ANALISIS PERILAKU PENGGUNA DIGILIB UNAIR DENGAN MENGGUNAKAN *CLUSTERING LOG DATA MINING*

Yunus Abdul Halim[1]

Abstracts

The analysis of user attitude digilib UNAIR (ADLN) is known through access log web ADLN which contains information about access of location, period of time, content of information and when user access it. All of that information has been saved at file log-access.txt.

This research used real data about 39.086 records in log web ADLN from May 2008 to 2009. Clustering data mining was used for knowing user attitude in this analysis. This clustering method used K-mean for clustering data.

According to analysis clustering result at access log data mining Airlangga Digital Library shows 92,24% user access information via ADLN service in Airlangga Library which takes time more than 10 minutes. Also, the analysis result shows 75.64% user access more than 10 minutes from 92.24%. Whereas, almost of all the user open the ADLN site at around 10.00 WIB to 15.00 WIB.

That condition should need attention if the library wants to develop ADLN as one of source richt file for valuation at webometrics degree. In the fact that richt file has contributed a big point in ADLN, however when we are looking at the access ADLN shows that there are lack of access of ADLN through internet rather than intranet, so it influences on valuation webometrics from accessible point.

Keywords: *digital library, Clustering, k-mean*

Latar Belakang

Perpustakaan Digital adalah sebuah sistem yang memiliki berbagai layanan dan obyek informasi yang mendukung akses obyek informasi tesebut melalui perangkat digital. Layanan ini diharapkan dapat mempermudah pencarian informasi di dalam koleksi obyek informasi seperti dokumen, gambar dan database dalam format digital dengan cepat, tepat, dan akurat.

Koleksi perpustakaan digital tidaklah terbatas pada dokumen elektronik pengganti bentuk cetak saja, ruang lingkup koleksinya malah sampai pada artefak digital yang tidak bisa digantikan dalam bentuk tercetak. Koleksi menekankan pada isi informasi, jenisnya dari dokumen tradisional sampai hasil penelusuran. Perpustakaan ini melayani mesin, manajer informasi, dan pemakai informasi. Semuanya ini demi mendukung manajemen koleksi, menyimpan, pelayanan bantuan penelusuran informasi.Gagasan perpustakaan digital ini diikuti Kantor Kementerian Riset dan Teknologi dengan program Perpustakaan Digital yang diarahkan memberi kemudahan akses dokumentasi data ilmiah dan teknologi dalam bentuk digital secara terpadu dan lebih dinamis. Upaya ini dilaksanakan untuk mendokumentasikan berbagai produk intelektual seperti tesis, disertasi, laporan penelitian, dan juga publikasi kebijakan.

[1] Korespondensi: Yunus Abdul Halim. Departemen Informasi dan Perpustakaan, FISIP, Universitas Airlangga. Jl. Airlangga 4-6 Surabaya, 60286, Indonesia. Telp. (031) 5011744. E-Mail: zero@unair.ac.id

Digitasi perpustakaan merupakan salah satu jawaban terhadap pelayanan sirkulasi dan pelayanan informasi yang selama ini dikeluhkan masyarakat pengguna jasa perpustakaan. Hal ini tentunya dapat mengeliminir image negatif terhadap perpustakaan beralih fungsi menjadi tempat nongkrong, gosip, dan sebagainya dan bukan tidak dapat memainkan perannya yang signifikan sebagai bagian dalam dunia informasi, baik yang bersifat ilmiah, edukatif, rekreatif, ataupun fungsi-fungsi lainnya. Beberapa keunggulan perpustakaan digital diantaranya yaitu *long distance service*, akses yang mudah, murah (*cost efective*), pemeliharaan koleksi secara digital, jawaban yang tuntas, jaringan global. Chapman dan Kenney (1996), mengemukakan empat alasan yaitu: institusi dapat berbagi koleksi digital, koleksi digital dapat mengurangi kebutuhan terhadap bahan cetak pada tingkat lokal, penggunaannya akan meningkatkan akses elektronik, dan nilai jangka panjang koleksi digital akan mengurangi biaya berkaitan dengan pemeliharaan dan penyampaiannya.

Di sisi lain, Internet sebagai media dimana bahan digital tersedia, standar dan teknologinya akan terus mengalami pertumbuhan dan perkembangan.Perpustakaan yang mengembangkan perpustakaan digital apabila infrastruktur dan peralatan yang diperlukan sudah tersedia. Langkah selanjutnya, pustakawan harus mampu mengidentifikasi sumberdaya yang tersedia di dalam sekolah terutama sumberdaya manusia yang dapat dijadikan mitra dalam pengembangan. Kolaborasi sebagai hubungan formal dalam proses pengembangan mulai dari formulasi ide, perancangan, pengujian produk hingga implementasi adalah sangat penting..

Sosialisasi program perpustakaan digital terhadap para anggota jaringan dan para pengguna itu penting. Dalam hal ini, perlu peningkatan kesadaran akan fungsi utama mereka, yaitu memberikan kemudahan akses pengguna terhadap informasi. Untuk mempermudah akses, pustakawan perlu mendorong pengguna perpustakaan digital untuk melek informasi (*information literate*). Pengguna perpustakaan yang seperti ini adalah mereka yang sadar kapan memerlukan informasi dan mampu menemukan informasi, mengevaluasinya, dan menggunakan informasi yang dibutuhkannya itu secara efektif dan beretika. Perpustakaaan digital bisa juga dikatakan sebagai *virtual library* (perpustakaan maya) karena semua pengguna perpusakaan ini harus menggunakan internet. Semua akses terhadap perpustakaan digital dapat dilihat pada *log access*, yang terdapat pada *apahce services*. Melalui *log access* tersebut perilaku pengguna perpstakaan digital datap diketahui.

Salah satu metode yang bisa digunakan untuk menganalisis *log access* yaitu data mining. Data mining merupakan sebuah analisa dari observasi data dalam jumlah besar untuk menemukan hubungan yang tidak diketahui sebelumnya dan metode baru untuk meringkas data agar mudah dipahami serta kegunaannya untuk pemilik data (David Hand *et al*, 2001), sedangkan *Clustering* adalah salah satu teknik *unsupervised learning* dimana kita tidak perlu melatih metode tersebut atau dengan kata lain, tidak ada *fase learning.* Melalui monitoring dan analisa *clustering log acces* pada site perpustakaan digital maka akan dapat diketahui pola atau perilaku pengguna perpustakaan digital.

Permasalaan

Pengguna perpustakaan digital merupakan pengguna internet, karena hanya melalui internet perpustakaan digital dapat diakses. Karena semuanya dilakukan di dunia maya maka sulit mengukur perilaku pengguna perpusakaan secara tradisional. Salah satu cara yang memungkinkan yaitu menganalisis *log acces* pada site perpustakaan digital, sehingga timbul permasalahan-permasalahan antara lain, yaitu:

- Sulitnya melakukan analisa pengguna yang efektif karena tidak adanya sistem yang dapat menyajikan data historis sehingga dapat memberikan *output* tentang berapa banyak jumlah *pengguna* yang dimiliki dan kelompok-kelompok *pengguna* yang aktif maupun tidak menurut frekuensi transaksinya, karena data yang ada masih berbentuk data manual dan belum dimanfaatkan secara maksimal, sehingga harus memanfaatkan data *log access*.
- Tidak diketahui dengan pasti jumlah *penguna* yang aktif dan yang kurang aktif dalam melakukan transaksi, sehingga sangat sulit untuk melakukan mengetahui posisi jumlah pengguna secara tepat.

Batasan Masalah

Untuk menghindari salah pengertian dan untuk lebih menfokuskan terhadap permasalahan, maka fokus permasalahan dititik beratkan pada memanfaatkan *data mining* dengan *metode clustering* menggunakan *algoritma hirarki divisive* untuk melakukan analisis perilaku pengguna perpustakaan digital.

Batasan perumusan masalah dalam penelitian yang dilakukan, meliputi:

- *Competitive intelligence* dalam penelitian ini hanya sebatas untuk melakukan pengelompokan pengguna berdasarkan data transaksi yang dilakukan saja tanpa melakukan proses *competitive intelligence* lainnya.
- Basis data yang akan digunakan dalam studi kasus ini adalah basis data pengguna dan transaksi yang dilakukan saja tanpa melibatkan basis data lainnya, yang kemudian akan diolah berdasarkan proses-proses yang ada dalam *data mining.*
- Kemiripan antar data dalam penelitian ini diterjemahkan sebagai jarak kedekatan antar data dengan titik pusat, sehingga menghasilkan klaster-klaster *customer* yang sesuai dengan tujuan dari penelitian.
- Penggunaan *metode clustering* untuk mengelompokan *customer* dengan menggunakan *algoritma hirarki divisive k-means*.
- Data penelitian yang digunakan yaitu log access perpustakaan digital Universitas Airlangga Surabaya selama kurun waktu minimal 3 tahun.

Tujuan

Maksud dan tujuan dari penelitian analisis perilaku pengguna perpustakaan digital dengan menggunakan *cluster log access data mining*, adalah:

a. Menerapkan proses *data mining* untuk pengolahan basis data pengguna dengan *metode clustering* menggunakan *algoritma hirarkis divisive k-means* untuk mengelompokan pengguna.
b. Mengetahui kemiripan karakteristik antar data dalam basis data pengguna berdasarkan transaksi yang dilakukan, guna membentuk kelompok – kelompok pengguna melalui *metode clustering* dan *algoritma hirarki divisive k-means.*
c. Menganalisis perilaku dan karakteristik pengguna perpustakaan digital Universitas Airlangga Surabaya.

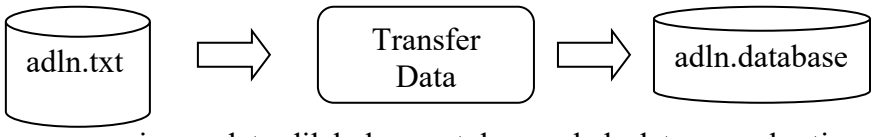
Data dan Metode

Penelitian ini menggunakan data primer berupa data log akses web adln.lib.unair.ac.id, yang berisikan karakteristik pengguna *Airlangga Digitall Library* dalam bentuk txt. File txt yang digunakan yaitu log akses satu tahun penuh selama satu tahun yakni bulan Mei 2008 sampai dengan Mei 2009 dengan atribut log akses terdiri

dari alamat pengakses, waktu akses, lama akses dan konten yang sedang diakses. Total log akses selama bulan Mei 2008 s/d Mei 2009 sebesar 39.086 *record*.

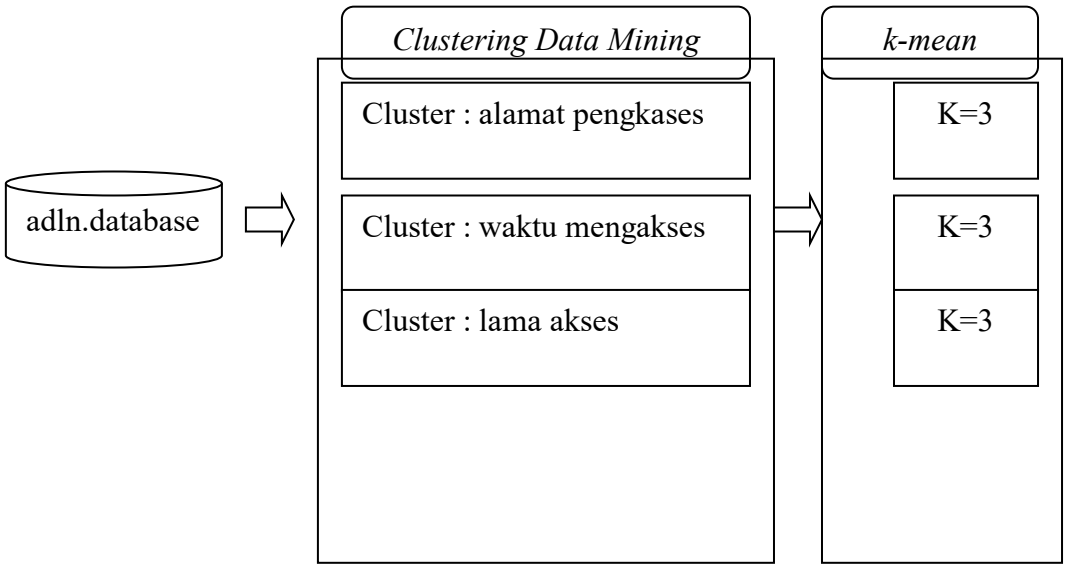
Metode penelitian yang digunakan yaitu penelitian deskriptif dengan menggunakan analisis *clustering data mining* dengan algoritma partisi yaitu k-mean. Adapun langkah langkah pengolahan *clustering data mining* sebagai berikut:

a. Penyiapan data



Proses penyiapan data dilakukan untuk merubah data yang bertipe txt menjadi database, berikut dengan melakukan *cleansing* terhadap data yang tidak dibutuhkan. *Field* data yang diambil yaitu alamat pengakses, waktu akses, lama akses dan konten yang diakses.

b. Pengolahan data



- *Clustering* alamat pengakses dikelompokkan menjadi 3 yaitu internal perpus, intranet Unair dan luar Unair.
- *Clustering* waktu mengakses dikelompokkan menjadi 3 yaitu jam 07.00 s/d 10.00, 10.00 s/d 15.00 dan di atas 15.00.
- *Clustering* lama akses dikelompokkan menjadi 3 yaitu <5 menit, 5 s/d 10 menit dan diatas 10 menit.

Struktur Database

Penelitian perilaku pengguna adln.lib.unair.ac.id hanya bersifat deskriptif terhadap pola akses yang meliputi lokasi pengkases, waktu akses dan lama akses. Berkenaan dengan hal tersebut maka struktur data yang digunakan menggunakan emapat tabel yaitu tabel log, cluster_akses, cluster_jam_akses dan cluster_waktu_akses.

Tabel 1. Tabel Log

Atribut	Jenis	Panjang
Alamat_pengakses	varchar	20
tgl_akses	date	default

awal_akses	time stamp	default
akhir_akses	time stamp	default
Konten	text	default

Tabel 2. Tabel Cluster_akses

Atribut	Jenis	Panjang
Alamat_pengakses	varchar	20
status_akses	varchar	1

Tabel 3. Tabel Cluster_jam_akses

Atribut	Jenis	Panjang
kelompok	varchar	5
jam_akses	time stamp	default

Tabel 4. Tabel cluster_waktu_akses

Atribut	Jenis	Panjang
kelompok	varchar	5
waktu_akses	time stamp	default

ICT dan Perpustakaan

Di era globalisasi ini memungkinkan banyaknya akses untuk mencari informasi dari segala penjuru dunia. Salah satunya adalah melalui perpustakaan. Dengan adanya perpustakaan kita dapat mencari, mengolah ataupun menyimpan data, yang kini telah berkembang dalam bentuk digital, atau yang dikenal dengan *perpustakaan digital*. Teknologi informasi atau *Information and Communication Technology* (ICT) telah membawa perubahan dalam berbagai sektor, termasuk perpustakaan. Perubahan penting dan mendasar bagi pengelolaan perpustakaan, baik dalam memberikan layanan maupun dalam menjalin hubungan antar lembaga, unit atau institusi.

Terjadinya perubahan pola pikir tentang perpustakaan, yaitu penyediaan koleksi yang dimiliki ke arah konsep *"tidak harus memiliki*" akan tetapi dapat *"memberikan informasi"*, telah menjadikan jalinan kerjasama antar perpustakaan dalam menampilkan koleksi yang dapat memudahkan penyampaian informasi, semakin mudah untuk diwujudkan, apalagi dengan adanya ICT. Maka konsep gedung yang besar dan mewah serta banyaknya koleksi bukan merupakan sesuatu yang ideal lagi.

Menurut Budi Sutedjo (2002:168) dan Rahayuningsih, Rochaety, Yanti, (2006:4). Informasi merupakan pemrosesan data yang diperoleh dari setiap elemen sistem menjadi bentuk yang mudah dipahami dan merupakan pengetahuan yang relevan dan dibutuhkan, dimana Informasi itu sendiri merupakan pernyataan yang menjelaskan suatu peristiwa sehingga manusia dapat membedakan antara satu dengan yang lainnya.

Implementasi ICT di perpustakaan perlu direncanakan secara matang karena memerlukan pendanaan yang tidak murah, apalagi perkembangan teknologi khususnya komputer terus berubah dengan sangat cepat. Hal ini untuk mengantisipasi kinerja aplikasi ICT dapat dioptimalkan. Kesia-siaan dapat terjadi karena perencanaan yang kurang baik yang dapat mengakibatkan pemborosan.

Adapun penerapan teknologi informasi di perpustakaan dapat difungsikan dalam bentuk:

1. *Automasi Perpustakaan*: Konsep Sistem Informasi Manajemen (SIM) perpustakaan, yang menekankan aplikasi ICT antar sub sistem informasi perpustakaan pengadaan, inventarisasi, katalogisasi, sirkulasi, pengelolaan anggota, dan statistik dalam bentuk terintegrasi.
2. *Perpustakaan Digital* : Penerapan teknologi informasi sebagai sarana untuk menyimpan, mendapatkan dan menyebarluaskan informasi, ilmu pengetahuan dan teknologi lokal secara *full text* dalam format digital seperti tugas akhir (skripsi, tesis, disertasi), laporan penelitian dan artikel majalah ilmiah.
3. *Publikasi e-books :* Publikasi buku elektronik untuk kepentingan lokal (internal), dimaksudkan untuk kemudahan dalam pencarian dan mendapatkan kembali secara utuh sesuai dengan format aslinya.

Ketiga fungsi penerapan teknologi informasi ini dapat terpisah maupun terintegrasi sebagai suatu sistem informasi, tergantung dari kemampuan software yang digunakan, sumber daya manusia, dan infrastruktur peralatan teknologi informasi yang mendukung ketiganya.

Peran Internet

Orang sudah tidak asing lagi untuk menggunakan internet dalam kehidupannya. Untuk itu perpustakaanpun harus dapat memberikan layanan melalui media ini. Melalui media web, perpustakaan memberikan informasi dan layanan kepada penggunanya. Selain itu, perpustakaan juga dapat menyediakan akses internet baik menggunakan *computer station* pribadi maupun *Access Poin Cyberlib* yang tersedia di perpustakaan pusat ITB dapat digunakan pengguna sebagai bagian dari layanan yang diberikan oleh perpustakaan. Perpustakaan juga bisa menggunakan fasiltas *web-conferencing* untuk memberikan layanan secara interaktif kepada pengguna perpustakaan. *Web-Conferencing* ini dapat juga dimanfaatkan oleh dosen dalam rangka kuliah jarak jauh. Awal tahun 2008, perpustakaan ITB mendapat hibah peralatan *Tele Conference* dari keduataan Amerika. *OPAC* atau *Online Catalog* merupakan bagian penting dalam sebuah perpustakaan.

Pustakawan harus dapat melayani keperluan pengguna seperti permintaan dengan akses yang lebih cepat ke informasi yang diperlukan dari dalam maupun luar perpustakaan. Dengan begitu, diharapkan agar para pustakawan mahir dalam penggunaan teknologi informasi sehingga mereka dapat membantu pengguna perpustakaan dalam menemukan informasi yang diperlukan. Fasilitas memungkinkan yaitu melalui sebuah portal web.

Melalui portal web semua aktifitas kegiatan akses terhadap *resource* perpustakaan kelihatan, seperti pola kunjungan akses perpustakaan digital dan *e-book*. Semua akses pengunjung melalui internet tercatat dalam log akses server. Log tersebut berisikan informasi data yang bisa dimanfaatkan untuk mengetahui pola perilaku pengunjung perpustakaan melalui internet.

Pengertian Data Warehouse

Menurut W.H. Inmon dan Richard D.H., *data warehouse* adalah koleksi data yang mempunyai sifat berorientasi subjek,terintegrasi,time-variant, dan bersifat tetap dari koleksi data dalam mendukung proses pengambilan keputusan management. Sedangkan Vidette Poe, *data warehouse* merupakan database yang bersifat analisis dan *read only* yang digunakan sebagai fondasi dari sistem penunjang keputusan. *Data*

warehouse juga bisa diartikan sebagai database relasional yang didesain lebih kepada *query* dan analisa dari pada proses transaksi, biasanya mengandung *history* data dari proses transaksi dan bisa juga data dari sumber lainnya.

Data warehouse memisahkan beban kerja analisis dari beban kerja transaksi dan memungkinkan organisasi menggabung/konsolidasi data dari berbagai macam sumber. Jadi, *data warehouse* merupakan metode dalam perancangan database, yang menunjang DSS(Decission Support System) dan EIS (Executive Information System). Secara fisik *data warehouse* adalah database, tapi perancangan data warehouse dan database sangat berbeda. Dalam perancangan database tradisional menggunakan normalisasi, sedangkan pada *data warehouse* normalisasi bukanlah cara yang terbaik.

Sedangkan Karakteristik data warehouse menurut Inmon, yaitu :

***a. Subject Oriented* (Berorientasi subject)**

Tabel 5. Data Operasiaonal dan Data Warehouse

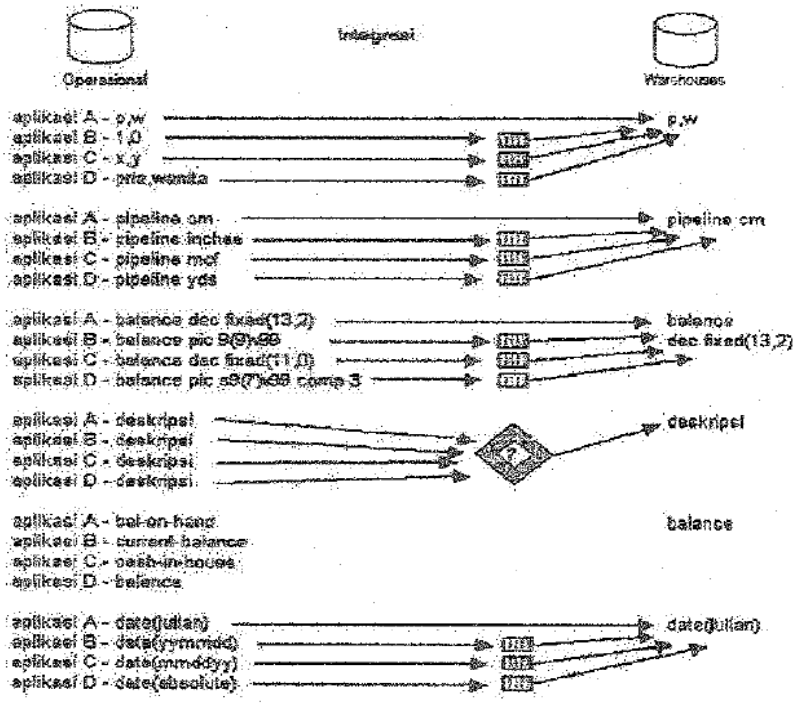
Data Operasional	Data Warehouse
Dirancang berorientasi hanya pada aplikasi dan fungsi tertentu	Dirancang berdasar pada subjek-subjek tertentu(utama)
Focusnya pada desain database dan proses	Focusnya pada pemodelan data dan desain data
Berisi rincian atau detail data	Berisi data-data history yang akan dipakai dalam proses analisis
Relasi antar table berdasar aturan terkini(selalu mengikuti rule(aturan) terbaru)	Banyak aturan bisnis dapat tersaji antara tabel-tabel

Data warehouse berorientasi subject artinya data warehouse didesain untuk menganalisa data berdasarkan subject-subject tertentu dalam organisasi,bukan pada proses atau fungsi aplikasi tertentu. Data warehouse diorganisasikan disekitar subjek-subjek utama dari perusahaan(customers,products dan sales) dan tidak diorganisasikan pada area-area aplikasi utama(customer invoicing,stock control dan product sales). Hal ini dikarenakan kebutuhan dari data warehouse untuk menyimpan data-data yang bersifat sebagai penunjang suatu keputusan, dari pada aplikasi yang berorientasi terhadap data. Jadi dengan kata lain, data yang disimpan adalah berorientasi kepada subjek bukan terhadap proses. Secara garis besar perbedaan antara data operasional dan data warehouse dapat dilihat pada Tabel 2.1.

***b.Integrated* (Terintegrasi)**

Data Warehouse dapat menyimpan data-data yang berasal dari sumber-sumber yang terpisah kedalam suatu format yang konsisten dan saling terintegrasi satu dengan lainnya. Dengan demikian data tidak bisa dipecah-pecah karena data yang ada merupakan suatu kesatuan yang menunjang keseluruhan konsep data warehouse itu sendiri. Syarat integrasi sumber data dapat dipenuhi dengan berbagai cara sepeti konsisten dalam penamaan variable,konsisten dalam ukuran variable,konsisten dalam struktur pengkodean dan konsisten dalam atribut fisik dari data. Contoh pada lingkungan operasional terdapat berbagai macam aplikasi yang mungkin pula dibuat oleh developer yang berbeda. Oleh karena itu, mungkin dalam aplikasi-aplikasi tersebut ada variable yang memiliki maksud yang sama tetapi nama dan format nya berbeda. Variable tersebut harus dikonversi menjadi nama yang sama dan format yang disepakati bersama. Dengan demikian tidak ada lagi kerancuan karena perbedaan nama, format

dan lain sebagainya. Barulah data tersebut bisa dikategorikan sebagai data yang terintegrasi karena kekonsistenannya.

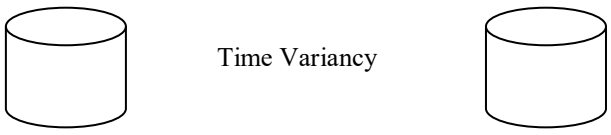


Gambar 1. Integrasi Data Warehouse

c. Time-variant (*Rentang Waktu*)

Seluruh data pada data warehouse dapat dikatakan akurat atau valid pada rentang waktu tertentu. Untuk melihat interval waktu yang digunakan dalam mengukur keakuratan suatu data warehouse, kita dapat menggunakan cara antara lain :

- Cara yang paling sederhana adalah menyajikan data warehouse pada rentang waktu tertentu, misalnya antara 5 sampai 10 tahun ke depan.
- Cara yang kedua, dengan menggunakan variasi/perbedaan waktu yang disajikan dalam data warehouse baik implicit maupun explicit secara explicit dengan unsur waktu dalam hari, minggu, bulan dsb. Secara implicit misalnya pada saat data tersebut diduplikasi pada setiap akhir bulan, atau per tiga bulan. Unsur waktu akan tetap ada secara implisit didalam data tersebut.
- Cara yang ketiga,variasi waktu yang disajikan data warehouse melalui serangkaian snapshot yang panjang. Snapshot merupakan tampilan dari sebagian data tertentu sesuai keinginan pemakai dari keseluruhan data yang ada bersifat read-only.

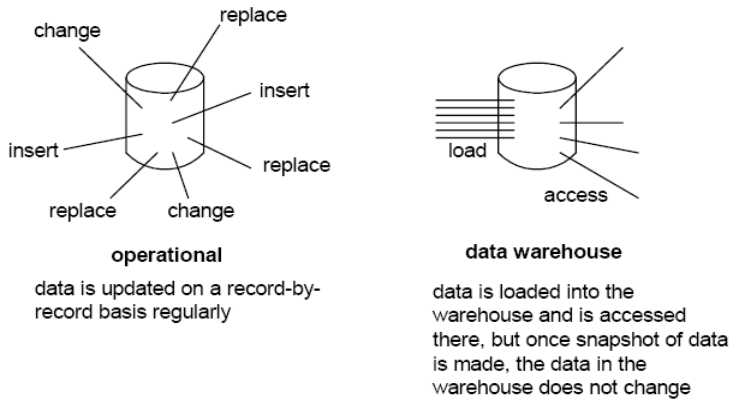


Operasional	Data Warehouse
Current value data:	Snapshot data:
- time horizon :60-90 days - key may or may not have an element of time - data can be update	- time horizon :5-10 years - key contain an element of time - once snapshot is created, record cannot be update

Gambar 2. Time Variancy Data Operasional dan Warehouse

d. Non-Volatile

Karakteristik keempat dari data warehouse adalah non-volatile,maksudnya data pada data warehouse tidak di-*update* secara *real time* tetapi di *refresh* dari sistem operasional secara reguler. Data yang baru selalu ditambahkan sebagai suplemen bagi database itu sendiri dari pada sebagai sebuah perubahan. Database tersebut secara kontinyu menyerap data baru ini, kemudian secara incremental disatukan dengan data sebelumnya. Berbeda dengan database operasional yang dapat melakukan update,insert dan delete terhadap data yang mengubah isi dari database sedangkan pada data warehouse hanya ada dua kegiatan memanipulasi data yaitu loading data (mengambil data) dan akses data (mengakses data warehouse seperti melakukan query atau menampilan laporan yang dibutuhkan, tidak ada kegiatan updating data).



Gambar 3. Non Volatile Data Warehouse

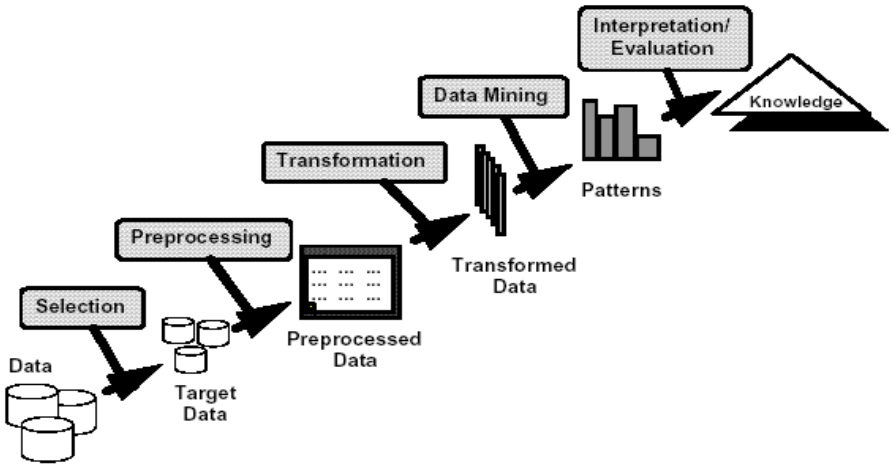
Data warehouse merupakan pendekatan untuk menyimpan data dimana sumber-sumber data yang heterogen(yang biasanya tersebar pada beberapa database OLTP) dimigrasikan untuk penyimpanan data yang homogen dan terpisah. Keuntungan yang didapatkan dengan menggunakan data warehouse tersebut dibawah ini (Ramelho).

Sedangkan kombinasi data mining verifikasi dan penemuan merupakan perkembangan *data mining* di masa depan akan mengkombinasikan pendekatan hipotesis dan penemuan. Perkembangan ini menggunakan penalaran yang sama yang mendasari konsep Sistem Pendukung Keputusan (*Decision Support System* – DSS). Konsep tersebut memungkinkan pemakai dan komputer bekerja sama untuk

memecahkan suatu masalah. Pemakai menerapkan keahliannya dalam hal masalah, dan komputer melakukan analisis data yang canggih untuk memilih data yang tepat dan menempatkannya dalam format yang tepat untuk pengambilan keputusan. Menurut Fayyad Usama (1996), proses KDD secara garis besar dapat dijelaskan sebagai berikut:

1. *Data Selection*
   Pemilihan (seleksi) data daru sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional.
2. *Pre-processing/ Cleaning*
   Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD.
   Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi).
   Juga dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.
3. *Transformation*
   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data
4. *Data mining*
   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.
5. *Interpretation/ Evaluation*
   Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut dengan *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

Proses KDD secara garis besar memang terdiri dari 5 tahap seperti yang telah dijelaskan sebelumnya. Akan tetapi, dalam proses KDD yang sesungguhnya, dapat saja terjadi iterasi atau pengulangan pada tahap tahap tertentu. Pada setiap tahap dalam proses KDD, seorang analis dapat saja kembali ke tahap sebelumnya. Sebagai contoh, pada saat *coding* atau *data mining*, analis menyadari proses *cleaning* belum dilakukan dengan sempurna, atau mungkin saja analis menemukan data atau informasi baru untuk "memperkaya" data yang sudah ada.



Gambar 4. Tahapan proses KDD

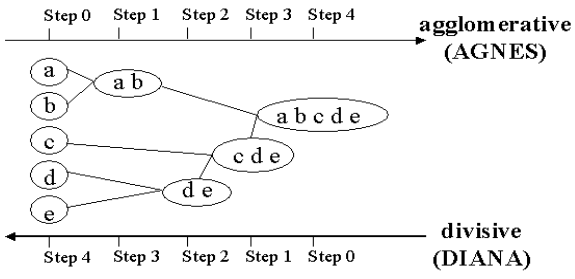
KDD mencakup keseluruhan proses pencarian pola atau informasi dalam basis data, dimulai dari pemilihan dan persiapan data sampai representasi pola yang ditemukan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. *Data mining* merupakan salah satu komponen dalam KDD yang difokuskan pada penggalian pola tersembunyi dalam basis data.

Hierarchical Clustering

Pada algoritma *clustering*, *data* akan dikelompokkan menjadi *cluster-cluster* berdasarkan kemiripan satu *data* dengan yang lain. Prinsip dari *clustering* adalah memaksimalkan kesamaan antar anggota satu *cluster* dan meminimumkan kesamaan antar anggota *cluster* yang berbeda. Kategori algoritma *clustering* yang banyak dikenal adalah *Hierarchical Clustering*. *Hierarchical Clustering* adalah salah satu algoritma *clustering* yang dapat digunakan untuk meng-*cluster* dokumen *(document clustering*). Dari teknik *hierarchical clustering*, dapat dihasilkan suatu kumpulan partisi yang berurutan, dimana dalam kumpulan tersebut terdapat:

a. *Cluster – cluster* yang mempunyai poin – poin individu. *Cluster – cluster* ini berada di level yang paling bawah.
b. Sebuah *cluster* yang didalamnya terdapat poin – poin yang dipunyai semua *cluster* didalamnya. *Single cluster* ini berada di *level* yang paling atas.

Hasil keseluruhan dari algoritma *hierarchical clustering* secara grafik dapat digambarkan sebagai *tree*, yang disebut dengan *dendogram*. *Tree* ini secara grafik menggambarkan proses penggabungan dari *cluster – cluster* yang ada, sehingga menghasilkan *cluster* dengan level yang lebih tinggi. Gambar 1 adalah contoh *dendogram*.



Gambar 5.. *Dendogram* (Han, 2001)

2.8.1. Agglomerative Hierarchical Clustering

Metode ini menggunakan strategi disain *Bottom-Up* yang dimulai dengan meletakkan setiap obyek sebagai sebuah *cluster* tersendiri (*atomic cluster*) dan selanjutnya menggabungkan *atomic cluster – atomic cluster* tersebut menjadi *cluster* yang lebih besar dan lebih besar lagi sampai akhirnya semua obyek menyatu dalam sebuah *cluster* atau proses dapat pula berhenti jika telah mencapai batasan kondisi tertentu (Han, 2001). Metode *Agglomerative Hierarchical Clustering* yang digunakan pada penelitian ini adalah metode *AGglomerative NESting* (AGNES). Cara kerja AGNES dapat dilihat pada gambar 2.11. Adapun ukuran jarak yang digunakan untuk menggabungkan dua buah obyek cluster adalah *Minimum Distance ,* yang dapat dilihat pada persamaan 1.

$$d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'| \quad \text{.....(1)}$$

dimana |p – p'| jarak dua buah obyek p dan p'.

2.9 K-Mean

K-Means adalah suatu metode penganalisaan data atau metode Data Mining yang melakukan proses pemodelan tanpa supervisi (unsupervised) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Metode k-means berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain. Dengan kata lain, metode ini berusaha untuk meminimalkan variasi antar data yang ada di dalam suatu cluster dan memaksimalkan variasi dengan data yang ada di cluster lainnya. *Objective function* yang berusaha diminimalkan oleh k-means adalah:

J (U, V) = SUM (k=1 to N) SUM (i=1 to c) (a_ik * (x_k, v_i)^2)

dimana:

- U : Matriks keanggotaan data ke masing-masing cluster yang berisikan nilai 0 dan 1
- V : Matriks centroid/rata-rata masing-masing cluster
- N : Jumlah data
- c : Jumlah cluster
- a_ik : Keanggotaan data ke-k ke cluster ke-i
- x_k : data ke-k
- v_i : Nilai centroid cluster ke-i

Prosedur yang digunakan dalam melakukan optimasi menggunakan k-means adalah sebagai berikut:

- Step 1. Tentukan jumlah cluster
- Step 2. Alokasikan data ke dalam cluster secara random
- Step 3. Hitung centroid/rata-rata dari data yang ada di masing-masing cluster.
- Step 4. Alokasikan masing-masing data ke centroid/rata-rata terdekat
- Step 5. Kembali ke Step 3, apabila masih ada data yang berpindah cluster atau apabila perubahan nilai centroid, ada yang di atas nilai threshold yang ditentukan atau apabila perubahan nilai pada objective function yang digunakan, di atas nilai threshold yang ditentukan. Centroid/rata-rata dari data yang ada di masing-masing cluster yang dihitung pada Step 3. didapatkan menggunakan rumus sebagai berikut:

  v_ij = SUM (k=0 to N_i) (x_kj) / N_i

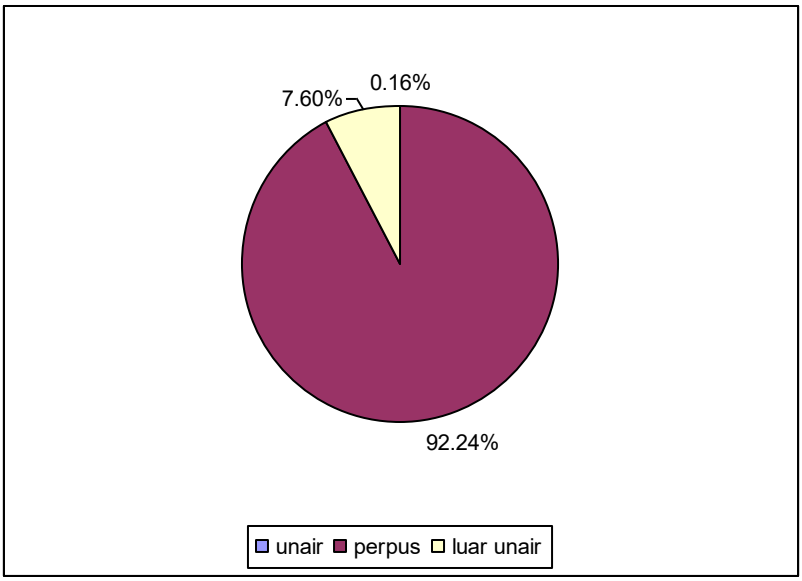
  dimana:

i,k : indeks dari cluster
j : indeks dari variable
v_ij : centroid/rata-rata cluster ke-i untuk variabel ke-j
x_kj : nilai data ke-k yang ada di dalam cluster tersebut untuk variabel ke-j
N_i : Jumlah data yang menjadi anggota cluster ke-i

Sedangkan pengalokasian data ke masing-masing cluster yang dilakukan pada Step 4. dilakukan secara penuh, dimana nilai yang memungkinkan untuk a_ik adalah 0 atau 1. Nilai 1 untuk data yang dialokasikan ke cluster dan nilai 0 untuk data yang dialokasikan ke cluster yang lain. Dalam menentukan apakah suatu data teralokasikan ke suatu cluster atau tidak, dapat dilakukan dengan menghitung jarak data tersebut ke masing-masing centroid/rata-rata masing-masing cluster. Dalam hal ini, a_ik akan bernilai 1 untuk cluster yang centroidnya terdekat dengan data tersebut, dan bernilai 0 untuk yang lainnya.

Clustering Alamat Pengakses

Berdasarkan pengolahan data mining dengan analisis clustering diperoleh hasil bahwa 92% pengguna adln.lib.unair.ac.id melakukan akses melalui perpustakaan layanan adln di Perpustakaan Unair. Fenomena ini terjadi karena jika mengakses melalui local adln atau melalui layanan adln yang berada diperpustakaan dapat memperoleh akses secara *full-text*. Sedangkan jika diakses dari luar Perpustakaan Unair tidak bisa melakukan akses secara *full-text*. Kebijakan ini memberikan dampak bahwa profil pengguna adln yang melakukan akses dari luar perpustakaan sebesar 8% saja dan hampir 0%, tepatnya 0,15% melakukan akses melalui jalur intranet Universitas Airlangga.

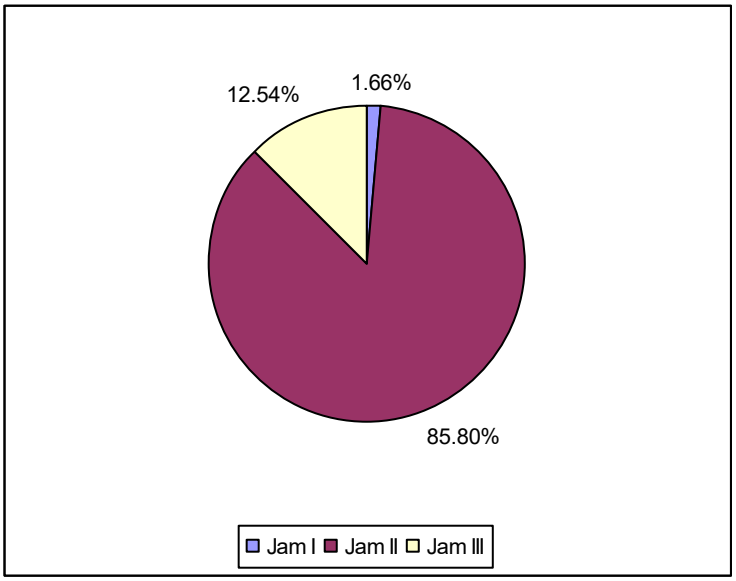


Gambar 4. Profil Lokasi Akses Pengguna ADLN.

Clustering Waktu Akses

Analisis terhadap waktu akses dikelompokkan menjadi tiga yaitu jam 07.00 s/d 10.00 (jam I), 10.00 s/d 15.00 (jam II) dan di atas 15.00 (jam III). Clustering waktu tersebut dipilih berdasarkan fenomena akses internet di Indonesia melalui *google analytic*, bahwa jam akses internet akan mengalami peningkatan jika di atas jam 15.00

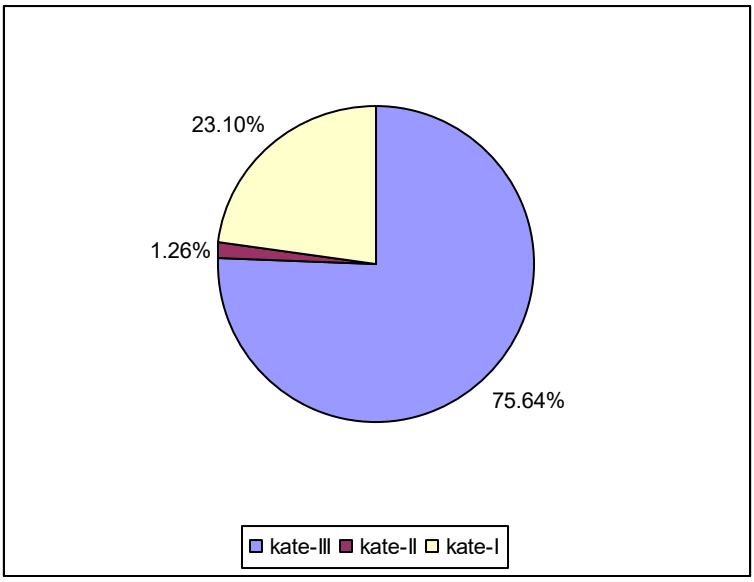
WIB dan jam padat akses inetrnet yaitu jam 10.00 WIB s/d 15.00 WIB. Hasil analisis menunjukan bahwa hampir 85,80% pengguna ADLN melakukan akses pada jam 10.00 WIB s/d 15.00 WIB. Karaketristik ini sesuai dengan jam layanan ADLN dan dipengaruhi oleh aktifitas pengguna ADLN itu sendiri, di mana ADLN hanya bisa diakses secara *full-text* melalui fasilitas ADLN Perpustakaan Unair saja. Jadi Pengguna melakukan akses terhadap ADLN ketika mereka berada di lingkungan kampus. Profil selengkapnya waktu akses dapat dilihat pada Gambar 5.2.



Gambar 5. Profil Waktu Akses Pengguna ADLN

Clustering Lama Akses

Lama akses pengguna ketika melakukan *surfing* dan *browsing* terhadap konten ADLN digolongkan menjadi tiga kategori, yakni <5 menit (Kate-1), 5 s/d 10 menit (Kate-2) dan diatas 10 menit (Kate-3). Berdasarkan hasil analisis diperoleh bahwa hampir 75.64% pengguna ADLN melakukan akses lebih dari 10 menit dan jika dilakukan *cross analysis* dengan clustering lokasi pengguna maka diperoleh hasil hampir 100% pengguna dengan lama akses lebih dari 10 menit melakukan akses ADLN melalui fasilitas layanan ADLN di Perpustakaan Universitas Airlangga. Sedangkan yang melakukan akses ADLN dari luar Universitas Airlangga sebanyak 91.5% terdeteksi hanya bertahan < 5 menit. Perilaku akses ini terjadi karena jika diakses dari luar tidak bisa *full-text* jadi hampir semua pengguna ADLN yang melakukan akses dari luar Perpustakaan atau Unair hanya bertahan <5 menit kemudian memutuskan untuk closing koneksi karena merasa ridak memperoleh informasi sesuai dengan yang diharapkan, yaitu akses secara *full-text* terhadap konten Airlangga Digital Library. Profil lengkap lama akses pengguna ADLN dapat dilihat pada gambar 5.3 dan tabel 5.1.



Gambar 6. Profil Lama Akses Pengguna ADLN

Tabel 10. *Cross Clustering Analysis* antara lama akses dan lokasi akses

		perpus	**unair**	**luar-unair**
kate-I	23.10 %	5.97%	2.53%	91.5%
kate-II	1.26 %	64.43%	10.23%	25.34%
kate-III	75.64 %	99.67%	0.14%	0.19%

Simpulan

Berdasarkan hasil analisis *clustering* pada data *mining* log akses Airlangga Digital Library menunjukkan bahwa 92.24% pengguna melakukan akses melalui layanan ADLN di Perpustakaan dengan lama akses lebih dari 10 menit. Hasil analisis juga memperlihatkan 75.64% pengguna melakukan akses lebih dari 10 menit dari 92.24% yang melakukan akses melalui layanan ADLN. Sedangkan jam berkunjung ke site ADLN, mayoritas pengguna melakukan akses pada jam 10.00 WIB s/d 15.00 WIB.

Kondisi tersebut perlu dicermati jika ingin mengembangkan ADLN sebagai salah satu sumber *richt file* pada penilian peringkat *webometrics*. Secara konten (*richt file*) memang secara nyata telah menyumbangkan point besar, namun melihat dari segi akses ADLN menunjukkan bahwa ADLN kurang diakses dari internet dan lebih banyak diakses secara intranet sehingga penilaian webometrics dari sisi *accessible* menjadi kurang.

Saran

Penggunaan analisis clustering dengan metode k-mean memang dapat memperlihatkan sebuah karakteristik dan pola akses dari situs web. Sebagai pembanding, disarankan untuk melakukan analisis *web mining* dengan menggunakan *google anlytics*.

Daftar Pustaka

Bezdek, J.C. 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York.

Chaturvedi, A.D., Green, P.E. and Carroll, J.D. (2001). *K-Modes Clustering. Journal of Classification*, 18, 35-56.

Fayyad, 1996. Usama. *Advances in Knowledge Discovery and Data Mining*. MIT Press.

Han, Jiawei and Micheline Kamber, 2001. *Data Mining: Concepts and Techniques,* Morgan Kaufmann.

Han,Jiawei. *Data Mining Concept and Techniques*. Presentation.http://www.cs.sfu.ca ~han/dmbook

J. A. Hartigan, 1997. *Clustering Algorithms*. Wiley.

J. A. Hartigan and M. A. Wong, 1979. *A K-Means Clustering Algorit*. Applied Statistics 28 (1): 100–108.

J. B. MacQueen, 1997. Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press.

Lee, Wenke. Salvatore, J. Stolvo. 2000. *A Framework for Constructing Features and Models for Intrusion Detection System*. ACM Transaction Information and System Security Vol.3,No.4, November 2000.

Lee, Wenke. Salvatore, J. Stolvo. Mok, Kui W. 1998 *Mining Audit Data to Build Intrusion Detection Models*. http://www.aaai.org.

Sundaram, Aurobindo. *An Introduction to Intrusion Detection*. whitepaper. 1996

Thearling, Kurt. 1998. *An Introduction To Data Mining*. Whitepaper. http://www3. shore.net/~kht/dmwhite/dmwhite.htm