

RESEARCH AND DEVELOPMENT
TECHNOLOGY AND INNOVATION

2005

TECHNOLOGY AND INNOVATION
RESEARCH AND DEVELOPMENT

2005

150/3051

1100043466



PERPUSTAKAAN
KOLEJ UNIVERSITI SAINS & TEKNOLOGI MALAYSIA
21030 KUALA TERENGGANU

1100043466		

Lihat sebelah

HAK MILIK
PERPUSTAKAAN KUSTEM

30
26
2005

**PERLOMBONGAN DATA : KAJIAN TERHADAP TEKNIK -TEKNIK UNTUK
MENGESAN DATA TERSISIH**

AKBAR AHMAD

**Tesis Ini Dikemukakan Bagi
Memenuhi Sebahagian Daripada Syarat Untuk
Memperolehi Sarjana Muda Teknologi Maklumat
Kolej Universiti Sains dan Teknologi Malaysia (KUSTEM)**

1100043465



**JABATAN SAINS KOMPUTER
FAKULTI SAINS DAN TEKNOLOGI**

**PENGAKUAN DAN PENGESAHAN LAPORAN
PROJEK PENYELIDIKAN II**

Adalah ini diakui dan disahkan bahawa laporan penyelidikan bertajuk:

**Pelombongann Data : Kajian Terhadap Teknik-Teknik Untuk Mengesan Data
Tersisih**

Oleh **Akbar Ahmad**, No.Matrik **UK 6482** telah diperiksa dan semua pembedaan yang disarankan telah dilakukan. Laporan ini dikemukakan kepada Jabatan Sains Komputer sebagai memenuhi sebahagian daripada keperluan memperoleh **Ijazah Sarjana Muda Teknologi Maklumat (Kejuruteraan Perisian)**, Fakulti Sains dan Teknologi, Kolej Universiti Sains dan Teknologi Malaysia.

Disahkan oleh :

Penyelia Utama

Pn. Zuriana Abu Bakar

Cop Rasmi: **ZURIANA ABU BAKAR**
Pensyarah
Jabatan Sains Komputer
Fakulti Sains dan Teknologi
Kolej Universiti Sains dan Teknologi Malaysia
(KUSTEM)
21030 Kuala Terengganu.

Tarikh: 04.04.05

Ketua Jabatan Sains Komputer

P.M Dr. Mustafa Mat Deris.

Cop Rasmi:

Tarikh: 04.04.05

Prof. Madya Dr. Mustafa Mat Deris
Ketua
Jabatan Sains Komputer
Fakulti Sains & Teknologi
Universiti Sains dan Teknologi Malaysia
21030 Kuala Terengganu.

PENGAKUAN

Dengan ini saya mengakui bahawa segala karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

.....
AKBAR AHMAD

4 APRIL 2005

Disahkan oleh :



.....
P.M Dr. Mustafa Mat Deris
Ketua
Jabatan Sains Komputer
Fakulti Sains dan Teknologi
Kolej Universiti Sains dan Teknologi Malaysia



.....
Pn. Zuriana Abu Bakar
Pensyarah
Jabatan Sains Komputer
Fakulti Sains dan Teknologi
Kolej Universiti Sains dan Teknologi Malaysia

PENGHARGAAN

Assalammualaikum dan Salam Sejahtera.

Pertama sekali ingin saya ungkapkan kesyukuran kehadiran ilahi dengan limpah kurnia-Nya, akhirnya dapat juga menyiapkan kajian ini tanpa mengalami sebarang masalah semasa melaksanakannya. Seterusnya jutaan terima kasih kepada kedua ibu bapa dan keluarga tercinta berkat doa dan dorongan mereka segala masalah yang dihadapi semasa membuat kajian ini dapat diatasi.

Jutaan penghargaan juga dirakamkan kepada penyelia projek Puan Zuriana binti Abu Bakar yang telah banyak meluangkan masa dengan memberikan tunjuk ajar, pendapat dan komen bagi menjayakan kajian ini.

Ribuan penghargaan juga diberikan kepada teman-teman dan mereka yang terlibat samada secara langsung atau tidak langsung sepanjang saya menyiapkan kajian ini. Akhir kata segala kerjasama yang diberikan amatlah dihargai dan dikenang selamanya.

DATA MINING : A STUDY ON TECHNIQUES FOR DETECTING OUTLIER DATA

ABSTRACT

The quality of air in an area is commonly agreed as one of the main factors that will influence its general condition. Hence, the process of recording a clean and useful data for the data mining needs an appropriate and effective methods. That is why the purpose of this research being studied is to determine the best technique to detect the outlier of the air pollution based on the data taken from Victoria Institution in 2002. The comparison is done based on the distance, which is more widely known as the Manhattan technique and also based on deviation which is a technique using either Control Charts or modeled from the Linear Regression Model. The results from the research shown that the outlier produced by the Manhattan technique is based on the value of the predetermined threshold. If the value of threshold is big, the number of outlier will be small. Meanwhile, for the deviation technique, it can be concluded that the technique using Control Graph is better than the technique modeled from Linear Regression Model because the number of outlier produced by the former is smaller.

ABSTRAK

Kualiti udara merupakan salah satu faktor utama yang mempengaruhi keadaan persekitaran di sesuatu kawasan. Bagi merekodkan data yang bersih dan berguna pada perlombongan data ia memerlukan kaedah yang sesuai. Oleh yang demikian, tujuan kajian ini dilaksanakan untuk mengenalpasti teknik terbaik untuk mengesan data tersisih pada data pencemaran udara di *Victoria Institution* pada tahun 2002. Perbandingan dilakukan pada kaedah yang berdasarkan kepada jarak (*distance*) iaitu teknik *Manhattan* dan kaedah berdasarkan lencongan (*deviation*) iaitu teknik Graf Kawalan (*Control Chart*) dan teknik bermodelkan Regresi linear (*Linear Regression Model*). Hasil daripada kajian yang dilakukan, data tersisih yang dikenalpasti melalui teknik *Manhattan* adalah bergantung kepada nilai *threshold* yang ditetapkan iaitu semakin besar nilai *threshold* maka, bilangan data tersisih yang dihasilkan adalah lebih kecil. Manakala, bagi kaedah berdasarkan lencongan (*deviation*), teknik Graf Kawalan adalah lebih baik dibandingkan dengan teknik model Regresi Linear kerana data tersisih yang dikenalpasti melalui teknik Graf Kawalan lebih kecil daripada hasil teknik bermodelkan Regresi Linear.