

**SKRIPSI**

**DATA MINING HISTORI PENCARIAN RUTE ANGKOT**



**JOVAN GUNAWAN**

**NPM: 2011730029**

**PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS  
UNIVERSITAS KATOLIK PARAHYANGAN**

**2014**



**UNDERGRADUATE THESIS**

**DATA MINING HISTORY SEARCHING ROUTE**



**JOVAN GUNAWAN**

**NPM: 2011730029**

**DEPARTMENT OF INFORMATICS  
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES  
PARAHYANGAN CATHOLIC UNIVERSITY  
2014**



**LEMBAR PENGESAHAN**

**DATA MINING HISTORI PENCARIAN RUTE ANGKOT**

**JOVAN GUNAWAN**

**NPM: 2011730029**

**Bandung, 14 September 2014**

**Menyetujui,**

**Pembimbing Tunggal**

**Pascal Alfadian, M.Com.**

**Ketua Tim Penguji**

**Anggota Tim Penguji**

**Plato**

**Euclid**

**Mengetahui,**

**Ketua Program Studi**

**Thomas Anung Basuki, Ph.D.**



## **ABSTRAK**

**Kata-kata kunci:** Skripsi, Tugas Akhir





*Dipersembahkan untuk diri sendiri*



# KATA PENGANTAR

Bandung, September 2014

Penulis



## DAFTAR ISI

<b>KATA PENGANTAR</b>	<b>xi</b>
<b>DAFTAR ISI</b>	<b>xiii</b>
<b>DAFTAR GAMBAR</b>	<b>xiv</b>
<b>DAFTAR TABEL</b>	<b>xv</b>
<b>1 PENDAHULUAN</b>	<b>1</b>
1.1 Latar Belakang . . . . .	1
1.2 Perumusan Masalah . . . . .	2
1.3 Tujuan . . . . .	2
1.4 Batasan Masalah . . . . .	2
1.5 Metode Penelitian . . . . .	2
1.6 Sistematika Pembahasan . . . . .	2
<b>2 LANDASAN TEORI</b>	<b>3</b>
2.1 <i>Data Mining</i> . . . . .	3
2.1.1 <i>Data Cleaning</i> . . . . .	3
2.1.2 <i>Data Integration</i> . . . . .	5
2.1.3 <i>Data Selection</i> . . . . .	5
2.1.4 <i>Data Transformation</i> . . . . .	6
2.1.5 <i>Data Mining</i> . . . . .	8
2.1.6 <i>Pattern Evaluation</i> . . . . .	8
2.1.7 <i>Knowledge Presentation</i> . . . . .	8
2.2 Log Histori KIRI . . . . .	8
<b>DAFTAR REFERENSI</b>	<b>11</b>

## DAFTAR GAMBAR

2.1 Tahap <i>Data Mining</i> , Sumber Data Mining Concepts and Techniques . . . . .	4
---	---

## DAFTAR TABEL





# BAB 1

## PENDAHULUAN

### 1.1 Latar Belakang

Pertumbuhan teknologi hingga saat ini telah menghasilkan banyak sekali data-data, namun sering kali pemilik data hanya menggunakan data tersebut seperlunya saja. Jika dilihat lebih rinci, sebenarnya jika data tersebut diolah lebih lanjut, dapat menghasilkan sesuatu yang lebih. Salah satu cara mengolah data tersebut adalah dengan menggunakan teknik *data mining*. Dengan menggunakan teknik *data mining* akan mempermudah menganalisa masalah, pengambilan kesimpulan, bahkan mempermudah konsumen dalam membeli jasa atau barang.

Tujuan utama dari *data mining* adalah *knowledge*. *Knowledge* merupakan suatu informasi yang berharga dan dapat dijadikan landasan untuk menganalisa atau membuat kesimpulan. Untuk mendapatkan *knowledge*, dapat dilakukan dengan cara melakukan pencarian *pattern* atau pola yang merupakan salah satu tahap dari *data mining*. Pola inilah yang akan memperlihatkan data manakah yang menarik dan dapat dijadikan *knowledge* yang akan digunakan untuk menganalisa data tersebut.

Pada penelitian *data mining* ini, penulis memiliki data *log* histori KIRI selama 1 bulan. Data tersebut akan diimplementasikan proses *data mining* untuk mendapatkan *pattern* dan *knowledge* yang terkandung pada data *log* KIRI. Data *log* tersebut memiliki 5 *field* untuk setiap *entry* sebagai berikut:

- *statisticId*, primary key dari entry
- *verifier*, mengidentifikasi sumber dari pencarian ini
- *timestamp*, waktu ketika pengguna KIRI mencari rute angkot
- *type*, tipe fungsi yang digunakan
- *additionalInfo*, mencatat koordinat awal, koordinat akhir, dan banyak rute yang ditemukan pada pencarian ini

Berdasarkan hal diatas, penulis ingin mendapatkan pola yang menarik dan menghasilkan *knowledge* yang berguna dan dapat dipakai baik untuk KIRI ataupun pemerintah.

## 1.2 Perumusan Masalah

Dengan mengacu pada uraian diskripsi diatas, maka permasalahan yang dibahas dan diteliti oleh penulis adalah

- Bagaimana cara mengolah pola yang diperoleh dari *data log histori* KIRI agar pola menjadi menarik dan bermakna?
- Bagaimana membuat perangkat lunak untuk melakukan data mining pada data log history?

## 1.3 Tujuan

Penelitian ini bertujuan untuk

- Mencari pola dan informasi yang menarik dari *log histori* KIRI
- Perangkat lunak dapat melakukan data mining dari *log histori* KIRI

## 1.4 Batasan Masalah

Penelitian *data mining* yang diatas akan ditentukan batasan masalah yang diteliti berupa :

- Penelitian ini dibatasi hanya pada permasalahan pada penerapan *data mining* pada *data log* KIRI
- Data log yang merupakan masukan akan dibatasi sebanyak 10000 buah data

## 1.5 Metode Penelitian

Berikut adalah Metode Penelitian yang digunakan :

- Melakukan studi literatur tentang algoritma-algoritma yang berkaitan dengan pemrosesan *data mining*
- Melakukan penelitian *data mining* yang diterapkan pada *log* KIRI
- Merancang dan mengimplementasikan algoritma untuk *data mining*
- Mengimplementasikan pembangkit pola *data mining*
- Melakukan pengujian dan eksperimen

## 1.6 Sistematika Pembahasan

Sistematika pembahasan dalam penelitian ini adalah: Bab 1: Pendahuluan, berisi latar belakang dari penelitian ini, rumusan masalah yang timbul, tujuan yang ingin dicapai, ruang lingkup atau batasan masalah dari penelitian ini, serta metode penelitian yang akan digunakan dan sistematika pembahasan dari penelitian ini. Bab 2: Landasan Teori, berisi dasar teori mengenai *data mining* dan *spatial and Spatiotemporal* Bab 3: Bab 4: Bab 5:

## BAB 2

### LANDASAN TEORI

#### 2.1 *Data Mining*

*Data mining* merupakan merupakan proses yang melakukan pengambilan inti sari atau penggalian *knowledge* dari data yang besar dan merupakan salah satu langkah dari *knowledge discovery*.

Menurut [1], *knowledge discovery* dapat dibagi menjadi 7 tahap (gambar 2.1):

1. *Data cleaning*
2. *Data integration*
3. *Data selection*
4. *Data transformation*
5. *Data mining*
6. *Pattern Evaluation*
7. *Knowledge presentation*

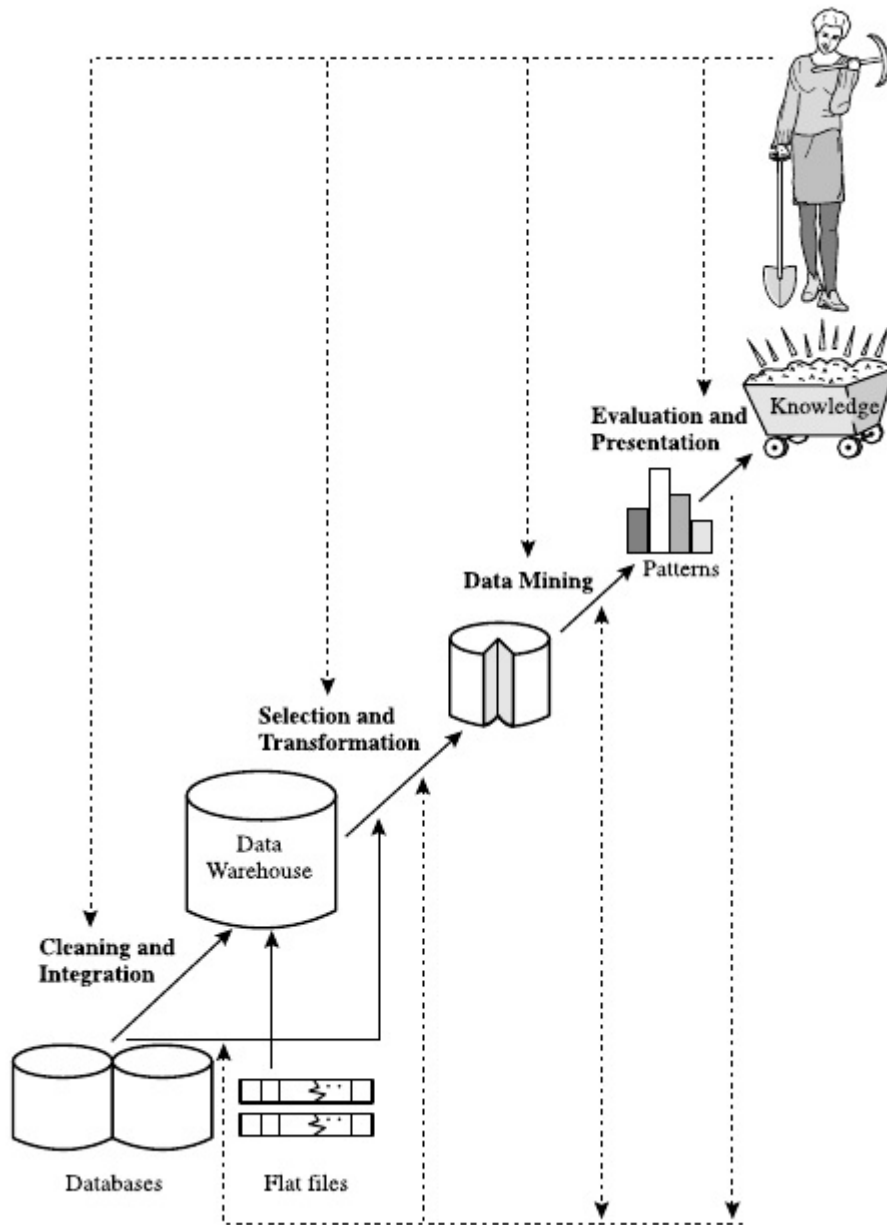
Tahap pertama hingga keempat merupakan bagian dari *data preprocessing*, dimana data-data disiapkan untuk dilakukan penggalian data. Tahap *data mining* merupakan tahap dimana melakukan penggalian data. Tahap keenam merupakan tahap pencarian pola yang merepresentasikan *knowledge*. Sedangkan tahap terakhir merupakan visualisasi dan representasi dari *knowledge* yang sudah diperoleh dari tahap sebelumnya.

##### 2.1.1 *Data Cleaning*

*Data cleaning* merupakan tahap *data mining* untuk menghilangkan *missing value* dan *noisy data*. Pada umumnya, *data* yang diperoleh dari *database* terdapat nilai yang tidak sempurna seperti nilai yang hilang, nilai yang tidak valid atau bahkan salah ketik. Nilai-nilai tersebut dapat diatasi dengan cara *smoothing techniques*. Atribut dari suatu *database* yang tidak relevan atau redundansi bisa diatasi dengan menghapus atribut tersebut.

##### *Missing Values*

*Missing values* akan mengganggu proses *data mining* pada komputer dan dapat menghasilkan nilai akhir yang tidak sesuai. Terdapat beberapa teknik untuk mengatasi *missing values* yaitu



Gambar 2.1: Tahap *Data Mining*, Sumber Data Mining Concepts and Techniques

- Membuang tuple yang terdapat nilai yang hilang
- Mengisi nilai yang hilang secara manual
- Mengisi nilai yang hilang dengan menggunakan nilai konstan yang bersifat umum
- Menggunakan nilai rata-rata dari suatu atribut untuk mengisi nilai yang hilang

### **Noisy Data**

*Noisy data* merupakan nilai yang berasal dari error atau tidak valid. *Noisy data* dapat dihilangkan dengan menggunakan teknik *smoothing*. Terdapat 3 metode untuk menghilangkan *noisy data* yaitu

- *Binning*, merupakan metode pengisian data sesuai dengan proses yang dilakukan pada data tersebut
- *Regression*, merupakan metode yang mencari persamaan atribut untuk memprediksikan suatu nilai
- *Clustering*, merupakan metode pengelompokan dimana ditemukan *outliers* yang dapat dibuang

#### **2.1.2 Data Integration**

*Data integration* merupakan tahap menggabungkan data dari berbagai sumber. Sumber tersebut bisa termasuk beberapa *database*, *data cubes*, atau bahkan *flat data*. Tahap ini harus dilakukan secara teliti terutama ketika dalam memasang nilai-nilai yang berasal dari sumber yang berbeda. Pada tahap ini, perlu dilakukan identifikasi data apakah data tersebut dapat diturunkan atau tidak agar data yang diperoleh tidak terlalu besar. *Data integration* yang baik merupakan integrasi yang dapat memaksimalkan kecepatan dan meningkatkan akurasi dari proses *data mining*.

#### **2.1.3 Data Selection**

Proses dimana data-data yang relevan dengan analisis akan diambil dari database dan data yang tidak relevan akan dibuang. Sebagai contoh kasus, jika akan dilakukan analisa mengenai nilai mahasiswa dalam satu semester, atribut pada tabel nilai sebagai berikut

- NPMMahasiswa
- NamaMahasiswa
- JenisKelamin
- Alamat
- MataKuliah
- NilaiART
- NilaiUTS

- NilaiUAS

Maka, atribut yang akan diambil adalah MataKuliah, NilaiART, NilaiUTS, NilaiUAS, sedangkan atribut yang akan dibuang adalah NPMMahasiswa, NamaMahasiswa JenisKelamin, dan Alamat karena tidak terlalu berhubungan dengan analisa.

### 2.1.4 Data Transformation

*Data transformation* merupakan tahap pengubahan data agar siap dilakukan proses *data mining*. *Data transformation* bisa melibatkan,

- *Smoothing*, proses untuk membuang *noise* seperti yang dilakukan pada tahap *data cleaning*
- *Aggregation*, proses mengganti nilai-nilai menjadi suatu nilai yang dapat mewakili nilai sebelumnya
- *Generalization*, proses dimana membuat suatu nilai yang bersifat khusus menjadi nilai yang bersifat umum
- *Normalization*, proses dimana suatu nilai dapat diubah skalanya menjadi nilai yang lebih kecil dan spesifik
- *Attribute construction*, proses membuat atribut baru yang berasal dari beberapa atribut untuk membantu proses data mining

—tanyakan!!!—>Proses *smoothing* merupakan proses yang sama seperti pada tahap *data cleaning*.<—tanyakan!!!—

### Normalization

Atribut dapat dinormalisasi dengan memberi skala pada nilainya sehingga nilai tersebut menjadi suatu range yang lebih spesifik dan kecil seperti 0,0 sampai 1,0. Dua teknik normalisasi yaitu, *min-max normalization* dan *z-score normalization*. *Min-max normalization* akan mengubah semua nilai menjadi nilai dengan skala tertentu. Dengan menggunakan rumus

$$\nu' = \frac{\nu - \min_A}{\max_A - \min_A}(\text{newMax}_A - \text{newMin}_A) + \text{newMin}_A$$

Contoh kasus, misalkan nilai minimum dan maximum dari suatu pendapatan adalah 12.000 dan 98.000, akan diubah menjadi berskala antara 0,0 sampai 1,0. Jika ada nilai pendapat yang baru, yaitu 73.600, maka akan menjadi

$$\frac{73.600 - 12.000}{98.000 - 12.000}(1,0 - 0) + 0 = 0,716$$

*z-score normalization* merupakan normalisasi berdasarkan nilai rata-rata dan standar deviasi dari nilai-nilai atribut dengan cara

$$\nu' = \frac{\nu - \bar{A}}{\sigma_A}$$

Contoh kasus, misal nilai rata-rata dan standar deviasi dari nilai-nilai atribut pendapatan adalah 54.000 dan 16.000. Dengan *z-score*, jika ada nilai pendapatan baru yaitu 73600, maka akan diubah menjadi

$$\frac{73.600 - 54.000}{16.000} = 1,225$$

### ***Attribute Construction***

*Attribute Construction* merupakan teknik menambahkan atribut baru yang berdasarkan dari atribut yang sudah ada guna menambah akurasi. Contoh kasus, dibuat atribut baru bernama area berdasarkan atribut panjang dan lebar.

### ***Data Reduction***

Proses *aggregation* dan *generalization* akan dilakukan dalam bentuk proses *data reduction* dan *Data Cube Aggregation*. *Data reduction* dilakukan untuk mendapatkan nilai yang representatif namun tetap menjaga keakuratan hasil *data mining*. Terdapat beberapa cara dalam mengimplementasikan *data reduction* yaitu

- *Data subset selection*
- *Dimensionality reduction*
- *Numerosity reduction*
- *Discretization and concept hierarchy generation*

### ***Attribute Subset Selection***

*Attribute subset selection* merupakan salah satu cara melakukan *data reduction* dengan menghilangkan atribut-atribut yang tidak relevan atau data yang redundansi. Hal ini dapat mempermudah pencarian pola dikarenakan banyak atribut yang muncul akan berkurangnya.

### ***Dimensionality Reduction***

*Dimensionality Reduction* merupakan metode pengurangan nilai secara acak dengan cara melakukan konversi data. Jika data original dapat dibuat ulang dari data yang sudah dikompresi tanpa kehilangan informasi, maka akan dikatakan *lossless*, namun jika hanya mendapatkan data pendekatannya saja, akan disebut *lossy* [1].

### ***Numerosity Reduction***

*Numerosity Reduction* merupakan metode dimana data diganti atau ditentukan dengan cara parametrik atau nonparametrik.

### ***Discretization and Concept Hierarchy Generation***

lewat dulu

### 2.1.5 *Data Mining*

### 2.1.6 *Pattern Evaluation*

*Pattern evaluation* merupakan tahap mengidentifikasi apakah *pattern* atau pola tersebut menarik dan merepresentasikan *knowledge* berdasarkan beberapa *interestingness measures*. Suatu *pattern* atau pola dapat dinyatakan menarik apabila

- mudah dimengerti oleh manusia
- valid untuk data percobaan maupun data yang baru
- memiliki potensi atau berguna
- merepresentasikan *knowledge*

### 2.1.7 *Knowledge Presentation*

*Knowledge presentation* merupakan tahap representasi dan visualisasi terhadap *knowledge* yang merupakan hasil dari *knowledge discovery*.

## 2.2 Log Histori KIRI

KIRI memiliki log histori yang melakukan pencatatan untuk setiap user ketika menggunakan KIRI. Log tersebut memiliki 5 *field* untuk setiap *entry* sebagai berikut:

- *logId*, primary key dari entry
- *APIKey*, mengidentifikasi sumber dari pencarian ini
- *Timestamp* (UTC), waktu ketika pengguna KIRI mencari rute angkot menggunakan waktu UTC / GMT
- *Action*, tipe log, untuk penelitian ini selalu berisi FINDROUTE
- *AdditionalData*, mencatat koordinat awal, koordinat akhir, dan banyak rute yang ditemukan pada pencarian ini

*LogId* merupakan *field* dengan tipe data int dengan batas 6 karakter yang digunakan sebagai *primary key* dari tabel tersebut. *LogId* diisi dengan menggunakan fungsi *increment integer*. *Increment integer* merupakan fungsi untuk pengisian data pada database dengan menambahkan nilai 1 dari nilai yang terakhir kali diisi. *APIKey* merupakan *field* dengan tipe data varchar yang digunakan untuk memeriksa pengguna KIRI ketika menggunakan KIRI. *Timestamp* (UTC) merupakan *field* dengan tipe data *timestamp* yang digunakan untuk mencatat waktu penggunaan KIRI oleh user, diisi dengan menggunakan fungsi *current time*. *Current time* merupakan fungsi untuk pengisian data pada database dengan mengambil waktu pada komputer ketika record dibuat. *Action* merupakan *field* dengan tipe data varchar yang digunakan untuk memeriksa fungsi apa yang dipanggil dari API KIRI. Terdapat beberapa tipe pada *field* ini, yaitu /



- *ADDAPIKEY*, *action* yang dicatat ke dalam log ketika fungsi pembuatan *API key* yang baru dipanggil.
- *FINDROUTE*, *action* yang dicatat ketika user melakukan pencarian rute
- *LOGIN*, *action* yang dicatat ketika developers melakukan login dengan menggunakan *API key*
- *NEARBYTRANSPORT*, *action* yang dicatat ketika user mencari transportasi di daerah rute sedang dicari
- *PAGELOAD*, *action* yang dicatat ketika user memasuki halaman KIRI
- *REGISTER*, *action* yang dicatat ketika developers melakukan pendaftaran pada KIRI *API key*
- *SEARCHPLACE*, *action* yang dicatat ketika user memanggil fungsi pencarian lokasi dengan menggunakan nama tempat
- *WIDGETERROR*, mencatat log tersebut ketika user menerima error dari *widget*
- *WIDGETLOAD*, mencatat log tersebut ketika user mengdownload widget

*AdditionalData*, merupakan *field* dengan tipe data *varchar* yang digunakan untuk mencatat informasi yang dibutuhkan sesuai dengan *field action*.



## DAFTAR REFERENSI