信用进行评估，它应用于信用卡征信审核阶段，通过申请人填写的有关个人信息，即可有效、快速地辨别和划分客户质量，决定是否审批通过并对审批通过的申请人核定初始信用额度，帮助发卡行从源头上控制风险。申请评分模型不依赖于人们的主观判断或经验，有利于发卡行推行统一规范的授信政策。

(2) 行为评分模型是针对已有持卡人，通过对持卡客户的行为进行监控和预测，从而评估持卡客户的信用风险，并根据模型结果，智能化地决定是否调整客户的信用额度，在

授权时决定是否授权通过，到期换卡时是否进行续卡操作，并对可能出现的逾期情况进行预警。

(3) 催收评分模型是申请评分模型和行为评分模型的补充，是在持卡人产生了逾期或坏账的情况下建立的。催收评分模型用于预测和评估对某一笔坏账所采取的措施的有效性，诸如客户对警告信件反应的可能性。这样，发卡行就可以根据模型的预测，对不同程度的逾期客户采取相应的措施进行处理。

以上 3 种评分模型在建立时所利用的数据主要是人口统计学数据和行为数据。人口统计学数据包括年龄、性别、婚姻状况、教育背景、家庭成员特点、住房情况、职业、职称、收入状况等。行为数据包括持卡人过去使用信用卡的表现信息，如使用频率、金额、还款情况等。由此可见，数据挖掘技术的使用可以使银行有效地建立事前、事中到事后的信用风险控制体系。

3. 运营管理

虽然数据挖掘在信用卡运营管理领域的应用不是最重要的，但它已为国外多家发卡公司在提高生产效率、优化流程、预测资金和服务需求、提供服务次序等方面的分析上取得了较大成绩。

4. 实例分析

下面以逻辑回归方法建立信用卡申请评分模型为例，说明数据挖掘技术在信用卡业务中的应用。申请评分模型设计可分为以下 7 个基本步骤。

1) 定义好客户和坏客户的标准

好客户和坏客户的标准根据适合管理的需要定义。按照国外的经验，建立一个预测客户好坏的风险模型所需的好、坏样本至少各有 1000 个。为了规避风险，同时考虑到信用卡市场初期，银行的效益来源主要是销售商的佣金、信用卡利息、手续费收入和资金的运作利差，因此，一般银行把降低客户的逾期率作为一个主要的管理目标。例如，将坏客户定义为出现过逾期 60 天以上的客户，将好客户定义为没有 30 天以上逾期且当前没有逾期的客户。

一般来讲，在同一样本空间内，好客户的数量要远远大于坏客户的数量。为了保证模型具有较高的识别坏客户的能力，取好、坏客户样本数的比率为 1∶1。

2) 确定样本空间

样本空间的确定要考虑样本是否具有代表性。一个客户是好客户，表明持卡人在一段观察期内用卡表现良好；而一个客户只要出现过“坏”的记录，就把他认定为坏客户。所以，一般好客户的观察期要比坏客户长一些，好、坏客户可以选择在不同的时间段，即不同的样本空间内。例如，好客户的样本空间为 2003 年 11 月至 2003 年 12 月的申请人，坏客户的样本空间为 2003 年 11 月至 2004 年 5 月的申请人，这样既能保证好客户的表现

期较长，又能保证有足够数量的坏客户样本。当然，抽样的好、坏客户都应具有代表性。

3）数据来源

在美国，由统一的信用局对个人信用进行评分，通常被称为“FICO 评分”。美国的银行、信用卡公司和金融机构在对客户进行信用风险分析时，可以利用信用局提供的个人数据报告。在我国，由于征信系统还不完善，建模数据主要来自申请表。随着我国全国性征信系统的逐步完善，未来建模的一部分数据可以从征信机构收集到。

4）数据整理

大量抽样数据要真正最后进入模型，必须经过数据整理。在数据处理时，应注意检查数据的逻辑性，区分“数据缺失”和“0”，根据逻辑推断某些值，寻找反常数据，评估是否真实。可以通过求最小值、最大值和平均值的方法，初步验证抽样数据是否随机、是否具有代表性。

5）变量选择

变量选择要同时具有数学统计的正确性和信用卡实际业务的解释力。Logistic 回归方法是尽可能准确地找到能够预测因变量的自变量，并给予各自变量一定权重。若自变量数量太少，则拟合的效果不好，不能很好地预测因变量的情况；若自变量太多，则会形成过拟合，预测因变量的效果同样不好。所以应减少一些自变量，如用虚拟变量表示不能量化的变量，用单变量和决策树分析筛选变量。与因变量相关性差不多的自变量可以归为一类，如地区对客户变坏概率的影响，假设广东和福建两省对坏客户的相关性分别为－0.381 和－0.380，可将这两个地区归为一类，另外，可以根据申请表上的信息构造一些自变量，例如结合申请表上的“婚姻状况”和“抚养子女”，根据经验和常识结合这两个字段，构造新变量“已婚有子女”，进入模型分析这个变量是否真正具有统计预测性。

6）模型建立

借助 SAS9 软件，用逐步回归法对变量进行筛选。这里设计了一种算法，分为 6 个步骤。

（1）求得多变量相关矩阵（若是虚拟变量，则＞0.5 属于比较相关；若是一般变量，则 0.7＜变量值＜0.8 属于比较相关）。

（2）旋转主成分分析（一般变量要求＞0.8 属于比较相关，虚拟变量要求＞0.6～0.7 属于比较相关）。

（3）在第一主成分和第二主成分分别找出 15 个变量，共 30 个变量。

（4）计算所有 30 个变量对好/坏的相关性，找出相关性大的变量加入步骤（3）得出的变量。

（5）计算 VIF（Variance Inflation Factor，方差膨胀因子）。若 VIF 数值比较大，则查看步骤（1）中的相关矩阵，并分别分析这两个变量对模型的作用，剔除相关性较小的

一个。

(6) 循环步骤(4)和步骤(5),直到找到所有变量,且达到多变量相关矩阵相关性强,而单个变量对模型的贡献作用大。

7) 模型验证

在收集数据时,把所有整理好的数据分为用于建立模型的建模样本和用于模型验证的对照样本。对照样本用于对模型总体的预测性、稳定性进行验证。申请评分模型的模型检验指标包括 K-S(Kolmogorov-Smirnov)值、ROC(Receiver Operating Characteristic,接受者操作特征)、AR(Association Rules,关联规则)等。虽然受到数据不干净等客观因素的影响,但是本例申请评分模型的 K-S 值已经超过 0.4,达到了可以使用的水平。

5. 数据挖掘在国内信用卡市场的发展前景

在国外,信用卡业务信息化程度较高,数据库中保留了大量的数据资源,运用数据技术建立的各类模型在信用卡业务中的实施非常成功。目前国内信用卡发卡银行首先利用数据挖掘建立申请评分模型,作为在信用卡业务中应用的第一步,不少发卡银行已经用自己的历史数据建立了客户化的申请评分模型。总体而言,在我国的信用卡业务中,由于数据质量问题,难以应用数据挖掘技术构建业务模型。

随着国内各家发卡银行已经建立或着手建立数据仓库,将不同操作源的数据存放到一个集中的环境中,并且进行适当的清洗和转换。这为数据挖掘提供了一个很好的操作平台,将给数据挖掘带来各种便利和功能。人民银行的个人征信系统也已上线,在全国范围内形成了个人信用数据的集中。在内部环境和外部环境不断改善的基础上,数据挖掘技术在信用卡业务中将具有越来越广阔的应用前景。

5.8　数据挖掘技术助力新冠病毒感染疫情防控

随着互联网的发展与普及,近年来网络与数据挖掘等技术成为推动社会发展及应对突发事件强有力的工具。新冠病毒感染疫情发生后,移动互联网以及物联网产生的海量数据在抗疫的诸多场景中发挥了显著作用,为疫情防控措施的有效实施提供了帮助。

1. 数字化技术全方位支援疫情防控

传染病流行的 3 个基本要素为传染源、传播渠道和易感人群。在本次抗疫过程中,我国采取了严格的管控措施、医疗检测与隔离手段:一是识别、定位早期症状人员,经筛查检测,尽快识别传染源;二是实施网格化管理,进行全民居家隔离,阻断相互接触的传染渠道,追溯并隔离近距离接触者群体,保护易感人群;三是举国驰援武汉,分类隔离疑似、轻症、重症人群,力争在最短时间内治愈确诊病例,切断传染源;四是积极研究病毒与疫苗,争取彻底解决病毒感染问题。

数字技术与数据应用在新冠疫情早期识别疑似病例的过程中，发挥了极大的作用。例如在湖北武汉实施拉网式全民体温筛查的同时，全国范围内公共场所的体温检测也被当作疫情防控的第一道防线予以贯彻。

（1）全国各大医疗机构、健康平台纷纷开放专门的线上疾病咨询渠道，方便群众在线问诊。

（2）武汉市狮南社区率先使用了以语音识别技术为基础的人工智能访谈系统，该系统通过拨打电话，针对联系人是否发热等基础问题实施居民调查。在精准识别、分析居民回应信息后，再对高风险人群进行人工访问和体温测量，迅速完成普筛工作。据统计，通过人工智能访谈系统，该社区仅用 6 小时便完成了对 3000 余户居民的识别与普筛工作。

（3）以无接触式体温感知技术为基础的体温感知系统，在办公楼宇、车站机场及交通要道等人流密集场所担负着对大批量人群进行体温筛查的工作，其应用大大减少了人工干预，实现了在高效获取数据的同时降低人群交叉感染的风险。

2. 助力人群定位与轨迹追踪

时空大数据的外延范围包括所有关于时间与位置的数据。以人的行为轨迹为对象的时空数据的应用在本次疫情防控过程中发挥了不可替代的作用。

（1）在疫情之初，中国联通、中国电信和中国移动三大电信运营商迅速以短信的形式为用户提供 14～30 天位置查询的服务，这也是健康码的雏形。而面向卫生防疫机构的人口迁移大数据则为疫情防控部门及时、准确地部署应对策略提供了基础数据。

（2）高铁、航空系统推出了通过输入电话号码查询一定时间段乘坐车次、搭乘航班的信息服务，结合政府不断公布的确诊人员的行动轨迹，实施确诊病例搭乘信息动态发布，方便群众判断是否曾与感染源有过接触，有效提高了易感人群的识别与隔离。

（3）描绘确诊病例分布的社区地图，为百姓及时了解周边疫情提供了直接有效的信息，是保证百姓生活节奏张弛有度的重要参考依据。

3. 助力医疗系统进行诊断与治疗

此次疫情呈现出不均匀分布的状态，大量病例的聚集地对医疗资源的供应提出了严峻挑战。除了医务人员驰援之外，数字化技术与数据挖掘应用也发挥了不可小觑的作用。

（1）在大部分新冠病毒感染定点收治医院，医疗服务智能机器人已承担起无人导诊、自动响应的发热问诊、初步诊疗、引领病人以及传送化验单和药物等多项辅助任务，隔离点的自动送餐机器人、消毒机器人等也极大地缓解了工作人员数量不足和情绪紧张的压力。

（2）通过信息技术手段调动全国各地，尤其是疫情较轻地区的医疗资源，以远程诊

断、多方会诊的方式帮助重灾区的医疗团队分担诊断、分析等工作，有效缓解了疫区医疗团队的工作负担。

4. 助力社区防控管理

疫情发生之后，全国范围内实行了严格的社区网格化管理，有效减少了人与人之间的相互接触。

(1) 适合不同应用场景的人员识别与登记系统。通过将居民微信或支付宝扫码与手机号、身份信息相结合，向社区工作人员提供特殊时期的特定人员的定位信息。同时通过无接触红外等测温手段完成对各类人员的体温监测。

(2) 基于微信、支付宝等开发出的“健康码”程序，有效提高了人员健康信息的记录与识别，助力加强高危人员的管控。

5. 助力中长期科研与药物研发

目前，国内很多科研院所以及从事人工智能医疗技术研究与应用的企业都在积极地开展药物与疫苗的研发、病毒的分析与测序等各项工作。以数据分析与机器学习为基础的数据技术在毒株的分析筛选、药物的分子结构和蛋白结构以及晶型分析等方面都将发挥无法替代的作用。阿里云宣布疫情期间向公共科研机构免费开放病毒疫苗和新药研发所需的一切人工智能算力，腾讯云为防治新冠病毒感染的药物筛选等工作提供免费的云超算服务，为病毒研究、药物筛选等科研工作提供了极大的算力保证。在现代化数据分析手段的协助下，军事科学院军事医学研究院陈薇院士领衔的科研团队已经成功研制出重组新冠病毒疫苗，并于 2020 年 3 月 16 日获批展开临床试验。

此外，复工复产、物流规划、物资调配和对国际客流的管理等很多基础性工作均需保持正常运行，释放出对数字化与数据应用最为直接的业务与管理的需求。

6. 数据深度应用前路漫长

新冠病毒感染疫情防控工作可谓是对我国数字化建设进行了一次全面而深刻的实战检验。通过分析发现，无论是时空数据的综合关联应用，还是基于移动互联网的其他应用，只有在技术相对成熟、应用场景简单完整，同时又无须大量额外基础设施投资的情况下才能真正地发挥作用。针对数据应用过程中面临的问题，笔者进行了初步总结，并从以下 5 个维度进行阐述。

1) 分隔管控措施造成数据挖掘有难度

从医学角度出发采取的一些有效措施并不利于数据的收集、分析与应用。例如，针对不同的人群采取的是不同的隔离管理措施：疑似与密切接触者进行居家或宾馆隔离，轻症患者在方舱医院进行管理观察，重症患者在专门的重症监护室或隔离病房抢救。此举能够最大化地发挥隔离、救治、防治交叉感染的效果。然而，这也导致了各种检验检测数据产生并存放在不同机构的异构数据集中，在缺乏信息共享联通机制的情况下，会给

数据分析研究带来很大的困难。

2）技术产品融入业务场景有难度

疫情之下，很多公司开展了人工智能自动或者辅助判读病情等系统研发工作。但是，医疗人员在如战场般的医疗现场和既定的医疗流程中争分夺秒，一个全新系统的植入将对既有业务流程造成冲击。而且，人与新工具的适应程度以及系统本身的不成熟等因素，均使得新型系统未能充分发挥真正有效的助力。

3）管理流程不适应战时需求

存放于不同位置、不同机构的数据安全管控措施严格，再加上异构数据的数据标准与格式存在巨大差异，使得数据很难实施有效的集成分析与应用。在紧急状态下，当数据融合成为必需的时候，便需要打破既有的壁垒，迅速形成综合的数据集合，提供最有效的数据支撑。然而，目前由于缺乏成熟的数据分享原则、授权管理机制，新冠病毒感染疫情防控期间的数据应用依然呈现比较分散、独立的态势。

4）数据的共享与分析存在壁垒

在医药研发领域，大部分大型公司与研究院所往往会进行独立研究。因此，在缺乏完善的共享机制的情况下，原始的检测数据、试验数据等会被各单位采取内部留存的方式管控，很难支撑产业的大协同研究与分析。

5）数据流动成为迫切需求

当前，我国的防疫压力得到缓解，但开始面临另外两个方面的压力：一方面是对内，有序地复工复产、经济重启均需要实现跨地域、跨行业的广泛信息的共享与连接；另一方面则是对外，输入病例防控工作显得尤为艰巨。因此，应急状态下的数据管理、管控与应用是未来需要深入研究分析的系统课题；促进数据研究成果与医学的深度融合，持续推进医工结合的数据研究与成果转化，是需要长期推进的艰巨使命；运用大数据与人工智能技术，对疫情所造成的损失进行有效评估，协助政府与企业制定并采取相应的经济恢复举措，逐步恢复正常的经济活动，防控输入病例、转阴复阳病例的二次传播，防止形成病毒的第二次大面积失控以及预防社会心理问题的集中爆发，都是需要我们认真思考的问题。

5.9 空间数据挖掘在地理信息系统中的应用

随着卫星和遥感技术以及其他自动数据采集工具的广泛应用，当前存储于空间数据库中的数据量迅速增长，海量的地理数据在一定程度上已经超过了人们能够处理的能力，特别是从这些海量的、多维的空间数据中提取有用的信息显得异常复杂，这就形成了“数据泛滥但信息匮乏”的尴尬局面。如何整理和解释这些数据，尽可能提取和发现地学信息，给当前地理信息系统(Geographic Information System，GIS)技术提出了新的挑战。

传统的GIS系统无法有效地发现大量数据中存在的关系和规则，很难把握数据背后隐藏的信息，而数据挖掘技术有望解决这一问题，它的出现为GIS组织、管理空间和非空间数据提供了新的思路，在一定程度上推动了地理信息系统的发展。

空间数据挖掘也称基于空间数据库的数据挖掘和知识发现(Spatial Data Mining and Knowledge Discovery)，是指从空间数据库中提取用户感兴趣的空间模式与特征、空间与非空间数据的普遍关系及其他一些隐含在数据库中的普遍的数据特征。空间数据挖掘是数据挖掘的一个新的分支。

空间数据挖掘系统大致为3层结构，如图5.1所示。

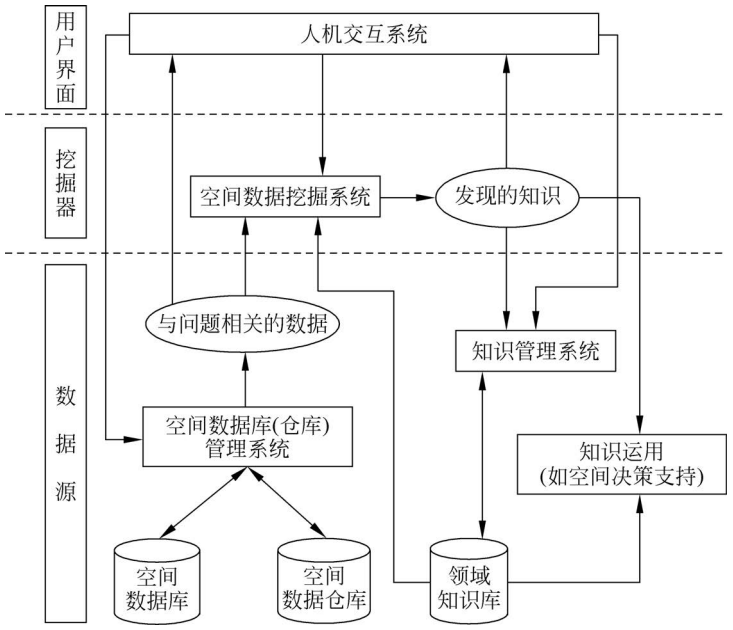

图5.1　空间数据挖掘系统

从图5.1可知，第一层是数据源，指利用空间数据库或数据仓库管理系统提供的索引、查询优化等功能获取和提炼与问题领域相关的数据，或直接利用存储在空间立方体中的数据，这些数据可称为数据挖掘的数据源或信息库。第二层是挖掘器，利用空间数据挖掘系统中的各种数据挖掘方法分析被提取的数据以达到用户的需求。第三层是用户界面，使用多种方式(如可视化工具)将获取的信息和发现的知识反映给用户，用户对发现的知识进行分析和评价，并将知识提供给空间决策支持使用，或将有用的知识存入领域知识库内。

常用的空间数据挖掘技术包括空间分析方法、统计分析方法、空间关联规则挖掘方法、聚类和分类方法、空间离群挖掘模式、时间序列分析、神经网络方法、决策树方法、粗糙集理论、模糊集理论、遗传算法、云理论等。

1. 空间分析方法

空间分析方法是利用GIS的各种空间分析模型和空间操作对空间数据库中的数据进行深加工，从而产生新的信息和知识。目前，常用的空间分析方法有综合属性数据分析、拓扑分析、缓冲区分析、密度分析、距离分析、叠置分析、网络分析、地形分析、趋势面分析、预测分析等，可发现目标在空间上的相连、相邻和共生等关联规则，或发现目标之间的最短路径、最优路径等辅助决策的知识。

空间分析方法常作为预处理和特征提取方法与其他数据挖掘方法结合使用。例如，探测性的数据分析（Exploratory Data Analysis，EDA）采用动态统计图形和动态链接技术显示数据及其统计特征，发现数据中非直观的数据特征和异常数据。Ester. Kriegel和Sander在空间数据库管理系统的基础上，基于邻图和邻径，提出了针对空间数据库的挖掘空间相邻关系的算法。邸凯昌把探测性的数据分析与空间分析相结合，构成探测性的空间分析（Exploratory Spatial Analysis，ESA），再次与面向属性的归纳（Attributed-Oriented Induction，AOI）结合，则形成探测性的归纳学习（Exploratory Inductive Learning，EIL），它们能在SDM（Spatial Data Mining）中聚焦数据，初步发现隐含在空间数据中的某些特征和规律。图像分析可直接用于含有大量图形图像数据的空间数据挖掘，也可作为其他知识发现方法的预处理手段。

2. 空间关联规则挖掘方法

空间关联规则挖掘方法（Spatial Association Rule Mining Approach）首先由Agrawal等提出，主要是从超级市场销售事务数据库中发现顾客购买多种商品时的搭配规律。最著名的空间关联规则挖掘算法是Agrawal提出的Apriori算法（R. Agrawal，1993），其主要思路是统计多种商品在一次购买中共同出现的频数，然后将出现频数多的搭配转换为关联规则。空间关联规则的形式是X－＞Y[S%，C%]，其中X、Y是空间或非空间谓词的集合，S%表示规则的支持度，C%表示规则的置信度。空间谓词的形式有3种：表示拓扑结构的谓词、表示空间方向的谓词和表示距离的谓词。各种各样的空间谓词可以构成空间关联规则。实际上，大多数算法都是利用空间数据的关联特性改进其分类算法的，这使得它适合挖掘空间数据中的相关性，从而可以根据一个空间实体而确定另一个空间实体的地理位置，有利于进行空间位置查询和重建空间实体等。

算法描述如下：

（1）根据查询要求查找相关的空间数据。

（2）利用邻近等原则描述空间属性和特定属性。

（3）根据最小支持度原则过滤不重要的数据。

（4）运用其他手段对数据进一步提纯。

（5）生成关联规则。

3. 聚类和分类方法

聚类是将地理空间实体或地理单元集合依照某种相似性度量原则划分为若干个类似地理空间实体或地理单元组成的多个类或簇的过程。类中实体或单元彼此间具有较高相似性，类间实体或单元彼此间具有较大差异性。常用的经典聚类方法有 K-Means、K-Medoids、ISODAIA 等。在空间数据挖掘中，R. Ng 等提出了基于面向大数据集的 CLARANS 算法；Ester 提出了 DBSCAN 算法；周成虎、张健挺等将信息熵的概念引入 SDM 中，提出了基于熵的时空一体化的地学数据分割聚类模型等。

分类就是假定数据库中的每个对象（在关系数据库中对象是元组）属于一个预先给定的类，从而将数据库中的数据分配到给定的类中。研究者根据统计学和机器学习提出了很多分类算法。大多数分类算法用的是决策树方法，它用一种自上而下分而治之的策略将给定的对象分配到小数据集中，在这些小数据集中，叶节点通常只连着一个类。许多研究者研究了空间数据的分类问题。Fayyad 等用决策树方法对恒星的影像数据进行了分类，总共有 3TB 的栅格数据。训练数据集由天文学家进行分类，在此基础上，建立了用于决策树分类的 10 个训练数据集，接着用决策树进行分类，发现模式，这个方法不适合 GIS 中的矢量数据。Ester 等利用邻近图提出了空间数据的分类方法，该方法是基于 ID3 而来的，它不但考虑了分类对象的非空间属性，而且考虑了邻近对象的非空间属性。K. Koperski 等用决策树进行空间数据分类，接着分析了空间对象的分类问题和数据库中空间对象之间的关系，最后提出了一个能处理大量不相关的空间关系的算法，并针对假设和真实数据进行了空间数据分类实验。

分类和聚类都是对目标进行空间划分，划分的标准是类内差别最小而类间差别最大。分类和聚类的区别在于分类事先知道类别数和各类的典型特征，而聚类则事先不知道。

4. 空间离群挖掘模式

离群点就是不同于邻近域属性值的目标对象，或者由于其特殊的应用价值，一些学者认为它是由某种特有的机制产生的。离群点的识别能够导致很多有意义知识的挖掘，其应用范围也很广，例如运动员体能分析、天气预报、计算机辅助设计等。从空间意义上来说，发现局部异常对象是极其重要的。空间离群点就是在空间上非空间属性显著不同于空间邻近域的目标对象。有时，空间离群点在整个数据集合上并不是那么显著的，但是对于局部而言就是一个不稳定点，挖掘空间离群点在很多程序中都有应用，如地理信息系统、交通运输等领域。

近来，为了在多维空间中挖掘目标对象，提出了许多双边分裂多维测试离群点算法。它们把自身的属性分为空间属性（地点、邻近域属性和距离等）和非空间属性（对象编号、对象的从属者以及名称等）。其中，空间属性用来定位对象之间的关系以及邻近域集合

的选择，而非空间属性用来比较目标对象与其邻近域集合。从空间统计学的角度出发可以把它分为两类：图形方法和定量测试方法。图形方法就是基于空间数据可视化来区分空间离群点的，例如变量云图方法、Morancsatterplots 方法、Shekhar 等提出的基于图形的空间数据挖掘算法等（S. Shekhar，2002）。而定量测试方法提供了一种在其邻近域中准确地挖掘目标对象的方法，如 Scatterplots（A. Luc，1995）以及 Chang-TienLu 的空间离群点定量测试算法。它们都从非空间属性差值的空间统计分布出发，对一维的非空间属性值进行统计判断，有效地提高了空间离群点判断的准确性。

GIS 发展的重要趋势是与遥感（Remote Sensing，RS）和全球定位系统（Global Positioning System，GPS）相结合，向集成化、自动化及智能化迈进。GIS 发展的另一个重要方向是智能化的决策支持系统，这都需要用到专家系统的知识。因此，知识的自动获取是建立智能化 GIS 的瓶颈。由于当前空间数据模型、空间数据结构及 GIS 数据库管理系统的多样性，致使三者的集成不那么轻而易举，但集成的关键是如何从 GIS 中获取样本数据。下面针对常用的扩展式 GIS 数据库管理系统提出 3 种集成模式。

1）松散耦合模式，也称外部空间数据挖掘模式

这种模式基本上将 GIS 当作一个空间数据库看待，在 GIS 环境外部借助其他软件或计算机语言进行空间数据挖掘。它与 GIS 之间采用数据通信的方式联系，而 GIS 只充当数据源的作用。由于这种模式是基于内存的，挖掘本身并不使用 GIS 数据库系统提供的数据结构和查询优化方法，因此，对于大数据集，松散耦合模式的系统很难获得可伸缩性和良好的性能。它的优点是集成能方便灵活地实现。图 5.2 为基于松散耦合模式的空间数据挖掘与 GIS 的集成框架图。

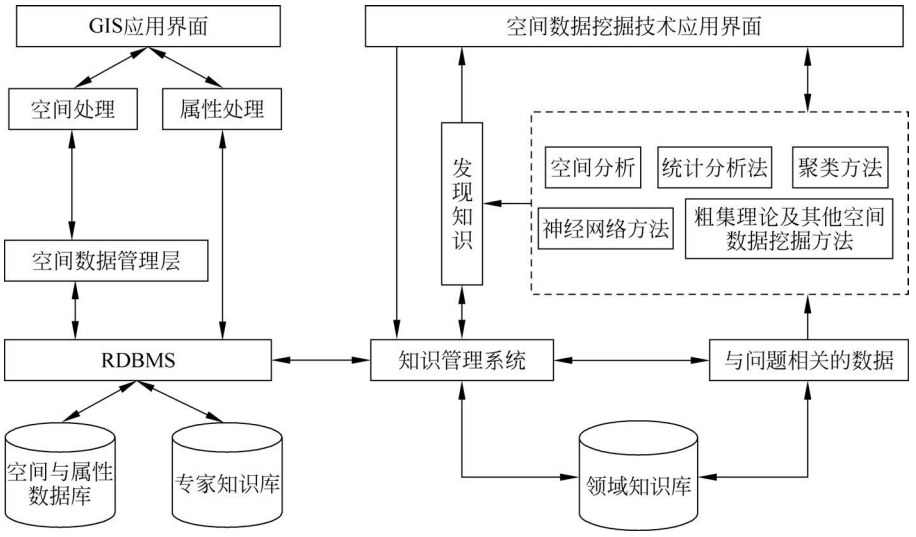

图 5.2　基于松散耦合模式的空间数据挖掘与 GIS 的集成框架图

2）内部空间数据挖掘模式

这种模式把数据挖掘子系统视为地理信息系统的一部分，就像GIS其他功能模块一样。SDM预处理模块的功能将并入GIS的数据库管理模块，数据挖掘的知识库成为GIS数据库管理模块下的一个子库，结果由系统界面显示与表达，数据挖掘方法库及管理模块形成类似于空间查询与空间分析的模块，通过把数据挖掘查询优化成循环的挖掘处理和检索过程，将二者结合起来，实现数据挖掘系统和GIS的紧密结合，融为一体，达到高层次上的集成，这也是完善和发展GIS的方向。图5.3为基于嵌入模式的空间数据挖掘与GIS的集成框架图。

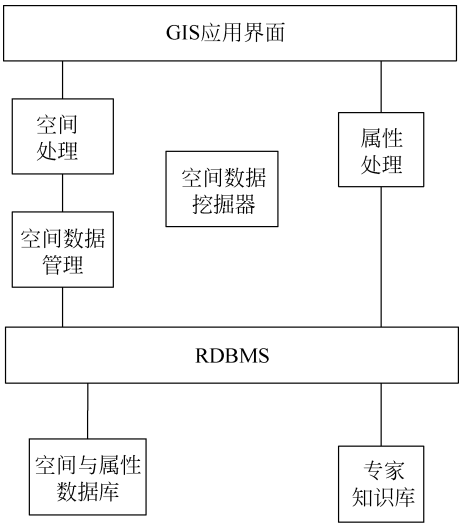

图5.3　基于嵌入模式的空间数据挖掘与GIS的集成框架图

3）混合型空间模型法

混合型空间模型法是前两种方法的结合，即尽可能利用GIS提供的功能，最大限度地减少用户自行开发的工作量和难度，又保持外部空间数据挖掘模式的灵活性。

5. 空间数据挖掘可发现的主要知识类型

GIS数据库是空间数据库的主要类型，从GIS数据库中发现的知识类型及知识发现方法可以涵盖其他类型的空间数据库。利用空间数据挖掘技术可以从空间数据库中发现如下几种主要类型的知识。

1）空间特征规则

空间特征规则是指对某类或几类空间目标的普遍特性的描述规则，即某类空间目标的共性。空间几何特征是指目标的位置、形态特征、走向、连通性、坡度等普遍的特征。空间属性特征指目标的数量、大小、面积、周长等定量或定性的非几何特性。这类规则是最基本的，是发现其他类型知识的基础。例如河流与山脉的走向、道路的连通性等。

2）空间分布规律

空间分布规律是指地理目标（现象）在地理空间的分布规律，分为水平向分布规律、垂直向分布规律、水平和垂直向的联合分布规律以及其他分布规律。水平向分布指地物（现象）在水平区域的分布规律，如不同区域农作物的差异、公用设施的城乡差异等；垂直向分布即地物沿高程带的分布，如高山植被沿坡度、坡向的分布规律。

3）空间聚类规则

空间聚类规则是指根据空间目标特征的集散程度将它们划分到不同的组中，组之间的差别尽可能大，组内的差别尽可能小，可用于空间目标信息的概括和综合，如精确农业中的作物产量图可聚类成高、中、低产区。

4）空间演变规律

若 GIS 数据库是时空数据库或 GIS 数据库中存有同一地区多个时间数据的快照，则可以发现空间演变规律。换言之，空间演变规律是指空间目标依时间的变化规律，如哪些地区易变，哪些地区不易变、怎么变，哪些目标固定不变等，人们可以利用这些规律进行预测预报。

5）空间分类规则

空间分类规则是指根据目标的空间或非空间特征，利用分类分析将目标划分为不同类别的规则。空间分类是有导师的，并且事先知道类别数和各类的典型特征。

6）空间序贯模式

空间序贯模式是指空间数据库中满足用户指定最低支持的最小的空间数据时间序列或属性数据时间序列。

7）空间混沌模式

空间混沌模式是指空间数据库的空间数据、属性数据中存在介于确定关系和纯随机关系间的混沌关系，是一种无序中的有序关系。

8）面向对象的知识

面向对象的知识是指由某类复杂对象的子类构成，并具有普遍特征的知识。可用的知识表达方法包括：特征表、谓词逻辑、产生式规则、语义网络、面向对象的表达方法、可视化表达方法等。

9）空间偏差型知识

空间偏差型知识是对空间目标之间的差异和极端特例的描述，揭示空间目标或现象偏离常规的异常情况，如空间聚类中的孤立点和空洞。这些知识和规则从信息内涵上讲是有区别的，但从形式上讲又是密切联系的。对于空间分布的图形描述既传递了空间分布信息，又传递了空间趋势和空间对比信息。例如从世界人口分布图上，我们既可以了解人口分布情况，又可以感受到人口分布的基本趋势，同时，各国之间的人口密度对比也反映得一清二楚。我们在不同的应用中可选择相应的知识表达方法，各种方法之间也可

以相互转换。

地理信息系统是空间数据库发展的主体，随着计算机科学、地理学、统计学、环境科学、遥感技术以及移动通信等综合信息技术的不断发展，GIS 技术将不断与新的领域结合产生新的集成，发挥更大的作用。同时，从空间数据库中挖掘出有意义的隐含的知识将越来越受到人们的重视，空间数据挖掘技术在 GIS 中的广泛应用，将使得 GIS 集成系统朝着网络化、智能化、标准化、全球化与大众化的方向发展，充分、恰当地发挥 GIS 的潜能，使之更好地为人类的生活服务。

5.10　数据挖掘技术在个性化推荐系统中的应用

个性化推荐系统是互联网和电子商务发展的产物，它是建立在海量数据挖掘基础上的一种高级商务智能平台，向顾客提供个性化的信息服务和决策支持。近年来已经出现了许多非常成功的大型推荐系统实例，与此同时，个性化推荐系统也逐渐成为学术界的研究热点之一。

1995 年 3 月，卡内基-梅隆大学的 Robert Armstrong 等在美国人工智能协会上提出了个性化导航系统 Web Watcher，斯坦福大学的 Marko Balabanovic 等在同一会议上推出了个性化推荐系统 LIRA。

1995 年 8 月，麻省理工学院的 Henry Lieberman 在国际人工智能联合大会(International Joint Conference on Artificial Intelligence，IJCAI)上提出了个性化导航智能体 Litizia。

1996 年，Yahoo 推出了个性化入口 My Yahoo。

1997 年，AT&T 实验室提出了基于协同过滤的个性化推荐系统 PHOAKS 和 Referral Web。

1999 年，德国 Dresden 技术大学的 Tanja Joerding 实现了个性化电子商务原型系统 TELLIM。

2000 年，NEC 研究院的 Kurt 等为搜索引擎 CiteSeer 增加了个性化推荐功能。

2001 年，纽约大学的 Gediminas Adoavicius 和 Alexander Tuzhilin 实现了个性化电子商务网站的用户建模系统 1:1Pro。

2001 年，IBM 公司在其电子商务平台 Websphere 中增加了个性化功能，以便商家开发个性化电子商务网站。

2003 年，Google 开创了 AdWords 广告模式，通过用户搜索的关键词来提供相关的广告。AdWords 的点击率很高，是 Google 广告收入的主要来源。2007 年 3 月开始，Google 为 AdWords 添加了个性化元素。不但关注单次搜索的关键词，而且对用户近期的搜索历史进行记录和分析，据此了解用户的喜好和需求，更为精确地呈现相关的广告

内容。

2007 年，Yahoo 推出了 SmartAds 广告方案。雅虎掌握了海量的用户信息，如用户的性别、年龄、收入水平、地理位置以及生活方式等，再加上对用户搜索、浏览行为的记录，使得雅虎可以为用户呈现个性化的横幅广告。

2009 年，Overstock（美国著名的网上零售商）开始运用 ChoiceStream 公司制作的个性化横幅广告方案在一些高流量的网站上投放产品广告。Overstock 在运行这项个性化横幅广告的初期就取得了惊人的成果，公司称："广告的点击率是以前的 2 倍，伴随而来的销售增长也高达 20%～30%。"

2009 年 7 月，国内首个个性化推荐系统科研团队北京百分点信息科技有限公司成立，该团队专注于个性化推荐、推荐引擎技术与解决方案，在其个性化推荐引擎技术与数据平台上汇集了国内外百余家知名电子商务网站与资讯类网站，并通过这些 B2C 网站每天为数以千万计的消费者提供实时智能的商品推荐。

2011 年 8 月，纽约大学个性化推荐系统团队在杭州成立载言网络科技有限公司，在传统协同滤波推荐引擎的基础上加入用户的社交信息和用户的隐性反馈信息，包括网页停留时间、产品页浏览次数、鼠标滑动、链接点击等行为，辅助推荐，提出了迄今为止最为精准的基于社交网络的推荐算法。团队专注于电商领域个性化推荐服务以及商品推荐服务社区——e 推荐。

2011 年 9 月，百度世界大会 2011 上，李彦宏将推荐引擎与云计算、搜索引擎并列为未来互联网重要战略规划以及发展方向。百度新首页将逐步实现个性化，智能地推荐出用户喜欢的网站和经常使用的 App。

个性化推荐最初的诞生是由于在逐渐信息过载的时代中，适当的筛选可以让用户高效地获得自己所需要的信息。后来才逐步应用于商业，尤其是成为电商行业的有效销售手段，还有一些文化、社交性的站点（比如豆瓣、知乎、网易云等）。

推荐系统是自动联系用户和物品的一种工具，它通过研究用户的兴趣爱好来进行个性化推荐。它与搜索引擎的不同在于，它不需要用户提供输入目标，而是基于历史记录自动推荐，是一种主动的机制。它能够通过分析用户的历史行为来对用户的兴趣进行建模，从而主动给用户推荐可满足他们兴趣和需求的信息。每个用户所得到的推荐信息都是与自己的行为特征和兴趣有关的，而不是笼统的大众化信息，因此称之为"个性化"。

关于推荐引擎的工作原理，首先它需要得到一些基本信息，主要包括：

（1）要推荐的内容的元数据，如关键字。

（2）用户的基本信息，如性别、年龄、职业。

（3）用户的偏好，偏好信息又可以分为显式用户反馈和隐式用户反馈。显式用户反馈是用户在网站上自然浏览或者使用网站以外，显式地提供的反馈信息，如用户对物品的评分或者对物品的评论等。

隐式用户反馈是用户在使用网站时产生的数据，隐式地反映了用户对物品的喜好，如用户购买了某物品，用户查看了某物品的信息，用户在某页面停留的时间等。推荐引擎通过对这些信息的统计分析关联，再给用户个性化地推荐相应物品或信息。

对于当前大部分的推荐机制可以进行以下分类：

（1）基于人口统计学的推荐，即根据用户个人的基本数据信息来发现用户的相关程度。

（2）基于内容的推荐，即根据不同内容的元数据进行内容相关性的分析。

（3）根据协同过滤的推荐，通过对用户偏好信息的过滤发现不同内容的相关性或者不同用户的相关性。

这些数据挖掘相关技术已经在很多领域取得了成就，譬如推荐系统应用的鼻祖Amazon，就是通过消费偏好对比以及一些混合手法来对用户进行精准的页面推荐，现在的淘宝、京东、天猫等电商平台显然也采用了这种方式进行个性化推荐。个性化的流量分配可以最大化流量的使用效率，这使得它们的获客成本居高不下。

电商领域的个性化推荐也面临以下挑战：由于推荐是基于已有信息对用户意图与心理进行的猜测，及时识别用户每个行为背后的真实意图，甚至每个页面、每个标题对用户心理的影响就十分重要，这些关键的影响因素可能是一张购物券、一张明星街拍、一个偶遇的促销活动，尤其是激情消费易发的当下。这里面涉及较为复杂的用户购物状态的推理和判定，如果不借助人工输入，比如通过产品设计提供用户筛选接口，让用户人工输入限制项，典型的比如过滤器、负反馈等，则对目前的机器算法来说是一个非常大的挑战。

还有一个问题是用户体验问题。这些平台，乃至个性化推荐的算法，本质上都是为了用户服务的。可以看到，常常被抱怨的体验问题包括买了还推，推荐商品品类单一，没有让人眼前一亮的商品能满足一下发现的惊喜等，不一而足。往往这些体验问题的解决都需要人工规范的干预，但凡有规则的介入，比如加入购买过滤、类目打散展示等策略，都会造成交易类指标的下降，平衡两者之间的关系对推荐系统是一个现实的挑战。

个性化推荐在其他领域的应用也面临着类似的问题。例如基于人口统计学的推荐机制基于用户的基本信息对用户进行分类的方法过于粗糙，尤其是对品位要求较高的领域，如图书、电影和音乐等领域，无法得到很好的推荐效果。基于内容的推荐需要对物品进行分析和建模，推荐的质量依赖于物品模型的完整和全面程度；对于物品相似度的分析仅依赖于物品本身的特征，而没有考虑人对物品的态度；因为是基于用户以往的历史做出推荐，所以对于新用户有“冷启动”的问题等。还有协同推荐的效果过于依赖用户历史偏好数据的多少和准确性；对于一些特殊品位的用户不能给予很好的推荐；由于以历史数据为基础，因此抓取和建模用户的偏好后，很难修改或者根据用户的使用进行演变，从而导致这个方法不够灵活。

当然，现在大多流行的是混合型推荐，可能把一种推荐机制的输出当作输入送入另

一种机制中，或者把不同机制得到的推荐结果都推荐给用户，这些也是能够有效提高推荐效果的。

随着推荐技术的研究和发展，其应用领域越来越多。例如，新闻推荐、商务推荐、娱乐推荐、学习推荐、生活推荐、决策支持等。推荐方法的创新性、实用性、实时性、简单性也越来越强。例如，上下文感知推荐、移动应用推荐、从服务推荐到应用推荐。下面分别分析几种技术的特点及应用案例。

1. 新闻推荐

新闻推荐包括传统新闻、博客、微博、RSS 等新闻内容的推荐，一般有以下 3 个特点：

（1）新闻的事件时效性很强，更新速度快。

（2）新闻领域的用户更容易受流行和热门的事件影响。

（3）新闻领域推荐的另一个特点是新闻的展现问题。

2. 电子商务推荐

电子商务推荐算法可能会面临各种难题，例如：①大型零售商有海量的数据、数以千万计的顾客以及数以百万计的登记在册的商品；②实时反馈需求，在半秒之内，还要产生高质量的推荐；③新顾客的信息有限，只能以少量购买或产品评级为基础；④老顾客信息丰富，以大量购买和评级为基础；⑤顾客数据不稳定，每次的兴趣和关注内容差别较大，算法必须对新的需求及时响应。

解决电子商务推荐问题通常有 3 个途径：协同过滤、聚类模型以及基于搜索的方法。

3. 娱乐推荐

音乐推荐系统的目标是基于用户的音乐口味向终端用户推送喜欢和可能喜欢但不了解的音乐。而音乐口味和音乐的参数设定是受用户群特征和用户个性特征等不确定因素影响的。例如对年龄、性别、职业、音乐受教育程度等的分析能够帮助提升音乐推荐的准确度。部分因素可以通过使用类似 FOAF 的方法来获得。

总而言之，个性化推荐是日常生活中最能体现数据挖掘的应用实例之一，人们对它的研究已经很多年了，而且还将基于社会文化的不断变迁继续发展下去。

5.11 数据挖掘技术在证券行业中的应用

在券商企业多年来的运营中，积累了大量投资者真实的第一手买卖金融产品数据，近年互联网金融的发展加速了各类运营数据的产生，也让数据真正成为价值的核心，数据成为数据资产。数据资产的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行分析和挖掘，找出其中蕴含的价值，助推证券行业的业务创新、服务创新、产品创新。本节在简要介绍数据挖掘技术的基础上，探讨证券数据挖掘的方法论和

挖掘方向，并结合华泰证券的数据挖掘实践证明，数据分析和挖掘确能给企业的业务发展提供有益的帮助。

证券市场是国家经济的晴雨表，国家经济的细微波动都会在证券市场及时地反映出来。因而证券业的经营对数据的实时性、准确性和安全性的要求都很高。在国内证券行业领域政策日趋开放的大环境下，证券业的竞争也越来越激烈。这就要求证券公司在做分析决策时不仅需要大量数据资料，更需要通过数据发掘其运行规律和未来走势。

数据挖掘技术在证券领域中的应用，就是将证券交易及证券活动中所产生的海量数据及时提取出来，通过清洗和变换，采用分类、聚类、关联分析等方法发现新知识，及时为证券从业人员提供参考咨询服务，分析客户交易行为，掌握企业经营状况，控制证券交易风险，从而帮助从业人员在证券交易中增强决策的智能性和前瞻性。

1. 证券数据挖掘方法论

1）证券数据的特点

与其他领域的数据相比较，证券数据具有很多特点。

（1）证券数据具有多样性。作为社会经济系统的一部分，证券系统的数据不仅受客户数据、交易数据、经济数据等的影响，而且受网络信息、心理行为信息的强烈影响，甚至一些主观数据的变化也会导致证券市场的剧烈波动。

（2）证券数据的关系复杂。证券市场是一个复杂系统，数据之间的关系有时很难用一个简单的数学公式或者线性函数来表示，呈现出高度的复杂性和非线性。

（3）证券数据具有动态性。证券市场随着时间的推移会发生剧烈变化，但仍受前期市场的影响，呈现出动态特征。

为了更好地研究证券市场，需要利用这些物理数据、网络信息及心理行为信息。由于这些信息是不断变化的，因此形成了一个巨大的数据仓库。证券数据的高度复杂性使得一般的数据建模方法在进行金融数据建模时失效，而数据挖掘方法具有灵活性、自适应性及非线性等特征，因此在处理证券数据时可以达到较好的应用效果。

证券行业的数据仓库是由证券交易过程中的基础数据（主要是数据库数据）组成的。证券行业的基础数据主要包括以下 4 部分。

（1）业务数据。业务数据包括结算数据、过户数据、交易系统数据。结算数据是由深圳和上海证券登记公司以交易席位为单位发布的证券公司当日资金、股份交收明细以及分红、送股、配股等数据。过户数据是由深圳和上海证券交易所以交易席位为单位发布的证券公司当日投资者买卖证券的过户明细数据。结算数据和过户数据由证券交易所通过地面和卫星网络系统发送到证券公司。交易系统数据是证券公司最重要和最实时的数据。它由交易系统在实时交易中产生，是进行数据挖掘、客户分析、构建 CRM 系统的主要基础数据。

（2）行情数据。行情数据是由深圳、上海证券交易所在开市期间发布的证券实时交

易的成交撮合数据，是进行股市行情分析的关键数据。

（3）证券文本数据。狭义的证券文本数据是指由证券交易所通过证券卫星发送的证券领域有关政策和各股资讯等实时信息。广义的证券文本数据是指由各种传媒方式发布的与证券相关的信息，主要包括卫星、电视、广播、因特网、移动互联网、书刊杂志等传媒方式，其中，因特网和移动互联网是涵盖信息量最多的传媒方式。

（4）用户和客户行为数据。移动互联网及互联网金融的发展使得证券服务的外延得到了很大的扩展，不但证券公司开户的用户能使用证券公司的服务，不在证券公司开户的用户也能通过多种形式（如证券软件、证券互联网、证券移动应用等）获取证券公司提供的部分产品服务。用户和客户在使用这些软件产品的过程中，会产生很多的行为数据，如浏览路径、浏览兴趣、停留时间等。

2）证券数据挖掘方向探索

根据证券业务与数据特点，可以实施的挖掘方向有：客户分析、客户管理、证券营销、财务指标分析、交易数据分析、风险分析、投资组合分析、用户行为分析等。下面简要介绍各个方向的思路。

（1）客户分析及营销。通过数据进行挖掘和聚类分析，可以清晰发现不同类型客户的特征，挖掘不同类型客户的特点，提供不同的服务和产品。反过来，如果我们知道了客户的特征与偏好，有针对性地设计新的产品和服务，势必能获得更好的推广效果。

通过对客户资源信息进行多角度挖掘，了解客户各项指标，掌握客户投诉、客户流失等信息，从而在客户离开券商之前捕获信息，及时采取措施挽留客户。

通过对客户交易行为的分析与挖掘，了解客户的交易行为、方式、风险偏好，从而提升交叉营销的成功率，同时结合挖掘结果，给客户提供更加贴心的服务，提升客户忠诚度。

（2）用户行为分析。通过对证券软件、证券互联网、证券移动终端开放用户使用行为的分析和挖掘，了解用户的兴趣点、访问规律，为用户转换为客户提供目标人群，提高用户转换为客户的成功率；同时，利用访问模型改进软件和网站的布局，提升软件和网站的人性化设计。

（3）市场预测。对股票的基本面、消息面、技术指标等数据进行聚类分析，从而将股票划分为不同的群体，预测板块轮动或未来走势。

根据采集行情和交易数据，结合行情分析，预测未来大盘走势，发现交易情况随着大盘变化的规律，并根据这些规律做出趋势分析，对客户进行针对性咨询。

（4）投资组合。利用数据挖掘技术不仅可以更好地刻画预期的不确定性，改进已有的投资组合模型，使之更加符合现实需求，同时可以为投资组合模型的求解提供更为精确的手段，从而为投资者提供更为精准的知识。

（5）风险防范。通过对资金数据的分析，可以控制营业风险，同时可以改变公司总部

原来的资金控制模式，并通过横向比较及时了解资金情况，起到风险预警的作用。

（6）经营状况分析。通过数据挖掘，可以及时了解营业状况、资金情况、利润情况、客户群分布等重要的信息，并结合大盘走势，提供不同行情条件下的最大收益经营方式。同时，通过对各营业部经营情况的横向比较，以及对本营业部历史数据的纵向比较，对营业部的经营状况做出分析，提出经营建议。

3）华泰证券数据挖掘实施业务流程

华泰证券数据挖掘实施业务流程如下：

（1）项目背景和业务分析需求提出。

针对需求收集相关的背景数据和指标，与业务方一起熟悉背景中的相关业务逻辑，并收集业务方对需求的相关建议、看法，这些信息对于需求的确认和思路的规划乃至后期的分析都是至关重要的。从数据分析的专业角度评价初步的业务分析需求是否合理，是否可行。

（2）制定需求分析框架和分析计划。

针对前面对业务的初步了解和需求背景的分析，我们制定了以下初步的分析框架和计划：分析需求转换成数据分析项目中目标变量的定义，分析思路的大致描述，分析样本的数据抽取规则，根据目标变量的定义选择一个适当的时间窗口，然后抽取一定的样本数据，潜在分析变量（模型输入变量）的大致圈定和罗列，分析过程中的项目风险思考和主要应对策略，项目落地应用价值分析和展望。

（3）抽取样本数据，熟悉数据，数据预处理。

根据前期讨论的分析思路和建模思路，以及初步圈定的分析字段（分析变量）编写代码，从数据仓库中提取分析、建模所需的样本数据；通过对样本数据的熟悉和摸底，找到无效数据、脏数据、错误数据等，并且对样本数据中存在的这些明显的数据质量问题进行清洗、剔除、转换，同时视具体的业务场景和项目需求，决定是否产生衍生变量，以及怎样衍生等。

（4）按计划初步搭建挖掘模型。

对数据进行初步的摸底和清洗之后，就进入初步搭建挖掘模型阶段了。在该阶段，包括3个主要的工作内容：进一步筛选模型的输入变量；尝试不同的挖掘算法和分析方法，并比较不同方案的效果、效率和稳定性；整理经过模型挑选出来的与目标变量的预测最相关的一系列核心输入变量，将其作为与业务方讨论落地应用的参考和建议。

（5）讨论模型的初步结论，提出新的思路和模型优化方案。

整理模型的初步报告、结论，以及对主要预测字段进行提炼，还要通过与业务沟通和分享，在此基础上讨论出模型的可能优化方向，并对落地应用的方案进行讨论，同时罗列出注意事项。

（6）按优化方案重新抽取样本并建模，提炼结论并验证模型。

在优化方案确定的基础上，重新抽取样本，一方面验证之前优化方向的猜想，另一方面尝试搭建新的模型提升效果。模型建好后，还不能马上提交给业务方进行落地应用，还必须用最新的实际数据来验证模型的稳定性。如果通过相关验证得知模型的稳定性非常好，那么无论对模型的效果还是项目应用的前景，都有比较充足的底气了。

（7）完成分析报告和落地应用建议。

在上述模型优化和验证的基础上，我们提交给业务方一份详细完整的项目结论和应用建议。该建议包括以下内容：

- 模型的预测效果和效率，以及在最新的实际数据中验证模型的结果，即模型的稳定性。
- 通过模型整理出来的可用作运营参考的重要自变量及相应的特征、规律。
- 数据分析师根据模型效果和效率提出的落地应用的分层建议，以及相应的运营建议，包括：预测模型打分应用基础上进一步的客户特征分层建议、相应细分群体运营通道的选择建议、运营文案的主题或噱头建议、运营引导方向和目的建议、对照组与运营组设置建议、效果监控方案等。

数据分析师进一步的相关建议如下。

① 制定具体的落地应用方案和评估方案。

与业务方讨论，确定最终的运营方案及评估方案。业务方实施落地应用方案并跟踪、评估效果，按照上述的运营和监控方案对运营组和对照组进行分层的精细化运营，取一段时间如一周的运营结论，主要从两个方面来衡量：一是预测模型的稳定性评测；二是运营效果。

② 落地应用方案在进行实际效果评估后，不断修正完善。

通过对第一次运营效果的评估和反思，从正反两个方面进行总结，如果模型稳定性好，有较好的预测效果，那么可以放心使用模型，优化运营方案。

③ 不同运营方案的评估、总结和反馈。

根据实际情况制定多种运营方案，监控不同运营方案的执行情况及效果。

2. 华泰证券数据挖掘实践

华泰证券一直重视数据资产的价值发现，在数据分析与挖掘方面做了很多的技术储备和实践。在对华泰证券某集合理财产品的销售数据分析中，我们通过数学方法结合数据挖掘软件建立了预测模型，验证了模型的有效性，并且通过模型获得了很好的预期提升效果。主要步骤如下。

1）数据准备

首先，确定合适的观察期。在从数据中心提取观察期内的原始数据后，进行数据预处理，例如剔除资产过小的客户、剔除长时间无主动交易的客户、剔除机构客户等，得到规模为50多万条记录的初始数据集。

2）变量分析与数据抽样

由于初始数据集是一个包含较多属性的宽表，因此，为了选取主要变量，舍弃无关变量，减少变量数目，以利于实施数据挖掘算法，我们进行了以下的变量分析处理：

（1）对属性定义一个被称为信息值（Information Value，IV）的变量，计算每个属性的信息值。该值越大，表示对结果的影响越大，该变量越重要；该值越小，则认为可舍弃该变量。

（2）为应用 Logistic 分析，将上述步骤中的连续性变量进行分段，再一次计算信息值并舍弃区分度不高的变量。

（3）利用 Stepwise Logistic 方法结合默认的概率值确定入选变量和剔除变量。

（4）对变量进行主成分分析，进一步挑选较少个数的重要变量。

（5）在确定入选变量后，将数据集按比例分为建模数据集与验证数据集，并对建模数据集进行过抽样，以减少建模记录数并提高事件率，验证数据集则用于对将要生成的模型进行验证。

3）建立模型

针对上述建模数据集，采用 Logistic 回归建模，将结果输出至结果集。

4）模型验证与结果展示

对验证集进行单因子非参数方差分析，即 npar1way 过程，得到 K-S 检验值 0.619，大于 0.05，则可认为验证集服从建模集的数据分布，即由建模集生成的模型是有效的。

随着互联网、移动互联网的发展，证券行业信息化的应用环境正在发生着深刻的变化，外部数据迅速扩展，企业应用和互联网应用的融合越来越快。互联网金融给证券行业带来的传统价值创造和价值实现方式的根本性转变，让数据分析和挖掘逐步走向证券业务发展和创新的前台。相信随着金融互联网的多样化，证券行业内外数据的不断完备，数据分析和挖掘将在证券行业的运用越来越广泛，并成为证券公司数据化运营的一部分。

5.12　数据挖掘技术在钢铁行业质量管理中的应用

钢铁行业是一个资源密集型、资金密集型行业，其生产过程主要呈现出生产流程长、自动化程度高、质量要求高的特点。一方面，我国钢铁行业由高速发展转向高质量发展阶段，在环保、质量等多方面均有更高的要求。另一方面，由于钢铁行业市场需求转变，生产特点逐渐由传统的大规模批量生产向多品种小批量定制开发模式转变。这导致钢铁企业的产品设计、原材料选择、质量把控等产品生产的质量管理周期进一步缩短，因此钢铁产品质量管理凸显出越来越重要的作用。

随着云计算、大数据、物联网等新兴技术的不断发展成熟，基于工业大数据构建全流

程质量管理体系，对生产过程中的质量数据进行收集、存储、分析、预警等，深入挖掘产品质量相关信息，发现隐匿在数据背后的一些规律性、趋势性关系，可以更加科学地指导产品生产、质量标准修订以及其他产品质量管理工作。因此，运用大数据技术推进企业全面质量管理日益受到钢铁行业的青睐。

1. 钢铁质量管理存在的问题

产品质量管理是随着现代化生产的发展而逐步形成和发展起来的，目前已经发展到全面质量管理阶段，由企业的全体人员参加，运用现代化科学和管理技术，预先把整个生产过程中影响产品质量的各种因素加以控制，从而保证和提高产品质量，使用户得到最满意的产品。在产品质量管理过程中，存在着大量未能充分挖掘的数据价值，造成钢铁行业质量大数据资源的浪费。其存在的问题主要有信息孤岛、数据质量低下、存储机制落后、数据价值利用低等。

1）数据采集信息孤岛问题严重

目前，大部分钢铁企业都在产线上加装了各类传感器、监测仪等相关设备，以实现生产过程的相关参数检测和数据存储。但是，由于受缺乏统筹、分步上线以及传统信息技术限制等因素的影响，各产线往往建有单独的数据收集系统，且各系统之间缺乏有效的相互沟通及数据共享，造成上下游数据孤岛问题严重。下游生产环节无法及时得到上游质量数据，造成生产成本、废品率居高不下。因此，质量数据的生产全流程管控、上下游数据充分共享对于建设大数据质量分析具有非常重要的作用。

2）数据质量控制面临巨大挑战

在数据质量方面，钢铁行业的质量大数据存在着大数据的普遍特征，即“二八定律”，也就是20%的结构化数据占有80%的价值，而80%的非结构化数据占有20%的价值。在工业大数据的收集、处理过程中，由于数采系统链路、硬件故障、人为因素等主客观因素的影响，数据质量问题广泛存在。这些数据质量问题可能导致大数据分析结果的偏差，从而不利于质量管理的有效分析。

3）海量数据无法确保有效存储

钢铁行业内各种状态监测仪器逐渐向多功能、系统化、智能化方向发展。随之产生的大量生产质量数据，如铁水含量、高炉温度、气体含量、加热温度、轧材规格、轧材硬度等，正以极快的速度迅速增长（以每秒为单位进行测量），传统的关系数据库和集中式文件管理效率已经无法实现对海量数据的有效保存与查询、计算的功能。因此，为了实现海量数据下的质量数据分析和挖掘，需要对钢铁行业传统的数据存储技术进行更新和优化。

4）数据价值尚未得到充分挖掘

智能制造时代，数据已成为企业最优价值的资产。虽然我国钢铁企业基础自动化水平较高，数据收集仪器较全，但是大部分企业只是将海量质量数据作为产品缺陷的追溯

基础，无法为自身带来可视的经济效益。由于对于数据的重要性并未得到充分的认识，导致大量数据遗失。此外，数据的利用也呈现出单一化、局部化的趋势，因此数据价值无法得到充分的挖掘。

2. 钢铁质量大数据相关技术

质量大数据是指具有能够反映质量特性的各类数据，钢铁行业质量大数据是在目前质量数据的基础上拓展到大数据范畴，范围涵盖产品研发、工艺设计、生产过程等产品的全生命周期。其主要特点及相关应用技术与工业大数据相似，具有数据量巨大(Volume)、数据处理速度快(Velocity)、数据多样性(Variety)和数据价值密度小(Veracity)的特征。此外，钢铁行业质量大数据的收集特点、数据结构等方面的特点还表现在：数采设备繁多，数据流通协议复杂；以结构化数据为主，声音、图像等非结构化数据较少；非正常数据较多，数据降噪困难。

大数据技术在钢铁行业质量管理方面应用的主要技术包括数据采集、数据处理、数据分析和数据展示。

数据采集是大数据应用的前提条件，钢铁行业质量大数据的采集涵盖产品设计、研发、采购、生产的全流程，需要对产品全生命周期的质量信息进行全面精细采集，以得到尽可能完整的数据云，从而为数据的处理提供必要基础。

数据处理针对数据采集信息的可用性及边际价值进行必要的清洗，及时准确地分辨信息是否与质量管理工作相关，并确定其相关程度，剔除一些无关紧要的数据，保证相关性较高的数据。

数据分析的主要目的是针对不同的分析目标，从多个方向、多个维度对产品质量数据进行分析，以得到相关设备、工艺、人为因素等多方面的分析结果，进而为数据展示提供必要的数据支撑。另外，还可根据不同的使用目的建立产品质量数据平台，对产品整体的质量控制过程进行相对真实和完善的还原，并进行可信度较高的信息预测。

数据展示主要是针对不同的用户利用 Highcharts、Tableau、JpGraph 等相关可视化工具进行关键指标、历史趋势、预测效果等维度的展示。

3. 数据挖掘在钢铁质量管理中的应用前景

质量大数据管理的应用能够有效提高全面质量管理水平，保证全流程生产管控的协调一致。大数据在钢铁行业质量管理中的应用主要体现在生产过程检测、趋势预测、原因分析等智能模型的应用上，做到事前质量异议风险预测、事中关键环节生产监控、事后质量数据全流程追溯。

1) 数据质量优化改进

质量大数据的质量控制体系是一项复杂的系统工程，涉及管理、技术和流程三大方面的因素。由于钢铁行业质量大数据呈现出生产过程复杂、采集设备繁杂、通信机制众

多、异常数据占比大的特点，因此对于钢铁行业质量大数据的质量提升是大数据处理的必然前提。可以从以下 3 个方面提高质量大数据的数据质量。

（1）高质量感知数据。

明确设备对质量数据检测的要求，减少多读、漏读、误读的情况。预先设定机器读取数据的标准，采用更加先进的检测设备，从而在设备读取质量数据的过程中减少冗余数据的产生。

（2）高效数据清洗机制。

在处理设备读取数据的过程中，可通过基于规则发现、关联分析、聚类分析、偏差检测等多种方式发现异常的质量数据，并通过机器学习、冲突数据检测、规则学习等方式删除、修复异常的数据。

（3）建立数据采集标准。

坚持以应用为导向，从数据质量定义、数据质量评价、数据质量分析及数据质量改进等方面进行闭环管理，从而达到数据质量管理的持续优化。

2）质量异议风险预测

无所不在的传感器、互联网技术的引入使得产品故障实时诊断变为现实，大数据应用与建模、仿真技术的结合则使得预测成为可能。利用大数据技术对产品生产过程中的质量数据进行实时评估，从而能够将事后质量管理转移至事前预测，从而有效地降低企业生产成本。质量异议风险评估主要包含两个部分，分别是质量异议严重程度预测和质量异议发生概率预测。质量异议严重程度是指针对预测可能发生的质量异议的风险程度的评估，是发生的质量异议不足以影响销售、产品降级为二级品、产品发生重大质量异议不能销售等量化后的结果。质量异议发生概率是指在以往产品生产过程中，生产的实际总数量与发生质量异议的产品的数量之间的比例的统计量。质量预测在炼钢生产过程中具有诸多应用场景，如转炉终点磷含量预测、设备状态监测与维护预测、基于聚类分析的新钢种变形抗力预测等。

产品质量异议的风险评估能够有效地帮助企业提高风险预警能力，将不合格品的数量有效地控制在较低范围内，从而提高企业的生产管理水平，降低生产成本。

3）关键环节生产监控

生产环节监控能够有效地掌握生产现状，起到质量管理的事中检测与管控的作用。一方面，通过运用大数据快速获取、处理、分析的能力，为生产管理人员提供可视化交互引擎、人机交互管控模式、可视化关键信息展示。另一方面，通过传感器网络将生产过程监控与企业运营联系起来，在加工过程中尽早发现存在的质量波动，并通过生产和企业运作的匹配尽早做出反应，实现对最优企业运作的预期并自动调整生产流程。

在钢铁制造过程中，各工序生产环节复杂，每个环节的工艺参数设置较多，造成生产过程中诸多产品缺陷的可能性，如擦伤、温度过高、边裂、划痕等。通过大数据挖掘构建

一个集成多方面的生产缺陷识别模型，利用图像处理、成分检测等技术分析缺陷类型及原因，及时发现不合格品。在此方面的应用已逐渐发展成熟，如智能缺陷系统检测技术、转炉炉衬侵蚀动态监视技术、转炉炼钢终点精准控制技术等。

4）全流程质量数据管理

全流程质量数据管理系统采用在线质量管理与离线质量管理相结合的方式，实现在线对生产过程工艺数据、性能数据的监控、质量决策，离线质量追溯和质量趋势分析。全流程质量数据管理涵盖铁水生产、炼钢、铸机、轧钢等钢铁企业生产的全流程的工艺数据、质量数据、生产数据的采集和存储等，实现对数据的抽取、集成、展示，从而为每一批次钢种的质量分析、质量追溯、质量决策提供有力的数据支撑。

习题

1. 网络数据挖掘有什么特点？
2. 数据挖掘如何应用于企业的 CRM 系统中？
3. 在电信业中应用数据挖掘技术可以挖掘哪些有价值的信息？
4. 数据挖掘如何在金融行业中进行风险评估？试举例说明。
5. 数据挖掘技术如何应用到交通领域？
6. 通过实例分析数据挖掘技术在信用卡业务中的应用。