School of Computing & Mathematical Sciences
Oxford Brookes University

Zoltan Papp

Dissertation Plan : Distributed data mining with message pasing

Monday, 29 April 2002

Dissertation plan

Supervisor's name : Chris Cox ccox@brookes.ac.uk
Student's name : Zoltan Papp 01255638@brookes.ac.uk
Course : Computer science Msc.

General area of investigation : Distributed database analysis.

Tentative project title : **Distributed data mining with message passing**

Abstract

Distributed programming have gained an unprecedented attention in the last decade. This change have primarily been fostered by the high demand for computation. High performance computing can be achieved by either Massively Parallel Processors and/or Network computing. I will be studying Network computing by applying message passing because this area provides a cheaper alternative to MPPs while delivering comparable performance. [1]
The industry needed a standard communication interface for the rapidly evolving distributing technologies. Among the many attempts, TCP/IP (Transfer Control Protocol/Internet Protocol), MPI (Message Passing Interface) and PVM (Parallel Virtual Machine) became significant in industrial terms.

TCP/IP is a transfer protocol, hence, provides efficient and error free data transfer across networks. MPI is a library that uses TCP/IP and can be used for data transfer, task management in a shared/distributed memory and/or multiprocessor environment as well as clusters. A version of MPI, MPICH-G is capable of operating in a GRID environment, that is, on a wide area network.[2] PVM is a highly portable and easy to use distribution tool that uses TCP/IP but provides dynamic task spawning and other high level task management operations. Hence, all 3 technologies are in active use today, depending on the target application.
These technologies are primarily present in applications running on local hosts rather than thin client web applications. Apart from binary applications, thin web clients also use a number of distributed technologies. Here, I will focus on production and computation oriented applications and distribution tools running on local workstations.

Hence, I will exploit the advantages of the standardization of distributed communication by a database application.

Corporate databases contain a considerable amount of information that can be extracted by data analysis. Therefore, counting, measuring and relating a corporation's operational factors is important to help supporting the following goals

- Management decisions and business strategy construction
- Eliminating useless work to gain more time → better focus → better quality and more control
- Eliminating unnecessary communication between departments

- Higher throughput of the company without adding and/or replacing people [3]

Hence, I will write a distributed data-mining program that analyse an SQL database and generates a report, based on user defined options. The program will analyse governmental taxation database.

Keywords
Data mining, SQL, C/C++, MPI, Win32 GUI/MFC, Taxation

Introduction

The program comprise of 3 parts.

(a) SQL database that holds data.
→ Passive data
(b) Distributed analysis to accept user specified report requests and generate report.
→ Passive information
(c) MS-Windows client as front-end application to enter requests and display results.
→ Active models and information delivery

This project will employ data mining and distributed database technology together with message passing software to develop a high level business data analysis capability. More specifically, I will analyse a taxation database based on the SA-100 form used by Inland Revenue department of the UK Government. [4]

Clearly, data processing is done in 3 phases. In the first phase, the largest amount of data is processed locally because (1) Data transfer overhead
(2) Security of data, by not transferring it over public networks
Data mining is done on client's machine since data mining uses pre-processed data
(1) Proportionately less than the original data being used
(2) Specific reference to individuals or confidential data is no longer present in the information. Data protection and privacy law applies, hence security is still a major concern.
User will interact with system via the Win32 GUI. User will have access to both statistical information and models created by data mining algorithms as its shown on the following figure.

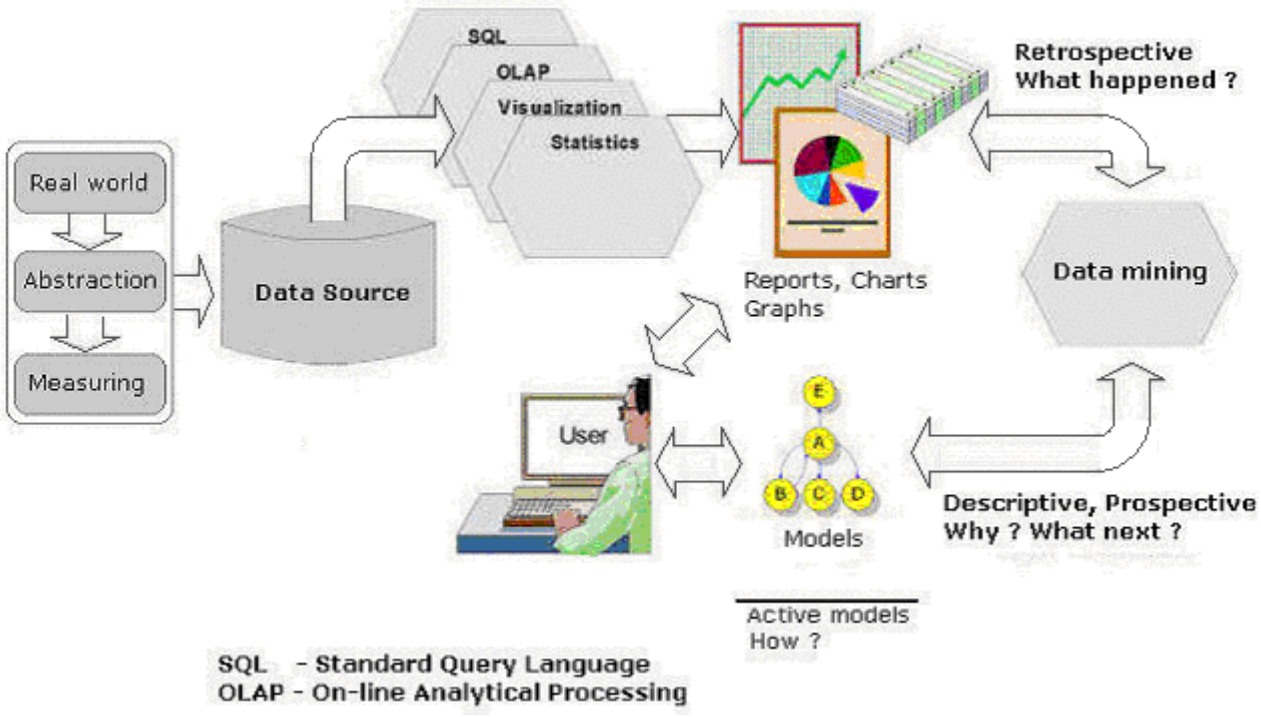

Figure 1. Operation of a computer aided management software

Rationale for choice of project

The choice of this project stems from both academic interests in mathematics and distributed processing. I have been using message passing software such as PVM [Ref: BSc Project - Distributed graph analysis]. I undertook a series of mathematical modules that relate to Discrete mathematics in any way, such as Discrete mathematics, Further discrete mathematics, Algebra, Linear Algebra, Graph theory and Geometry. I am interested in discrete mathematics and its use in real world problem solving. Graph theory and linear algebra in particular. Data mining requires model construction and models can be represented as graphs, etc. linked states of a system.

Embedded programming in a real time environment have also become an important factor because of the increasing number of intelligent electronic devices such as fridge with internet connection, car navigation etc. Hence, I will write the program as interoperable and easy to use in embedded development as much as possible. [5]

I was working for the Oxford office of Inland Revenue (IR), UK. In the UK, people are able to submit their tax forms online, vote online and a series of upcoming options will be availabe. The UK government is in the frontier of revolutionizing governmental operation by applying information technology. [6]
IR records nearly all of its electronic activities including the number of hours of its employees working in different areas and functionality. Data gathered at local offices then being transferred to the central office once a month, where it is analysed mainly for human resources.
I will apply and test the efficiency of the program on tax records, namely the SA-100 self assessment record database. This record is used for self assessing businesses and individuals across the UK. Hence, an important source of information on the country's economical operation and perhaps equally important, income for the public government body that, in turn is spent on the public. The SA-100 form holds information on the business owner's name, address and most financial and taxation information of the business. I will introduce further analysis and data mining techniques to perform a sharper financial analysis and selection for examination.

Objectives

- I have been studying the network and software technology being used at the IR Oxford office. I have now full information on the type of information BMSD (Business Management System Department) records and the client programs that every IR employee uses on a daily basis to record her or his weekly schedule and how it is been used. I have also been studying the technology of Data-mining Inc. [Ref: www.datamining.com]

- Both hardware and software technology is chosen based on the lowest common denominator of availability and cost. This result in the use of :
  - o Ethernet networking
  - o MS-Windows® and Unix hosts
  - o C/C++ as implementation language
  - o MPI (Message passing interface) and TCP/IP as distribution tool

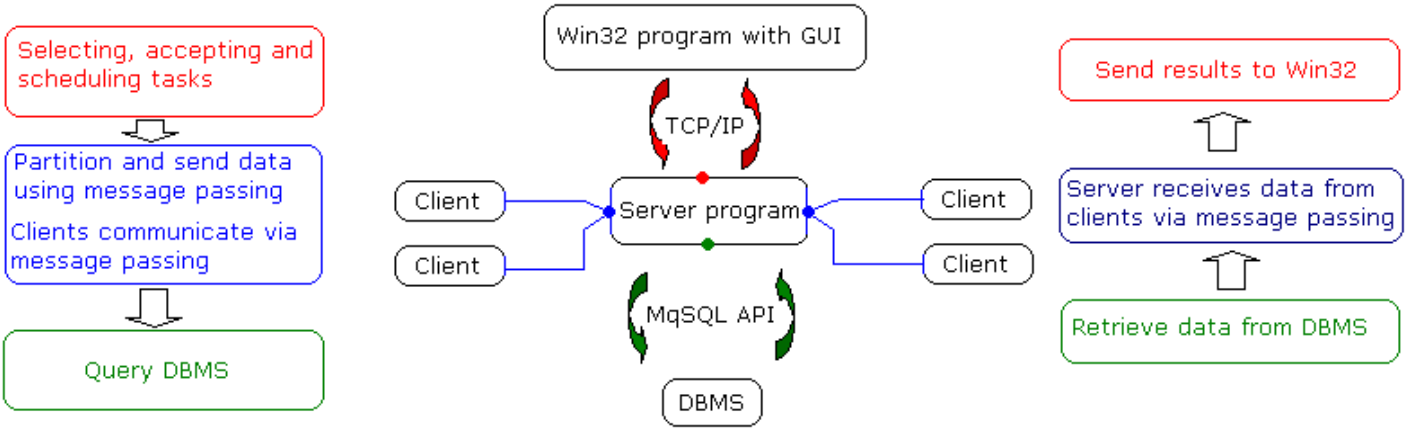

Figure 2. Operational plan of my software

- Complete system will comprise of a
  - o Win32 GUI(Graphical User Interface) executable
  - o Server program and client binaries with database API(Application Programming Interface)
  - o Documentation of project

  Required environment will be
  - o MS-Windows® 98/ME/NT/XP
  - o Microsoft .NET® Framework
  - o MPI/Pro®
  - o MySQL server(s) on a Unix or MS-Windows® host
  - o TCP/IP network with SSL encryption

  I will enclose a CD-ROM with the project material and all required free tools and user manuals.

- Basic modular operation of the software is as follows :

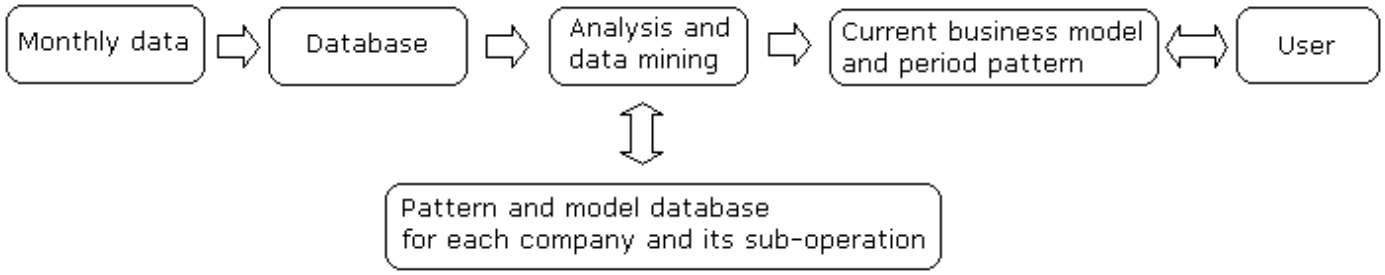

Figure 3. Data processing flow diagram

Patterns are used to determine processes. This scheme is analogous to forecasting weather on current and past weather conditions.

Method

- Background material

  - o The algorithms and software design I plan to apply is mine, since competitive data mining companies do not publish their software's internal operation. I found Data mining Inc.'s website informative and applicable for my project. Basic overview of Data mining Inc's processing technology can be found at (Ref : http://www.datamining.com/dmsuite.htm) I also find my discrete math knowledge valuable. For first phase analysis, I will use statistical functions such as average, median, standard deviation, cumulative beta distribution etc. In the second phase I expect to work on already established domains generated by the 1st phase. I also plan to be able to analyse complete databases, that is determine the parameters of a corporate database. Statistics is obtained by taking samples of a complete population as it is been shown in the following example.

Cost of collection (pence per £ collected)	1995/96	1996/97	1997/98	1998/99	1999/00
Income Tax	2.05	1.98	1.81	1.64	1.23
Corporation tax	0.73	0.7	0.62	0.71	0.76
Petroleum Revenue Tax	0.36	0.18	0.34	0.64	0.24
Capital Gains Tax	4.1	2.68	2.3	1.61	1.49
Inheritance Tax	2.5	2.05	1.75	1.57	1.46
Stamp Duties	0.35	0.25	0.17	0.13	0.11
National Insurance Contributions (NICs)	-	-	-	-	0.57
All taxes and NICs	1.7	1.57	1.41	1.33	1.1

Figure 4. Raw data in table format

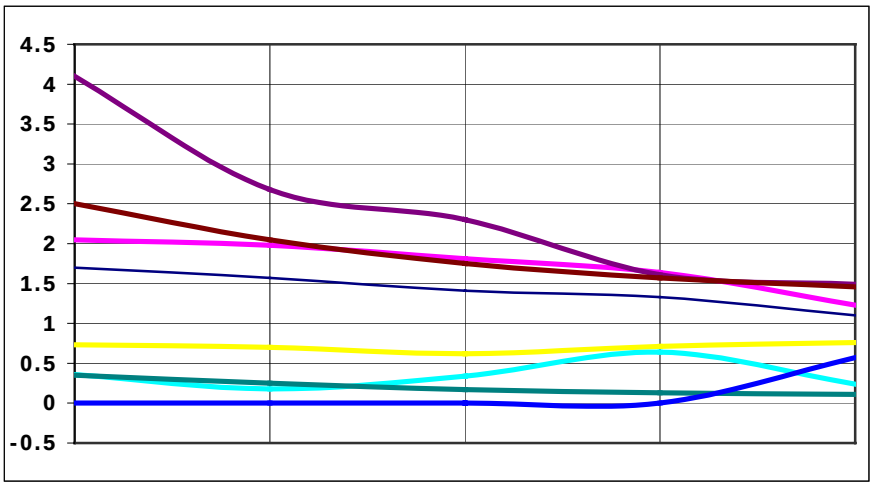

Figure 5. Pre-processed data

$$\overline{\mathrm{X}} = \frac{\mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_n}{n}$$

Average

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} (\mathrm{X}_i - \overline{\mathrm{X}})^2}{n-1}}$$

Standard deviation

Statistical formulas are then used to process raw data and extract valuable information. Complexity of functions increases as sophistication of information discovery increases. Only computationally complex functions are worth to partition since data transfer overhead would be too high.

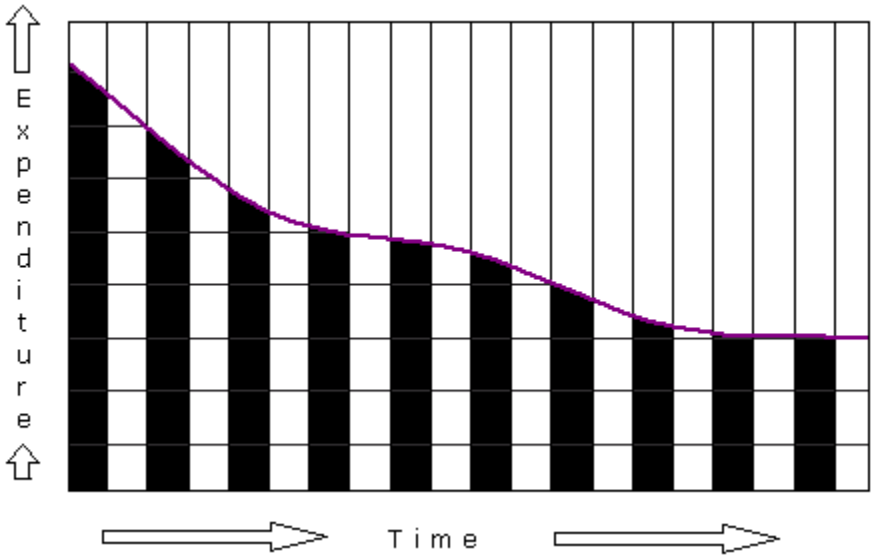

Figure 6. Sample to perform Fourier and other analysis

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx)$$

Fourier series

Once data have been pre-processed, Fourier and other algorithms will be used to further process data and identify domains and patterns in the processes. Such properties can be related by constructing graphs to represent and further study the information.

$$\mathrm{B}(x; n, p) = \sum_{y=0}^{x} b(y; n, p)$$

Cumulative beta distribution

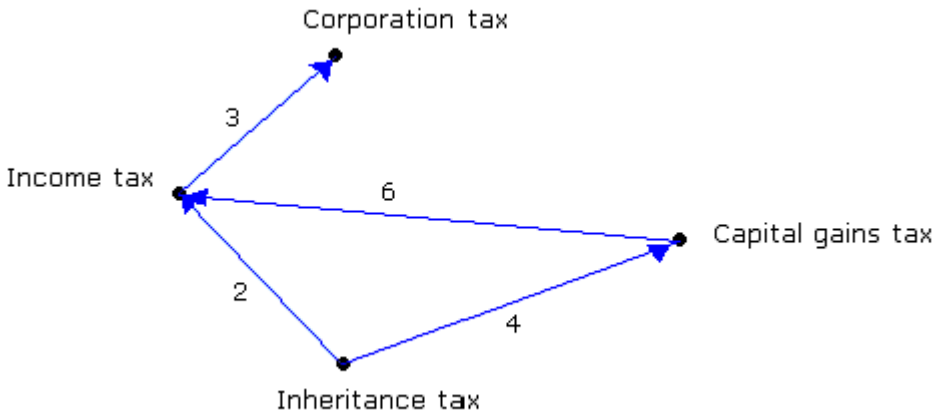

Figure 7. Sample financial model

Hence, will apply graph theory and linear algebra algorithms such as shortest path, simplex algorithm and graph analysis.

- Process of project creation

  - I intend to work by myself with the supervision of Chris Cox. I also have the opportunity to obtain further material from IR.
    Faye Mitchell of Brookes University assembled a Linux/SunSparc (6) cluster that I will also have the opportunity to use with MPI.

  - The development of the server and the Win32 application is interdependent since they communicate. I will use MS-Visual Studio® C++ 6 with Microsoft Foundation Classes® (MFC) for the development of the Win32 application. I will be using the "Getting started with MS-VC++ 6" book from Deitel & Deitel Inc. and the Microsoft Developer Network's library.
    I will first develop the Win32 application user interface and test TCP/IP communication between a remote server and Win32. If that works, I will implement the server and the clients and than the full Win32 application. Relational database will be first specified and designed on paper and I will test and analyse it using MS-Access®.

- Implementation plan in 3 levels

  - Core level, Database management and statistical analysis
  - Operational level, same as above plus MS-Windows client
  - Extended level, same as above plus other remote interfaces, report delivery methods, and task scheduling
  - Documentation, testing, checking, post-processing

Resources

- Online resources

  - Software development links
    http://shinji.brookes.ac.uk/~zoltanp/Technology.html

  - MySQL user's guide
    http://www.mysql.com/documentation/mysql/alternate.html

  - Win32 GUI programming
    http://www.winprog.org/tutorial/

  - MPI/PVM programming
    http://www.jics.utk.edu/documentation.html

- Experts

  - Chris Cox, MSc supervisor, project, ideas, message passing
    ccox@brookes.ac.uk

  - Faye Mitchell, BSc supervisor, cluster environment, systems programming
    frmitchell@brookes.ac.uk

  - John Roberts, Business team manager IR/Oxford, taxation, risk analyzis
    +44 1865 788556

- Budgets

All commercial software tools are included in the Microsoft® Computational Clustering Technical Preview (CCTP) [7] as an evaluation pack free of charge.
The LAN I built at home is based on a $86 USD Netgear® RP114 integrated router and switch and 3 PCs that I bought on my money of which some parts such as monitor, keyboard, house have been disposed by Brookes university. I also obtained free computers from Cambridge University Computer Preservation Society that I have exchanged to other computer peripherals. I have also been given an old SunSparc® SLC by Cognitive science department of University of Edinburgh that I, again sold and exchanged to computer peripherals.
I will not have to spend on hardware because the University network provides and excellent medium for development. I spent about 25 GBP on "Parallel programming" (Pren.Hall).
Other books such as "Getting started with VC++ 6" Deitel & Deitel are available in the library.

Schedule

Phase	Task name	Duration	Start	Finish
1	Identify task	1 day	1-Apr	2-Apr
2	Gather resources	2 weeks	2-Apr	16-Apr
3	Specify project	1 week	16-Apr	23-Apr
4	Database and its connectivity	2 weeks	23-Apr	7-May
5	Database analysis	3 weeks	7-May	28-May
6	Data mining Win32 GUI user interface	3 weeks	28-May	18-Jun
7	Data mining algorithms	2 weeks	18-Jun	2-Jul
8	Operational data mining	2 weeks	2-Jul	16-Jul
9	Testing overall efficiency and security	1 week	16-Jul	23-Jul
10	Improving algorithms	2 weeks	23-Jul	6-Aug
11	Documentation	3 weeks	6-Aug	27-Aug
12	Post processing of documentation	2 weeks	27-Aug	10-Sep

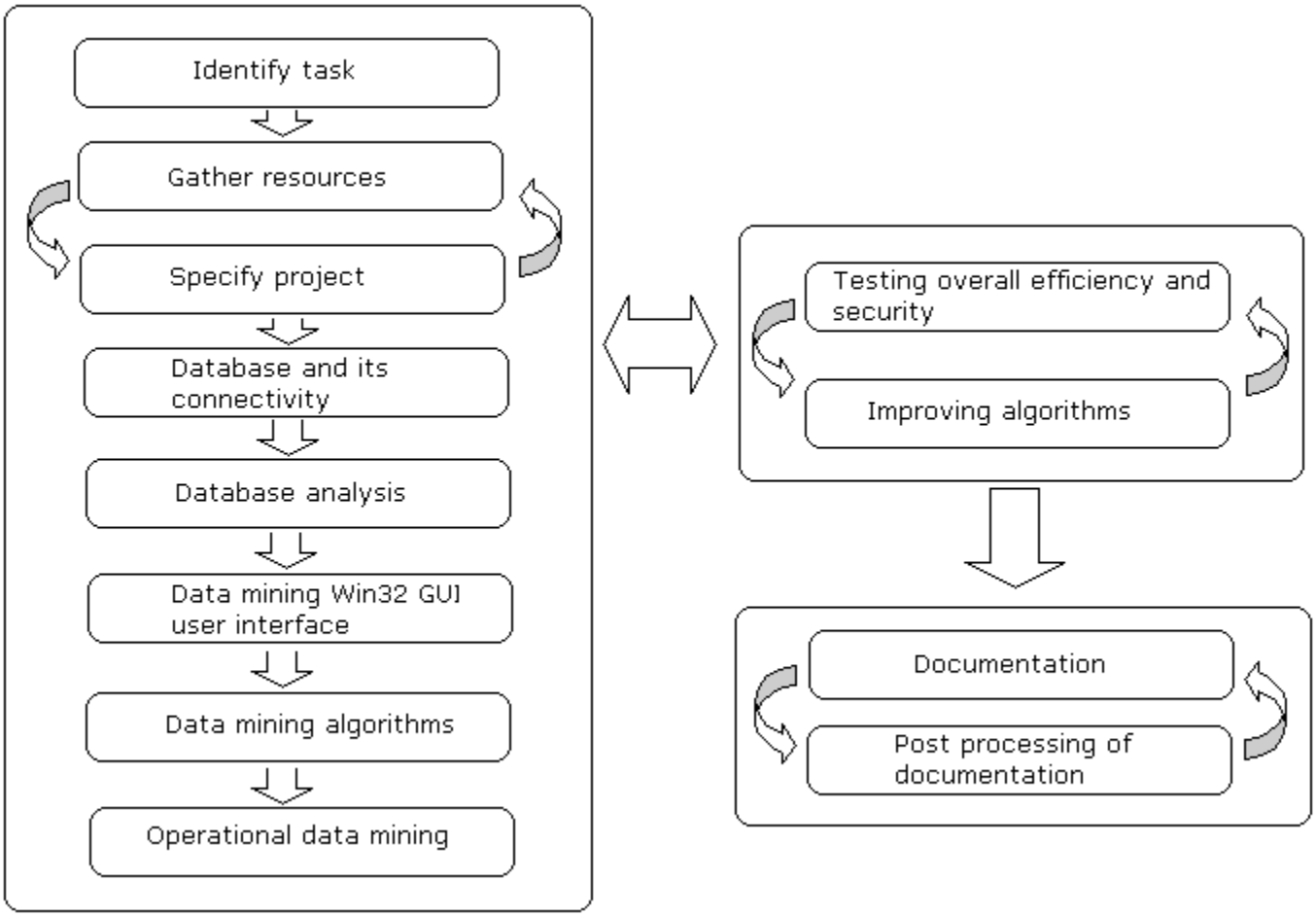

Figure 4. Project implementation plan

References

[1] John Stone, Software developer at Beckman Institute
Tachyon -- Parallel / Multiprocessor Ray Tracing Software
http://jedi.ks.uiuc.edu/~johns/raytracer/

[2] Globus Project, MPI implementation for GRID computing
http://www.globus.org/documentation/incoming/gempi.pdf

[3] Andy S. Grove, co-founder and ex CEO of Intel
"Only the paranoid survive" - March 16, 1999. Book on technology industry management
http://www.amazon.com/exec/obidos/ASIN/0385483821/qid=1019228891/sr=1-1/ref=sr_1_1/102-8734771-7453726

[4] SA-100 form used with permission by Inland Revenue, UK
http://www.inlandrevenue.gov.uk/pdfs/2001_02/tax_return/sa100.pdf

[5] First International Workshop on Embedded Software, October, 8-10th, 2001
http://link.springer-ny.com/link/service/series/0558/tocs/t2211.htm

[6] UK Gateway, Government portal for information processing
http://www.gateway.gov.uk/

[7] Computational Clustering Technical Preview
http://www.microsoft.com/windows2000/hpc/toolkit.asp

Bibliography

- C++ How to program, Deitel & Deitel — ISBN :0-13-117334-0
- Getting started with MS-Visual C++ 6, Deitel/Nieto/Strassberger — ISBN :0-13-013249-7
- Internetworking with TCP/IP volume ii, Comer/Stevens — ISBN :0-13-134677-6
- Parallel programming, Wilkinson/Allen — ISBN :0-13-671710-1
- Elementary linear algebra, Anton — ISBN :0-471-17055-0
- Graphs, an introductory approach, Wilson/Watkins — ISBN :0-471-61554-4
- Statistics for business and economics, Anderson/Sweeney/Williams — ISBN :0-314-01244-3

- Internet Engineering Task Force, Request for Further Comments
  http://www.ietf.org/rfc.html
- Internet Society
  http://www.isoc.org/internet/
- Fortune magazine
  http://www.fortune.com/
- IEEE Computer magazine
  http://computer.org/computer/
- IEEE Distributed Systems Online
  http://dsonline.computer.org/index.htm
- Wired magazine
  http://www.wired.com/wired
- Cyber geography
  http://www.cybergeography.org/
- Federation of American Scientists
  http://www.fas.org/
- Netcraft, IT survey
  http://www.netcraft.com/
- Computing dictionary
  http://wombat.doc.ic.ac.uk/
- Online research aids
  http://www.gci275.com/searches.shtml