

数据分析需求与实现

易 定

(深圳职业技术学院 计算中心, 广东 深圳 518000)

摘 要:数据分析是从海量数据中发现隐含信息或知识的过程。基于一个公安破案辅助数据分析系统,深入研究数据分析任务的需求与实现,提出首先规划分析思路、细化分析功能,然后用多视角数据透视和智能分析两种手段,从微观与宏观、定量与定性等不同角度互为补充地使系统具有完备的分析功能。该研究对如何开发具有实用价值的的分析系统有普遍的指导意义。

关键词:数据分析;分析思路;分析功能;分析手段;数据透视;智能分析

中图分类号:TP311.13

文献标识码:A

文章编号:1673-629X(2006)09-0112-03

Requirement and Implementation of Data Analysis

YI Ding

(Computer Center, Shenzhen Polytechnic, Shenzhen 518000, China)

Abstract: Data analysis is the process of mining information or knowledge from the data sea. Based on the auxiliary data analysis system of criminal investigation, purports to study in depth the requirement and implementation of data analysis, puts forward the stages to facilitate the system with full function of analysis. First, marks out the analysis track and details the function of analysis. Second, it gives supplementary use of the means of multi data perspective and intelligent analysis, from perspectives of microscope and macroscope, quan and qual. This research would be of general guidance to the R and D (programming) of data analysis system with practical value.

Key words: data analysis; track of analysis; function of analysis; means of analysis; data perspective; intelligent analysis

0 引言

数据分析的目的是要从数据中发现信息或知识,与此相关的前沿的理论研究是数据挖掘,它们在分析算法方面的研究已经取得了一些成果,但是有关数据挖掘的理论基础研究还没有成熟,没有一致的理论框架指导数据分析任务的实现。另一方面,分析任务往往都与特定的应用密不可分,需要用分析对象所涉及领域的知识来指导分析过程,因此,数据分析任务还具有跨学科、跨领域的特点^[1]。鉴于以上特点,当设计一个具有实用价值的的分析任务时,会遇到许多问题。例如如何合理设计分析思路和分析功能?这个问题与需求相关,它明确数据分析的目标;采用什么分析手段来体现分析思路和分析功能,这是分析任务的人机接口方面的设计,文中将从这个角度进行探讨。

公安破案是从涉案人员的行为中找出侦破线索,这是一个典型的数据分析过程。笔者曾与十几位经验丰富的刑警一道反复研究,并参与了几起真实案件的通信信息的处理与分析,开发了“通信信息智能分析系统”。下文将基于该项目深入探讨数据分析任务的需求与实现。

1 问题研究

1.1 分析思路与分析功能

大量的数据分析需求都与特定的应用相关,需要相关领域知识的支持。通用的数据挖掘工具在处理特定应用问题时有其局限性,常常需要开发针对特定应用的分析系统^[2]。因此数据分析系统设计的第一步是对特定应用的业务进行深入地分析与研究,总结归纳分析思路并细分出所需的分析功能。

分析思路的设计是宏观地对分析对象所需要分析的几大目标进行归纳,也就是说需要从哪几个大的方面来进行分析;分析功能的设计是将每一分析目标,即分析思路,细分为若干小的分析主题。好的分析思路和分析功能应该较全面地覆盖被分析的对象(见图1、图2)。

这一项工作非常重要,它决定着分析系统的优劣;这项工作也很难做好,因为它对系统分析师的要求非常高。它不仅要求系统分析师具备良好的系统分析与系统设计的素质、丰富的实践经验以及知识发现与数据挖掘的相关理论和技术,它还要求具备特定应用领域的背景知识,深入理解该应用的业务。因此要求系统分析师积极地与该学科或领域的专家沟通,应具备较强的与人沟通的能力和新技术的领悟能力。

1.2 分析手段

数据挖掘有大量文献研究挖掘方法和算法,但是提供

收稿日期:2005-11-24

基金项目:深圳市科技局资金资助项目(001CJC008)

作者简介:易 定(1965-),女,湖南岳阳人,高工,主要从事智能软件的研究与开发。

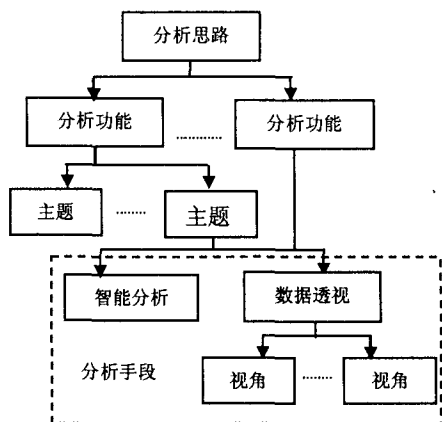


图1 分析思路、分析功能与分析手段

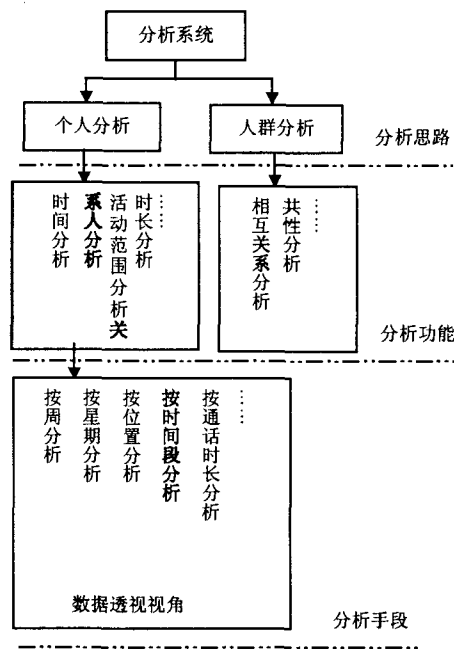


图2 实例详解分析思路、功能与手段

什么分析手段来体现分析思路和分析功能,使用户从数据中获得信息或知识,这是数据分析任务人机交互的接口,这个问题被数据挖掘的研究者们忽略了。在研制上述公安项目的过程中,对这个问题进行了深入的研究,提出采用多视角数据透视和智能分析两种手段来实现数据分析功能,分别从微观与宏观、定量与定性等不同角度互为补充地进行分析,使数据分析系统具有完备的分析功能。

(1)用多视角数据透视手段实现数据分析。

对于特定的分析功能可以细分多种不同的数据透视角。采用数据透视手段实现数据分析,使系统具有以下特点(见图3、图4):

* 灵活方便的多视角、多层次(概念分层)的高性能数据分析界面,使得数据背后隐藏的含义显露出来;

* 数据详略展开,即方便地查看汇总数据和生成汇总数据的明细数据。展开时可深入仔细地分析,折叠时可以进行宏观分析。也可以说是在不同概念层次上自由伸缩,来对数据进行定量透视。

(2)智能分析。

智能分析以数据挖掘思想为指导,结合领域知识,对数据的全局或者特定主题进行分析,较宏观地、全面地或定性地给出描述或某种推测、评价,主动找出数据背后隐藏的含义,并用图表和文字等多种形式充分地展现出来^[3,4]。

通话日期 按月		呼叫类型	电话类型			
28		全部	全部			
对方号码	时段		总计			
	0	2	3	4	5	次数
次数	次数	次数	次数	次数	次数	次数
13823219398	1					1
13923831151	1	2	2			5
07555633999	1					1
13823138888				1		1
114					1	1
13602583989				1		1
13802294725					1	1
总计	3	2	2	2	2	11

双击,展开

总计	755	主叫	2000-11-28	0:12:59	119
	755	被叫	2000-11-28	0:25:54	27
	755	主叫	2000-11-28	0:41:28	35

图3 凌晨零点机主的三次电话通话记录

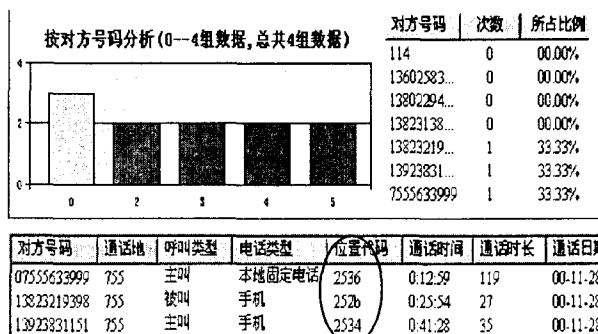


图4 机主的位置变化

因此,采用智能分析手段实现数据分析,使系统具有以下特点:

* 提供全局的、宏观的、面向主题的分析结果;

* 好的系统还使数据、分析过程和分析结果可视化。

数据可视化和智能分析可从以下方面融合:数据可视化、分析结果可视化、分析过程可视化、交互式的可视化分析。“可视化”的引入将使系统更加具有吸引力,这是前沿研究的一个方向。

数据挖掘的研究成果可以指导智能分析,但是数据分析任务的跨学科、跨领域特点使得智能分析没有固定的设计模式。

(3)两种分析手段的比较。

多视角数据透视手段,实现不同概念层次上的分析视角的伸缩与变换,使数据背后隐藏的含义显露出来,同时可以极其方便地进行数据宏观和微观分析;智能分析手段以数据挖掘思想为指导,结合领域知识进行推理,主动找出数据背后隐藏的含义,给出某种推测或评价。

多视角数据透视手段注重数据的定量分析,可以对数据在不同概念层次上进行宏观或微观的分析;智能分析手段更注重对数据的全局或者特定主题进行研究,提供更整

体的、定性的描述,或有关问题的推测。这两种分析手段互为补充。

综上所述,针对数据分析任务的需求与实现,笔者提出首先规划分析思路,细化分析功能,然后用多视角数据透视和智能分析两种手段,从微观与宏观、定量与定性等不同角度互为补充地使系统具有完备的分析功能(见图 1)。

2 实例详解

2.1 分析思路与分析功能

公安破案需要分析涉案人员的通信往来,从中找出侦破线索,这是一个典型的数据分析过程。传统的人工比对方式既耗费大量的人力,又难以提供及时准确、有效的信息。“通信信息智能分析系统”结合公安人员实际工作经验,采用最新计算机信息处理技术来分析比对涉案人员的电话往来,全方位地挖掘出对侦破工作有意义的信息,是一个分析功能和分析方法完备且实用的数据分析系统。

深入细致地归纳和总结破案思路,设计了两种分析思路:

(1)对一个涉案人员的情况进行分析,实现对个人通信情况的全方位的详尽分析;

(2)对多个涉案人员的情况同时分析,实现对人群中人物间相互关系和共性的研究。

然后,针对上述分析思路逐一细分分析功能。例如,思路(1)细分为时间分析、关系人分析、活动范围分析、时长分析等。这些分析功能从多个角度发掘其主要关系人、关系人的亲疏度、其本人的活动规律、日常活动范围和主要停留地等信息。对于特定的分析功能可以细分多种不同的数据透视视角,这些视角各有其特别的内涵。例如“关系人分析”设计下列 6 个数据透视视角:月、周、星期、时间段、位置、通话时长(如图 2 所示)。

2.2 分析手段

用一个真实案件数据来展示分析手段的特点。

2.2.1 数据透视实例详解

设计的数据透视视角各有其特别的内涵。以关系人分析中按时间段分析为例,如图 3 是机主 11 月 28 日清晨 0 点至 4 点的电话的通话汇总。机主在凌晨 0 点有三次通话,两次主叫,一次被叫。清晨 0 点至 4 点通常是休息时间,关系一般的人不会在这段时间打扰机主,所以这段时间与机主通话的人与机主的关系特别,或这段时间的电话是因为有特别的事情。

如图 4 所示,从凌晨 0:12 至 0:41,机主从地点 2536 经过 252b 到达另一地点 2534。随后在凌晨 2 点、3 点、4 点、5 点的详细记录表示,机主的位置都在 2534。因此初步推测机主在 0 点时正在往地点 2534 的路上,然后在 2534 处休息。

数据透视视角采用 Microsoft Office Web Component (MS OWC)中数据透视表技术实现^[5]。

2.2.2 智能分析与数据透视技术结合

对于特定的分析功能结合公安侦破领域知识和经验,可以细分多种不同的智能分析主题。例如,找出多人间的相互关系——相互通话,它反映多人间的联系与亲疏程度(见图 5);找出多人间的共性——相同联系人,它反映多人间的社交圈重叠情况等等(见图 6)。

图 5 非常清楚地显示出四位通信人之间的电话往来关系。有连线的通信人之间有相互通话;连接线的粗细对应通话次数的多少。13923867168 与其他三位都有联系,似乎是四人中的核心人物。

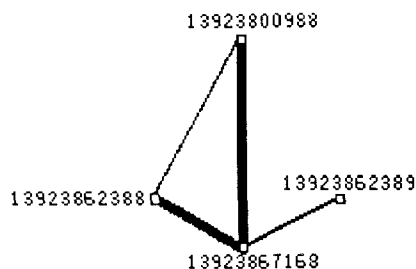


图 5 四位通信人之间的电话往来关系

图 6 所示相同联系人是与多人之间都有交往的人,图中圈出的二人之间有 44 位相同联系人,社交圈重叠度很高。这些结论是宏观且定性的。图 7 展示二人与这 44 人的详细通话情况,这是微观与定量的。

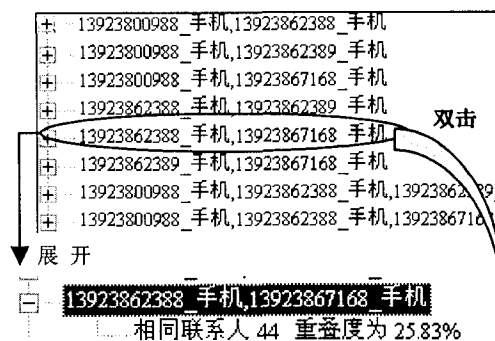


图 6 多人间的相同联系人

通话日期	按周	呼叫类型	全部	全部	全部
全部	全部	全部	全部	全部	全部
本机号码	全部	全部	全部	全部	全部
13923862388	13923867168	总计			
对方号码	次数	次数	次数		
13501596308	1	174	175		
13923800988	1	170	171		
07555633999	37	71	108		
13602623603	20	81	101		
13025412121	1	81	82		
13910000400	21	30	51		
07555633666	15	31	46		
13828756436	1	44	45		
13823219398	34	3	37		
13823138888	22	14	36		
13025491543	2	26	28		
13809895558	4	22	26		

图 7 二人与 44 位相同联系人的电话往来关系

智能分析的实现采用了数据挖掘理论和统计分析、集合论、模糊数学、排列组合等数学原理。图 5~图 7 融合了多视角数据透视和智能分析两种分析手段,从微观与宏观、定量与定性等不同角度互为补充地进行了多人之间的

(下转第 116 页)

End Sub

这段程序的功能是:类方法程序 readpass1 接受用户程序传递来的数据,据此判断是否满足条件,若不满足则激活事件 password。其中命令 RaiseEvent 是十分关键的,它的含义就是激活事件!由此也可以看出,所谓类的事件的发生,其实于类仅仅是激活事件而已。

4) 用户窗体中的按钮事件:

```
Private Sub button1_Click()  
    Set mycx2 = New myClass2  
    mycx2.readpass1(Text1.Text)  
    Set mycx2 = Nothing  
End Sub
```

该程序从功能上看很简单,不外乎将文本框 Text1.Text 的输入内容传递给上面介绍的类方法程序 readpass1。但其中所用命令确是十分有必要了解清楚的。

首先,Set mycx2 = New myClass2 命令是将变量 mycx2 直接设置为类 myClass2,这与在窗体声明中对变量 mycx2 所作的声明的概念不同,因为此时类 myClass2 的所有属性、方法和事件均可作为变量 mycx2 直接调用。

其次,mycx2.readpass1(Text1.Text) 命令是将文本 Text1 输入的数据传递给类 myClass2 提供的方法 readpass1 去处理,目的是让 readpass1 根据提供的数据决定是否激活用户自定义的事件。

最后,Set mycx2 = Nothing 命令是取消变量 mycx2 的作用。它的作用是关闭类。

此三条命令体现了类的标准使用方法,即通过变量打开类、使用类、关闭类的过程,这在类的使用中十分重要。

5) 用户窗体代码中的自定义事件程序。

程序清单如下:

```
Private Sub mycx2_password()  
    MsgBox "wish you bad luck! and event response!", vbOKOnly, "class event"  
End Sub
```

这段程序提供了如下综合信息:

(上接第 114 页)

相互关系与共性分析,提供了极有价值的侦破线索。

3 结束语

基于“公安项目”,文中深入研究了数据分析的需求与实现,提出首先规划分析思路、细化分析功能。分析思路是宏观地归纳分析大目标,分析功能是将每一分析思路细分为若干小的分析目标。对于特定的分析功能又可以细分为多种不同的数据透视视角和智能分析主题。这两种分析手段,分别从微观与宏观、定量与定性等不同角度互为补充,使系统具有完备的分析功能。该研究对如何开发具有实用价值的数据分析系统有普遍的指导意义。好的数据分析系统还应该在数据表示、分析过程和分析结果可视化方面努力。图 5~图 7 用图和表结合一目了然地展示

(1)事件的激活源自于类,而事件本身由用户提供,它的代码对于使用者是开放的。

(2)事件程序名的命名规则是:用户定义的类变量在前,类定义的事件名在后,中间以下划线联之,犹如 mycx2_password,而且用户定义的类变量必须设置成带有事件声明的类的变量,只有如此当事件发生时该程序才可以被激活。

(3)发生事件的条件由用户程序向类提供,其方法是通过类方法的调用将载有事件产生的条件参数传送到类库中。

(4)当事件的产生条件一旦成立,RaiseEvent 命令将激活指定的事件。

以上是利用 Visual Basic 对可视化程序设计关于如何建立用户自定义的实践的方法讨论。相关程序均运行通过。

4 结论

通过上述的讨论,可以清晰地看出作为控件的事件的类设计,必须从两个方面考虑:首先是设计者对控件的事件的类的通用性和封装性的考虑;其次是用户正确的使用。相信通过此文读者能够在今后的可视化图形界面程序设计中自如地应用类的事件了。

参考文献:

- [1] Stevens A. Wiley's Teach C++ [M]. 北京:电子工业出版社,2004.
- [2] 龚沛曾. Visual Basic 程序设计教程[M]. 北京:高等教育出版社,2000.
- [3] 张左营. 基于 VB6.0 的工控机数据采集系统的开发[J]. 微计算机信息,2004(11):66-67.
- [4] 王玉斌. VB 程序中动态设置 ODBC 数据源方法[J]. 华南金融电脑,2004(11):46-49.
- [5] 王明军. 基于 VC 的 AHS 车辆自动驾驶仿真实现[J]. 计算机工程与应用,2004(4):202-205.

人群中人物间相互关系、亲疏程度、共性。“可视化”的引入将使数据分析系统更加直观易用^[2]。

参考文献:

- [1] 胡运发. 数据与知识工程导论[M]. 北京:清华大学出版社,2003.
- [2] 刘同明. 数据挖掘技术及其应用[M]. 北京:国防工业出版社,2001.
- [3] 俞瑞钊,陈 奇. 智能决策支持系统实现技术[M]. 杭州:浙江大学出版社,2001.
- [4] 施伯乐,朱扬勇. 数据库与智能数据分析:技术、实践与应用[M]. 上海:复旦大学出版社,2003.
- [5] Microsoft Visual Studio. Net MSDN[Z]. 2000.