



DISERTASI [TE143597]

TWO-STAGES CLUSTERING UNTUK SEGMENTASI PENGUNJUNG WEB PADA WEB USAGE MINING

**YUHEFIZAR
2210 301 010**

**DOSEN PEMBIMBING
Dr. Ir. Yoyon Kusnendar Suprpto., M.Sc.
Dr. I Ketut Eddy Purnama., S.T., M.T**

**PROGRAM DOKTOR
JURUSAN TEKNIK ELEKTRO
FAKULTAS TEKNOLOGI INDUSTRI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2016**



DISSERTATION [TE143597]

TWO-STAGES CLUSTERING FOR SEGMENTING WEB VISITOR IN WEB USAGE MINING

YUHEFIZAR
2210 301 010

SUPERVISOR :
Dr. Ir. Yoyon Kusnendar Suprpto, M.Sc.
Dr. I Ketut Eddy Purnama, S.T., M.T

DOKTORAL PROGRAM
DEPARTMENT OF ELECTRICAL ENGINEERING
FACULTY OF INDUSTRIAL TECHNOLOGY
INSTITUT TEKNOLOGI SEPULUH
NOPEMBER SURABAYA
2016

LEMBAR PENGESAHAN DISERTASI

Disertasi disusun untuk memenuhi salah satu syarat memperoleh gelar Doktor (Dr)
di Institut Teknologi Sepuluh Nopember

Judul :

Two-Stages Clustering Untuk Segmentasi Pengunjung Web
Pada Web Usage Mining

Oleh:

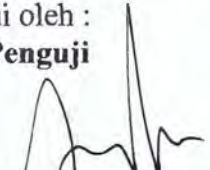
Yuhefizar


NRP : 2210 301 010

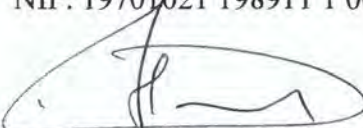
Tanggal Ujian : 27 Januari 2016


Periode Wisuda : Maret 2016

Disetujui oleh :
Dosen Penguji

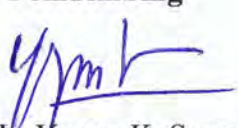

1. Prof. Budi Santosa, M.S., Ph.D
NIP. 19690512 19942 1 001



2. Dr. Anto Satriyo Nugroho, M.Eng
NIP. 19701021 198911 1 001


3. Mochammad Hariadi, S.T., M.Sc, Ph.D
NIP. 19691209 199703 1 002


4. Dr. Surya Sumpeno, S.T., M.Sc
NIP. 19690613 199702 1 003

Dosen Pembimbing


1. Dr. Ir. Yoyon K. Suprpto, M.Sc
NIP. 19540925 197803 1 001


2. Dr. I Ketut Eddy Purnama, S.T.,M.T
NIP. 19690730 199512 1 001

Direktur Program Pascasarjana,




Prof. Ir. Djaubar Manfaat, M.Sc, Ph.D

NIP. 19601202 198701 1 001

***TWO-STAGES CLUSTERING* UNTUK SEGMENTASI PENGUNJUNG WEB PADA WEB USAGE MINING**

Nama mahasiswa : Yuhefizar
NRP : 2210 301 010
Pembimbing : Dr. Yoyon Kusnendar Suprapto., M.Sc
Dr. I Ketut Eddy Purnama., S.T., M.T

ABSTRAK

Web Usage Mining (WUM) berhubungan dengan ekstraksi *knowledge* dari *data web log*, salah satu tujuannya adalah untuk segmentasi pengunjung web. *Data web log* sebagai data utama dari WUM memiliki banyak item data yang tidak relevan untuk dilakukan proses penambangan lebih lanjut, sehingga perlu dilakukan tahapan-tahapan untuk menghapus data tersebut agar hasil akhir segmentasi pengunjung web lebih baik.

Oleh karena itu, dalam penelitian ini dilakukan tahapan pra-pemrosesan lebih detail dan mengajukan pendekatan baru untuk tujuan segmentasi pengunjung web yang disebut dengan pendekatan klasterisasi bertahap (*two-stages clustering*). Klasterisasi tahap pertama dilakukan pada data yang berbentuk *frequently access* (frekuensi kunjungan) menggunakan metoda klaster hirarki dan non hirarki, kemudian dilanjutkan dengan klasterisasi tahap kedua pada data yang berbentuk *user access pattern* (pola kunjungan *user*). Pada klasterisasi tahap kedua digunakan kombinasi metode klaster hirarki dan non hirarki.

Dari penerapan metode ini berhasil mereduksi *data web log* sebesar 98.38% dan memperoleh klaster-klaster/segmentasi pengunjung web beserta profilnya yang dapat dijadikan acuan untuk tujuan personalisasi web, modifikasi web dan kepentingan lainnya dalam lingkup WUM.

Kata Kunci : *two-stages clustering*, klaster hirarki, klaster non hirarki, k-means, web usage mining, data web log, *web mining*.

Halaman ini sengaja dikosongkan

TWO-STAGES CLUSTERING FOR SEGMENTING WEB VISITOR IN WEB USAGE MINING

Name : Yuhefizar
NRP : 2210301010
Supervisor : Dr. Yoyon Kusnendar Suprapto., M.Sc
Dr. I Ketut Eddy Purnama, S.T., M.T

ABSTRACT

Web Usage Mining (WUM) is associated with knowledge extraction of web log data. One of the purposes is for web visitor's segmentation. Web log data as the primary data of WUM has many irrelevant data item for further mining process, therefore some stages need be done to reduce the data in order to make a better result of web visitor segmentation.

For this purpose, this research conducted a more detail pre-processing stage and proposed a new approach for web visitor's segmentation called two-stages clustering. First stage clustering is conducted on data with frequently access form by using hierarchical and non-hierarchical method which is followed by the second stage clustering for the data with user access pattern form. On the second stage clustering the combined method of hierarchical and non-hierarchical was used.

Application of the method was successful in reducing web log data for 98.38% and gained clusters of segments of web visitor and its profile that can be used as reference for web personalization, web modification, and other purposes within the WUM scope.

Keywords : two-stages clustering, hierarchical cluster, non-hierarchical cluster, k-means, web usage mining, web log data, web mining.

Halaman ini sengaja dikosongkan

KATA PENGANTAR

Alhamdulillah, segala puji syukur penulis sampaikan kehadirat Allah, SWT, Tuhan Yang Maha Esa yang telah melimpahkan rahmat, hidayah, inayah serta karunia-Nya yang tak terhingga kepada penulis, sehingga penulisan disertasi ini dapat diselesaikan. Disertasi ini disusun untuk memenuhi salah satu syarat akademik dalam menyelesaikan studi Doktorat di Program Studi Teknik Elektro, Fakultas Teknologi Industri, Institut Teknologi Sepuluh Nopember, Surabaya.

Selesainya penulisan disertasi ini tidak terlepas dari bantuan banyak pihak. Untuk itu, dengan segala kerendahan hati, penulis menyampaikan banyak terima kasih dan penghargaan yang sebesar-besarnya atas bantuan, bimbingan dan dukungan, baik moril dan materil kepada :

1. Pemerintah Republik Indonesia melalui Kemristekdikti yang telah memberikan bantuan beasiswa BPPS.
2. Bapak Aidil Zamri, S.T.,M.T, Direktur Politeknik Negeri Padang yang telah memberikan ijin tugas belajar, beserta segenap unsur pimpinan Politeknik Negeri Padang yang terus mensupport penulis.
3. DP2M DIKTI, yang telah memberikan dana hibah Doktor.
4. Bapak Dr. Ir. Yoyon K. Suprpto., M.Sc dan Bapak Dr. I Ketut Eddy Purnama., S.T.,M.T, selaku dosen pembimbing (promotor) yang telah banyak meluangkan waktunya untuk memberikan bimbingan, bantuan, dan masukan serta motivasi kepada penulis untuk mampu menyelesaikan studi ini. Begitu banyak kendala yang penulis hadapi, namun selalu ada solusi yang diberikan oleh dosen pembimbing.
5. Bapak Prof. Budi Santosa., M.S.,Ph.D, Bapak Dr. Anto Satriyo Nugroho., M.Eng, Bapak Mochammad Hariadi., S.T.,M.Sc.,Ph.D dan Bapak Dr. Surya Sumpeno., S.T.,M.Sc sebagai tim penguji yang telah banyak memberikan saran perbaikan terhadap penulisan disertasi ini.

6. Bapak Prof. Dr. Ir. Mauridhi Hery Purnomo., M.Eng, terima kasih banyak pak atas bantuan dan dukungan Bapak dalam usaha penyelesaian studi penulis, semoga bapak selalu dilimpahi nikmat kesehatan dan keberkahan dalam hidup ini, Aamiin.
7. Bapak Ir. Djoko Purwanto., M.Eng.,Ph.D selaku Koordinator Program Pasca Sarjana Jurusan Teknik Elektro ITS.
8. Segenap pengelola Program Pascasarjana (PPs) ITS, dosen dan karyawan PPs Jurusan Teknik Elektro ITS yang telah memberikan dukungan dan bantuan selama menempuh program pendidikan ini.
9. Keluarga yang terus mendukung tanpa henti, ter-khusus kepada Papa(alm) dan Mama (almh), terima kasih atas didikan dan kasih sayangnya, semoga Allah, SWT menempatkan di Surga-Nya, Aamiin. Terima kasih juga untuk Papa dan Mama mertua yang terus mensupport dan terima kasih tak terhingga untuk yang selalu mendampingi dikala suka dan duka, istri tercinta, Arrita Elfiyanti, dan anak-anak tersayang Dzaky, Daffa dan Ariq yang terus memberikan keceriaan dan semangat untuk segera menyelesaikan studi.
10. Teman-teman seperjuangan mahasiswa S3, Bu Atik, Pak Max Mulyono, Pak Ruri, Pak Rusmono, Pak Arif dan rekan-rekan lainnya yang tidak bisa penulis sebutkan satu persatu, terima kasih atas bantuan, dukungan, kebersamaan dan motivasi serta doanya. Spesial untuk buk Atik yang sangat banyak membantu penulis dalam proses penyelesaian studi ini, semoga Buk Atik diberi kemudahan dan segera mampu menyelesaikan studi ini.
11. Dihari-hari penentuan, terima kasih banyak ke pada Pak Nyoman Sukajaya, Pak Dewa Made Wiharta dan Pak Ida Bagus Gede Manuaba atas bantuan, saling memotivasi dan *sharingnya* yang luar bisa.
12. Dan pihak-pihak lain yang tidak dapat penulis sebutkan satu persatu. Terimalah ini sebagai bentuk tanggung jawab dan terima kasih penulis atas segala bantuannya, baik moril maupun materil.

Semoga Allah, SWT, Tuhan Yang Maha Esa membalas amal kebaikan Bapak dan Ibu semuanya serta selalu melimpahkan karunia-Nya kepada kita semua. Akhirnya, semoga studi ini membawa banyak manfaat untuk orang banyak, Aamiin YRA.

Penulis menyadari bahwa dalam penulisan disertasi ini masih jauh dari sempurna, oleh karena itu, kritikan dan saran yang membangun dari semua pihak demi perbaikan dan penyempurnaan buku disertasi ini sangat penulis harapkan.

Surabaya, Februari 2016
Penulis,

Yuhefizar

Halaman ini sengaja dikosongkan

DAFTAR ISI

HALAMAN JUDUL	i
LEMBAR PENGESAHAN	iii
PERNYATAAN KEASLIAN DISERTASI	v
ABSTRAK	vii
KATA PENGANTAR	xi
DAFTAR ISI	xv
DAFTAR GAMBAR	xix
DAFTAR TABEL	xxi
DAFTAR NOTASI	xxiii

BAB 1 : PENDAHULUAN

1.1 Latar Belakang Penelitian	1
1.2 Perumusan Masalah	6
1.3 Tujuan dan Manfaat Penelitian	6
1.3.1 Tujuan Penelitian	6
1.3.2 Manfaat Penelitian	6
1.4 Batasan Masalah	7
1.5 Susunan Penelitian	7
1.6 Peta Penelitian	8

BAB 2 : KAJIAN PUSTAKA

2.1 Penambangan Data	9
2.2 Perlakuan Terhadap Data	10
2.3 Penambangan Web	11
2.4 Web Usage Mining	13
2.4.1 Data Web Log	14
2.4.2 Lokasi Data Web Log	17

2.5 Analisis Faktor	17
2.5.1 Analisis Komponen Utama	19
2.5.2 Model Analisis Faktor	20
2.5.3 Jenis-Jenis Analisis Faktor	22
2.5.4 Tahapan Analisis Faktor	23
2.6 Analisis Klaster	26
2.7 Segmentasi Pengunjung Web	31

BAB 3 : PRA-PEMROSESAN dan TAHAPAN KLASTERISASI DATA

WEB LOG

3.1 Pra-Pemrosesan Data Web Log	35
3.1.1 Pembersihan Data	35
3.1.2 Identifikasi Halaman Web	38
3.1.3 User Identification	28
3.1.4 Klasterisasi	30
3.2 Tahapan Klasterisasi	41
3.2.1 Klasterisasi Tahap Pertama	42
3.2.2 Klasterisasi Tahap Kedua	42

BAB 4 : SINGLE-STAGE CLUSTERING PADA DATA WEB LOG YANG BERBENTUK *FREQUENTLY ACCESS* UNTUK SEGMENTANSI PENGUNJUNG WEB

4.1 Pendahuluan	45
4.2 Tahapan Penelitian	47
4.3 Dataset	48
4.4 Pra-Pemrosesan	48
4.5 Metoda Klaster	48
4.6 Validasi Klaster	49
4.7 Analisis dan Hasil	49
4.8 Segmentasi Pengunjung Web	53
4.9 Kesimpulan	54

**BAB 5 : SINGLE-STAGE CLUSTERING PADA DATA WEB LOG YANG
BERBENTUK *USER ACCESS PATTERN* UNTUK
SEGMENTANSI PENGUNJUNG WEB**

5.1	Pendahuluan	58
5.2	Tinjauan Pustaka	58
5.3	Tahapan Penelitian	60
5.4	Dataset	61
5.5	Pra-Pemrosesan	61
5.6	Penerapan Algoritma	63
5.7	Analisis dan Hasil	64
5.8	Profil Klaster untuk Segmentasi Pengunjung Web	65
5.9	Kesimpulan	67

**BAB 6 : TWO-STAGES CLUSTERING UNTUK SEGMENTASI
PENGUNJUNG WEB**

6.1	Pendahuluan	69
6.2	Format Data Web Log dan Dataset	71
6.3	Pra-Pemrosesan	72
6.4	Analisis Faktor	73
6.4.1	Menyiapkan Data	73
6.4.2	Menentukan Metoda Analisis Faktor	74
6.4.3	Membuat Skor Faktor	80
6.5	Klasterisasi	80
6.5.1	Klasterisasi Tahap Pertama	80
6.5.2	Klasterisasi Tahap Kedua	82
6.6	Segmentasi Pengunjung Web	84
6.7	Kesimpulan	86

BAB 7 : KESIMPULAN DAN SARAN

7.1	Kesimpulan	87
7.2	Saran	88

Daftar Pustaka	89
Lampiran Identifikasi Halaman Web	97
Curriculum Vitae	99

DAFTAR GAMBAR

1. Gambar 1.1 : Jumlah Pengguna Internet Dunia	3
2. Gambar 1.2 : Peta Penelitian	8
3. Gambar 2.1 : Tahapan Penambangan Data	10
4. Gambar 2.2 : Klasifikasi Penambangan Web	12
5. Gambar 2.3 : Tahapan Web Usage Mining	14
6. Gambar 2.4 : Diagram Analisa Kluster	27
7. Gambar 2.5 : Contoh Dendogram	29
8. Gambar 3.1 : Tahapan Pra-Pemrosesan	37
9. Gambar 3.2 : Grafik Perbandingan Jumlah Data	38
10. Gambar 3.3 : Contoh User Navigation	41
11. Gambar 3.6 : Statistik Kunjungan	31
12. Gambar 4.1 : Tahapan Penelitian 1	47
13. Gambar 5.1 : Tahapan Penelitian 2	60
14. Gambar 5.2 : Struktur Database	63
15. Gambar 5.2 : Tabel dari Dataset	50
16. Gambar 5.3 : Diagram Kluster	64
17. Gambar 6.1 : Tahapan Penelitian 3	72
18. Gambar 6.2 : Scree Plot Hasil Faktorisasi	76
19. Gambar 6.3 : Komponen Plot	79

DAFTAR TABEL

1. Tabel 1.1 : Jumlah Pengguna Internet Dunia	3
2. Tabel 2.1 : Common Log Format	15
3. Tabel 2.2 : Extended Log Format	16
4. Tabel 2.3 : Contoh Data Web Log	16
5. Tabel 3.1 : Sebagian Daftar Kode Respon HTTP	36
6. Tabel 3.2 : Contoh Perbandingan Jumlah Data	38
7. Tabel 3.3 : Contoh Output Identifikasi Pengunjung Web	39
8. Tabel 3.4 : Contoh Hasil Identifikasi Pengunjung Web 1	39
9. Tabel 3.5 : Contoh Hasil Identifikasi Pengunjung Web 2	40
10. Tabel 3.6 : Contoh Hasil Identifikasi Pengunjung Web 3	40
11. Tabel 3.7 : Pola Frequently Access	42
12. Tabel 3.8 : Pola User Access Pattern	43
13. Tabel 4.1 : Perbandingan Jumlah Data	49
14. Tabel 4.2 : Matriks Vektor	50
15. Tabel 4.3 : Penamaan Variabel User dan Halaman Web	50
16. Tabel 4.4 : Persentase Keterwakilan Variable dalam Faktor	51
17. Tabel 4.5 : Validitas Dengan Silhouette	52
18. Tabel 4.6 : Jumlah Keanggotaan Kluster	52
19. Tabel 4.7 : Pusat Kluster Akhir	53
20. Tabel 4.8 : Halaman Web yang diakses	54
21. Tabel 5.1 : Contoh Pola Perilaku Pengunjung Web	61
22. Tabel 5.2 : Tabel dari Dataset	62
23. Tabel 5.3 : Penamaan Variabel	62
24. Tabel 5.4 : Perbandingan Jumlah Data	64
25. Tabel 5.5 : Populasi Kluster	65
26. Tabel 5.6 : Contoh Profil User	66
27. Tabel 5.7 : Tingkat Probability Akses	56
28. Tabel 6.1 : Hasil Pengujian KMO dan Bartlett's	74

29. Tabel 6.2 : Variabel dengan MSA < 0,5	74
30. Tabel 6.3 : Communalities.....	75
31. Tabel 6.4 : Nilai Eigenvalues	75
32. Tabel 6.5 : Nilai Faktor Loading Pada Komponen Matriks	77
33. Tabel 6.6 : Rotasi Komponen Matriks	78
34. Tabel 6.7 : Faktor Dengan Variabel	79
35. Tabel 6.8 : Agglomeration Schedule Tahap 1	80
36. Tabel 6.9 : Klaster Tahap 1	81
37. Tabel 6.10 : Anggota Klaster yang Tidak Terpilih	81
38. Tabel 6.11 : Informasi Data Web Log 2	82
39. Tabel 6.12 : Agglomeration Schedule Tahap 2	82
40. Tabel 6.13 : Hasil Uji SSE	83
41. Tabel 6.14 : Klaster Tahap 2 dan Jumlah Anggotanya	83
42. Tabel 6.15 : Pusat Klaster Akhir	83

DAFTAR NOTASI

p	: variabel untuk halaman web
u	: variabel untuk pengunjung web (<i>user</i>)
v	: variabel untuk faktor
n	: jumlah data
r	: matriks korelasi
μ_1	: rata-rata variabel i
ε_i	: faktor spesifik ke -1
F_j	: <i>common</i> faktor ke $-j$
l_{ij}	: <i>loading</i> dari variabel ke $-i$ pada faktor ke $-j$
m	: banyak faktor
t	: transaksi
F_i	: faktor ke- i
W_i	: bobot atau koefisien nilai faktor ke- i
k	: banyaknya variabel
a_{ji}	: koefisien variabel asal ke- i untuk komponen utama ke- j
λ_j	: <i>eigen value</i> untuk komponen utama ke- j
r_{ij}	: Koefisien korelasi antara variabel i dan j
a_{ij}	: Koefisien korelasi parsial antara variabel i dan j
$ R $: nilai determinan
d_{ij}	: jjarak antara objek ke- i dan objek ke- j
X_{ik}	: data dari objek ke- i pada variabel ke- k
X_{jk}	: data dari objek ke- j pada variabel ke- k
z_0	: keadaan <i>state</i> 0

BAB 1

PENDAHULUAN

Data *mining* merupakan suatu proses dalam menganalisis data dari berbagai perspektif yang berbeda dan merangkumnya menjadi informasi yang berguna. Oleh karena itu, data yang digunakan dalam proses ini sangat menentukan kualitas informasi yang dihasilkannya. Saat ini penelitian di bidang data *mining* terus berkembang, salah satunya pada kajian penambangan web (*web mining*). Topik penelitian di bidang *web mining* terdiri atas 3 (tiga) kategori, yaitu *web structure mining*, *web content mining* dan *web usage mining*. Penelitian dalam disertasi ini fokus dalam topik *web usage mining* dengan menggunakan data *web log* sebagai data utamanya untuk keperluan segmentasi pengunjung *web*.

1.1 Latar Belakang Penelitian

Data merupakan komponen utama dari lahirnya suatu informasi. Informasi yang berkualitas berasal dari data yang valid dan diproses menurut aturan tertentu yang terukur. Data adalah kumpulan angka, fakta, fenomena atau keadaan yang merupakan hasil dari pengamatan, pengukuran atau pencacahan terhadap karakteristik atau sifat dari objek yang dapat berfungsi untuk membedakan objek yang satu dengan yang lainnya pada sifat yang sama.

Data yang handal (*reliable*) dan dapat berguna sebagai input dalam pengambilan keputusan, memiliki syarat utama sebagai berikut:

1. Objektif, data harus sesuai dengan keadaan yang sebenarnya,
2. Representatif, data harus dapat mewakili dari objek yang diteliti, dan
3. Mempunyai tingkat kesalahan baku yang kecil, artinya kesalahan hipotesis dari data sampel harus seminimal mungkin.
4. Relevan, artinya data yang digunakan berhubungan dengan persoalan yang dibahas.

Alvin Toffler[1] dalam bukunya yang berjudul *The Third Wave* mengatakan bahwa sejak akhir tahun 50-an, masyarakat dunia dalam proses transisi menuju

era informasi (*information age*). Pada masa sebelum era informasi, Alvin menyebutnya revolusi agraria dan revolusi industri, kegiatan mencari dan memperoleh data merupakan pekerjaan yang tidak mudah/sulit di masa itu.

Pada era informasi sekarang ini, tidak dapat dipungkiri bahwa peran teknologi internet telah memicu terjadinya ledakan data, yaitu data tersedia sangat banyak dan dapat diperoleh dengan mudah. Di awal tahun 90-an, kehadiran teknologi komputer diprediksi mampu memicu terjadinya ledakan data, maka kehadiran internet, tidak saja menimbulkan ledakan data tapi lebih besar lagi dari itu yang disebut dengan *tsunami* data serta didukung dengan perkembangan ilmu pengetahuan dalam mengelola dan mengakses data itu sendiri.

Pemicu lain terjadi ledakan data di era informasi ini adalah lahirnya layanan di internet yang disebut dengan media sosial yang berbasis *website* diikuti dengan berkembangnya teknologi komunikasi *mobile*, sehingga komunikasi dan pertukaran data, tidak hanya melalui media komputer saja, namun dapat dilakukan melalui berbagai media *mobile* dan dapat dilakukan dimana saja serta kapan saja. Hal ini juga menjadi pemicu munculnya pengguna internet baru yang berimbas kepada bertambahnya sumber *input* data di internet.

Berdasarkan informasi dari www.internetworldstats.com[2] yang diakses pada tanggal 2 Juni 2015, jumlah pengguna internet dunia terus meningkat secara signifikan yaitu sebesar 753%, (perhatikan Tabel 1.1). Hal ini menunjukkan tingginya peningkatan pengguna internet, maka akan berkorelasi dengan semakin tinggi pula pertukaran data di internet yang akan menimbulkan data-data baru.

Dari Tabel 1.1 diperoleh informasi bahwa telah terjadi peningkatan pengguna internet yang signifikan berdasarkan data tahun 2000 dibandingkan dengan estimasi tahun 2015, dan ini diperkirakan akan terus meningkat signifikan.

Sebuah lembaga riset yang fokus dalam bidang pemasaran media sosial, *we are social*, pada bulan November 2015 merilis hasil penelitian tentang tingkat pengguna internet dunia, seperti terlihat pada Gambar 1.1.

Terlihat dari hasil riset dua lembaga tersebut membuktikan bahwa tingkat pertumbuhan pengguna internet meningkat tajam setiap tahunnya. Kehadiran teknologi *mobile* dan aplikasi sosial media telah menjadi bagian penting terhadap peningkatan jumlah pengguna internet tersebut.

Tabel 1.1 Jumlah Pengguna Internet Dunia versi www.internetworldstats.com

Lokasi	Estimasi 2015	Desember 2000	Pertumbuhan 2000-2015
Afrika	1,158,353,014	4,514,400	6,958.2 %
Asia	4,032,654,624	114,304,000	1,129,3 %
Eropa	827,566,464	105,096,093	454.2 %
Timur Tengah	236,137,235	3,284,800	3,358.6 %
Amerika Utara	357,172,209	108,096,800	187.1 %
Amerika Latin	615,588,127	18,068,919	1,684,4 %
Australia	37,157,120	7,620,480	251.6 %
TOTAL	7,264,623,793	360,985,492	753.0 %



Gambar 1.1 Jumlah Pengguna Internet Dunia versi www.wearesocial.sg

Di sisi lain, tingginya pertumbuhan pengguna internet didominasi oleh peran dari layanan yang berbasis *website*, sebagai media dalam menyebarkan berbagai macam informasi dan komunikasi. Hal ini menandakan bahwa peran *website* sangat penting.

Website sebagai pemasok data utama dan terbesar di internet, mempunyai banyak karakteristik, di antaranya :

- a. Jumlah data/informasi yang terdapat pada *website* sangat besar dan terus berkembang dengan cakupan informasi yang sangat luas dan beragam.
- b. Semua jenis data tersedia pada *website*, seperti tabel yang terstruktur, teks yang tidak terstruktur, file multimedia (gambar, audio, video) ataupun halaman *web* yang semi terstruktur.
- c. Informasi yang ditampilkan bersifat heterogen.
- d. Bersifat dinamis.
- e. Aktifitas pengguna selama mengakses *website* (*clickstream*) direkam oleh *server* sebagai data *web log*.

Semua karakteristik tersebut merupakan peluang dan tantangan untuk melakukan data *mining*/penambangan data untuk menemukan informasi dan pengetahuan dari *web* di antaranya dalam melakukan segmentasi pengunjung *web* itu sendiri.

Keberadaan *website* e-bisnis maupun *website* portal dan sosial media telah menjadi bagian penting saat ini. Hal ini tidak terlepas dari kemudahan pemenuhan kebutuhan hidup masyarakat global yang ditawarkannya. Mulai dari layanan *e-banking*, *e-learning*, *e-shopping*, *e-commerce* dan lainnya, sehingga tingkat kunjungan/akses ke sebuah *website* sangat tinggi. Otomatis hal ini meninggalkan data *web log* pengunjung *website* (*web user*) yang sangat besar pada sisi *web server*. Data *web log* merupakan rekam jejak (*track record*) atas aktifitas pengunjung *web* dalam berinteraksi pada sebuah *website*.

Ukuran data *web log* tersebut dapat mencapai hingga hitungan *Giga Byte* (GB) perbulannya tergantung kepada tingkat kunjungan terhadap *website* tersebut. Informasi yang terkandung dalam data *web log* ini dapat digunakan untuk berbagai tujuan, di antaranya untuk melakukan segmentasi pengunjung *web*, menganalisa perilaku pengunjung *web* sehingga dapat digunakan untuk memprediksi perilaku pengunjung atau untuk keperluan *website* yang adaptif, dan untuk keperluan tertentu lainnya.

Point penting lainnya dari kemajuan teknologi internet ini, khususnya layanan *website* adalah data yang melimpah tersebut dapat dengan mudah

diperoleh, sehingga hal ini menjadi bahan baku para peneliti, khususnya di bidang *web usage mining*, untuk mengkaji lebih dalam, pengetahuan apa yang dapat dihasilkan dari data *web log* tersebut.

Namun, data *web log* yang melimpah tersebut memiliki banyak data yang tidak diperlukan (data “sampah”) dalam proses penambangan data di bidang *web usage mining* ini. Yang dimaksud dengan data “sampah” dari data *web log* ini adalah data yang tidak merujuk langsung ke halaman *web* yang diakses oleh pengunjung *web*, sehingga data tersebut tidak menggambarkan pola perilaku pengunjung *web*. Data tersebut dapat berupa data gambar, suara, video, file .css, .txt, kode respon dan metode dari HTTP dan lainnya.

Banyaknya data “sampah” tersebut menyebabkan kesulitan dalam menemukan pola perilaku sesungguhnya dari pengunjung *web*. Oleh karena itu, proses membersihkan data *web log* dari data-data “sampah” menjadi bagian pembahasan penting dalam penelitian ini. Sehingga dihasilkan data *web log* yang baik digunakan untuk proses penambangan data lebih lanjut, dalam hal ini digunakan untuk tujuan segmentasi pengunjung *web*.

Segmentasi pengunjung *web* merupakan proses mengelompokkan pola-pola pengunjung *web* berdasarkan kriteria tertentu sehingga terbentuk kelompok-kelompok pengunjung *web* yang mempunyai ciri khas tersendiri, sehingga pengunjung *web* yang berada dalam kelompok yang sama akan memiliki pola perilaku yang sama (homogenitas), namun akan mempunyai pola perilaku yang berbeda dengan anggota di kelompok lainnya (heterogenitas).

Oleh karena ini, disertasi ini mencoba meneliti dan mengangkat topik di bidang *web usage mining*, khusus dari sudut pandang bagai mana mereduksi data *web log* (membersihkan data *web log* dari data “sampah”) sebagai tahapan pra pemrosesan dalam *web usage mining* dan kemudian diimplementasikan untuk melakukan proses segmentansi pengunjung *web*. Dalam *web usage mining*, pra-pemrosesan merupakan tahap yang penting, kompleks dan hampir 80% waktu penambangan data berada pada tahap ini, oleh karena itu pembahasan pada tahap tersebut mendapat bagian penting dalam disertasi ini, kaitannya dalam upaya memperoleh data yang bersih untuk dilakukan proses segmentansi pengunjung *web*.

1.2 Perumusan Masalah

Secara garis besar permasalahan yang dibahas pada penelitian ini adalah :

1. Data *web log* memiliki data “sampah” yang sangat banyak.
2. Data *web log* mengandung informasi pola perilaku pengunjung *web* yang sulit diketahui.
3. Sulit untuk melakukan segmentasi pengunjung *web* karena banyaknya data “sampah”.

1.3 Tujuan dan Manfaat Penelitian

1.3.1 Tujuan Penelitian

Tujuan dari kegiatan penelitian ini adalah untuk :

1. Membersihkan data *web log* dari data “sampah” sehingga layak untuk diproses *mining* lebih lanjut.
2. Melakukan segmentasi untuk mendapatkan kelompok-kelompok pengunjung *web* berdasarkan *frequently access* dan *user access pattern* menggunakan metode *two-stages clustering*.

1.3.2 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah untuk memberikan *input* dalam proses atau kegiatan penelitian yang berhubungan dengan segmentasi pengunjung *web* terutama untuk bidang :

1. Personalisasi *web*, adalah sebuah respon yang diberikan oleh *web* kepada pengunjung dalam bentuk konten yang sesuai dengan kebutuhan pengunjung tersebut.
2. Modifikasi *web*, adalah upaya untuk memperbaiki tampilan dan konten *website* sehingga sesuai dengan kebutuhan pengunjung.
3. Kajian perilaku pengunjung *web* berguna untuk berbagai kebutuhan, misalnya untuk pemasaran produk, iklan, segmentasi pasar, bisnis dan sebagainya.

1.4 Batasan Masalah

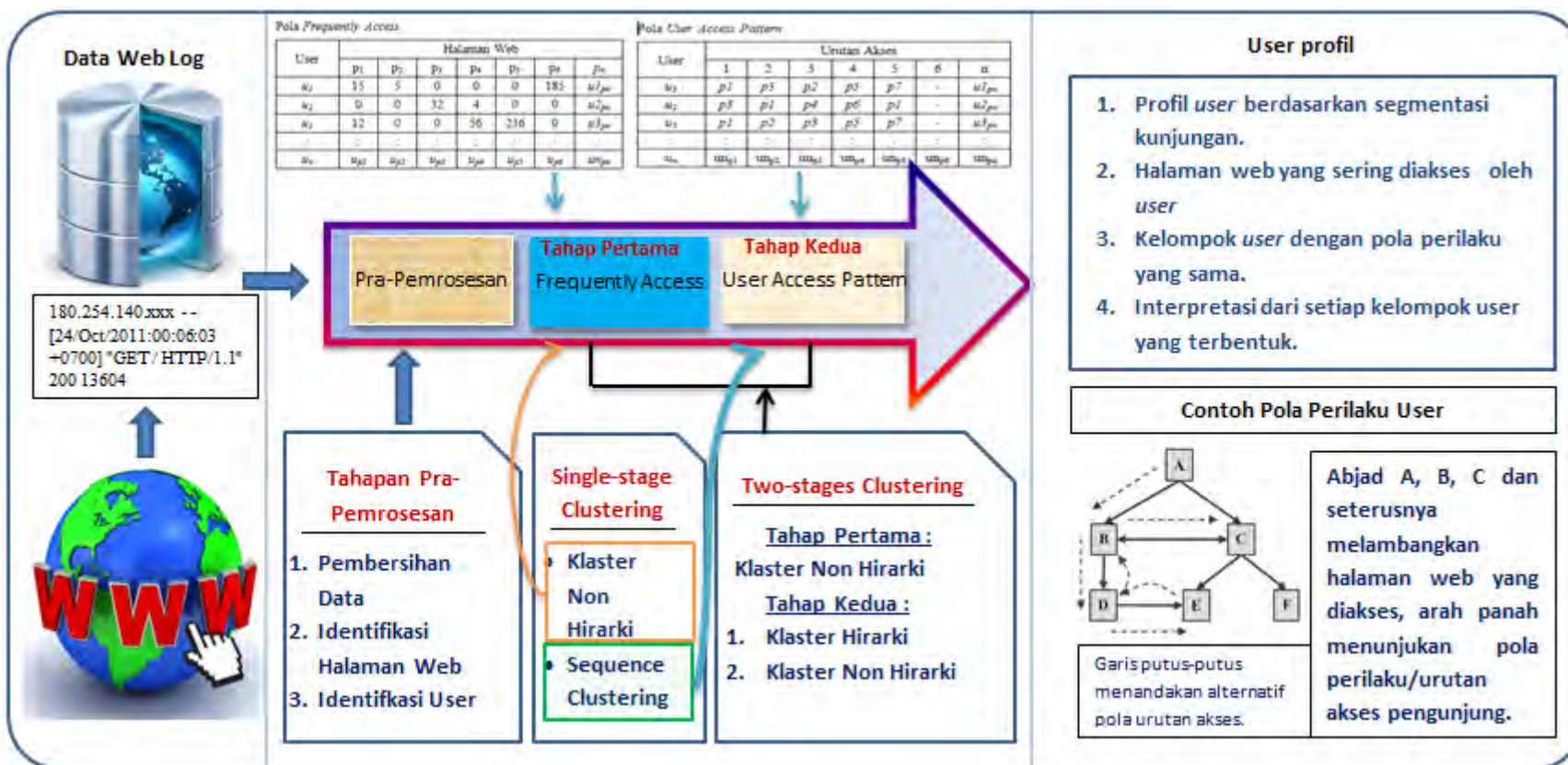
Agar penelitian lebih sistematis dan terarah, ditentukan batasan masalah sebagai berikut

1. Topik penelitian di bidang *web usage mining* dengan data *web log* sebagai sumber data utamanya.
2. Data yang digunakan dalam penelitian ini adalah data *web log public* dan *non public*.

1.5 Susunan Penelitian

Penulisan buku disertasi ini terdiri atas tujuh bab. Bab 1 membahas tentang pendahuluan yang menjelaskan tentang latar belakang, perumusan masalah, tujuan dan manfaat penelitian, batasan masalah, susunan dan peta jalan penelitian. Kajian teori yang terkait dengan topik penelitian dibahas pada Bab 2. Bab 3 membahas tentang pra-pemrosesan (*pre processing*) data secara lebih khusus di bidang *penambangan web* untuk kemudian diterapkan pada proses segmentasi pengunjung *web*. Pada bab 4 di bahas uji coba segmentasi pengunjung *web* menggunakan satu tahap klasterisasi pada data *web log* yang berbentuk *frequently access*, terlebih dahulu diawali dengan pra-pemrosesan untuk reduksi data. Kemudian dilanjutkan dengan pembahasan klasterisasi untuk segmentasi dengan metode *sequence* pada bab 5 terhadap data yang berbentuk *user access pattern*. Pada bab 6 dilakukan penerapan dua tahap klasterisasi (*two-stages clustering*) untuk mendapatkan data *web log* yang baik dan dilakukan proses segmentasi pengunjung *web* serta pembuatan profilisasi *user* dari klaster yang terbentuk. Bab 7 berisi kesimpulan dan saran.

1.6 Peta Penelitian



Gambar 1.2 Peta Penelitian

BAB 2

KAJIAN PUSTAKA

Bab ini menjelaskan kajian-kajian yang terkait dengan topik penelitian di bidang *web usage mining*. Diawali sekilas tentang penambangan data dan *penambangan web*, metoda pra-pemrosesan dan klasterisasi atau segmentasi data *web log*.

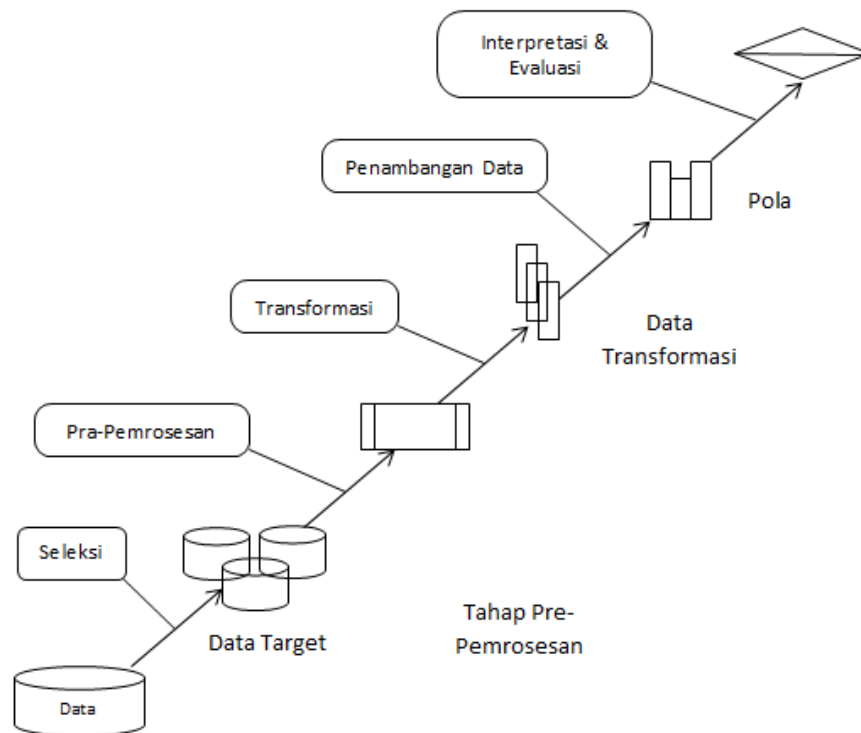
2.1 Penambangan Data

Data *mining* (penambangan data) adalah proses penemuan pola yang berguna atau pengetahuan dari sumber data, seperti basis data, teks, gambar, *web* dan lainnya. Penambangan data terdiri atas multi disiplin ilmu, seperti statistik, kecerdasan buatan, pencarian kembali informasi, visualisasi dan sebagainya.

Secara umum, terdapat tiga langkah utama dalam proses penambangan data [3], yaitu :

- 1 **Pra-pemrosesan.** Data mentah biasanya tidak dapat langsung dilakukan proses penambangan karena mengandung beberapa item data yang tidak berguna, sehingga perlu ditata terlebih dahulu. Tahapan inilah yang disebut dengan pra-pemrosesan.
- 2 **Penambangan data.** Pada tahap ini, data telah siap untuk diolah dengan menggunakan algoritma penambangan data untuk menghasilkan pola atau pengetahuan baru.
- 3 **Pasca-pemrosesan.** Dalam banyak kasus, tidak semua pola/pengetahuan baru yang ditemukan itu berguna sehingga diperlukan proses evaluasi dan teknik visualisasi untuk pendukung dalam mengambil keputusan.

Dari tiga tahapan umum di atas, tahapan penambangan data dapat lebih detail seperti terlihat pada Gambar 2.1.



Gambar 2.1 Tahapan Penambangan Data.

Diawali dari seleksi data, kemudian dilakukan pra-pemrosesan, transformasi data, lalu dilakukan penambangan data untuk mendapat pola atau pengetahuan baru dan terakhir dilakukan interpretasi dan evaluasi.

2.2 Perlakuan Terhadap Data

Penambangan data pada intinya adalah melakukan analisis statistik terhadap data yang berjumlah besar. Proses standar yang dilakukan terhadap data yang besar tersebut adalah mengelompokkan dan atau membaginya menjadi data latihan (*training data*) sebagai data untuk latihan berbagai model analisis dan data uji (*test data*) untuk menguji model yang dikembangkan. Dengan membagi data dan menggunakan sebagian dari data tersebut untuk pengembangan model, serta mengujinya melalui data uji yang terpisah diharapkan dapat memperoleh akurasi data yang lebih baik.

Berikut ini beberapa teknik dalam penambangan data yang dapat digunakan:

1. **Asosiasi**, adalah teknik penambangan data untuk menemukan aturan assosiatif atau hubungan antara suatu transaksi item data dengan hal lain dalam transaksi yang sama untuk mendapatkan sebuah pola/*pattern*.
2. **Klasifikasi**, yaitu metoda yang ditujukan untuk pembelajaran fungsi-fungsi yang berbeda dan memetakan masing-masing data terpilih ke dalam salah satu dari kelompok kelas yang telah ditetapkan sebelumnya. Tujuan dari metoda ini dapat memprediksi secara otomatis kelas dari data lain yang belum diklasifikasikan.
3. **Klasterisasi**, yaitu metoda untuk mengelompokkan data berdasarkan tingkat kemiripan antar item datanya. Proses klasterisasi ini tidak membutuhkan kelompok pembelajaran (*unsupervised*).
4. **Prediksi**, metode ini berusaha untuk menemukan hubungan antara variabel bebas dan terikat dari item data, hubungan antara variabel bebas dengan variabel lainnya. Penerapan metode-metode ini berhubungan dengan teknik regresi.

2.3 Penambangan Web

Penambangan web merupakan topik khusus dari ranah penambangan data (data *mining*) untuk menemukan pengetahuan atau ekstraksi pola-pola penting dan bermanfaat yang tersimpan secara implisit pada kumpulan data yang relatif besar pada layanan *world wide web*.

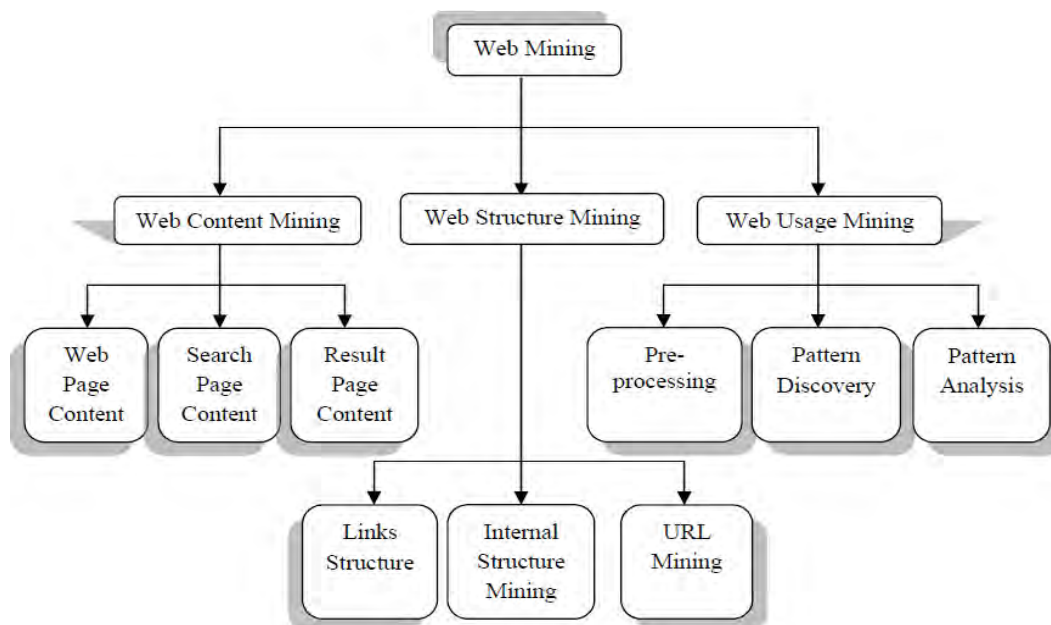
Penambangan web terdiri atas tiga kategori [4], perhatikan Gambar 2.2, yaitu :

1. *Web Content Mining*
2. *Web Structure Mining*
3. *Web Usage Mining*

Web Content Mining adalah proses untuk menemukan pengetahuan dari konten sebuah *website*, yaitu dari teks, gambar, data audio, data video maupun data lainnya seperti *metadata* dan *hyperlink*. Konten *web* ini tersedia dalam bentuk data tidak terstruktur, semi terstruktur seperti file HTML atau terstruktur seperti tabel dan basis data yang dihasilkan oleh HTML.

Pada prinsipnya teknik ini mengekstraksi kata kunci yang terkandung pada dokumen. Konten *web* antara lain dapat berupa teks, gambar, suara, video, metadata, dan *hyperlink*. Ada dua strategi yang umum digunakan: pertama langsung melakukan penambangan terhadap data, dan kedua melakukan pencarian serta meningkatkan hasil pencarian seperti layaknya mesin pencari (*search engine*).

Dengan *web content mining* dapat melakukan klasifikasi halaman *web* sesuai dengan topik tertentu, menemukan pola-pola pada halaman *web* untuk mengekstrak data yang berguna, seperti deskripsi sebuah produk, konten pada forum untuk sentimen konsumen dan lainnya.



Gambar 2.2 Klasifikasi Penambangan web

Web Structure Mining adalah kegiatan untuk menemukan pengetahuan dari *hyperlink/link* yang mewakili struktur dari sebuah *website*. Salah satu manfaatnya adalah untuk menentukan *pagerank* pada suatu halaman *web*. Sebagai contoh, dari sebuah *link* dapat ditemukan halaman *web* yang penting berdasarkan peran dari mesin pencari, dapat juga menemukan komunitas pengguna yang mempunyai

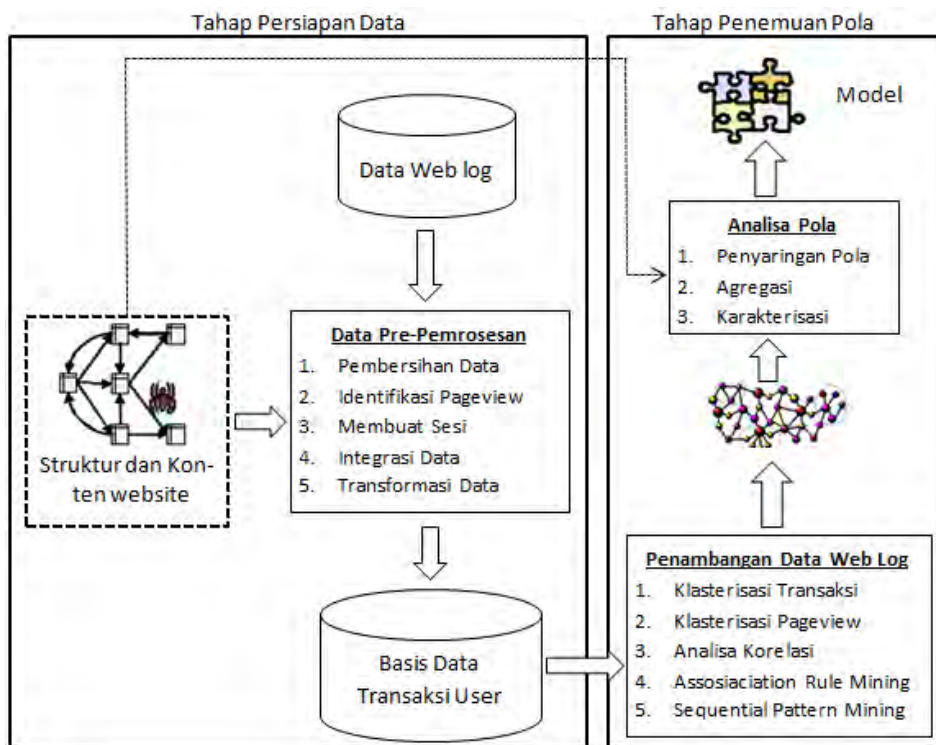
minat yang sama karena sering mengakses halaman *web* tertentu dan lain sebagainya.

Web Usage Mining merupakan kegiatan untuk memperoleh pengetahuan dari data *web log* sebuah *website* atau sering disebut juga dengan teknik untuk mengenali perilaku pengunjung, segmentasi pengunjung dan struktur *web* melalui informasi yang diperoleh dari *log*, *click stream*, *cookies*, dan *query*. Manfaat *web usage mining* adalah untuk kustomisasi halaman berdasarkan profil pengguna (personalisasi *web*), segmentasi pengunjung *web*, menentukan ketertarikan pengunjung/pelanggan (*web user*) terhadap produk tertentu, modifikasi *web*, dan menentukan target market yang sesuai (perilaku pengunjung).

2.4 Web Usage Mining

Tahapan *web usage mining* dibagi menjadi tiga kelompok utama yaitu pra-pemrosesan, penemuan pola dan analisis, serta interpretasi dan evaluasi. Perhatikan Gambar 2.3. Tahapan pra-pemrosesan meliputi data *cleaning*, data *integration*, data *transformation* dan data *reduction*. Pada tahap penemuan pola dan analisis diterapkan sejumlah formulasi statistik antara lain teknik klusterisasi, asosiasi, dan klasifikasi dan atau prediksi. Pada tahap evaluasi dilakukan analisis lebih lanjut untuk mengolah hasil penambangan data dari tahapan sebelumnya. Hal ini perlu dilakukan sebab sering sekali hasil yang diperoleh pada tahap penemuan pola dan analisis tidak memberikan sesuatu yang dapat digunakan secara langsung, sehingga diperlukan teknik lainnya seperti visualisasi grafik dan analisis statistik lainnya.

Tahapan pra-pemrosesan merupakan tahapan penting dalam *web usage mining*. Pada tahap ini, item data yang akan diolah diproses dengan berbagai cara dengan tujuan untuk membuang item data yang tidak perlu, hanya item data yang mempunyai relevansi kuat yang akan diproses, sehingga efisiensi *space* dan waktu dapat dicapai dan dihasilkan data yang lebih baik.



Gambar 2.3 Tahapan *Web Usage Mining*

2.4.1 Data Web Log

Data *web log* berisi informasi transaksi/rekam jejak atas setiap klik (*clickstream*) yang dilakukan pengunjung *web* terhadap *link* pada suatu halaman *website*. Data transaksi ini tersimpan secara otomatis pada *web* server, proxy server atau browser *log*.

Secara garis besar, data *web log* terdiri atas dua format, yaitu :

1. *Common Log Format (CLF)*, yaitu format file teks standar yang digunakan oleh *web* server saat membuat file log server (data *web log*). Format ini kebanyakan digunakan oleh *web* server Apache. Setiap baris dalam file CLF memiliki format berikut:

Host - ident - authuser - datetime - request - respond - bytes

Keterangan :

- *Host*, merujuk ke nama komputer/IP Address yang mengakses halaman *web*.
- *Ident/identifier*, merujuk ke RFC931, www.ietf.org/rfc/rfc931.txt.
- *Authuser*, merujuk ke nama user yang mengakses halaman *web*.

- *Datetime*, berisi data tanggal dan jam mengakses halaman *web*.
- *Request*, berisi metode HTTP dan *URL* dari halaman *web* yang diakses beserta protocol yang digunakan.
- *Respond*, berisi kode respon dari HTTP berupa angka [5], seperti :
 - Kode 200, berarti http berhasil merespon permintaan dengan baik.
 - Kode 301, berarti halaman *web* yang diminta sudah berganti URL (*moved permanently*)
 - Kode 404, berarti halaman *web* yang diminta tidak ditemui (*not found*).
 - Dan seterusnya,
http://en.wikipedia.org/wiki/List_of_HTTP_status_codes
- *Bytes*, yaitu jumlah bytes yang digunakan untuk mengakses halaman *web*.

Contoh data *web log* dalam bentuk CLF dapat dilihat pada Tabel 2.1.

Tabel 2.1 Common Log Format.

Log	Keterangan
172.21.13.45 - Microsoft\JohnDoe [08/Apr/2001:17:39:04 -0800] "GET /scripts/iisadmin.php HTTP/1.0" 200 3401	
172.21.13.45	Nama host/IP Address
-	RFC931 tidak tersedia
Microsoft\John Doe	Authuser
[08/Apr/2001:17:39:04 -0800]	Tanggal dan jam akses
GET /scripts/iisadmin/ism.dll?http/ser v HTTP/1.0	Metode akses, URL dan protocol yang digunakan
200	http kode
3401	Ukuran pengiriman file

2. *Extended Log Format (ELF)*, adalah format file teks standar seperti CLF yang digunakan oleh *web server* saat membuat file *log*, tetapi file ELF memberikan informasi yang lebih fleksibel. Format ini kebanyakan digunakan oleh *web server* IIS. Perhatikan Tabel 2.2

Tabel 2.2 Extended Log Format

Log	Keterangan
date	Tanggal akses
time	Jam akses
c-ip	IP Address dari klien.
cs-username	User name yang mengakses <i>web</i> .
s-sitename	ISP
s-computername	Nama <i>web</i> server
s-ip	IP Address dari server
s-port	Nomor port yang digunakan.
cs-method	Metoda akses yang digunakan, Metoda GET/Post.
cs-uri-stem	Halaman <i>web</i> yang diakses
sc-status	Kode status HTTP
sc-win32-status	Kode status windows
sc-bytes	Jumlah byte yang dikirim oleh server
cs-bytes	Jumlah byte yang diterima oleh server
time-taken	Jumlah waktu yang dibutuhkan untuk meloading halaman <i>web</i> .
cs-version	Versi protocol yang digunakan klien.
Dan seterusnya.	

Tabel 2.3 Contoh Data *Web* Log

180.254.140.182 - - [24/Oct/2011:00:06:03 +0700] "GET / HTTP/1.1" 200 13604 "http://www.facebook.com/" "Mozilla/5.0 (Windows NT 5.1) AppleWebKit/535.1 (KHTML, like Gecko) Chrome/13.0.782.220 Safari/535.1"
180.254.140.182 - - [24/Oct/2011:00:06:04 +0700] "GET /style.css HTTP/1.1" 200 1738 "http://e-tokobuku.com/" "Mozilla/5.0 (Windows NT 5.1) AppleWebKit/535.1 (KHTML, like Gecko) Chrome/13.0.782.220 Safari/535.1"
180.254.140.182 - - [24/Oct/2011:00:06:05 +0700] "GET /images/tm_left.gif HTTP/1.1" 200 380 "http://e-tokobuku.com/" "Mozilla/5.0 (Windows NT 5.1) AppleWebKit/535.1 (KHTML, like Gecko) Chrome/13.0.782.220 Safari/535.1"
180.254.140.182 - - [24/Oct/2011:00:06:05 +0700] "GET /images/tm_right.gif HTTP/1.1" 200 366 "http://e-tokobuku.com/" "Mozilla/5.0 (Windows NT 5.1) AppleWebKit/535.1 (KHTML, like Gecko) Chrome/13.0.782.220 Safari/535.1"

2.4.2 Lokasi Data *Web Log*

Data *web log* dapat ditemukan pada 3 (tiga) lokasi, yaitu :

1. *Web Server*, data *web log* yang tersimpan pada pada *web server* berisi seluruh data transaksi pengunjung dari sebuah *website*.
2. *Proxy Server*, hanya menyimpan data *web log* yang berada di proxy servernya saja.
3. *Client Browser*, hanya menyimpan data *web log* dari user yang menggunakan client browser itu saja,

Dalam konteks topik penelitian ini, digunakan data *web log* yang berasal dari *web server*.

2.5 Analisis Faktor

Analisis faktor (*factor analysis*) merupakan salah satu metode statistik multivariat yang mencoba menerangkan hubungan antara sejumlah variabel-variabel yang saling independen antara satu dengan yang lain sehingga bisa dibuat satu atau lebih kumpulan variabel yang lebih sedikit dari jumlah variabel awal, yang disebut dengan faktor [6].

Terdapat dua metode ekstraksi faktor pada analisis faktor, yaitu:

- Analisis komponen utama (*principal component analysis*), dan
- Analisis faktor (*Common factor analysis*).

Kedua analisis tersebut bertujuan untuk menerangkan struktur variabel melalui kombinasi linear dari variabel-variabel pembentuknya. Sehingga dapat dikatakan bahwa faktor atau komponen adalah variabel bentukan (variabel laten) bukan variabel asli.

Secara umum analisis faktor atau analisis komponen utama bertujuan untuk mereduksi data dan menginterpretasikannya sebagai suatu variabel baru yang berupa variabel bentukan. Pada dasarnya analisis faktor atau analisis komponen utama mendekati data pada suatu pengelompokan atau pembentukan suatu variabel baru yang berdasarkan adanya keeratan hubungan antar dimensi pembentuk faktor atau adanya konfirmatori sebagai variabel baru atau faktor.

Meskipun dari p buah variabel awal dapat diturunkan atau dibentuk sebanyak p buah faktor atau komponen untuk menerangkan keragaman total dari variabel, namun sering kali keragaman tersebut dapat diterangkan dengan lebih baik hanya oleh sejumlah kecil faktor yang terbentuk, katakanlah oleh sebanyak v buah faktor atau komponen yang terbentuk, di mana $v < p$; umpamanya dari sejumlah variabel p yaitu sebanyak 15 dimensi, dari 15 dimensi tersebut terbentuk sebanyak $v = 2$ buah faktor atau komponen yang dapat menerangkan kesepuluh dimensi awal. Maka akan diperoleh sebagian besar informasi tentang struktur dari p buah variabel asal yang dapat diterangkan oleh v buah faktor atau komponen yang terbentuk. Dalam hal ini v buah faktor atau komponen utama dapat mewakili p buah variabel asalnya, sehingga lebih sederhana.

Data asli yang dianalisis dalam analisis faktor dinyatakan dalam bentuk matriks berukuran $n \times p$ (di mana n jumlah sampel dan p variabel pengamatan), yang dapat direduksi ke dalam matriks yang berukuran lebih kecil dan mengandung sejumlah n pengukuran pada v buah komponen utama atau faktor, sehingga matriks yang terbentuk berukuran $n \times v$ (n jumlah sampel dan v komponen utama atau faktor), dan $v < p$.

Analisis faktor sering kali digunakan sebagai langkah awal dalam kebanyakan analisis statistik yang bersifat lebih besar atau lebih kompleks, seperti dalam analisis kluster di mana faktor atau variabel baru yang terbentuk dipergunakan sebagai *input* untuk melakukan analisis kluster terhadap suatu set data.

Penggunaan analisis faktor bertujuan untuk :

a. Identifikasi Faktor

yaitu untuk mengidentifikasi faktor yang mendasari dari sekumpulan besar variabel. Dengan mengelompokkan sejumlah besar variabel ke dalam jumlah yang lebih kecil dari kumpulan yang homogen dan membuat variabel baru yang disebut faktor yang mewakili sekumpulan variabel tersebut dalam bentuk yang lebih sederhana, maka akan lebih mudah untuk diinterpretasikan.

b. Penyaringan Variabel

Yaitu untuk penyaringan variabel untuk disertakan dalam penelitian statistik selanjutnya, seperti analisis klaster.

c. Meringkas Data

Penerapan analisis faktor selanjutnya adalah untuk mengekstrak sedikit atau banyak faktor sesuai yang diinginkan dari satu set variabel.

d. Memilih Variabel

Penggunaan teknik analisis faktor selanjutnya adalah untuk memilih sekelompok kecil perwakilan variabel yang representatif, walaupun sebagian besar variabel berkorelasi, hal ini bertujuan untuk memecah berbagai masalah praktis.

e. Pengelompokkan Objek

Selain mengidentifikasi kesamaan antara variabel, analisis faktor dapat digunakan untuk mengelompokkan objek.. Dalam prosedur ini, sering disebut analisis faktor sebagai *inverse*, sebuah sampel individu diukur pada sejumlah variabel acak, dan dikelompokkan ke dalam kelompok yang homogen berdasarkan antar-korelasinya.

2.5.1 Analisis Komponen Utama

Analisis faktor merupakan perluasan dari teknik analisis komponen utama, yang menjelaskan struktur hubungan di antara banyak variabel antarketergantungan dalam suatu sistem yang sering dinyatakan dengan keeratan hubungan. Metode ini telah dipergunakan secara luas dalam berbagai bidang ilmu pengetahuan.

Analisis faktor yang diturunkan berdasarkan metode komponen utama (PCA), data harus berbentuk matriks, maka model analisis faktor dapat diturunkan dari matriks varians-kovarians (Σ) yang diduga berdasarkan matriks varians-kovarians sampel (S^2) atau matriks korelasi (r).

Apabila semua variabel yang diamati mempunyai satuan pengukuran yang sama, maka analisis faktor dapat diturunkan dari matriks koefisien korelasi ρ yang diduga berdasarkan matriks koefisien korelasi sampel r .

Berdasarkan metode komponen utama, dapat ditentukan banyaknya faktor yang perlu dilibatkan dalam analisis lanjutan. Banyaknya faktor yang terbentuk adalah sebanyak variabel asal = p . Penentuan banyaknya faktor atau komponen yang dilibatkan dalam analisis lanjutan tergantung pada struktur datanya dan hasil analisis faktor dengan komponen varians yang lebih besar dari pada satu.

Dalam situasi tertentu, apabila v buah faktor yang dilibatkan dalam analisis cukup banyak, maka terdapat kesulitan dalam menginterpretasikan hasil analisis faktor. Hal ini dikarenakan adanya tumpang tindih variabel-variabel X_j yang dapat diterangkan oleh v buah faktor bersama tersebut. Untuk mengatasi hal ini, maka dilakukan rotasi faktor (*factor rotation*) [6]. Rotasi faktor tidak lain merupakan transformasi ortogonal dari faktor yang telah terbentuk agar tidak terjadi keadaan variabel yang tumpang tindih dalam menerangkan faktor bersama atau komponen bersama yang dapat dilihat dari nilai loading faktornya.

2.5.2 Model Analisis Faktor

Model analisis faktor dapat ditulis sebagai berikut :

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \dots + l_{1m}F_m + \varepsilon_1 \quad (2-1)$$

$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \dots + l_{pm}F_m + \varepsilon_p \quad (2-2)$$

dengan :

μ_i = rata-rata variabel i

ε_i = faktor spesifik ke $- i$

F_j = *common* faktor ke $- j$

l_{ij} = *loading* dari variabel ke $- i$ pada faktor ke $- j$

m = banyak faktor

Faktor yang unik tidak berkorelasi dengan sesama faktor yang unik dan juga tidak berkorelasi dengan *common factor*. *Common factor* sendiri bisa dinyatakan sebagai kombinasi linear dari variabel-variabel yang terlihat.

$$F_i = W_{i1} X_1 + W_{i2} X_2 + W_{i3} X_3 + \dots + W_{ik} X_k, \quad (2-3)$$

$i = 1, 2, 3, \dots, p$ dan $k = 1, 2, 3, \dots, p$.

Dengan :

F_i = Perkiraan faktor ke- i (didasarkan pada nilai variabel X dengan koefisien W_i)

W_i = bobot atau koefisien nilai faktor ke- i

k = banyaknya variabel

Dimungkinkan untuk memilih timbangan (*weight*) atau koefisien nilai faktor (*factor score coefficient*) sehingga faktor yang pertama menjelaskan sebagian besar porsi seluruh varian atau menyerap sebagian besar varian seluruh variabel. Kemudian set timbangan kedua dapat dipilih, sehingga faktor yang kedua menyerap sebagian besar sisa varian, setelah diambil faktor pertama, dengan syarat bahwa faktor yang kedua tidak berkorelasi (*orthogonal*) dengan faktor pertama. Prinsip yang sama dapat dipergunakan untuk memilih faktor selanjutnya, sebagai faktor tambahan, yaitu faktor ketiga. Jadi, faktor bisa diperkirakan/diestimasi sehingga nilai faktor yang satu tidak berkorelasi dengan faktor lainnya. Faktor yang diperoleh merupakan variabel baru yang tidak berkorelasi antara satu faktor dengan faktor lainnya, artinya tidak terjadi *multi collinearity*. Banyaknya faktor lebih sedikit dari banyaknya variabel asli yang dianalisis faktor, sebab analisis faktor memang mereduksi jumlah variabel yang banyak menjadi variabel baru yang jumlahnya lebih sedikit.

Bagian dari varian variabel ke- i dari m *common factor* disebut h_i^2 komunalitas ke- i yang merupakan jumlah kuadrat dari loading variabel ke- i pada m *common factor*, dengan rumus :

$$h_i^2 = l_{i1}^2 + l_{i2}^2 + \dots + l_{im}^2 \quad (2-4)$$

Hubungan antara varians variabel asal dengan, varians faktor dan varians *error* adalah sebagai berikut :

$$\begin{aligned} \text{var}(X_i) &= \text{variens yang dijelaskan oleh faktor untuk variabel asal ke-}i + \text{var}(\text{error}) \\ &= \text{communality} + \text{specific variance} \\ &= h_i^2 + \psi_i \end{aligned}$$

Besarnya bobot l_{ij} menggunakan metode komponen utama ataupun kemungkinan maksimum (*maximum likelihood*). Metode komponen utama terbagi menjadi dua metode yaitu non-iteratif dan iteratif. Nilai dugaan c_{ij} yang diperoleh dengan metode non-iteratif adalah :

$$l_{ij} = \frac{a_{ji} \sqrt{\lambda_j}}{s_{xi}} \quad (2-5)$$

l_{ij} adalah bobot (*loading*) dari variabel asal ke- i pada faktor ke- j

a_{ji} adalah koefisien variabel asal ke- i untuk komponen utama ke- j

λ_j adalah *eigen value* untuk komponen utama ke- j

s_{xi} adalah simpangan baku (*standard of deviation*) variabel asal ke- i

Untuk kepentingan interpretasi, seringkali diperlukan untuk memberi nama masing-masing faktor sesuai dengan besar harga mutlak bobot l_{uj} .

Diharapkan setiap variabel asal hanya dominan di salah satu faktor saja (Nilai harga mutlak bobot variabel asal mendekati 1 di salah satu faktor dan mendekati 0 untuk faktor lainnya). Harapan ini kadang-kadang tidak dapat dipenuhi, untuk mengatasi hal ini diperlukan rotasi dari matriks bobot **I**. Beberapa macam teknik rotasi adalah *varimax*, *quartimax*, *equamax*, *parsimax*. [6]

2.5.3 Jenis-Jenis Analisis Faktor

Terdapat dua jenis analisis faktor, yaitu :

1. Analisis faktor eksploratori

Analisis faktor eksploratori sering digunakan sebagai analisis awal dalam rangkaian suatu penelitian, terutama untuk tujuan reduksi data dari variabel asal menjadi variabel baru atau faktor yang jumlahnya lebih kecil dari variabel awal. Analisis ini juga disebut analisis komponen utama/*principal component analysis*, yaitu suatu teknik analisis faktor untuk menemukan hubungan antar variabel baru atau faktor yang terbentuk yang saling independen sesamanya, sehingga bisa dibuat satu atau beberapa kumpulan faktor yang lebih sedikit dari jumlah variabel awal yang tidak berkorelasi sesamanya.

2. Analisis faktor konfirmatori

Analisis faktor konfirmatori berawal dari teori dan konsep yang diketahui atau ditentukan sebelumnya, maka dibentuk sejumlah faktor, serta variabel apa saja yang termasuk ke dalam masing-masing faktor yang dibentuk yang sudah pasti tujuannya. Tujuan analisis ini adalah untuk mengidentifikasi adanya hubungan antar variabel dengan melakukan pengujian korelasi dan atau untuk menguji validitas serta reliabilitas dari instrumen penelitian yang digunakan.

2.5.4 Tahapan Analisis Faktor

Berikut ini adalah tahapan dalam melakukan analisis faktor :

Tahap 1 : Menyiapkan Data

Tahapan awal dalam analisis faktor ini melakukan proses uji kecukupan data dan identifikasi korelasi antar variabel dengan metode *Measure of Sampling Adequacy (MSA)* pada persamaan (2-6), *Kaiser-Meyer-Olkin (KMO)* pada persamaan (2-7) dan *Bartlett's Test* pada persamaan (2-8).

$$MSA_i = \frac{\sum_{i=1}^p r_{ij}^2}{\sum_{i=1}^p r_{ij}^2 + \sum_{j=1}^p a_{ij}^2} \quad (2-6)$$

$$KMO = \frac{\sum_{i=1}^p \sum_{j=1}^p r_{ij}^2}{\sum_{i=1}^p \sum_{j=1}^p r_{ij}^2 + \sum_{i=1}^p \sum_{j=1}^p a_{ij}^2} \quad (2-7)$$

Dimana :

$g = 1, 2, 3, \dots, p$ dan $h = 1, 2, 3, \dots, p$

r_{ij} = Koefisien korelasi antara variabel i dan j

a_{ij} = Koefisien korelasi parsial antara variabel i dan j

$$Bartlett's Test = -\ln |R| \left[n - 1 - \frac{2p+5}{6} \right] \quad (2-8)$$

Dimana :

$|R|$ = nilai determinan

n = jumlah data

p = jumlah variabel

Berdasarkan metode ini, sekelompok data dikatakan memenuhi asumsi kecukupan data dan memenuhi asumsi berkorelasi jika nilai *MSA*, *KMO* lebih besar dari 0.5 dan dengan nilai signifikansi dari uji *Bartlett* < 0.06 . Variabel dengan nilai *MSA* < 0.5 dikeluarkan dari analisis.

Tahap 2 : Menentukan Metode Analisis Faktor

Ada dua metode dalam analisis faktor, khususnya untuk menghitung koefisien skor faktor, yaitu *principal components analysis* dan *common factor analysis*. Dalam *principal components analysis*, jumlah varian dalam data dipertimbangkan. Dalam disertasi ini menggunakan PCA.

Tahap 3 : Penentuan Banyaknya Faktor

Penentuan jumlah faktor yang diinginkan merupakan bagian penting dalam analisis faktor, karena salah satu tujuan melakukan analisis faktor adalah mencari variabel baru yang disebut faktor yang saling tidak berkorelasi, bebas satu sama lainnya, lebih sedikit jumlahnya daripada variabel asli, akan tetapi bisa menyerap sebagian besar informasi yang terkandung dalam variabel asli atau yang bisa memberikan sumbangan terhadap varian seluruh variabel. Cara tersebut antara lain :

a. Penentuan Berdasarkan Apriori

Metode ini merujuk kepada pengalaman sebelumnya, peneliti sudah tahu berapa banyaknya faktor sebenarnya, dengan menyebutkan suatu angka, misalnya 3 atau 4 faktor yang harus disarikan dari variabel atau data asli.

b. Penentuan Berdasarkan *Eigenvalues*

Dalam pendekatan ini, hanya faktor dengan *eigenvalues* lebih besar dari 1 (satu) yang dipertahankan, kalau lebih kecil dari satu, faktornya tidak diikutsertakan dalam model. Suatu *eigenvalues* menunjukkan besarnya sumbangan dari faktor terhadap varian seluruh variabel asli.

c. Penentuan Berdasarkan Scree Plot

Scree plot merupakan suatu plot dari *eigenvalue* sebagai fungsi banyaknya faktor, dalam upaya menentukan banyaknya faktor yang bisa ditarik (*factor extraction*).

Tahap 4 : Rotasi Faktor

Rotasi faktor bertujuan untuk menyederhanakan struktur faktor, sehingga mudah untuk diinterpretasikan. Rotasi faktor digunakan jika metode ekstraksi faktor belum menghasilkan komponen faktor utama yang jelas. Ada dua metode rotasi yang berbeda yaitu *orthogonal and oblique rotation* [6].

Rotasi disebut *orthogonal rotation* kalau sumbu dipertahankan tegak lurus sesamanya (bersudut 90 derajat). Metode rotasi yang banyak dipergunakan ialah *varimax procedure*. Prosedur ini merupakan metode *orthogonal* yang berusaha meminimumkan banyaknya variabel dengan muatan tinggi (*high loading*) pada satu faktor, dengan demikian memudahkan pembuatan interpretasi mengenai faktor. Rotasi *orthogonal* menghasilkan faktor – faktor yang tidak berkorelasi satu sama lain (*uncorrelated each other*).

Sebaliknya rotasi dikatakan: *oblique rotation* kalau sumbu tidak dipertahankan harus tegak lurus sesamanya (bersudut 90 derajat) dan faktor – faktor tidak berkorelasi. *Oblique rotation* digunakan kalau faktor dalam populasi berkorelasi sangat kuat.

Tahap 5 : Interpretasi Faktor dan Skos Faktor

Interpretasi dipermudah dengan mengenali/mengidentifikasi variabel yang muatannya (*loading*) besar pada faktor yang sama. Faktor tersebut kemudian bisa diinterpretasikan, dinyatakan dalam variabel yang mempunyai *high loading* padanya. Manfaat lainnya di dalam membantu untuk membuat interpretasi ialah melalui plot variabel, dengan menggunakan *factor loading* sebagai koordinat [6].

Kalau suatu faktor tidak bisa dengan jelas didefinisikan dinyatakan dalam variabel aslinya, seharusnya diberi label sebagai faktor tidak terdefinisikan atau faktor umum (*undefined or a general factor*). Variabel-variabel yang berkorelasi kuat (nilai *factor loading* yang besar) dengan faktor tertentu akan memberikan inspirasi nama faktor yang bersangkutan.

2.6 Analisis Klaster

Analisis klaster merupakan salah satu metoda dalam mengelompokkan data menjadi beberapa kelompok berdasarkan tingkat kemiripannya. Analisis ini diawali dengan pemahaman bahwa sejumlah data tertentu mempunyai kemiripan diantara anggota kelompok yang lain. Secara logika, kelompok yang baik adalah kelompok yang mempunyai :

- a. Homogenitas (kesamaan) yang tinggi antar anggota dalam satu kelompok (*within klaster*).
- b. Heterogenitas (perbedaan) yang tinggi antar kelompok yang satu dengan kelompok yang lain (*between klaster*).

Dari konsep di atas dapat disimpulkan bahwa sebuah klaster yang baik adalah klaster yang mempunyai anggota-anggota yang semirip mungkin satu dengan yang lain, namun sangat tidak mirip dengan anggota-anggota di klaster yang lain. Dengan demikian, konsep analisis klaster pada dasarnya adalah mencari/mengelompokkan data yang mirip (*similarity*). „Mirip“ diartikan sebagai tingkat kesamaan karakteristik antara dua data.

Langkah-Langkah Analisis Klaster

Proses analisis klaster dilakukan secara bertahap [7], yaitu :

Tahap 1: Mengukur kesamaan antar objek (*similarity*)/Menetapkan ukuran jarak antar data.

Pada tahap ini dilakukan pengukuran seberapa jauh ada kesamaan antar objek. Pada analisis klaster terdapat tiga ukuran untuk mengukur kesamaan antar objek, yaitu ukuran asosiasi, ukuran korelasi dan ukuran kedekatan.

a) Ukuran Asosiasi

Ukuran asosiasi hanya dipakai untuk mengukur data berskala non metrik (nominal atau ordinal), dengan cara mengambil bentuk-bentuk dari koefisien pada tiap objeknya, dengan memutlakkan korelasi-korelasi yang bernilai negatif.

b) Ukuran Korelasi

Ukuran korelasi digunakan untuk mengukur data skala matriks. Kesamaan antar objek dapat diketahui dari koefisien korelasi antar pasangan objek yang diukur dengan menggunakan beberapa variabel.

c) Ukuran Kedekatan

Pengukuran kedekatan objek ini ada beberapa metode, diantaranya :

- Metode *Euclidean Distance*, mengukur jumlah kuadrat perbedaan nilai pada masing-masing variabel. Lihat persamaan 2-1.

$$d_{ij} = \sqrt{\sum_{k=1}^p (X_{ik} - X_{jk})^2} \quad (2-9)$$

Dengan :

d_{ij} = jarak antara objek ke-i dan objek ke-j

p = jumlah variabel

X_{ik} = data dari objek ke-i pada variabel ke-k

X_{jk} = data dari objek ke-j pada variabel ke-k

- Metode *Squared Euclidean Distance*, variasi dari *Euclidean Distance*, Lihat persamaan 2-2.

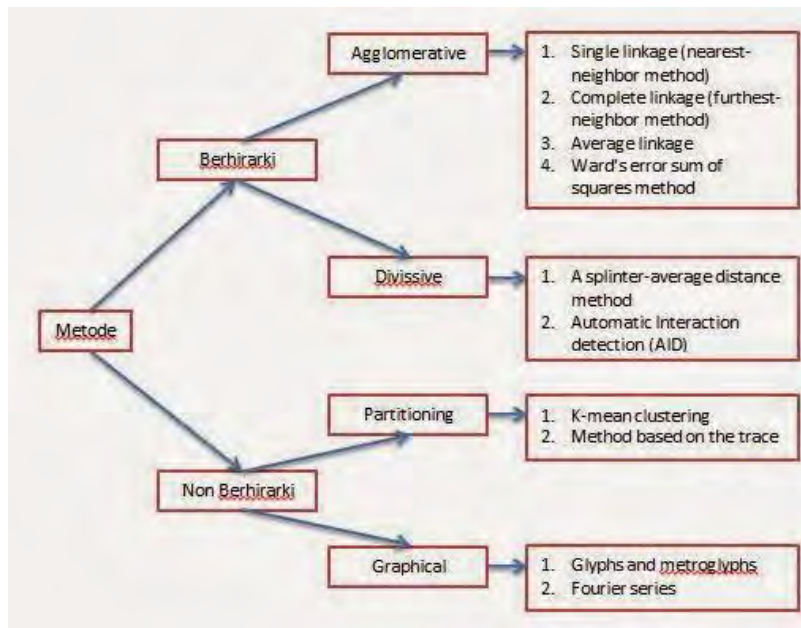
$$d_{ij} = \sum_{k=1}^p (X_{ik} - X_{jk})^2 \quad (2-10)$$

Tahap 2: Membuat Klaster.

Metode dalam membuat klaster ini dapat dilihat pada gambar 2.4.

Diantaranya :

- a. Metode Hirarki, Metode ini memulai pengelompokan dengan dua atau lebih objek yang mempunyai kesamaan paling dekat. Kemudian proses diteruskan ke objek lain yang mempunyai kedekatan kedua. Demikian seterusnya sehingga klaster akan membentuk seperti “pohon”, yang memiliki hirarki (tingkatan) yang jelas antar objek, dari yang paling mirip sampai paling tidak mirip.



Gambar 2.4 Diagram Analisis Kluster

Dalam metode hirarki ini, kluster terdapat dua tipe dasar yaitu :

- a. *agglomerative* (pemusatan) dan
- b. *divisive* (penyebaran).

Dalam metode pemusatan ini, setiap objek atau observasi dianggap sebagai sebuah kluster tersendiri. Dalam tahap selanjutnya, dua kluster yang mempunyai kemiripan digabungkan menjadi sebuah kluster baru demikian seterusnya.

Dalam metode Pemusatan ada lima metode yang sering digunakan, yaitu:

- *Single Linkage*, metode ini mengelompokkan dua objek yang mempunyai jarak terdekat terlebih dahulu. Jarak antar kelompok (i, j) dengan k dapat di lihat pada persamaan 2-11.

$$d_{ij} = \min(d_{ik}, d_{jk}) \quad (2-11)$$

Jika dua objek terpisah oleh jarak yang pendek/terdekat maka kedua objek tersebut akan digabung menjadi satu kluster, demikian seterusnya.

- *Complete Linkage*, berlawanan dengan *Single Linkage* prosedur ini pengelompokkannya berdasarkan jarak terjauh. Jarak antar kelompok (i,j) dengan k dapat di lihat pada persamaan 2-12.

$$d_{ij} = \max(d_{ik}, d_{jk}) \quad (2-12)$$

- *Average Linkage*, prosedur ini hampir sama dengan *Single Linkage* maupun *Complete Linkage*, namun kriteria yang digunakan adalah rata-rata jarak seluruh individu dalam suatu klaster dengan jarak seluruh individu dalam klaster yang lain. Jarak antar kelompok (i,j) dengan k dapat di lihat pada persamaan 2-13.

$$d_{ij} = \text{avg}(d_{ik}, d_{jk}) \quad (2-13)$$

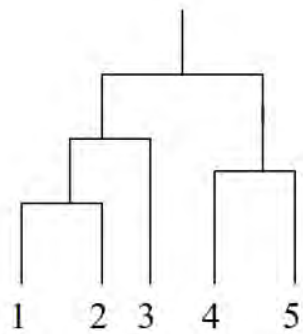
- *Ward's Method*, jarak antara dua klaster dalam metode ini berdasarkan *total sum of square* dua klaster pada masing-masing variabel.
- *Centroid Method*, jarak antara dua klaster dalam metode ini berdasarkan jarak *centroid* dua klaster yang bersangkutan. Jarak antar kelompok (i,j) dengan k dapat di lihat pada persamaan 2-14.

$$d_{ij} = \text{median}(d_{ik}, d_{jk}) \quad (2-14)$$

Dalam metode penyebaran, berawal dari sebuah klaster besar yang terdiri dari semua objek atau observasi. Selanjutnya, objek atau observasi yang paling tinggi nilai ketidakmiripannya dipisahkan demikian seterusnya.

Keuntungan penggunaan metode hirarki dalam analisis klaster adalah mempercepat pengolahan dan menghemat waktu karena data yang diinputkan akan membentuk hirarki atau membentuk tingkatan tersendiri sehingga mempermudah dalam penafsiran, namun kelemahan dari metode ini adalah seringkali terdapat kesalahan pada data *outlier*, perbedaan ukuran jarak yang digunakan, dan terdapatnya variabel yang tidak relevan.

Hasil akhir dari analisis hirarki disajikan dalam bentuk struktur pohon yang disebut dengan dendogram. Dendogram adalah representasi visual dari langkah-langkah dalam analisis klaster yang menunjukkan bagaimana klaster terbentuk. Perhatikan Gambar 2.5.



Gambar 2.5 Contoh Dendrogram

Gambar 2.5 menunjukkan bahwa, jika diinginkan 2 klaster, maka anggota klaster 1 adalah 1,2 dan 3, sedangkan 4 dan 5 masuk kedalam klaster 2. Jika diharapkan 3 klaster, maka anggota klaster 1 adalah 1 dan 2, anggota klaster 2 adalah 3 dan anggota klaster 3 adalah 4 dan 5.

Sedang metode non-hirarki memiliki keuntungan dapat melakukan analisis sampel dalam ukuran yang lebih besar dengan lebih efisien. Selain itu, hanya memiliki sedikit kelemahan pada data *outlier*, ukuran jarak yang digunakan, dan variabel tak relevan atau variabel yang tidak tepat. Sedangkan kelemahannya adalah untuk titik bakal random lebih buruk dari pada metode hirarkhi.

- b. Metoda Non Hirarki atau partisi, Berbeda dengan metode hirarki, metode ini justru dimulai dengan terlebih dahulu menentukan jumlah klaster yang diinginkan (dua klaster, tiga klaster atau yang lain). Setelah jumlah klaster diketahui, baru proses klaster dilakukan tanpa mengikuti proses hirarki. Metode ini biasa disebut dengan *K-Means Klaster*.

Tahap 3. Melakukan validasi dan profiling klaster.

Validasi merupakan proses untuk menilai hasil algoritma klaster yang terbentuk, menjamin bahwa solusi klaster yang dihasilkan dapat menggambarkan populasi yang sebenarnya. Ada 3 pendekatan untuk melakukan validasi ini, yaitu :

- a. Validasi eksternal, pada pendekatan ini data dibagi menjadi dua bagian, lalu hasil klasternya dibandingkan.
- b. Validasi internal, pada pendekatan ini, solusi klaster dibandingkan hasil metoda klaster non hirarki.

- c. Validasi relatif, dilakukan perbandingan menggunakan algoritma yang sama dengan parameter yang berbeda.

Validasi menggunakan *Root Mean Square Standard Deviation*, yaitu melihat simpangan baku gabungan dari semua variabel yang membentuk kluster. Perhatikan rumus 2.15.

$$RMSSTD = \sqrt{\frac{\sum_{i=1}^n \sum_{j=1}^d (x_{ij} - \bar{x}_i)^2}{\sum_{i=1}^n (d_i - 1)}} \quad (2-15)$$

Nilai RMSSTD berada pada interval $(0, \infty)$, dan nilai RMSSTD yang kecil berarti semakin mirip, mengindikasikan adanya homogenitas yang tinggi dalam kluster.

Validasi dengan *Sum of Square Error (SSE)*, yaitu menjumlahkan nilai kuadrat dari jarak data dengan pusat kluster. SSE dinyatakan dengan rumus 2-16.

$$SSE = \sum_{i=1}^n (d)^2 \quad (2-16)$$

Dengan d adalah jarak antara data dengan pusat kluster. Semakin kecil nilai SSE menandakan bahwa data semakin dengan pusat kluster, dan hal tersebut mengindikasikan adanya homogenitas yang tinggi dalam kluster.

Kesimpulannya bahwa validasi akan memberikan informasi tentang ketepatan jumlah kluster yang dipilih.

Proses profiling berguna untuk menjelaskan karakteristik setiap kluster berdasarkan profil tertentu, dalam disertasi ini digunakan untuk keperluan segmentasi pengunjung web.

2.7 Segmentasi Pengunjung Web

Segmentasi merupakan proses membentuk kelompok-kelompok dari suatu objek yang heterogen menjadi kelompok-kelompok dengan kriteria kesamaan (homogen) tertentu. Anggota yang berada dalam kelompok sama akan memiliki tingkat kesamaan yang tinggi sesuai dengan kriteria yang ditentukan dan akan mempunyai tingkat perbedaan yang tinggi dengan anggota di kelompok lainnya. Dapat juga diartikan bahwa segmentasi adalah suatu proses membagi objek yang heterogen ke dalam kelompok-kelompok pelanggan/pengunjung web potensial

dengan kesamaan karakteristik yang menunjukkan adanya kesamaan perilaku kunjungan.

Dalam kaitannya dengan segmentasi pengunjung web, maka sesungguhnya hal ini adalah upaya untuk meneliti perilaku pengunjung web itu sendiri, sehingga pengunjung yang memiliki pola perilaku dan intensitas kunjungan yang sama atau hampir sama akan bergabung dengan kelompok yang sama dan pengunjung dengan pola yang lain akan bergabung dengan kelompok yang lain pula. Pola-pola perilaku pengunjung web ini direkam secara otomatis yang disebut dengan data *web log*.

Komponen penting untuk analisa perilaku pengunjung ini yang terdapat di data *web log* adalah :

- IP Address/Host untuk identitas pengunjung,
- Tanggal dan jam akses untuk identifikasi waktu kunjungan, dan
- Halaman web yang diakses

Tujuan dari segmentasi pengunjung web ini adalah menganalisa pola perilaku pengunjung web itu sendiri sehingga dapat menjadi acuan dalam strategi menentukan perbaikan layanan/service terhadap pengunjung web (personalisasi web), perbaikan konten, perbaikan tampilan web (modifikasi web), prediksi perilaku pengunjung web ataupun untuk strategi pemasaran web secara lebih luas.

Dengan segmentasi ini, pemilik web akan mengetahui :

- Pengunjung web yang loyal (sering mengunjungi web),
- Halaman web yang sering diakses oleh pengunjung web.
- Pola perilaku pengunjung web.
- Halaman web yang jarang dikunjungi (tidak diminati).
- Dapat membedakan antara kelompok yang satu dengan kelompok lainnya.
- Dapat digunakan untuk mengetahui sifat masing-masing kelompok.
- Dapat digunakan untuk mencari kelompok mana yang potensinya paling besar.
- Dapat digunakan untuk memilih kelompok mana yang akan dijadikan sasaran untuk promosi dan lainnya.

Oleh karena itu, penelitian-penelitian di bidang segmentasi pengunjung web ini sangat aktif dan terus berkembang, terutama segmentasi pada website *e-commerce*, *e-news*, portal, sosial media, web perusahaan yang mengelola banyak pengunjung web (user).

Tingginya tingkat pengunjung web tidak terlepas dari cara memahami dan mendalami kebutuhan pengunjung web, sehingga kemampuan sebuah website untuk menampilkan yang dibutuhkan oleh pengunjung web secara personal merupakan sebagian dari hasil segmentasi pengunjung web.

Personalisasi web merupakan upaya pengelola website untuk memberikan layanan dan keleluasaan kepada pengunjung web untuk menentukan konten yang diinginkan, tampilan dan layout webnya sendiri, sehingga setiap pengunjung web yang sama akan memiliki tampilan halaman web yang berbeda.

Beberapa manfaat dari segmentasi pengunjung web dalam hal personalisasi halaman web diantaranya :

- *Web e-Commerce* : Hal mendasar dari personalisasi *web e-commerce* ini adalah kemampuan *web* memprediksi keinginan pengunjung, prediksi diambil dari pola perilaku pengunjung sebelumnya ataupun melalui survei. Sehingga, diantara ribuan produk yang mereka jual, web personalisasi mampu menawarkan langsung produk keinginan pengunjung atau berdasarkan pola perilaku pengunjung lainnya. *Website* mampu mengidentifikasi kepentingan setiap pengunjung, sehingga hanya menampilkan produk, iklan, konten dan lainnya yang sesuai dengan kebutuhan pengunjung tersebut. Dengan kata lain, setiap pengunjung yang mengakses halaman web yang sama, dapat memiliki konten, tampilan, fitur dan iklan yang berbeda.
- *Web Travel* : Website perjalanan dapat mengidentifikasi pengunjung yang sebelumnya telah pernah mengunjungi *web* mereka dengan desain yang relevan, banner, gambar, promosi dan artikel informasi, berdasarkan pada tujuan, musim dan jenis liburan di mana mereka tertarik.
- *Web e-News/Portal berita* : *Web* berita online dan portal dapat menampilkan iklan yang sangat relevan berdasarkan kebutuhan

pengunjung untuk meningkatkan pendapatan dari iklan, serta berita dan artikel yang relevan, berdasarkan kategori favorit pengunjung dan upaya untuk meningkatkan durasi kunjungan, jumlah artikel dibaca, serta persentase pengunjung kembali.

- Web Sosial, dan lainnya.

BAB 3

PRA-PEMROSESAN dan TAHAPAN KLASTERISASI DATA *WEB LOG*

Data *web log* yang tersimpan dalam jumlah yang sangat besar pada *web server* dalam bentuk *single file* memiliki sangat banyak item data yang tidak relevan (tidak valid), sehingga pra-pemrosesan terhadap data *web log* sangat penting dilakukan sebelum dilakukan proses *web mining* lebih lanjut, karena hampir 80% dari keseluruhan proses penambangan data berada pada tahap pra-pemrosesan ini.

Salah satu indikator data *web log* dikatakan valid adalah data tersebut merujuk langsung ke sebuah halaman *website* yang mengandung informasi, ditandai dengan ekstensi file *.php*, *.html*, *.jsp* dan lainnya. Sedangkan item data yang terkandung dalam halaman *website*, yang otomatis direkam sebagai *log* sewaktu pengunjung mengakses halaman *web* tersebut seperti file gambar, audio, video, respon dari HTTP (*HyperText Transfer Protocol*) yang berawalan selain 2 (1xx, 3xx, 4xx, 5xx) dan lainnya disebut data tidak valid (*irrelevant data*) dalam penelitian ini.

3.1 Pra-Pemrosesan Data *Web Log*

Kegiatan pada pra-pemrosesan data *web log* terdiri atas beberapa tahap dan diuraikan dibawah ini. Data *web log* yang telah dikumpulkan dikonversi menjadi sebuah basis data untuk dilakukan tahapan pra-pemrosesan. Prosedur yang dilakukan terlihat pada Gambar 3.1.

3.1.1 Pembersihan Data

Data *cleaning* adalah proses membersihkan data *web log* dari *item data* yang tidak memberikan informasi berguna dalam proses penambangan data selanjutnya. Item data yang dihapus didasarkan pada :

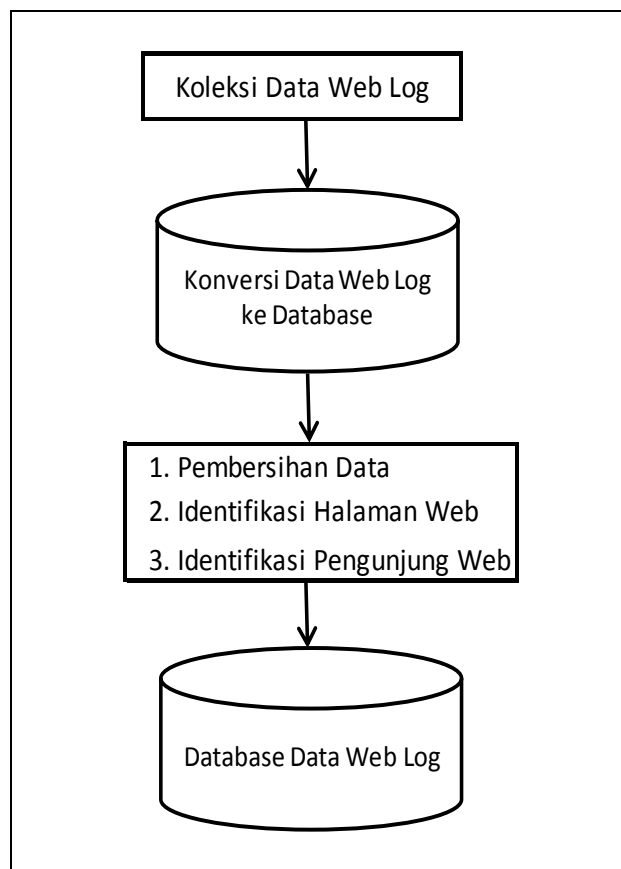
- a. **Ekstensi file**, ekstensi file yang diterima adalah .html, .php, .jsp, asp dan ekstensi lainnya yang merujuk langsung ke sebuah halaman *web*. Sedangkan ekstensi file yang merujuk ke file gambar, video, suara, style seperti : .jpg, .gif, .ico, .bmp, .cgi, .swf, .css, .txt, tidak menggambarkan perilaku pengunjung *web* sehingga item data ini dihapus.
- b. **Kode respon HTTP** [5][8] dari *web* server, kode status *web* server dengan angka 200 menandakan permintaan akses sebuah halaman *web* diterima dan ditampilkan oleh *web* server. Oleh karena itu, item data *web log* dengan kode selain dari 200 akan dihapus. Perhatikan Tabel 3.1.

Tabel 3.1 Sebagian Daftar Kode Respon HTTP

Kode	Makna	Kode	Makna
100	Continue	400	Bad Request
101	Switching Protocols	401	Unauthorized
200	Ok	402	Payment Required
201	Created	403	Forbidden
202	Accepted	404	Not Found
203	Non-Authoritative Information	405	Method Not Allowed
204	No Content	406	Not Acceptable
205	Reset Content	407	Proxy Authentication
206	Partial Content	408	Request Timeout
207	Multi-Status	409	Conflict
300	Multiple Choices	410	Gone
301	Moved Permanently	500	Internal Server Error
302	Found	501	Not Implemented
303	See Other	502	Bad Gateway
304	Not Modified	503	Service Unavailable
305	Use Proxy	504	Gateway Timeout
306	(Reserved)	505	HTTP Version Not Supported
307	Temporary Redirect	507	Insufficient Storage

- c. **Metode HTTP** [9]. Hanya akses dengan metode GET yang menandakan perilaku pengunjung *web*. Item data dengan metode akses lainnya, seperti HEAD, POST, OPTIONS, PUT dan lainnya akan dihapus.

Item data tersebut tidak memberikan informasi yang bermanfaat terhadap analisis pola perilaku/segmentasi dari pengunjung *website*, dan kriteria lainnya yang ditetapkan oleh peneliti.

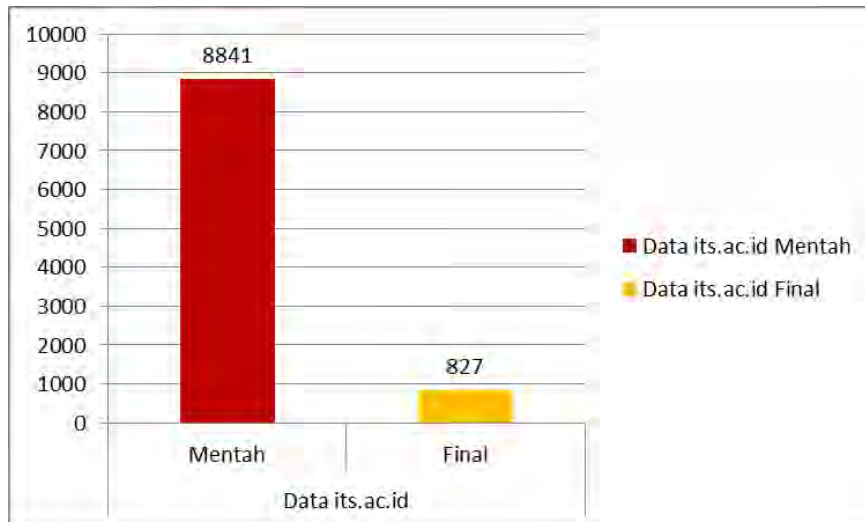


Gambar 3.1 Tahapan Pra-pemrosesan yang dilakukan

Tabel 3.2 dan Gambar 3.2 merupakan contoh rekapitulasi data *web log* setelah dilakukan pembersihan data pada sebuah dataset. Dari Tabel 3.2 tersebut dapat dinyatakan bahwa 90,65 % dari data mentah ternyata tidak digunakan dalam proses penambangan data, sehingga tahap ini memberikan andil yang sangat besar dalam menentukan keakuratan dan kecepatan proses penambangan *web* berikutnya.

Tabel 3.2 Contoh Perbandingan Jumlah Data Sebelum dan Sesudah dilakukan Pembersihan Data

	Jumlah Data	
	Sebelum <i>Cleaning</i>	Setelah <i>Cleaning</i>
Jumlah Data	8841 data	827 data
Persentase	100%	9,35%



Gambar 3.2 Grafik Perbandingan Jumlah Data Sebelum dan Sesudah dilakukan Proses *Cleaning*

3.1.2 Identifikasi Halaman Web

Identifikasi halaman web adalah proses menentukan halaman *web* mana saja yang diakses dan siapa saja yang mengakses halaman *web* tersebut. Dapat dimodelkan dengan notasi berikut :

$$P = p_1, p_2, p_3, \dots, p_n$$

Dengan P adalah halaman web yang diakses dan $p_1, p_2, p_3, \dots, p_n$ merupakan halaman *web* ke 1, ke 2 dan seterusnya. Sedangkan untuk transaksi pengunjung *web* dapat dimodelkan dengan notasi berikut :

$$T = t_1, t_2, t_3, \dots, t_m$$

Dengan T adalah transaksi pengunjung *web* dan $t_1, t_2, t_3, \dots, t_m$ merupakan jumlah transaksi pengunjung *web*. Dengan setiap t , ET adalah bagian dari P .
Tabel 3.3 merupakan contoh hasil dari tahap ini.

Tabel 3.3 Contoh Output Identifikasi Pengunjung Web

Halaman web	Jumlah Akses	Jumlah Pengunjung Unik
Halaman web 1	78	73
Halaman web 2	22	21
Halaman web 3	10	10
Halaman web 4	49	46
Halaman web 5	38	34
Halaman web 6	258	198
:	:	:
Halaman web n

3.1.3 Identifikasi Pengunjung Web

Identifikasi pengunjung web adalah proses menentukan aktifitas pengunjung web serta membedakan interaksi antar pengunjung web, karena seorang pengunjung web dapat mengunjungi *website* lebih dari satu kali dan dengan pola perilaku yang berbeda. Proses identifikasi pengunjung web ini dapat disaring berdasarkan alamat IP, tanggal akses, referensi, agent, rentang waktu jam akses atau kombinasi diantaranya.

Tabel 3.4 Contoh Hasil Identifikasi Pengunjung Web 1

Jam Akses	IP Address	URL Target	Referer
00:6:3	180.254.140.182	/index.html	http://www.facebook.com/
00:7:50	180.254.140.182	/website.html	http://e-tokobuku.com/
00:7:50	180.254.140.182	/database.html	http://e-tokobuku.com/
00:8:1	180.254.140.182	/katalog.html	http://e-tokobuku.com/
00:8:2	180.254.140.182	/bukubar.html	http://e-tokobuku.com/
06:54:2	110.137.107.244	/index.html	http://http://www.facebook.com/
06:54:27	110.137.107.244	/www.ephi.web.id	http://e-tokobuku.com/
06:54:38	110.137.107.244	/www.gramediaishop.com	http://e-tokobuku.com/
06:54:39	110.137.107.244	/www.lintau.info	http://e-tokobuku.com/
06:54:47	110.137.107.244	/bukubar.html	http://e-tokobuku.com/
06:54:59	110.137.107.244	/website.html	http://e-tokobuku.com/bukubar.html
06:55:2	110.137.107.244	/tentangwebsite.html	http://e-tokobuku.com/website.html
06:55:10	110.137.107.244	/penulis.html	http://e-tokobuku.com/tentangwebsite.html
06:56:1	110.137.107.244	/katalog.html	http://e-tokobuku.com/penulis.html

Tabel 3.4 dan Tabel 3.5 memperlihatkan contoh dari tahap identifikasi pengunjung web.

Tabel 3.5 Contoh Hasil Identifikasi Pengunjung Web 2

Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE6;WinXP;SP1
0:12	2.3.4.5	B	C	IE6;WinXP;SP1
0:15	2.3.4.5	E	C	IE6;WinXP;SP1
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE6;WinXP;SP1
0:22	1.2.3.4	A	-	IE6;WinXP;SP2
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE6;WinXP;SP2
0:33	1.2.3.4	B	C	IE6;WinXP;SP2
0:58	1.2.3.4	D	B	IE6;WinXP;SP2
1:10	1.2.3.4	E	D	IE6;WinXP;SP2
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE6;WinXP;SP2
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	
0:09	1.2.3.4	B	A	
0:19	1.2.3.4	C	A	
0:25	1.2.3.4	E	C	
1:15	1.2.3.4	A	-	
1:26	1.2.3.4	F	C	
1:30	1.2.3.4	B	A	
1:36	1.2.3.4	D	B	

Time	IP	URL	Ref	Agent
0:10	2.3.4.5	C	-	
0:12	2.3.4.5	B	C	
0:15	2.3.4.5	E	C	
0:22	2.3.4.5	D	B	

Time	IP	URL	Ref	Agent
0:22	1.2.3.4	A	-	
0:25	1.2.3.4	C	A	
0:33	1.2.3.4	B	C	
0:58	1.2.3.4	D	B	
1:10	1.2.3.4	E	D	
1:17	1.2.3.4	F	C	

Hingga tahap ini, telah memberikan gambaran pola perilaku pengunjung *website*. Data ini dapat diolah lebih lanjut dengan membuat *session by time*, seperti terlihat pada Tabel 3.6.

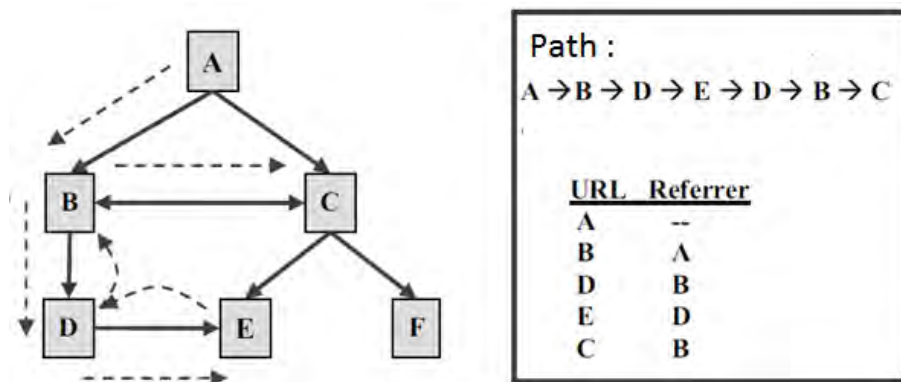
Tabel 3.6 Contoh Hasil Identifikasi Pengunjung Web 3

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C

Time	IP	URL	Ref
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Hasil dari identifikasi pengunjung web ini dikonversi menjadi *user navigation* untuk memudahkan dalam proses penambangan berikutnya. Contoh *user navigation* dapat dilihat pada Gambar 3.3.



Gambar 3.3 Contoh User Navigation

Berdasarkan Gambar 3.3, pada bagian *path* menunjukkan pola kunjungan dari pengunjung web, yaitu pengunjung web pertama kali mengakses halaman web A, kemudian berpindah ke halaman web B, D, E, D, B dan C. Pola-pola perilaku seperti ini akan dibagi berdasarkan session (*by time*) untuk dilakukan analisis berikutnya (lihat Tabel 3.5).

3.2 Tahapan Klasterisasi

Dalam penambangan data *web log* ini pada disertasi ini, klasterisasi dilakukan pada 2 model bentuk data *web log* yang dilakukan secara berurut, pertama dilakukan pada data *web log* yang berbentuk *frequently access* (berdasarkan jumlah kunjungan) dan yang kedua pada data *web log* yang berbentuk *user access pattern* (berdasarkan pola kunjungan terhadap web). Pembahasan tentang klasterisasi dapat dilihat pada sub bab 2.5.

Klasterisasi dilakukan dua tahap, dengan hipotesa untuk mendapatkan segmentasi pengunjung web yang lebih baik dibandingkan dari penggunaan metode satu tahap. Untuk itu, dalam penelitian pada disertasi ini diawali dengan klasterisasi satu tahap untuk data *web log* yang berbentuk *frequently access* (berdasarkan jumlah kunjungan) dan klasterisasi pada data *web log* yang berbentuk *user access pattern* (berdasarkan pola kunjungan).

3.2.1 Klasterisasi Tahap Pertama

Data *web log* yang telah melewati tahap pembersihan data (*data cleaning*) dan identifikasi pengunjung web (*user identification*) diubah formatnya menjadi berbentuk *frequently access* seperti terlihat pada Tabel 3.7. Pada format ini akan terlihat frekuensi kunjungan *user* dan perlu dilakukan proses penghitungan statistik terhadap frekuensi kunjungan tersebut.

Tabel 3.7 Pola *Frequently Access*

User	Halaman Web						
	p_1	p_2	p_3	p_4	p_5	p_6	p_n
u_1	15	5	0	0	0	185	u_{1p_n}
u_2	0	0	32	4	0	0	u_{2p_n}
u_3	12	0	0	56	236	0	u_{3p_n}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
u_n	u_{p1}	u_{p2}	u_{p3}	u_{p4}	u_{p5}	u_{p6}	u_{np_n}

Dengan p_1, p_2, \dots, p_n adalah variabel yang mewakili halaman *web* (misal variabel p_1 mewakili halaman *web* *index.html* dan seterusnya), dan u_1, u_2, \dots, u_n merupakan *user*/pengunjung *web*. Dari Tabel 3.7 diatas dapat dinyatakan bahwa u_1 telah mengakses halaman *web* p_1 sebanyak 15 kali, halaman *web* p_2 sebanyak 5 kali dan seterusnya.

Dari data dan pola ini diterapkan algoritma klasterisasi untuk mendapatkan kelompok-kelompok pengunjung *web* yang mempunyai karakteristik sama untuk setiap kelompoknya berdasarkan frekuensi kunjungannya. Tujuan klasterisasi pada tahap pertama ini adalah untuk memilih data *web log* yang akan diproses lebih lanjut. Kelompok yang memiliki data terbanyak akan digunakan sebagai data masukan pada klaster tahap kedua.

3.2.2 Klasterisasi Tahap Kedua

Data yang terpilih dari tahap klasterisasi pertama diubah formatnya menjadi berbentuk pola kunjungan *user* (*user access pattern*), seperti terlihat pada Tabel 3.8.

Tabel 3.8 Pola *User Access Pattern*

User	Urutan Akses						
	1	2	3	4	5	6	n
u_1	p_1	p_3	p_2	p_5	p_7	-	$u_{1_{pn}}$
u_2	p_3	p_1	p_4	p_6	p_1	-	$u_{2_{pn}}$
u_3	p_1	p_2	p_3	p_5	p_7	-	$u_{3_{pn}}$
:	:	:	:	:	:	:	:
u_n	$u_{n_{p1}}$	$u_{n_{p2}}$	$u_{n_{p3}}$	$u_{n_{p4}}$	$u_{n_{p5}}$	$u_{n_{p6}}$	$u_{n_{pn}}$

Angkat 1, 2, 3,... n, melambangkan urutan akses dari pengunjung web. $u_1, u_2, u_3, \dots u_n$ merupakan user/pengunjung web, sedangkan $p_1, p_2, \dots p_n$ adalah variabel yang mewakili halaman *web* (misalnya, variabel p_1 mewakili halaman web *index.php*, variabel p_2 mewakili halaman web *academic.php* dan seterusnya). Dari Tabel 3.8, dapat dinyatakan bahwa u_1 pertama kali mengakses halaman web p_1 , kemudian berpindah ke halaman web p_3 , halaman web p_2 dan seterusnya, sehingga terlihat pola urutan akses dari setiap pengunjung *web*.

Pola kunjungan inilah yang akan dilakukan klasterisasi tahap kedua untuk mendapatkan segmentasi pengunjung web, sehingga didapatkan profil serta interpretasinya.

Halaman ini sengaja dikosongkan

BAB 4

SINGLE-STAGE CLUSTERING PADA DATA *WEB LOG* YANG BERBENTUK *FREQUENTLY ACCESS* UNTUK SEGMENTASI PENGUNJUNG WEB

Klasterisasi merupakan bagian dari statistik multivariat yang bertujuan untuk mengelompokkan objek berdasarkan kesamaan karakteristik yang dimiliki sehingga data dalam satu klaster akan memiliki homogenitas yang tinggi, dan sebaliknya akan memiliki heterogenitas yang tinggi antar klasternya. Dalam penelitian ini diimplementasikan teknik klaster dengan menggunakan metode K-Means untuk mendapatkan klaster pengunjung sebuah *web* berdasarkan frekuensi kunjungan (*frequently access*) sehingga diperoleh klaster-klaster pengunjung *web* dengan pola frekuensi kunjungan yang sama. Penentuan jumlah klaster menggunakan metode koefisien *silhouette*.

Dengan penerapan metode klasterisasi ini, maka pemilik *web* dapat melakukan segmentasi pengunjung *web* untuk peningkatan layanan maupun modifikasi *website* dan personalisasi *web*.

Dari pengujian yang dilakukan terhadap data *web log* www.its.ac.id, diperoleh 6 klaster pengunjung *web* dan klaster ke-3 mempunyai jumlah anggota terbesar. Hal ini menjadi masukan bagi pengelola *web* untuk lebih memperhatikan pola perilaku anggota klaster ke-3 tersebut untuk peningkatan layanan/modifikasi *web*.

4.1 Pendahuluan

Pesatnya penyebaran informasi saat ini tidak terlepas dari peran internet, terutama melalui layanannya yang disebut dengan *world wide web (web)*. Sebuah *website* mampu menyebarkan informasi dalam format teks, gambar, video, suara dan atau multimedia[10]. Menurut laporan lembaga survei *netcraft*[11] pada bulan Juli 2012, menyatakan bahwa terdapat 665,916,461 *web* yang aktif. Jika dilihat dari tingkat pengguna internet, menurut laporan *internet world stats*[12], estimasi pengguna internet dunia pada tahun 2015 adalah 7.264.623.793 pengguna internet.

Hal ini berarti interaksi antara pengguna internet dengan *web* sangat tinggi dan *web* merekam setiap aktifitas pengunjung tersebut dalam bentuk file (*web log*). Hingga saat ini, *web log*[13] menjadi bagian terpenting dalam bidang *web mining* untuk menambang data pengunjung *web*, terutama dalam menemukan pola akses/perilaku pengunjung.

Bagian dari *web mining*[14] yang khusus melakukan penambangan pada data *web log* adalah *Web Usage Mining* (WUM). Khasawneh[15] menyebutkan bahwa WUM adalah penerapan teknik data *mining* untuk menemukan pola interaksi pengunjung dari sebuah *web* melalui data *web log*. Penambangan pada data *web log* berguna dalam berbagai bidang, di antaranya untuk keperluan segmentasi pengunjung *web* [16][17], personalisasi *web* (*web personalization*)[18][19] dan modifikasi *web* (*web modification*)[20].

Metode penambangan pada WUM meliputi analisis statistik[21], *association rules*[22][23], pola berurut[24][25], klasifikasi[26][27] dan klasterisasi[28][29][30].

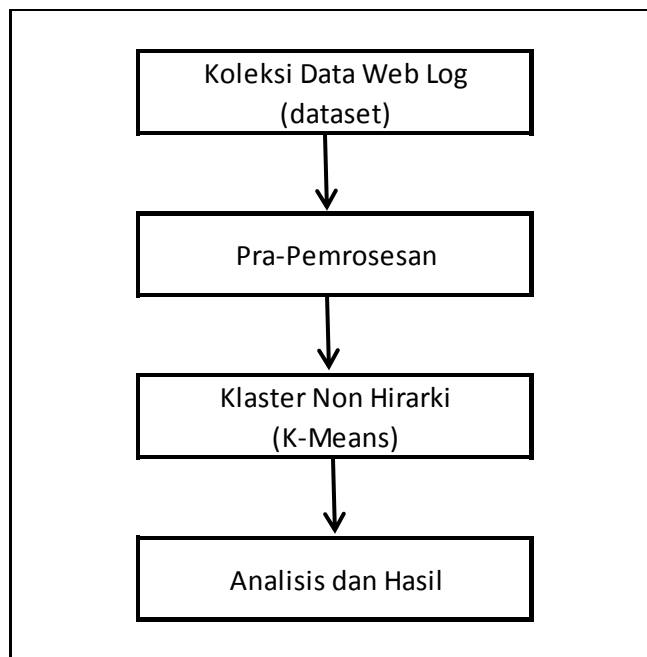
Klasterisasi merupakan salah satu topik penting pada WUM dalam segmentasi pengunjung *web* berdasarkan pola akses ataupun frekuensi kunjungan. Xie dan Phoha[31] menggunakan metode *belief function* dalam melakukan klasterisasi terhadap data *web log* pengunjung *web*. Mereka membagi pengunjung *web* dalam kelompok-kelompok yang berbeda dan menemukan pola akses yang umum untuk setiap anggota kelompok. Namun, pendekatan ini masih membutuhkan identifikasi sesi akses sehingga kurang efisien dari sisi pra-pemrosesan. Jin Hua Xu[32], melakukan klasterisasi pengunjung *web* dengan metode K-Means, namun belum disertai validasi dari hasil klasternya.

Salah satu permasalahan utama dalam klasterisasi adalah menentukan jumlah klaster yang optimal. Beberapa penelitian terdahulu menyebutkan bahwa penentuan jumlah klaster bersifat subjektif, sesuai kebutuhan peneliti, oleh karena itu dalam penelitian ini digunakan metode koefisien *silhouette* dalam menentukan jumlah klaster yang akan dibentuk guna menghindari subjektif tersebut. Sebuah klaster dikatakan baik jika anggota dalam satu klaster (*within cluster*) mempunyai tingkat kesamaan (*homogenitas*) yang tinggi dan mempunyai perbedaan (*heterogenitas*) yang tinggi dengan anggota klaster (*between cluster*) lainnya.

Menurut Chaofeng[30], klasterisasi pada sesi *web* mencakup tiga tahapan, yaitu pra-pemrosesan, pengukuran kesamaan (*similarity measure*) dan penerapan algoritma klaster. Pada penelitian ini dilakukan klasterisasi berdasarkan frekuensi kunjungan dari setiap pengunjung *web* dalam perioda waktu tertentu dengan mengabaikan sesi *web* sehingga lebih efisien dari sisi pra-pemrosesan dan dilakukan klasterisasi menggunakan metode K-Means yang disertai dengan validasi dari klaster yang terbentuk, sehingga hasil penelitian ini dapat bermanfaat dalam segmentasi pengunjung *web*. Penelitian ini adalah pengembangan lebih lanjut atas penelitian Jin Hua Xu[32] dan Chaofeng[33].

4.2 Tahapan Penelitian

Secara garis besar tahapan penelitian terlihat pada gambar 4.1.



Gambar 4.1 Tahapan Penelitian 1

4.3 Dataset

Dataset adalah sekumpulan data yang digunakan dalam penelitian, yaitu data *web log* dari *website* Institut Teknologi Sepuluh Nopember (ITS) Surabaya dengan alamat *web* www.its.ac.id. Perioda pengambilan data dari tanggal 15 hingga 16 juli 2012 dengan data mentah sejumlah 163.281.

4.4 Pra-Pemrosesan

Pra-Pemrosesan atau *pre-processing* adalah tahap membersihkan data dari item data yang tidak diperlukan sehingga data layak untuk dilakukan klusterisasi. Pembahasan tentang ini dapat dilihat pada sub bab 3.1.1.

Hasil akhir dari pra-pemrosesan ditampilkan dalam bentuk matriks vektor [32], lihat matriks vektor (4-1).

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ X_{N1} & X_{N2} & \cdots & X_{Nn} \end{bmatrix} \quad (4-1)$$

Dengan N merupakan jumlah pengunjung *web*, n merupakan jumlah halaman *web* dan X merupakan matriks vektor pengamatan.

4.5 Metode Kluster

Metode kluster yang digunakan adalah metode K-Means[31][35]. K-Means merupakan salah satu algoritma klusterisasi yang banyak digunakan di berbagai bidang karena mudah diimplementasikan, memiliki kemampuan untuk melakukan kluster pada data yang besar.

Algoritma untuk melakukan K-Means *clustering* adalah sebagai berikut :

- 1) Menentukan jumlah k , yaitu banyaknya kluster yang akan dibentuk. Hal ini juga bertujuan untuk mewakili centroid awal.
- 2) Mengalokasikan data ke dalam kluster secara random berdasarkan centroid terdekat.
- 3) Menghitung ulang posisi *centroid* k .
- 4) Mengulangi langkah 2 dan 3 hingga tidak ada lagi pemindahan objek antar kluster.

Proses pengelompokan data ke dalam suatu *cluster* dapat dilakukan dengan cara menghitung jarak terdekat dari suatu data ke sebuah titik *centroid* dengan metode *euclidean*.

4.6 Validasi Klaster

Setelah klaster terbentuk, maka dilakukan validitas menggunakan koefisien *silhouette*[35][36], lihat persamaan (4-2).

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (4-2)$$

a_i adalah rerata jarak objek dalam klaster, dan b_i adalah rerata jarak minimum untuk objek di klaster yang lain. Nilai koefisien *silhouette* berkisar antara -1 hingga 1. Nilai negatif menandakan rerata jarak antar objek jauh.

4.7 Analisis dan Hasil

Setelah dilakukan pra-pemrosesan terhadap 163.281 dataset, diperoleh 3.733 data pengunjung *web* dengan 57 variabel (halaman *web* yang diakses). Perhatikan Tabel 4.1, memperlihatkan perbandingan jumlah data sebelum dan sesudah dilakukan pra-pemrosesan.

Tabel 4.1 Perbandingan Jumlah data

	Data Mentah	Data Final	% Data Tereduksi
Total Record	163,281	3,733	97,71
Jumlah Web page/var	6,753	63	98,56
Jumlah user unik	5,468	165	97,56

Berdasarkan Tabel 4.1, terlihat hanya 2,29% data yang dapat diolah lebih lanjut, artinya terdapat 97,71% data tidak digunakan. Hal ini menandakan pentingnya peran pra-pemrosesan dalam penambangan data web log.

Kemudian data di olah berdasarkan matriks vektor (4-1). Hasilnya terlihat pada Tabel 4.2. Dengan $p1, p2, p3, pn$ merupakan variabel untuk halaman *web*, misalnya $p1$ adalah halaman *web* dengan nama *index.php*, $p2$ adalah halaman *web* dengan nama *academi.php*, dan seterusnya. $u1, u2, u3, uN$ merupakan variabel untuk identifikasi pengunjung *web*, misalnya $u1$ adalah pengunjung *web* dengan alamat IP, 66.249.69.xxx, penamaan variabel ini dapat dilihat pada Tabel 4.3.

Tabel 4.2 Matriks Vektor

User	Halaman Web											
	<i>p1</i>	<i>p2</i>	<i>p3</i>	<i>p4</i>	<i>p5</i>	<i>p6</i>	<i>p7</i>	<i>p8</i>	<i>p9</i>	<i>p10</i>	...	<i>pn</i>
<i>u1</i>	6	9	0	0	0	0	0	0	5	20	...	<i>u1n</i>
<i>u2</i>	0	0	0	35	0	35	0	0	0	0	...	<i>u2n</i>
<i>u3</i>	0	11	0	0	0	0	0	0	14	9	...	<i>u3n</i>
<i>u4</i>	0	1	4	3	2	3	0	4	4	1	...	<i>u4n</i>
<i>u5</i>	0	84	0	0	0	0	0	0	0	0	...	<i>u5n</i>
<i>u6</i>	5	5	5	0	1	5	4	2	6	0	...	<i>u6n</i>
<i>u7</i>	1	37	0	0	3	0	0	1	9	1	...	<i>u7n</i>
<i>u8</i>	2	21	3	0	4	0	2	1	7	0	...	<i>u8n</i>
:	:	:	:	:	:	:	:	:	:	:	:	:
<i>uN</i>	<i>uNp1</i>	<i>uNp2</i>	<i>uNp3</i>	<i>uNp4</i>	<i>uNp5</i>	<i>uNp6</i>	<i>uNp7</i>	<i>uNp8</i>	<i>uNp9</i>	<i>uNp10</i>	...	<i>uNpn</i>

Dari Tabel 4.2 dapat dinyatakan bahwa pengunjung dengan variabel *u1* telah mengakses halaman web *p1* sebanyak 6 kali, halaman web *p2* sebanyak 9 kali dan seterusnya.

Tabel 4.3 Penamaan Variabel User dan Halaman Web

Var	Halaman Web	Var	Host/IP Address
<i>p1</i>	GET /semuaberita.php	<i>u1</i>	72.233.72.xxx
<i>p2</i>	GET /personal/homebase.php	<i>u2</i>	208.86.227.xxx
<i>p3</i>	GET /en/academic.php	<i>u3</i>	157.56.94.xxx
<i>p4</i>	GET /en/index.php	<i>u4</i>	86.145.228.xxx
<i>p5</i>	GET /en/about.html	<i>u5</i>	10.170.17.xxx
<i>p6</i>	GET /en/services.php	<i>u6</i>	66.249.69.xxx
<i>p7</i>	GET /index.php	<i>u7</i>	202.46.129.xxx
<i>p8</i>	GET /en/admission.php	<i>u8</i>	202.46.129.xxx
..
<i>p63</i>	GET /en/tekkim.php	<i>u165</i>	222.124.169.xxx

Dari hasil pra-pemrosesan, ternyata perbandingan antara data dengan variabel tidak seimbang, sehingga dilakukan analisis faktor untuk mereduksi jumlah variabel tersebut. Proses ekstraksi data pada analisis faktor digunakan metode *principal component*. Proses ini berhasil mereduksi 57 variabel menjadi

14 faktor. 14 faktor ini kemudian diberi penamaan variabel dengan $f1$ untuk faktor 1, $f2$ untuk faktor ke-2, $f3$ untuk faktor 3 dan seterusnya. Dengan persentase keterwakilan variabel dalam faktor terlihat pada Tabel 4.4. Hasil akhir dari tahap ini diperoleh skor faktor.

Tabel 4.4 Persentase Keterwakilan Variabel dalam Faktor

$f1$	$f2$	$f3$	$f4$	$f10$...	$f14$
$p10(67.8\%)$	$p40(81\%)$	$p4(89.7\%)$	$p19(79.3\%)$	$p23(74\%)$...	$p2(71,8\%)$
$p12(88.4\%)$	$p42(76.9\%)$	$p6(96.8\%)$	$p31(83.7\%)$	$p29(53.7\%)$...	$p9(55,8\%)$
$p15(87.8\%)$	$p44(73.6\%)$	$p27(98.4\%)$	$p38(87.7\%)$	$p41(71.6\%)$...	
$p16(78.2\%)$	$p55(71.8\%)$	$p28(98.6\%)$	$p57(92.2\%)$	$p43(77.1\%)$	
$p17(92.7\%)$	$p59(75\%)$					
$p20(82.8\%)$	$p60(81.1\%)$					
$p21(87.7\%)$	$p62(66.4)$					
$p22(90.4\%)$						
$p24(92.4\%)$						
$p25(94.8\%)$						
$p49(81.1\%)$						

Tabel 4.4 menunjukkan bahwa halaman web yang berada pada $f1$ adalah halaman web $p10$, $p12$, $p15$, $p16$, $p17$, $p20$, $p21$, $p22$, $p24,25$ dan $p49$ dan persentase keterwakilan variabel dalam faktor dinyatakan disebelah kanan variabel tersebut. Hal yang sama juga dapat diketahui untuk $f2$, $f3$ dan seterusnya.

Tahap berikutnya dilakukan klasterisasi menggunakan metode K-Means dan melakukan validitas menggunakan koefisien *silhouette* dan *Sum of Squared Error* (SSE) (lihat rumus 2-8), dengan hasil seperti terlihat pada Tabel 4.5.

Tabel 4.5 Validitas Dengan Koefisien *Silhouette*

Jumlah Klaster	Rata-Rata Koefisien Silhouette	Nilai SSE
2	0.2882	2,187
3	0.3382	2,012
4	0.3630	1,890
5	0.4180	1.778
6	0.4499	1,659

Terlihat dari hasil validitas pada Tabel 4.5, jumlah klaster yang baik adalah 6 klaster, karena nilai rata-rata koefisien *silhouettenya* lebih tinggi dibandingkan dengan yang lain, semakin mendekati angka 1 semakin baik. Namun berbeda dengan validitas dengan SSE, semakin rendah nilainya berarti tingkat kemiripan antar anggota tinggi, dan kedua alat validitas ini menyatakan bahwa jumlah 6 klaster adalah yang terbaik.

Berdasarkan hasil validitas tersebut, diterapkan algoritma K-Means dengan 6 klaster. Tabel 4.6 dan 4.7 menampilkan hasil klaster.

Tabel 4.6 Jumlah Keanggotaan Klaster

Klaster	Jumlah Anggota
1	2
2	1
3	143
4	13
5	3
6	3
Valid data	165

Berdasarkan dari Tabel 4.6, dapat dinyatakan bahwa dari 165 pengunjung *web*, klaster 1 mempunyai 2 anggota, klaster 2 mempunyai 1 anggota, klaster 3 mempunyai 143 anggota, klaster 4 mempunyai 13 anggota, klaster 5 dan 6 masing-masingnya mempunyai 3 anggota, dengan pusat klaster akhir seperti pada Tabel 4.7.

4.8 Segmentasi Pengunjung Web

Dari Tabel 4.7 dapat dinyatakan bahwa setiap klaster mempunyai karakteristik tertentu dibandingkan dengan klaster yang lain. Hal ini tergambar dari nilai pusat klaster akhir pada setiap variabel/faktor, tanda positif (+, tidak ditampilkan secara langsung) menandakan bahwa bernilai di atas rata-rata dan tanda negatif (-) sebaliknya. Perhatikan nilai pada *f1*, bernilai positif pada klaster 1 (7.8920) dan bernilai negatif pada klaster lainnya, hal ini berarti bahwa halaman

web yang terdapat dalam *f1* (faktor 1) lebih banyak dikunjungi oleh anggota klaster 1 dibandingkan dengan anggota klaster lainnya. Begitu juga halnya dengan halaman *web* yang terdapat dalam *f2* (faktor 2) lebih banyak diakses oleh anggota klaster 3 dibandingkan dengan klaster lainnya. Jika dilihat berdasarkan klasternya, maka klaster 1 adalah pengunjung yang dominan mengakses halaman *web* yang berada dalam *f1*(7.8920) dan *f14*(1.0855), klaster 2 terdiri atas pengunjung yang dominan mengakses halaman *web* yang berada dalam *f3*(12.5911), dan seterusnya

Tabel 4.7 Pusat Klaster Akhir

Var	Klaster					
	1	2	3	4	5	6
<i>f1</i>	7.8920	-0.1505	-0.0995	-0.0432	-0.1427	-0.1389
<i>f2</i>	-0.1438	-0.3583	0.0373	-0.1643	-0.7247	-0.1245
<i>f3</i>	0.0392	12.5911	-0.0831	-0.0401	-0.1606	0.0741
<i>f4</i>	-0.0205	-0.1989	-0.1335	-0.0335	0.1303	6.4598
<i>f5</i>	0.0268	-0.3946	-0.0478	0.6646	-0.4289	-0.0569
<i>f6</i>	0.0098	-1.0626	-0.0094	0.1741	-0.6088	0.6489
<i>f7</i>	-0.0848	-0.0775	-0.1525	1.8302	-0.4769	-0.1047
<i>f8</i>	0.1868	-0.6947	-0.0027	-0.0751	0.6826	-0.1194
<i>f9</i>	-0.1958	-0.1536	-0.0368	0.3420	0.2661	0.1870
<i>f10</i>	-0.0712	-0.0759	0.0155	-0.1154	-0.2552	0.0903
<i>f11</i>	-0.3702	-0.0115	-0.1570	1.9501	-0.2282	-0.4895
<i>f12</i>	-0.0822	-0.0095	-0.1332	0.1811	5.6028	0.0211
<i>f13</i>	0.2036	0.0454	-0.0408	0.4048	0.1472	-0.1077
<i>f14</i>	1.0855	0.1113	-0.0415	0.2392	0.0181	0.1608

Interpretasi

Dari hasil klaster yang terbentuk, maka pihak pengelola website dapat lebih memperhatikan *user* yang berada pada klaster 3, yaitu kelompok *user* dengan keanggotaan terbanyak, dan berdasarkan dari pusat klaster akhir, anggota klaster 3 ini dominan mengakses halaman web yang terdapat pada *f2*, yaitu halaman web *p40*, *p42*, *p44*, *p55*, *p59*, *p60*, *p62* dan *f10* yaitu halaman web *p23*,

p29, p41, p43. Perhatikan Tabel 4.5. Tabel 4.8 menjelaskan halaman web yang diakses.

Tabel 4.8 Halaman Web yang diakses oleh Anggota Kluster 3

<i>f</i>2	Halaman Web
<i>p40(81%)</i>	<i>GET /en/administrasi.php</i>
<i>p42(76.9%)</i>	<i>GET /en/baak.html</i>
<i>p44(73.6%)</i>	<i>GET /en/scs.php</i>
<i>p55(71.8%)</i>	<i>GET /en/hotspot.php</i>
<i>p59(75%)</i>	<i>GET /en/cultural.php</i>
<i>p60(81.1%)</i>	<i>GET /en/facilities.html</i>
<i>p62(66.4)</i>	<i>GET /en/staff.html</i>
<i>f</i>10	Halaman Web
<i>p23(74%)</i>	<i>GET /personal/index.php</i>
<i>p29(53.7%)</i>	<i>GET /personal/priv/index.php</i>
<i>p41(71.6%)</i>	<i>GET /personal/priv/publikasi.add.php</i>
<i>p43(77.1%)</i>	<i>GET /personal/priv/publikasi.edit.php</i>

Tabel 4.8 menjelaskan bahwa halaman web *p40* adalah halaman web *administrasi.php* yang berisi informasi tentang layanan administrasi dengan tingkat keterwakilannya dalam faktor sebesar 81%. Hal yang sama untuk halaman web *p42, p44, p55* dan seterusnya.

4.9. Kesimpulan

Dari hasil pra-pemrosesan terhadap data *web log* (www.its.ac.id), hanya 2,29% data yang dapat digunakan untuk diproses lebih lanjut, sedangkan 97,71% tereliminasi pada tahap ini. Dari sisi halaman web, tereliminasi sebesar 98,85%, dan dari sisi pengunjung unik tereliminasi sebesar 97,56%. Hal ini membuktikan pentingnya pra pemrosesan dalam menyiapkan data yang digunakan sekaligus menentukan kualitas akhir dari proses penambangan data *web log*.

Dari penerapan algoritma K-Means terhadap data *web log* www.its.ac.id yang berbentuk *frequently access* serta menggunakan validasi dengan koefisien

silhouette dan *sum of squared error* (SSE) dapat disimpulkan bahwa metode ini mampu memberikan informasi baru bagi pengelola *web* yaitu dari 165 pengunjung *web*, 143 pengunjung berada pada klaster yang sama (klaster 3), hal ini berarti, bahwa 143 pengunjung ini memiliki frekuensi kunjungan yang hampir sama (*homogenitas*) dan halaman-halaman *web* yang mereka akses harus menjadi perhatian bagi pengelola *web* untuk meningkatkan layanan pada web, personalisasi web ataupun modifikasi web.

Disamping itu, halaman web tentang administrasi (*p40*) dan fasilitas kampus (*p60*) merupakan halaman web yang dominan diakses. Tabel 4.9 memberikan penjelasan lebih detail terhadap halaman web yang dikunjungi oleh anggota klaster 3 serta persentasenya. Dapat disimpulkan bahwa anggota klaster 3 ini ingin mendapat informasi tentang layanan administrasi, fasilitas kampus, karya ilmiah (publikasi), informasi BAAK, dan seterusnya.

Halaman ini sengaja dikosongkan

BAB 5

SINGLE-STAGE CLUSTERING PADA DATA *WEB LOG* YANG BERBENTUK *USER ACCESS PATTERN* UNTUK SEGMENTASI PENGUNJUNG WEB

User Access Pattern merupakan pola akses pengunjung website berdasarkan urutan aksesnya terhadap halaman web. Misalnya seorang *user*, sewaktu mengakses sebuah website pertama kali mengakses halaman web A, kemudian berpindah ke halaman web C, terus berpindah ke halaman web B dan seterusnya. Pola-pola urutan (*sequence*) seperti inilah yang disebut dengan *user access pattern* yang dapat menggambarkan pola perilakunya.

Teknik *sequence clustering* telah banyak digunakan pada bidang ilmu bio-informatika dalam menambang data pada struktur DNA/protein, namun saat ini, di bidang penambangan web juga telah menjadi topik penelitian yang aktif. Pada penelitian ini diterapkan algoritma *sequence clustering* terhadap data *web log* ITS dengan tujuan untuk mendapatkan segmentasi pengunjung *web* berdasarkan pola perilaku kunjungan (*user access pattern*).

Tahapan penelitian diawali dengan pra-pemrosesan hingga terbentuk database *web log* dalam bentuk data *sequence (user sequence)* kemudian diterapkan algoritma *sequence clustering*, yaitu kombinasi algoritma *markov model* dengan algoritma *expectation-maximization*.

Dari hasil uji coba diperoleh 12 klaster, dengan 6 klaster memiliki tingkat kesamaan (pola perilaku kunjungan) yang tinggi. Pola perilaku dari pengunjung *web* yang berada pada 6 klaster tersebut menjadi masukan bagi pengelola *web* untuk keperluan segmentasi, prediksi, personalisasi dan ataupun modifikasi *web* ITS.

5.1 Pendahuluan

Klasterisasi telah menjadi topik riset yang semakin penting saat ini karena manfaatnya yang sangat besar dalam mengelompokkan data, sehingga data yang berada dalam kelompok yang sama akan memiliki tingkat kesamaan (*homogenitas*) yang tinggi dan memiliki perbedaan (*heterogenitas*) yang tinggi dengan data pada kelompok lainnya[37].

Teknik klasterisasi telah digunakan dalam berbagai disiplin ilmu, salah satunya dalam bidang penambangan web. Teknik klasterisasi pada penambangan web dilakukan dengan menambang data transaksi (*click stream*) seorang pengunjung *web*. Data transaksi ini disebut dengan data *web log*. Urutan-urutan klik yang dilakukan pengunjung pada sebuah *web*, menjadi kumpulan data berurut (*data sequence*) yang dapat ditambang untuk keperluan segmentasi pengunjung *web* ataupun untuk prediksi berdasarkan pola perilaku kunjungan (*user access pattern*) tersebut.

Penelitian terkait klasterisasi untuk data berurut ini telah banyak dilakukan pada disiplin ilmu bio-informatika [38] [39] [40] [41], penambangan web [42] [43] [33], analisa perilaku konsumen [44] [45] [46] dan lainnya.

Fokus dalam penelitian ini adalah penerapan algoritma *sequence clustering* [47] untuk segmentasi pengunjung *web* berdasarkan pola perilaku pengunjung dengan studi kasus pada *website* ITS versi bahasa Inggris dengan alamat www.its.ac.id/en atau www.its.ac.id/index/en

5.2 Tinjauan Pustaka

Sequence clustering pada awalnya merupakan topik riset yang aktif pada bidang bio-informatika, namun prinsip-prinsip kerjanya dapat digunakan pada bidang lainnya yang memiliki data berurut (*data sequence*), seperti transaksi pengunjung (*clickstream*) pada *web* [47].

Secara umum, *sequence clustering* dirancang untuk menganalisis populasi yang berasal dari data/kejadian yang berurutan dalam kelompok kasus yang sama atau tidak. Dengan kata lain, *sequence clustering* adalah sekumpulan metode yang

bertujuan untuk mempartisi sejumlah urutan ke dalam kelompok yang lebih bermakna atau kelompok dengan urutan yang sama [48].

Algoritma *sequence clustering* dapat didasarkan kepada algoritma *markov model* orde pertama [46] [47]. Dalam algoritma ini, kondisi saat ini hanya bergantung pada keadaan sebelumnya. Misalnya $z = \{z_0, z_1, z_2, z_3, \dots, z_{n-1}\}$, panjang n dapat dinyatakan dengan persamaan 5-1.

$$p_k(z|c_k) = p_k(z_0, c_k) \cdot \prod_{i=1}^{i=n-1} p(z_i|z_{i-1}, c_k) \quad (5-1)$$

Dengan $p(z_0, c_k)$ adalah probabilitas dari z_0 sebagai keadaan (*state*) pertama dalam markov model yang berhubungan dengan c_k dan $p(z_i|z_{i-1}, c_k)$ adalah probabilitas transisi dari keadaan z_{i-1} ke keadaan z_i dalam markov model yang sama.

Dalam mengimplementasikan *markov model* harus memenuhi persyaratan :

- a) Jumlah probabilitas transisi untuk suatu keadaan awal dari sistem sama dengan 1.
- b) Probabilitas-probabilitas tersebut berlaku untuk semua partisipan dalam sistem.
- c) Sistem harus berkarakter *lack of memory*, di mana kondisi sistem di masa mendatang tidak dipengaruhi (*independent*) oleh kondisi sebelumnya. Artinya kondisi sistem saat evaluasi tidak dipengaruhi oleh kondisi sebelumnya, kecuali kondisi sesaat sebelum kondisi saat ini.
- d) Sistem harus *stationery* atau *homogen*, artinya perilaku sistem selalu sama di sepanjang waktu atau peluang transisi sistem dari satu kondisi ke kondisi lainnya akan selalu sama di sepanjang waktu.

Algoritma markov model ini kemudian dikombinasikan dengan algoritma *Expectation-Maximization* (EM) [49][50][51] untuk proses *sequence clustering* pada penelitian ini.

Algoritma EM adalah salah satu algoritma klasterisasi yang berdasarkan pada model (*model-based clustering*) dengan metode iteratif. Tujuannya adalah

untuk menemukan nilai estimasi *maximum likelihood* (ML) dari parameter dalam sebuah model probabilistik, dan sebuah model juga tergantung kepada variabel laten yang belum diketahui. Variabel laten adalah variabel yang tidak terobservasi (*missing data*, *hidden variable*, *missing measurement*) secara langsung.

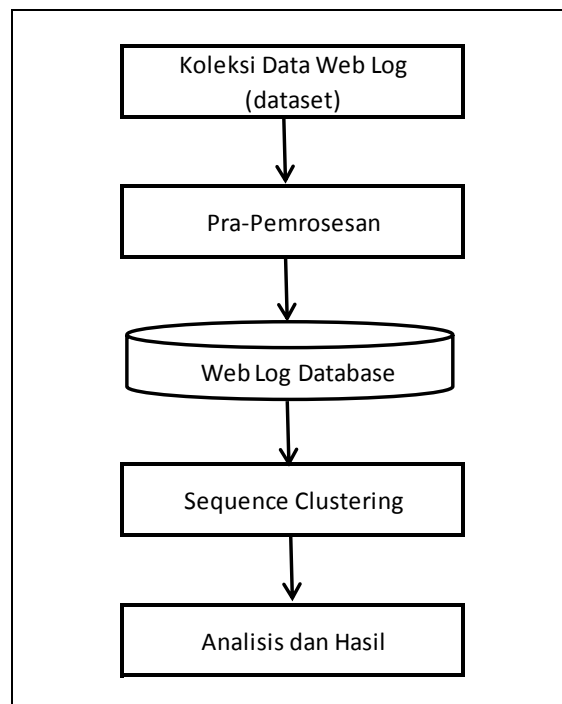
Pada algoritma ini, dalam setiap iterasinya terdiri dari dua tahap, yaitu :

- a) *Expectation-Step*, pada tahap ini dihitung nilai probabilitas bersyarat dari fungsi *loglikelihood* (kemiripan) data lengkap menggunakan estimasi parameter.
- b) *Maximization-Step*, pada tahap ini dihitung parameter yang memaksimalkan nilai probabilitas dari fungsi *loglikelihood* yang di peroleh pada *expectation-step*.

Kedua tahap tersebut dilakukan berulang-ulang sampai hipotesis dari *converge* (terpusat) mencapai nilai yang *stationer*.

5. 3. Tahapan Penelitian

Tahapan penelitian pada bab ini terlihat pada Gambar 5.1.



Gambar 5.1 Tahapan Penelitian 2

5.4 Dataset

Dataset yang digunakan dalam penelitian ini adalah data *web log* dari *website* Institut Teknologi Sepuluh Nopember (ITS) surabaya, khususnya untuk halaman *web* berbahasa Inggris dengan alamat *web* www.its.ac.id/index/en. Periode pengambilan data dari tanggal 3 hingga 16 Juli 2012 sejumlah 1.725 data mentah. Atribut dari *web* yang digunakan dalam penelitian ini adalah kategori academic, admission, research, library, services, news, map dan home (index).

5.5 Pra-Pemrosesan

Pra-Pemrosesan adalah tahap menyiapkan data *web log* agar layak dilakukan proses penambangan data. Pada tahap ini dilakukan proses membersihkan data *web log* dari item data yang tidak diperlukan. Pembahasan tentang hal ini dapat dilihat pada sub bab 3.1.1.

Hasil dari tahap ini pra-pemrosesan ini dilanjutkan dengan :

- Membuat *user sequence* yaitu data dikonversi menjadi data *sequence* berdasarkan perilaku *user* (pengunjung *web*). Perhatikan Tabel 5.1.
- Konversi ke database. Data hasil *user sequence* dikonversi ke database dalam bentuk 2 tabel. Tabel pertama (a) berisi profil user dan tabel kedua (b) berisi halaman *web* yang diakses serta urutan aksesnya. Perhatikan Tabel 5.2.

Tabel 5.1 Contoh Pola Perilaku Pengunjung

User	Urutan akses					
	1	2	3	4	5	6
u_1	academic	admission	academic			
u_2	news	services	library	admission	academic	services
u_3	news	services	map	library	admission	
u_4	academic	admission	news			
u_5	academic	admission	library			
u_6	admission	academic	services	admission	academic	
u_7	admission	academic	library	services		
u_8	academic	news	services	library		
u_n	academic	admission	library	services	News	map

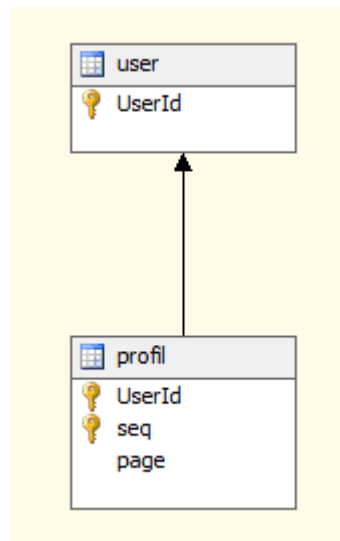
Tabel 5.2 Tabel dari Dataset

userID	userID	Page	Seq
<i>u1</i>	<i>u1</i>	index.php	1
<i>u2</i>	<i>u2</i>	academic.php	1
<i>u3</i>	<i>u2</i>	admission.php	2
<i>u4</i>	<i>u2</i>	academic.php	3
<i>u5</i>	<i>u2</i>	academic.php	4
<i>u6</i>	<i>u3</i>	news.php	1
<i>u7</i>	<i>u3</i>	services.php	2
<i>u8</i>	<i>u3</i>	library.php	3
...
<i>u108</i>	<i>u108</i>	academic.php	5
a	b		

Untuk memudahkan dalam proses analisis, dilakukan penamaan variabel untuk kategori halaman web yang di akses dengan *p1*, *p2*, *p3* dan seterusnya, dengan ketentuan seperti terlihat pada tabel 5.3, dan kemudian dikonversi menjadi database dengan struktur seperti terlihat pada Gambar 5.2.

Tabel 5.3 Penamaan Variabel Kategori Halaman Web

Variabel	Web Kategori
<i>p1</i>	Index (home)
<i>p2</i>	Academic
<i>p3</i>	Admission
<i>p4</i>	Research
<i>p5</i>	Library
<i>p6</i>	Services
<i>p7</i>	News
<i>p8</i>	Map



Gambar 5.2 Struktur Database

Tabel *user* berisi satu *field* yaitu *userid* yang berguna untuk menampung data *user* (pengunjung web), sedangkan tabel *profil* berisi 3 *field* yang menggambarkan pola urutan dari pengunjung. *Field UserId* digunakan untuk merelasikan dengan tabel *user*, sedangkan *field seq* berisi nomor urutan akses dan *field page* berisi halaman web yang diakses. Kedua tabel ini direlasikan secara *one to many*.

5.6 Penerapan Algoritma

Algoritma *sequence clustering* yang digunakan adalah algoritma markov model dan *Expectation-Maximization* (EM) [50][51] yang kemudian dikembangkan oleh Igor Cades [49] dan algoritma ini juga telah diimplementasikan pada Microsoft SQL server sebagai algoritma *sequence clustering*.

Konsep kerja algoritma ini dapat digambarkan sebagai berikut :

- 1) Inialisasi model parameter $p(x_0, c_k)$ dan $p(x_i|x_{i-1}, c_k)$ secara acak sesuai dengan algoritma markov model.
- 2) Menetapkan urutan untuk setiap klaster yang mempunyai probabilitas yang lebih tinggi berdasarkan persamaan 1.

- 3) Gunakan hasil langkah 2 untuk melakukan estimasi ulang setiap model parameter. yaitu menghitung ulang keadaan probabilitas transisi dari setiap markov model berdasarkan urutan-urutan yang dimiliki oleh klaster tersebut.
- 4) Ulangi langkah 2 dan 3 hingga tidak ada perubahan.

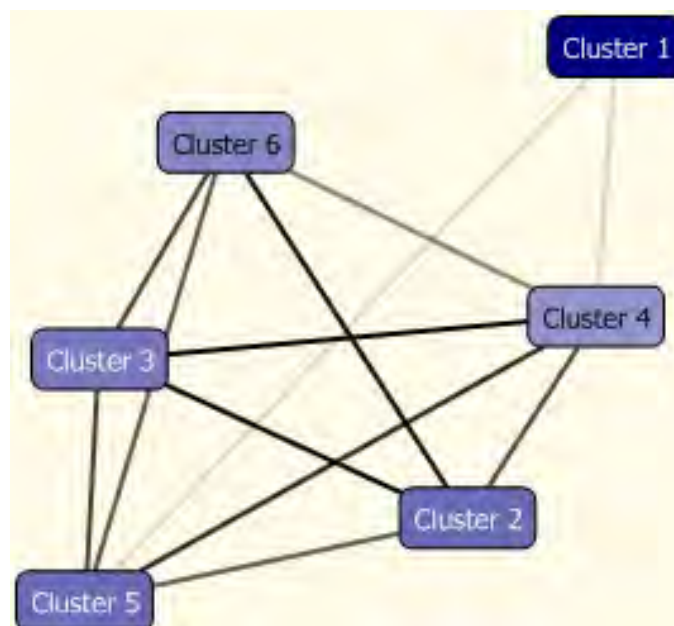
5.7 Analisis dan Hasil

Setelah dilakukan pra-pemrosesan terhadap 1.725 *user* sebagai dataset, diperoleh 108 data yang siap untuk diolah lebih lanjut. Perhatikan Tabel 5.4, perbandingan jumlah data sebelum dan sesudah dilakukan pra-pemrosesan.

Tabel 5.4 Perbandingan Jumlah data

	Sebelum pra-pemrosesan	Setelah pra-pemrosesan
Jumlah user	1.725	108
Persentase	100%	6%

Terlihat dari Tabel 5.4, hanya 6% data yang dapat digunakan lebih lanjut dalam proses penambahan data. Kemudian terhadap data 6% ini diterapkan algoritma *sequence clustering*, dengan hasil seperti terlihat pada Gambar 5.3.



Gambar 5.3 Diagram Klaster

Dari Gambar 5.3 diperoleh informasi bahwa algoritma ini otomatis membentuk 6 klaster. Ketebalan garis penghubung antar klaster pada gambar tersebut merepresentasikan tingkat kemiripannya, seperti klaster 4 dengan klaster 3, klaster 6 dengan klaster 2 dan seterusnya. Sebaliknya, semakin tidak jelas garis penghubung antar klaster menunjukkan semakin berkurangnya tingkat kemiripannya, seperti antara klaster 1 dengan klaster 5, klaster 1 dengan klaster 6 dan seterusnya.

Namun, jika dilihat dari sisi jumlah populasi, klaster 1 memiliki populasi terbanyak hal ini ditandai dengan warna latar belakangnya yang gelap. Populasi untuk setiap klaster terlihat pada Tabel 5.5.

Tabel 5.5 Populasi Klaster

Klaster	Jumlah
1	19
2	11
3	10
4	08
5	11
6	9

5.8 Profil Klaster untuk Segmentasi Pengunjung *Web*

Berdasarkan hasil dari klasterisasi didapatkan pola perilaku setiap klaster untuk keperluan segmentasi pengunjung web. Perhatikan contoh profil *user* pada Tabel 5.6, yang memperlihatkan pola kunjungan *user* pada setiap klasternya. Perhatikan Tabel 5.3 untuk melihat makna variabel *p1*, *p2*, *p3* dan seterusnya.

Tabel 5.6 Contoh Profil *User*

Klaster 2	Klaster 5
<p>p2 p3 p6 p7 p2 p3 p2</p> <hr/> <p>p7 p2 p3 p5 p6 p7 p8</p> <hr/> <p>p3 p2 p5 p6 p1</p> <hr/> <p>p3 p2 p1 p7 p1</p> <hr/> <p>p2 p3 p5 p6 p7 p8</p> <hr/> <p>p3 p2 p5 p6 p1</p> <hr/> <p>p3 p2 p3 p5 p6 p7 p2 p2</p> <hr/> <p>p2 p3 p5 p6 p7 p8</p> <hr/> <p>p2 p3 p6 p7 p2 p3 p2</p> <hr/> <p>p7 p8 p2 p7 p8 p8</p> <hr/> <p>p3 p2 p5 p6 p1</p>	<p>p2 p3 p6 p5 p7 p8 p1</p> <hr/> <p>p2 p7 p3 p6 p5</p> <hr/> <p>p2 p1 p5 p2 p1</p> <hr/> <p>p2 p3 p6 p2 p1</p> <hr/> <p>p7 p6 p8 p5 p3 p2 p6 p1 p8 p7 p6</p> <hr/> <p>p2 p1 p5 p2 p1</p> <hr/> <p>p6 p2 p7 p6 p5 p6 p5 p1</p> <hr/> <p>p7 p6 p8 p5 p3 p2 p6 p1 p8 p7 p6</p> <hr/> <p>p2 p3 p6 p5 p7 p8 p1</p> <hr/> <p>p2 p7 p6 p5 p1 p1</p> <hr/> <p>p2 p1 p5 p2 p1</p>
Klaster 3	Klaster 4
<p>p3 p1 p3 p3 p2 p3 p2 p3 p1 p2 p3</p> <hr/> <p>p1 p2 p3 p5 p6 p7 p8 p7 p6 p5 p2 p1 p2 p3</p> <hr/> <p>p2 p3 p2 p1 p1 p2 p3</p> <hr/> <p>p2 p3 p5 p6 p7 p8 p7 p8 p7 p6 p5 p1 p2 p1</p> <hr/> <p>p2 p2 p1 p2 p3 p5 p7 p8</p> <hr/> <p>p6 p5 p3 p2 p8 p7</p> <hr/> <p>p1 p3 p5 p7 p8 p6 p5 p1 p2 p3 p5 p6 p7 p8</p> <hr/> <p>p2 p2 p1 p2 p3 p5 p7 p8</p> <hr/> <p>p2 p3 p2 p1 p1 p2 p3</p> <hr/> <p>p6 p5 p3 p2 p8 p7</p>	<p>p3 p8 p2 p5 p2 p2 p3 p5 p8</p> <hr/> <p>p2 p3 p5 p6 p7 p7 p1 p1 p1 p1 p2 p1 p1</p> <hr/> <p>p3 p5 p6 p7 p8 p1 p2</p> <hr/> <p>p2 p3 p1 p1 p3 p7 p1 p2 p3 p5 p6 p7 p8 p1 p2</p> <hr/> <p>p8 p1 p2 p2 p5 p2 p1 p2</p> <hr/> <p>p3 p5 p6 p7 p8 p1 p2</p> <hr/> <p>p1 p3 p5 p6 p8 p2 p5 p8 p1 p2 p5 p6 p7</p> <hr/> <p>p8 p1 p2 p2 p5 p2 p1 p2</p>
Klaster 1	
<p>p3 p1 p6 p1 p1</p> <hr/> <p>p2 p2 p1 p3 p6 p1 p2 p1</p> <hr/> <p>p2 p7 p6 p1 p2 p1 p2 p1</p> <hr/> <p>p8 p1 p8 p2 p5 p1</p> <hr/> <p>p2 p3 p5 p7 p6 p8</p> <hr/> <p>p8 p8 p8 p8 p8 p8 p8 p8 p8</p> <hr/> <p>p2 p2 p3 p1 p8 p6 p3 p1 p2 p1 p2 p7 p5 p3 p3 p2 p1 p2 p2</p> <hr/> <p>p2 p3 p5 p7 p6 p8</p> <hr/> <p>p2 p1 p3 p2 p3 p3 p2 p2 p2 p1 p2 p2 p3 p2</p> <hr/> <p>...</p>	

Karakteristik perilaku pengunjung berdasarkan kluster secara umum dapat terlihat pada Tabel 5.7.

Tabel 5.7 Tingkat *Probability* Akses

Variables	Values	Probability
Page.Transitions	[Start] -> p2	
Page	p2	
Page	p1	
Page.Transitions	[Start] -> p3	
Page	p6	
Page	p3	
Page.Transitions	[Start] -> p1	
Page.Transitions	[Start] -> p7	
Page	p7	
Page	p5	
Page.Transitions	p6, p1	
Page	p8	

Berdasarkan Tabel 5.7, terlihat bahwa *probability* halaman web *p2*, *p1*, *p3* dan *p6* lebih dominan diakses oleh para pengunjung web.

5.9 Kesimpulan

Dari 1.725 dataset, hanya 108 data yang siap untuk diolah lebih lanjut, artinya terdapat 94% data yang tereliminasi pada tahap pra-pemrosesan. Hal ini membuktikan pentingnya proses pra-pemrosesan dalam upaya mereduksi data dan meningkatkan *output* yang dihasilkan dari penambangan data tersebut.

Dengan penerapan algoritma *sequence clustering* terhadap data *web log* ITS ini diperoleh kesimpulan bahwa metode ini dapat memberikan *input* kepada pengelola *web* dalam memahami pola perilaku pengunjung *web* untuk keperluan segmentasi pengunjung web, hal ini terlihat dari 6 kluster yang terbentuk, kluster 4 dengan kluster 3, kluster 2 dengan kluster 6 mempunyai tingkat kemiripan dalam pola perilaku kunjungan (*clickstream*) serta kluster 1 mempunyai jumlah populasi yang paling besar, yaitu sebesar 28%.

Dari karakteristik pola perilaku pengunjung web dari setiap klasternya, terlihat halaman web *p2*(*academic.php*) menjadi halaman yang sering dikunjungi, hal ini menandakan bahwa pengunjung web lebih dominan ingin mendapat

informasi tentang akademik, kemudian halaman web *p3* (*admission.php*) dan *p6* (*services.php*).

Dari segi urutan pola kunjungannya, terlihat *probability* urutan kunjungan, mulai dari halaman web *p2(academic)*, kemudian beralih ke halaman web *p1* (*index/home*), ke halaman web *p3 (Admission)* dan ke halaman web *p6 (Services)*.

Berdasarkan hasil tersebut, maka pola kunjungan dari klaster yang mempunyai tingkat kemiripan tinggi, populasi yang besar serta pola kunjungan dan halaman web yang sering diakses menjadi perhatian khusus bagi pengelola web, terutama untuk keperluan segmentasi pengunjung web.

Dalam penerapan algoritma ini, disarankan agar sangat memperhatikan pada tahap pra-pemrosesan, karena tahap ini sangat menentukan hasil dari proses penambangan data.

BAB 6

TWO-STAGES CLUSTERING UNTUK SEGMENTASI PENGUNJUNG WEB

Klasterisasi merupakan salah satu bagian penting dalam *web usage mining* untuk keperluan segmentasi pengunjung. *Web Usage Mining* (WUM) berhubungan dengan ekstraksi *knowledge* dari data *web log*. Data *web log* memiliki banyak item data yang tidak relevan untuk dilakukan proses WUM, sehingga perlu dilakukan tahapan-tahapan untuk membersihkan data agar hasil penambangan data memberikan *output* yang baik, dalam hal ini untuk segmentasi pengunjung web. Hasil akhir dari WUM sangat tergantung kepada kualitas data input yang digunakan. Oleh karena itu, dalam penelitian ini diajukan pendekatan baru dalam mengatasi permasalahan data dan segmentasi pengunjung web tersebut yang disebut dengan pendekatan klasterisasi bertahap (*two-stages clustering*).

Tahapan analisis klaster didahului dengan pra-pemrosesan data dan analisis faktor. Klasterisasi dilakukan pada data yang berbentuk *frequently* akses (tingkat keseringan kunjungan). Tahap pertama menggunakan metode *cluster non hirarki* dengan tujuan untuk reduksi data sekaligus untuk memilih data yang akan digunakan untuk klasterisasi tahap berikutnya, kemudian dilanjutkan dengan klasterisasi tahap kedua untuk segmentasi pengunjung web. Pada klasterisasi tahap kedua digunakan kombinasi metode klaster, yaitu metode klaster hirarki dan non hirarki.

Dari penerapan metode klasterisasi bertahap ini dapat mereduksi data *web log* hingga 98.38% dan menghasilkan 5 klaster untuk keperluan segmentasi pengunjung web.

6.1 Pendahuluan

Perkembangan internet saat ini tidak terlepas dari pemanfaatan website sebagai media dalam penyebaran informasi. Setiap akses yang dilakukan

pengunjung terhadap sebuah web otomatis di simpan oleh *web server* sebagai data *web log* [54].

Data *web log* berisi kumpulan transaksi (*clickstream*) yang dilakukan oleh pengunjung web, sehingga menggambarkan pola perilaku pengunjung tersebut. Semakin tinggi jumlah akses terhadap sebuah web, maka semakin besar jumlah data *web log* yang terekam [54]. Data *web log* tersebut telah menjadi bagian terpenting dalam bidang penelitian *Web Usage Mining* (WUM) [55] untuk diteliti lebih lanjut terutama dalam hal memahami pola perilaku pengunjung web [56][3][13].

Pada web *e-business/e-commerce* misalnya, keberadaan data *web log* sangat penting guna memperoleh informasi detil terhadap pola perilaku pengunjung *web*, mengetahui histori transaksi dan bahkan memprediksi transaksi berikutnya. Begitu juga halnya pada *web e-learning, e-news, e-banking, e-gov* dan lainnya. Pola-pola akses pengunjung tersebut dapat digunakan untuk berbagai kepentingan. Oleh karena itu, data *web log* menjadi sumber data utama dalam penelitian pada bidang WUM [57].

Beberapa peneliti sebelumnya telah mengajukan beberapa metode dalam upaya mereduksi data *web log* ini. Referensi [58] melakukan pra-pemrosesan berdasarkan data *cleaning* dan data *reduction*. Mereka mengajukan 2 algoritma untuk data *cleaning* dan data *reduction* tersebut Referensi [59], mengajukan 4 tahapan, dengan terlebih dahulu mengidentifikasi *unique user* dan *user sessions* dan kemudian menggunakan metode *association rules*..

Referensi [60], mengajukan pendekatan baru untuk menemukan frekuensi kunjungan dari user menggunakan teori *rough set* yang dapat mengekstrak aturan asosiasi untuk setiap klaster yang homogen dari rekaman data transaksi dan hubungan antar klaster yang berbeda. Dalam penelitiannya, mereka menggunakan algoritma pengurangan biner untuk mengurangi jumlah *dataset* yang besar untuk menemukan aturan asosiasi yang valid.

Referensi [61] menyajikan pra-pemrosesan data *web log*, yaitu menggunakan skema kerja LDAP, empat modul utama yang terlibat yaitu pembersihan data, penataan data, penyaringan data dan kesimpulan dari frekuensi kunjungan. Referensi [62][63] mengajukan 2 teknik pra-pemrosesan data *web log*,

yaitu data *cleaning* dengan menghapus item data gambar dari *log* dan *user identification* dengan menggunakan 3 atribut (*IP Address*, *Operating System* dan *User Agent*).

Wahab [64] menggunakan pendekatan *programming* untuk memindahkan data *log* menjadi sebuah database dan menghapus item data *web log* yang tidak diperlukan (*irrelevant* data). Metode hampir sama juga dilakukan oleh Theint Aye [13] dengan penekanan untuk reduksi data *web log* pada *field* ekstraksi, serta menghapus data yang tidak diperlukan.

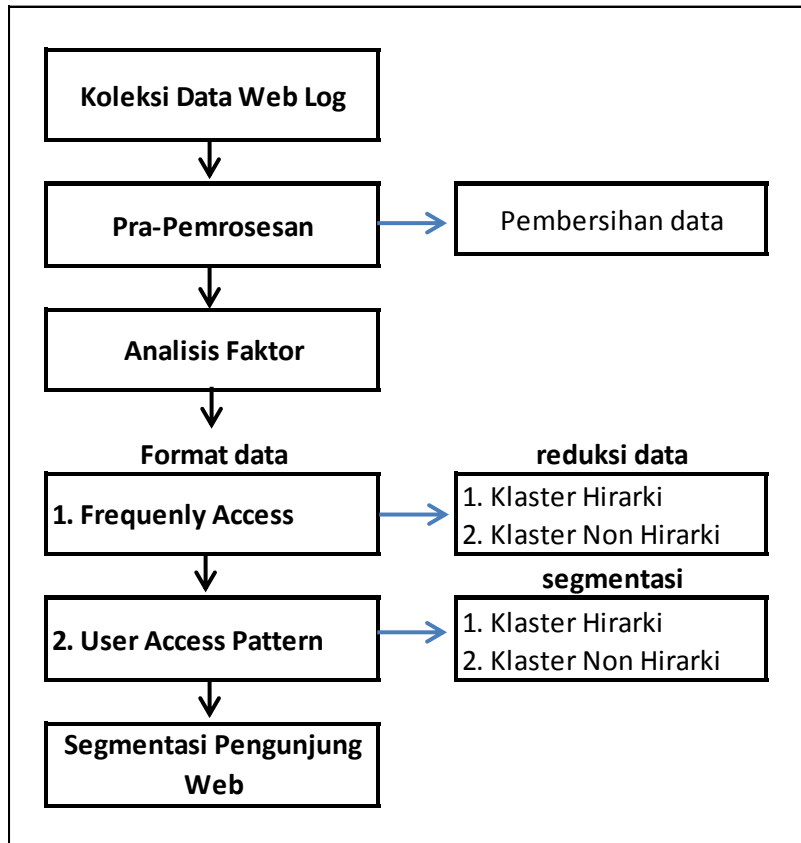
Dari penelitian-penelitian tersebut di atas, terlihat bahwa para peneliti masih dominan melakukan reduksi data *web log* dengan cara menghapus item data yang tidak diperlukan berdasarkan format baku dari data *web log* (perhatikan Tabel 2.1), sementara data *web log*, dapat dilihat dari dua bentuk, yaitu dari bentuk frekuensi kunjungan (*frequently access*) dan bentuk pola perilaku kunjungan (*user access patten*). Hal inilah yang mendasari penulis untuk mencoba melakukan reduksi data *web log* dan segmentasi berdasarkan pada dua sudut pandangan tersebut.

Dalam penelitian ini, diajukan metode baru untuk mereduksi data dan segmentasi dalam proses WUM, yaitu dengan mengusulkan klasterisasi bertahap (*two-stages clustering*). Klasterisasi tahap pertama dilakukan pada data yang berbentuk *frequently access* (lihat sub bab 3.2.1) menggunakan metode klaster non hirarki. Pada tahap ini, klaster yang mempunyai anggota terendah akan dipisahkan dari dataset. Berikutnya dilakukan klasterisasi tahap kedua. Pada tahap ini diterapkan kombinasi metode klaster, yaitu metode klaster hirarki dan non hirarki. Hasil akhir dari klasterisasi tahap kedua ini diperoleh segmen-segmen pengunjung web. Tahapan penelitian dapat di lihat pada Gambar 6.1.

6.2 Format Data *Web Log* dan Dataset

Format file data *web log* yang digunakan adalah *Common Log Format* (CLF) [13][14][19], yaitu format standar yang digunakan oleh *web server* pada saat membuat *log* file. Setiap baris dari CLF berisi informasi *IP Address* atau *DNS*, *User Identifier*, *Date & Time*, *method*, *http request*, *http respond code* dan *Transfer volume/size*, seperti terlihat pada Tabel 2.1.

Dataset yang digunakan dalam penelitian ini adalah data *web log* dari *web* Institut Teknologi Sepuluh Nopember (ITS) Surabaya dengan alamat *web* www.its.ac.id. Periode pengambilan data dari tanggal 3 hingga 16 juli 2012.



Gambar 6.1 Tahapan Penelitian 3

6.3 Pra-pemrosesan

Pada tahap ini dilakukan proses pembersihan (*cleaning*) data *web log* dari item data yang tidak diperlukan (*irrelevant* data). Proses pembersihan data dilakukan berdasarkan sub bab 3.1.1. Pada tahap ini berhasil mereduksi data hingga 97,71%, perhatikan Tabel 4.1, artinya hanya 2,29% yang dapat diolah lebih lanjut. Hal ini menandakan pentingnya peran pra-pemrosesan dalam penambahan data *web log*. Kemudian data di ubah formatnya seperti terlihat pada Tabel 4.2.

Pada tahap ini dilakukan penamaan variabel, yaitu variabel $u_1, u_2, u_3 \dots u_n$ untuk mewakili *user*/pengunjung web dan variabel $p_1, p_2, p_3 \dots p_n$ untuk mewakili halaman web, perhatikan Tabel 4.3.

Kemudian data *web log* hasil pra-pemrosesan ini diubah formatnya menjadi *frequently access* dalam bentuk matriks vektor seperti terlihat pada Tabel 4.2. Tabel tersebut memperlihatkan tingkat kunjungan user terhadap halaman web yang ada.

6.4 Analisis Faktor

Dari hasil pra-pemrosesan, ternyata perbandingan antara data dengan variabel tidak seimbang, (165 data dengan 63 variabel), sehingga perlu dilakukan analisis faktor untuk mereduksi jumlah variabel. Analisis faktor adalah metode multivariat yang digunakan untuk menggambarkan pola hubungan antar variabel dengan tujuan untuk menemukan variabel independen yang mempengaruhi objek. Dalam hal ini, analisis faktor bertujuan untuk mereduksi variabel-variabel menjadi beberapa set indikator yang disebut faktor [52], dengan tanpa kehilangan informasi yang berarti dari variabel awal. Pembahasan teori tentang analisis faktor dapat dibaca pada sub bab 2.5.

Terdapat tiga tahapan utama dalam analisis faktor, yaitu :

1. Menyiapkan data
2. Menentukan metode analisis faktor dan pembuatan faktor
3. Membuat Skor Faktor

6.4.1 Menyiapkan Data

Maksud dari menyiapkan data ini adalah melakukan pengujian terhadap data, apakah data dan variabel layak untuk dilakukan analisis faktor. Pengujian dilakukan dengan mengukur korelasi antar variabel, jika korelasi cukup kuat, maka variabel tersebut akan mengelompok dan membentuk faktor. Untuk mengukur korelasi ini digunakan metode *Kaiser-Meyer-Olkin (KMO)* pada persamaan (2-7), *Measure of Sampling Adequacy (MSA)* pada persamaan (2-6), dan *Bartlett's Test* pada persamaan (2-8).

Tabel 6.1 Hasil Pengujian KMO dan Bartlett's

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0,757
Bartlett's Test of Sphericity	Approx. Chi-Square	9872,112
	Df	1596
	Sig.	0,000

Nilai yang ditolerir dalam pengujian KMO adalah $>0,5$ dan dengan signifikansi $<0,05$. Dari Tabel 6.1, terlihat bahwa nilai pengujian dari KMO dan Bartlett's adalah 0,757 ($>0,05$) dengan signifikansi 0,000 ($<0,05$). Hal ini menandakan bahwa data secara umum sudah layak untuk dilakukan analisis selanjutnya.

Selanjutnya dilakukan penghitungan terhadap nilai *MSA* dengan ketentuan, jika nilai *MSA* $> 0,5$, variabel layak untuk dianalisis, namun jika *MSA* $<0,5$ variabel tersebut tidak layak dan dikeluarkan dari proses berikutnya. Perhatikan Tabel 6.2.

Tabel 6.2 Variabel dengan Nilai *MSA* $< 0,5$

Variabel	Nilai <i>MSA</i>
<i>p1</i>	0,401
<i>p7</i>	0,394
<i>p30</i>	0,491
<i>p33</i>	0,471
<i>p36</i>	0,416
<i>p50</i>	0,415

Sehingga jumlah variabel setelah dilakukan uji ini adalah $163 - 6 = 57$ variabel.

6.4.2 Menentukan Metode Analisis Faktor dan Pembuatan Faktor

Setelah data dinyatakan layak berdasarkan pengujian sebelumnya, maka dilanjutkan tahap proses ekstraksi terhadap sekumpulan variabel (proses faktorisasi), sehingga terbentuk satu atau beberapa faktor. Untuk ekstraksi variabel digunakan metode *principal component analysis* (PCA).

Langkah pertama pada metode ini adalah melihat kemampuan variabel awal dapat dijelaskan oleh faktor. Perhatikan Tabel 6.3.

Tabel 6.3 Communalities

Variabel	Ekstraksi
<i>p2</i>	71,8%
<i>p3</i>	67,3%
<i>p4</i>	89,7%
<i>p5</i>	66,5%
<i>p6</i>	96,8%
<i>p8</i>	69,0%
<i>p9</i>	55,8%
:	:
<i>p63</i>	34,0%
Metode Ekstraksi : PCA	

Dari Tabel 6.3 dapat dinyatakan bahwa variabel *p2* dapat dijelaskan oleh faktor yang terbentuk nantinya sebesar 71,8% (0.718). Semakin tinggi nilai pada kolom ekstraksi semakin kuat relasinya dengan faktor yang terbentuk. Hal yang sama juga dapat dijelaskan untuk variabel yang lainnya.

Jumlah faktor yang terbentuk dapat dijelaskan melalui total nilai eigenvalues. Perhatikan Tabel 6.4.

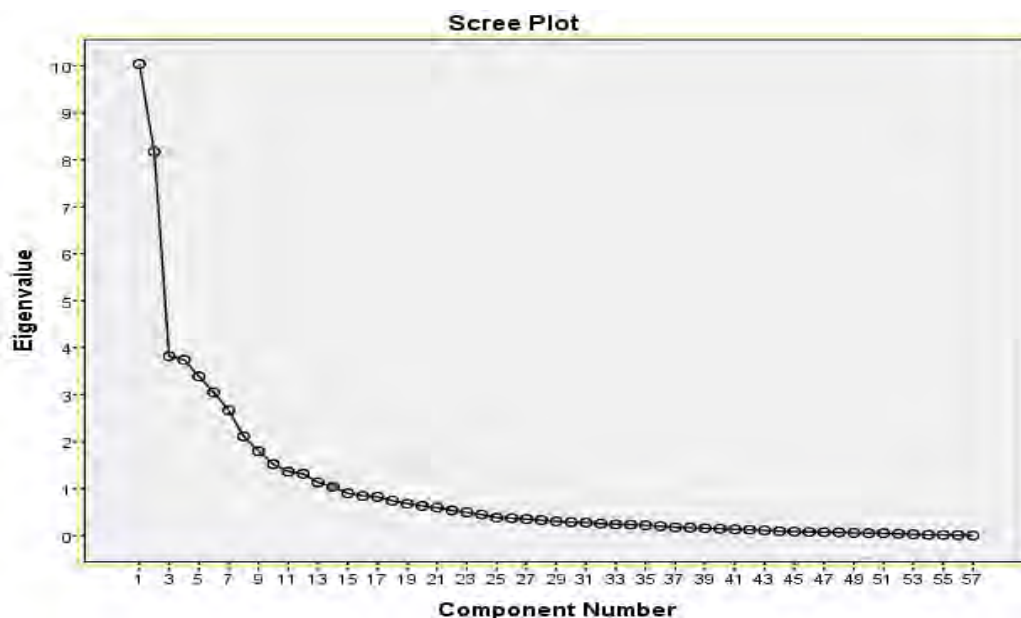
Tabel 6.4 Nilai Eigenvalues

Komponen/ Faktor	Eigenvalues	
	Total	% Varians
1	10,037	17,608
2	8,170	14,333
3	3,817	6,696
4	3,740	6,561
5	3,389	5,945
6	3,048	5,347
7	2,669	4,682
8	2,114	3,708

Tabel 6.4 Nilai Eigenvalues (Lanjutan)

Komponen/ Faktor	Eigenvalues	
	Total	% Varians
9	1,796	3,151
10	1,520	2,667
11	1,361	2,388
12	1,316	2,309
13	1,133	1,987
14	1,038	1,821
15	0,900	1,578
:	:	:
57	0,001	0,001
Metode Ekstraksi : PCA		

Terlihat pada Tabel 6.4 bahwa terbentuk 14 faktor dari 57 variabel, yaitu *eigenvalue* dengan nilai ≥ 1 [34], sedangkan *eigenvalue* dengan nilai dibawah 1 tidak digunakan dalam proses pembentukan faktor. Hasil yang terlihat pada Tabel 6.4 juga digambarkan dengan grafik melalui Scree Plot. Perhatikan Gambar 6.2



Gambar 6.2 Scree Plot Hasil Faktorisasi

Terlihat pada gambar diatas proses pembentukan faktor, dari satu faktor ke terbentuknya faktor 2, 3 dan seterusnya dengan melihat arah garis menurun pada bagian *componen number* di sumbu Y, dan pada eigenvalue dengan nilai = 1 dan dengan nilai komponen = 14 dinyatakan proses faktorisasi berhenti, karena nilai < 1 tidak mampu membentuk faktor dengan baik.

Langkah berikutnya adalah menentukan variabel mana saja yang akan bergabung dengan 14 faktor yang terbentuk. Hal ini dijelaskan dengan komponen matriks yang menunjukkan distribusi ke 57 variabel pada 14 faktor yang terbentuk. Semakin besar nilai *factor loading* pada komponen matriks menunjukkan besarnya korelasi antara suatu variabel dengan faktor 1, faktor 2 dan seterusnya. Perhatikan Tabel 6.5.

Tabel 6.5 Nilai Faktor Loading Pada Komponen Matrik

Komponen Matriks														
Var.	Faktor													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
<i>p2</i>	-,165	,067	-,144	,051	,097	-,098	,148	,102	,051	,478	,097	-,008	,595	,129
<i>p3</i>	,384	,074	,100	-,002	-,253	,515	,039	,261	-,294	-,001	,152	,013	-,036	,037
<i>p4</i>	,178	-,018	,749	,473	,222	,135	-,023	,107	,008	-,005	,016	-,007	-,007	,031
<i>p5</i>	,211	,022	,077	-,142	-,164	,336	,009	,245	-,439	,167	,123	,329	,077	,210
<i>p6</i>	,182	,016	,681	,624	,277	,028	-,016	,053	,026	,013	,004	,006	,007	-,003
<i>p8</i>	,497	,150	,165	-,064	-,378	,336	-,019	,256	-,074	-,025	,164	-,143	-,024	,115
<i>p9</i>	-,203	,394	-,207	,193	,046	-,022	,024	,000	-,053	,375	,065	-,048	,351	,076
<i>p10</i>	-,559	,529	,044	-,011	-,028	,019	-,018	-,083	-,084	-,222	-,076	-,051	,084	-,044
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
<i>p63</i>	,271	,254	-,039	-,170	,284	,166	-,034	-,021	-,042	,003	,079	-,214	-,022	-,085

Metode Ekstraksi : Principal Component Analysis.

Proses penentuan variabel mana yang akan masuk ke faktor tertentu dilakukan dengan membandingkan besarnya korelasi yang terbentuk pada setiap barisnya. Perhatikan variabel *p6* pada Tabel 6.5, nilai *factor loading* tertinggi

berada pada faktor 3, yaitu 0,681, hal ini membuktikan bahwa korelasi antara variabel $p6$ dengan faktor 3 kuat dibandingkan dengan korelasi dengan faktor yang lain dibaris yang sama, sehingga variabel $p6$ akan berada pada faktor 3.

Jika dalam proses diatas terdapat beberapa nilai *factor loading* $>0,5$ (korelasi kuat) atau $<0,5$ (korelasi lemah) pada baris yang sama, maka dilakukan proses rotasi untuk memperjelas posisi suatu variabel tergabung ke dalam faktor tertentu. Rotasi dilakukan dari hasil faktorisasi sebelumnya. Perhatikan Tabel 6.6 yang memperlihatkan hasil rotasi komponen matriks.

Tabel 6.6 Rotasi Komponen Matriks

Rotasi Komponen Matriks														
Var.	Faktor													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$p2$,069	,000	-,028	-,017	-,064	-,043	,010	-,039	-,053	,082	-,024	-,014	,042	,832
$p3$	-,091	-,023	,045	,077	,044	,208	,143	,723	,205	-,072	-,064	-,071	-,070	-,083
$p4$	-,051	-,012	,923	-,058	,023	,105	,013	,142	,008	-,042	,003	-,011	-,034	-,065
$p5$	-,082	-,068	-,032	-,049	,099	-,044	-,157	,688	,314	-,052	-,087	,099	,036	,145
$p6$	-,042	,031	,971	,089	,052	,106	-,005	,007	,007	-,026	,004	-,001	-,014	-,016
$p8$	-,095	,116	,019	-,040	,171	,388	,194	,652	-,028	-,075	,043	-,031	-,073	-,094
$p9$,365	,032	-,029	,223	,020	-,024	,057	-,017	,024	-,051	,086	-,028	,095	,590
$p10$,781	-,003	-,039	-,026	-,057	-,137	-,010	-,113	,006	-,069	-,021	,004	-,154	-,031
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:
$p63$	-,003	,174	,002	-,015	,081	-,095	,462	,095	,179	-,027	,190	-,037	,000	-,032

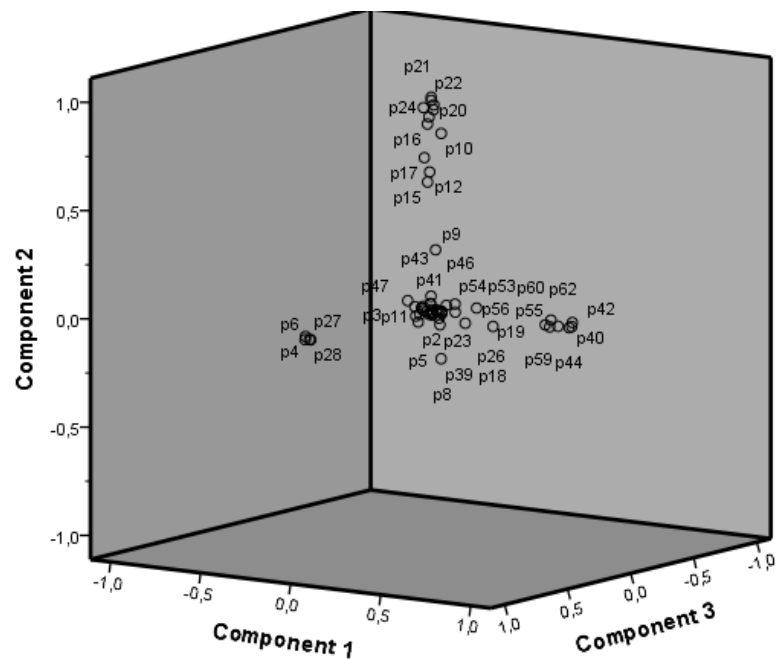
Metode Ekstraksi : Principal Component Analysis.
Metode Rotasi : Varimax

Berdasarkan informasi di Tabel 6.6, maka dapat ditentukan posisi sebuah variabel akan bergabung ke dalam faktor tertentu berdasarkan tingkat korelasinya. Hasilnya ditampilkan pada Tabel 6.7.

Merujuk ke Tabel 6.7, dengan demikian jelas bahwa faktor 1 merupakan gabungan dari variabel $p10, p12, p15, p16, p17, p20, p21, p22, p24, p25$ dan $p49$, demikian seterusnya, sehingga jelas posisi variabel dalam suatu faktor. Posisi keanggotaan dalam faktor dapat digambar seperti terlihat pada Gambar 6.3.

Tabel 6.7 Faktor dengan Variabel

Faktor	Keanggotaan Variabel						
1	<i>p10</i>	<i>p12</i>	<i>p15</i>	<i>p16</i>	<i>p17</i>	<i>p20</i>	<i>p21</i>
	<i>p22</i>	<i>p24</i>	<i>p25</i>	<i>p49</i>			
2	<i>p40</i>	<i>p42</i>	<i>p44</i>	<i>p55</i>	<i>p59</i>	<i>p60</i>	<i>p62</i>
3	<i>p4</i>	<i>p6</i>	<i>p27</i>	<i>p28</i>			
4	<i>p31</i>	<i>p38</i>	<i>p57</i>				
5	<i>p32</i>	<i>p34</i>	<i>p35</i>	<i>p37</i>			
6	<i>p11</i>	<i>p13</i>	<i>p14</i>	<i>p19</i>	<i>p26</i>		
7	<i>p45</i>	<i>p51</i>	<i>p61</i>	<i>p63</i>			
8	<i>p3</i>	<i>p5</i>	<i>p8</i>	<i>p18</i>			
9	<i>p46</i>	<i>p48</i>	<i>p53</i>	<i>p54</i>			
10	<i>p23</i>	<i>p29</i>	<i>p41</i>	<i>p43</i>			
11	<i>p52</i>	<i>p58</i>					
12	<i>p47</i>	<i>p56</i>					
13	<i>p39</i>						
14	<i>p2</i>	<i>p9</i>					



Gambar 6.3 Komponen Plot

Terlihat pada Gambar 6.3, bahwa variabel-variabel yang tergabung dalam faktor yang sama akan berada dalam posisi yang berdekatan. Perhatikan faktor 3 dengan anggota *p4*, *p6*, *p27* dan *p28*.

6.4.3 Membuat Skor Faktor

Setelah faktor terbentuk, maka dapat di buat skor faktor, yaitu memberikan skor terhadap faktor yang terbentuk untuk menggantikan nilai dari variabel asli dan memberikan penamaan variabel $f1$ untuk faktor 1, $f2$ untuk faktor 2 dan seterusnya. Dalam kasus ini, terdapat 14 faktor yang akan diberikan skor. Pembuatan skor digunakan sebagai input untuk analisis klaster. Hingga pada tahap ini, posisi dataset berubah menjadi 165 data dengan 14 faktor (variabel).

6.5 Klasterisasi

Pada tahap ini dilakukan klasterisasi berdasarkan pada data *web log* yang berbentuk *frequently access*, dengan tujuan untuk memilih kelompok-kelompok data yang akan diproses selanjutnya.

Kemudian dilanjutkan dengan klaster tahap kedua. Pada tahap kedua ini, digunakan kombinasi metode klaster hirarki *single linkage* dan metode klaster non hirarki. Metode *single linkage* melakukan pengelompokkan data berdasarkan jarak terdekat antar objek dan validasi dengan metode *Sum of Squared Error*.

6.5.1 Klasterisasi Tahap Pertama

Data hasil analisis faktor, diubah bentuknya seperti terlihat pada Tabel 4.2. Kemudian diterapkan klasterisasi tahap pertama menggunakan metode non hirarki. Selain bertujuan untuk mereduksi data dan mendapatkan kelompok-kelompok user dengan tingkat kunjungan yang sama, tahap ini juga berguna untuk memilih data yang akan digunakan pada klaster tahap kedua. Penentuan jumlah klaster yang dibentuk menggunakan metode *elbow rule* yaitu melihat perbedaan lonjakan nilai koefisien pada *agglomeration schedule*. Perhatikan Tabel 6.8

Terlihat pada Tabel 6.8, lonjakan perpindahan terbesar terjadi pada stage 158, hal ini berarti, bahwa dari 165 data dikurangi 158, maka jumlah klaster yang optimal adalah 7 klaster. Hasil klasterisasi dengan 7 klaster menggunakan metode klaster non hirarki terlihat pada Tabel 6.9.

Tabel 6.8 Agglomeration Schedule Tahap 1

Stage	Koefisien	
...
158	7,313	2,660
159	9,973	0,719
160	10,691	0,093
161	10,784	0,746
162	11,530	0,589
163	12,119	1,108

Tabel 6.9 Klaster Tahap 1 dan Jumlah Anggotanya

Klaster	1	1,000
	2	1,000
	3	3,000
	4	1,000
	5	2,000
	6	3,000
	7	154,000
Data valid		165,000

Pada tahap ini, klaster yang memiliki anggota terbanyak (klaster 7) dipilih untuk diproses pada klasterisasi tahap kedua. Data anggota klaster 1 sampai dengan klaster 6 tidak diproses lebih lanjut. *User* yang tergabung ke dalam klaster tersebut terlihat pada Tabel 6.10.

Tabel 6.10 Anggota Klaster Yang Tidak Terpilih

Klaster	User	Jumlah Kunjungan				
		<i>p1</i>	<i>p2</i>	<i>p3</i>	<i>p4</i>	...
1	<i>u1</i>	6	9	0	0	...
2	<i>u2</i>	0	0	0	35	...
3	<i>u10</i>	0	1	3	3	...
	<i>u29</i>	0	1	1	1	...
	<i>u65</i>	0	0	0	0	...

Tabel 6.10 Anggota Klaster Yang Tidak Terpilih (Lanjutan)

Klaster	User	Jumlah Kunjungan				
		<i>p1</i>	<i>p2</i>	<i>p3</i>	<i>p4</i>	...
4	<i>u4</i>	0	1	4	3	...
5	<i>u78</i>	1	0	0	0	...
	<i>u80</i>	0	1	0	1	...
6	<i>u6</i>	5	5	5	0	...
	<i>u16</i>	2	3	1	1	...
	<i>u31</i>	2	2	1	0	...

6.5.2 Klasterisasi Tahap Kedua

Data hasil klasterisasi tahap pertama dijadikan input klasterisasi tahap kedua. Sehingga informasi data berubah seperti terlihat pada Tabel 6.11.

Tabel 6.11 Informasi Data *Web log* ke 2

	Data Mentah	Data Final	% Data Tereduksi
Total <i>Record</i>	163,281	2,640	98,38
Jumlah <i>Web page/var</i>	6,753	57	99,16
Jumlah user unik	5,468	154	97,18

Terlihat pada Tabel 6.11 terjadi perubahan data yang signifikan, hanya 1.62% data yang diproses lebih lanjut, artinya tereliminasi sebanyak 98,38% dari tahap-tahap sebelumnya.

Pada tahap ini, klasterisasi dilakukan dengan mengkombinasikan metode klaster hirarki dan non hirarki untuk tujuan segmentasi pengunjung web. Diawali dengan menggunakan metode klaster hirarki. Hasil dari klasterisasi hirarki di uji dengan metode *elbow rules* dan *sum of squared error* untuk menentukan jumlah klaster yang optimal. Perhatikan Tabel 6.12.

Tabel 6.12 Agglomeration Schedule Tahap 2

Stage	Koefisien	
...
148	5,992	0,276
149	6,268	0,441
150	6,709	0,267
151	6,977	0,243
152	7,220	0,064
153	7,284	0,276

Terlihat pada stage ke 149 terjadi peningkatan/lompatan koefisien terbesar dibandingkan dengan stage yang lainnya. Berdasarkan *elbow rules*, maka jumlah kluster yang optimal adalah $154 - 149 = 5$, (5 kluster). Hasil uji dengan SSE terlihat pada Tabel 6.13

Tabel 6.13 Hasil Uji SSE

Jumlah Kluster	Nilai SSE
2	1,342
3	1,216
4	1,111
5	990

Kemudian hasil uji ini dijadikan input untuk klusterisasi menggunakan metode non hirarki, sehingga diperoleh hasil kluster seperti Tabel 6.14.

Tabel 6.14 Kluster Tahap 2 dan Jumlah Anggotanya

Cluster	1	5,000
	2	7,000
	3	9,000
	4	119,000
	5	14,000
Valid		154,000
Missing		0,000

Berdasarkan data dari Tabel 6.14, terlihat seluruh data berhasil di proses yang ditandai dengan nilai *missing*=0,000.

6.6 Segmentasi Pengunjung Web / Profiling User

Setiap klaster yang terbentuk mempunyai karakteristik tertentu dibandingkan dengan klaster yang lain. Hal ini tergambar dari nilai pusat klaster akhir, perhatikan Tabel 6.15.

Tabel 6.15 Pusat Klaster Akhir

Var/ Faktor	Klaster				
	1	2	3	4	5
<i>f1</i>	-0,14302	0,51931	-0,03075	-0,09381	-0,16654
<i>f2</i>	-0,11612	-0,03347	0,19022	-0,06189	-0,76314
<i>f3</i>	-0,10348	-0,12232	-0,14576	-0,08088	-0,06744
<i>f4</i>	-0,23206	-0,24023	-0,06301	-0,12663	-0,06569
<i>f5</i>	-0,11551	-0,19644	-0,13104	-0,31807	2,64206
<i>f6</i>	-0,39000	-0,31502	-0,38052	0,04298	-0,10582
<i>f7</i>	-0,06502	0,21856	-0,13370	0,01608	-0,21342
<i>f8</i>	-0,33749	-0,07229	-0,35506	0,07770	-0,08958
<i>f9</i>	-0,08250	-0,24050	1,13708	-0,05348	-0,32219
<i>f10</i>	5,25092	-0,27236	-0,19076	-0,17367	-0,16356
<i>f11</i>	-0,08643	-0,23092	-0,13389	0,01500	0,15548
<i>f12</i>	-0,03026	-0,21193	-0,21209	-0,09211	-0,12254
<i>f13</i>	-0,04055	0,26929	2,99888	-0,21120	-0,10504
<i>f14</i>	-0,21503	3,73817	-0,25141	-0,17834	-0,15155

Pusat klaster akhir dengan nilai positif menandakan tingkat kunjungan diatas rata-rata total kunjungan, sebaliknya tanda minus (-) menandakan tingkat kunjungan berada di bawah rata-rata. Berdasarkan Tabel 6.15, dapat dinyatakan bahwa *user*/pengunjung web yang berada pada klaster 1, dominan mengakses halaman web yang berada di faktor 10 (*f10*). Dalam faktor 10 ini terdapat halaman web *p23*, *p29*, *p41* dan *p43*, lihat Tabel 6.7. Identitas halaman web tersedia pada lampiran A, sehingga dapat dibuatkan profil klaster sebagai berikut :

Profil Klaster 1 : Member Web ITS

Pengunjung pada kelompok ini dominan mengakses halaman web yang berada di faktor 10, yaitu halaman web yang berhubungan dengan layanan personal (*p23*, *p29*, *p41* dan *p43*), dapat disimpulkan pengunjung di klaster ini adalah pengunjung yang sudah terdaftar sebagai *member* di web ITS.

Profil Klaster 2 : Pengunjung Baru Web ITS

Pengunjung pada kelompok ini dominan mengakses halaman web yang berada di faktor 14, yaitu halaman web yang berhubungan dengan halaman personal dan semua aktifitas personal (*p2* dan *p9*), dapat disimpulkan pengunjung di klaster ini adalah pengunjung baru yang ingin mendapatkan informasi tentang layanan keanggotaan di web ITS.

Profil Klaster 3 : Pengunjung dari Luar Negeri

Pengunjung pada kelompok ini dominan mengakses halaman web yang berada di faktor 13 dan faktor 9, yaitu halaman akses untuk mendapatkan kumpulan SK Rektor (*p39*) dan mengakses halaman web berbahasa Inggris untuk mendapatkan informasi tentang jurusan di ITS, seperti Jurusan Despro (*p46*), Jurusan Sistem Informasi (*p48*), Jurusan Arsitektur (*p53*) dan Jurusan Elektro(*p54*). Dapat disimpulkan, kemungkinan besar pengunjung di klaster ini berasal dari luar negeri yang ingin melanjutkan studi di ITS.

Profil Klaster 4 : Calon Mahasiswa Baru ITS

Pengunjung pada kelompok ini dominan mengakses halaman web yang berada di faktor 8, yaitu halaman web yang berhubungan dengan akademik (*p3*), informasi tentang ITS (*p5*), serta prosedur mendaftar ke ITS (*p8* dan *p18*). Dapat disimpulkan bahwa pengunjung yang berada di klaster ini merupakan calon-calon mahasiswa baru ITS yang ingin mendapatkan informasi tentang sistem akademik dan prosedur mendaftar sebagai mahasiswa baru ITS. Halaman web yang diakses oleh pengunjung di klaster ini harus mendapat perhatian khusus dari pengelola web, mencakup *layout* (tata letak, pewarnaan, dll) serta informasi yang disampaikan mudah dipahami, karena jumlah pengunjung di klaster ini mempunyai anggota terbanyak, yaitu 119 anggota.

Profil Klaster 5 : Pengunjung Umum

Pengunjung pada kelompok ini dominan mengakses halaman web yang berada di faktor 5, yaitu halaman web berbahasa Inggris yang berhubungan dengan halaman contact (p32), visi ITS (p34), Sejarah ITS (p35). Dapat disimpulkan bahwa pengunjung yang berada di klaster menguasai bahasa Inggris dan ingin mendapatkan informasi tentang ITS.

6.7 Kesimpulan

Berdasarkan hasil uji coba metode pra pemrosesan dan klasterisasi bertahap terhadap data *web log* ITS, dapat disimpulkan bahwa metode ini mampu mereduksi data *web log* hingga 98.38%., dari 163.281 *record* data mentah, hanya 2.640 *record* yang dapat diolah lebih lanjut (1.62 %). Jika dilihat dari sisi *user*, dari 5.468 *user*, hanya 154 *user* yang dapat mewakili untuk diproses lebih lanjut (2.82%). Hal yang sama untuk jumlah halaman web, hanya 0,84% yang akan diproses lebih lanjut. Hal ini membuktikan bahwa pendekatan pra-pemrosesan yang digunakan mampu mengeliminasi data-data yang tidak relevan secara signifikan, serta membuktikan pentingnya peran pra-pemrosesan pada WUM.

Merujuk ke Tabel 6.14, dapat disimpulkan bahwa dari 5 klaster yang terbentuk, klaster 4 merupakan klaster dengan anggota terbesar (119 anggota), maka pola perilaku pengunjung web pada klaster ini harus menjadi perhatian khusus bagi pengelola web sebagai input dalam meningkatkan kualitas informasi yang disajikan pada halaman web yang sering mereka akses. Dari profil 5 klaster pada sub bab 6.6, dapat dijadikan rujukan dalam membuat segmentasi pengunjung web dengan lebih baik.

BAB 7

KESIMPULAN DAN SARAN

Setelah melakukan serangkaian penelitian terhadap topik yang diangkat, maka dapat diambil kesimpulan dan saran perbaikan dimasa mendatang.

7.1 Kesimpulan

Berdasarkan hasil uji coba metode klasterisasi bertahap terhadap data *web log* ITS, dapat disimpulkan bahwa metode ini mampu mereduksi data *web log* hingga 98.38%, dari 163.281 *record* data mentah, hanya 2.640 *record* yang dapat diolah lebih lanjut (1.62 %). Jika dilihat dari sisi *user*, dari 5.468 *user*, hanya 154 *user* yang dapat mewakili untuk diproses lebih lanjut (2.82%). Hal yang sama untuk jumlah halaman web, hanya 0,84% yang akan diproses lebih lanjut. Hal ini membuktikan bahwa pendekatan pra-pemrosesan yang digunakan mampu mengeliminasi data-data yang tidak relevan secara signifikan, serta membuktikan pentingnya peran pra-pemrosesan pada WUM.

Pada proses klasterisasi untuk segmentasi pengunjung web, diperoleh 5 kelompok *user* dengan karakteristik masing-masing. Klaster ke 4 dengan jumlah anggota terbanyak menjadi perhatian khusus bagi pengelola web. Pengunjung pada kelompok ini dominan mengakses halaman web yang berhubungan dengan akademik, informasi tentang ITS, serta prosedur mendaftar ke ITS. Dapat disimpulkan bahwa pengunjung yang berada di klaster ini merupakan calon-calon mahasiswa baru ITS yang ingin mendapatkan informasi tentang sistem akademik dan prosedur mendaftar sebagai mahasiswa baru ITS. Halaman web yang diakses oleh pengunjung di klaster ini harus mendapat perhatian khusus dari pengelola web, mencakup *lay out* (tata letak, pewarnaan, dll) serta informasi yang disampaikan, karena jumlah pengunjung di klaster ini mempunyai anggota terbanyak, yaitu 119 anggota.

Untuk dataset www.its.ac.id/en, setelah dilakukan pra-pemrosesan diperoleh kesimpulan bahwa hanya 6% data yang dapat digunakan untuk diproses lebih lanjut, sedangkan 94% tereduksi pada tahap pra-pemrosesan.

Dari penerapan metoda klaster terhadap data yang berbentuk *user access pattern* ini diperoleh kesimpulan bahwa terdapat 4 kelompok user yang mempunyai kemiripan dalam pola perilaku kunjungan yaitu klaster 2, 3, 4 dan 6, namun jika dilihat dari jumlah populasi, maka klaster 1 mempunyai populasi terbesar, yaitu 28%.

Ditinjau dari karakteristik pola perilaku pengunjung web dari setiap klasternya, terlihat halaman web tentang akademik menjadi halaman yang sering dikunjungi, hal ini menandakan bahwa pengunjung web lebih dominan ingin mendapat informasi tentang akademik, kemudian halaman web tata cara/prosedur mendaftar di ITS dan layanan-layanan di ITS.

Berdasarkan hasil tersebut, maka pola kunjungan dari anggota klaster yang mempunyai tingkat kemiripan tinggi, populasi yang besar serta pola distribusi kunjungan dan halaman yang sering diakses menjadi perhatian khusus bagi pengelola *web* untuk keperluan segmentasi pengunjung *web* dan meningkatkan kualitas layanan web sehingga lebih informatif.

7.2 Saran

Setelah dilakukan penelitian terhadap topik ini, maka dapat diberikan beberapa saran sebagai berikut :

- a. Belum adanya metoda penelitian serupa dengan penerapan *two stage clustering* ini untuk data *web log*, maka perlu dilakukan kajian-kajian lebih lanjut.
- b. Hasil penerapan metoda ini sebaiknya diuji coba pada bidang prediksi perilaku pengunjung web disamping untuk segmentasi pengunjung web.

DAFTAR PUSTAKA

1. Toffler, Alvin, "The Third Wave", Bantam Books (USA), ISBN : 0-517-32719-9, 1980.
2. Miniwatts Marketing Group, Maret 2011, *Internet Usage Statistics: The Internet Big Picture World Internet User and Population Stats*. URL : <http://www.internetworldstats.com/stats.htm>, diakses tanggal 2 Juni 2015.
3. Olson, David; Shi, Yon; "Introduction to Business Data Mining", Penerbit Salemba Empat/Mc Graw Hill, ISBN : 978-979-691-442-5, 2008.
4. Bing Liu, *Web Data Mining : Exploring Hyperlinks, Contents, and Usage Data*, Springer, Berlin, 2007.
5. – Kode Status HTTP,
URL : http://en.wikipedia.org/wiki/List_of_HTTP_status_codes, diakses tanggal 7 Juni 2015.
6. Raykov T, Marcoulides AG. *An Introduction to Applied Multivariate Analysis*. New York. Taylor & Francis Group. 2008.
7. – Analisis Klaster, http://www.statistikian.com/2014/03/analisis-cluster_27.html., diakses tanggal 7 Juni 2015
8. -- HTTP Status Codes, http://www.restpatterns.org/HTTP_Status_Codes, diakses tanggal 12 Desember 2015.
9. -- HTTP Method, http://www.restpatterns.org/HTTP_Methods, Tanggal diakses tanggal 12 Desember 2015.
10. Yuhefizar.: *10 Jam Menguasai Internet : Teknologi dan Aplikasinya*. PT. Elex Media Komputindo, Jakarta, November 2008. Nomor ISBN : 978-979-27-3470-6.
11. Web Server Survey. <http://news.netcraft.com/archives/2012/07/03/july-2012-web-server-survey.html>. Diakses terakhir tanggal 30 Juli 2012.
12. Internet Usage Statistic. <http://www.internetworldstats.com/stats.htm>, diakses terahir tanggal 30 Juli 2012.

13. Aye TT.: *Web Log Cleaning For Mining of Web Usage Patterns*. International Conference on Computer Research and Development (ICCRD). Shanghai. 2011; 2 : 490-494.
14. Hussain T, Asghar S, Masood N.: *Web Usage Mining: A Survey on Preprocessing of Web Log File*. International Conference on Information and Emerging Technologies (ICIET). Karachi. 2010: 1 – 6.
15. Khasawneh N, Chan C-C.: *Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining*. International Conference on Web Intelligence, IEEE/WIC/ACM. Washington 2006: 325 – 328.
16. Houqun Y, Jingsheng L, Fa F.: *An Approach of Multi-path Segmentation Clustering Based on Web Usage Mining*. International Conference on Fuzzy Systems and Knowledge Discovery. 2007:644-648.
17. Slaninović M, Dolžan M, Mikusić M, Martinović M, Šnecelj V.: *User Segmentation Based on Finding Communities with Similar Behavior on the Web Site*. International Conference on Web Intelligence and Intelligence Agent Technology. 2010:75 – 78.
18. Oemarjadi C S, Maulidevi N U.: *Web Personalization in Used Cars Ecommerce Site*. International Conference on Electrical Engineering And Informatic. 2011:1 – 4.
19. Wang X-G, Li Yue.: *Web Mining Based On User Access Pattern for Web Personalization*. International Colloquium on Computing, Communication, Control and Management. 2009:194 - 197.
20. Kumar R.: *Mining Web Logs: Applications and Challenges* DD'09 Proceedings of the 15th ACM SIGKDD, International Conference on Knowledge discovery and data mining. New York. 2009.
21. Srivastava J, Cooley R, Deshpande M, Tan P.-N.: *Web Usage Mining: Discovery and Applications of Usage Patterns From Web Data*. SIGKDD Explorations. 2000; 1(2): 12-23.
22. Jian L, Yan-Qing W.: *Web Log Data Mining Based on Association Rule*. International Conference on Fuzzy System and Knowledge Discovery (FSKD). Shanghai. 2011; 3: 1855 – 1859.

23. Nagi M, ElSheikh A, Sleiman I, Peng P, Rifaie M, Kianmehr K, Karampelas P, Ridley M, Rokne J, Alhajj R.: *Association Rules Mining Based Approach for Web Usage Mining*. IEEE International Conference on Information Reuse and Integration (IRI). Las Vegas, NV. 2011: 166 – 171.
24. Gaol FL.: *Exploring the Pattern of Habits of Users Using Web logs Sequential Pattern*. Second International Conference on Advances in Computing, Control and Telecommunication Technologies (ACT). Jakarta. 2010: 161 – 163.
25. Wu H-Y, Zhu J-J, Zhang X-Y.: *The Explore of the Web-Based Learning Environment Base in Web Sequential Pattern Mining*. International Conference on Computational Intelligence and Software Engineering (CISE). Wuhan. 2009: 1 – 6.
26. Gui-ling L.: *The Study on Web Data Mining on Belief Rough Set Classification*. International Conference on Communication Software and Networks (ICCSN). IEEE 3rd Xi'an 2011: 673 – 677.
27. Trabelsi S, Elouedi Z, Lingras P.: *Belief Rought Set Classification For Web Mining Based on Dynamic Core*. International Conference on Intelligent Systems Design and Application (ISDA). Cairo. 2010: 403 – 408.
28. Suresh K, MadanaMohana R, RamaMohan Reddy A, Subramanyam A.: *Improved FCM Algorithm for Clustering on Web Usage Mining*. International Conference on Computer and Management (CAMAN). Wuhan. 2011: 1 – 4.
29. Sudhamathy G, Venkateswaran JC.: *Web Log Clustering Approaches – A Survey*. International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397. 2011; 3(7): 2896 – 1903.
30. Shi P.: *An Efficient Approach for Clustering Web Access Patterns from Web Logs*. International Journal of Advanced Science and Technology. 2009; 5 : 1 – 14.
31. Xie Y, Phoha VV.: *Web User Clustering From Access Log Using Belief Function*. in: Proceedings of the ACM K-CAP'OI, First International Conference on Knowledge Capture, Victoria. 2001: 202-208.

32. Xu HJ, Liu H.: *Web User Clustering Analysis Based on KMeans Algorithm*. International Conference on Information, Networking and Automation (ICINA). Kunming. 2010; 2 : V2-6 – V2-9.
33. Chaofeng L.: *Research on Web Session Clustering*. Journal of Software. Academy Publisher. 2009; 4(5): 460 – 468.
34. Johnson RA, Wichern DW.: *Applied Multivariate Statistical Analysis*, Sixth Edition. New Jersey: Pearson Prentice Hall. 2007.
35. Aranganayagi S, Thangavel K.: *Clustering Categorical Data Using Silhouette Coefficient as a Relocating Measure*. International Conference on Computational Intelligence and Multimedia. 2007: 13 - 17.
36. Santosa B. *Data Mining Terapan dengan Matlab*. Edisi Pertama. Yogyakarta, Indonesia: Graha Ilmu. 2007.
37. Lammport, L., 1994, *LaTeX: A Document Preparation System*, Second Edition, Addison Wiley, Canada.
38. Chen, W-B., 2007, *Biological Sequence Clustering and Classification with a Hybrid Method and Dynamic Programming*, International Conference on AINAW. Vol. 1, pages 684 – 689.
39. Vijaya PA, 2006, *Efficient Bottom-up Hybrid Hierarchical Clustering Techniques For Protein Sequence Classification*, Pattern Recognition, Elsevier Journal, Vol. 39 Issue 12, pages 2344 – 2355.
40. Han SI, dkk, 2006, *CLAGen: A Tool For Clustering and Annotating Gene Sequences Using a Suffix Tree Algorithm*, Elsevier Journal, Biosystems, Vol. 84, Issue 3, pages 175-182.
41. Ferles C, 2008, *Sequence Clustering With The Self-Organization Hidden Markov Model Map*, International Conference on BIBE, pages 1 – 7.
42. Santhisree K, Damodaram A, 2011, *CLIQUE” Clustering Based on Density on Web Usage Data: Experiments and Test Results*, International Conference on ICECT, Vol. 4, pages 233-236.
43. Azimpour KM, Azmi R, 2011, *A Webpage Similarity Measure For Web Sessions Clustering Using Sequence Alignment*, International Symposium on AISP, pages 20 – 24.

44. Hong T, Kim E, 2012, *Segmenting Customers in Online Stores Based on Factors That Affect The Customer's Intention to Purchase*. Journal Expert Systems with Application, Volume 39, Issue 2, pages 2127 – 2131.
45. Indranil B, Xi C, 2011, *A Fuzzy Clustering Based Analysis of Migratory Customer Behavior*, International Conference on ICCIS, pages 480-483.
46. Qiuru C, dkk, 2012, *Telecom Customer Segmentation Based On Cluster Analysis*, International Conference on CSIP, pages 1179 – 1182.
47. Ferreira D, 2009, *Applied Sequence Clustering Technique for Process Mining*, Handbook of Research Business Process Modelling, pages 492-513.
48. Ferreira D, 2007, *Approaching Process Mining with Sequence Clustering : Experimentals and Findings*, Lecture Notes in Computer Science, Vol. 4714, pages 360-374.
49. Cadez I, dkk, 2003, *Model-Based Clustering and Visualization of Navigation Patterns on a Web Site*, Data Mining and Knowledge Discovery, 7(4): 399-424.
50. Rabiner LR, 1989, *A Tutorial on Hidden Markov Models and Selected Applications Inspeech Recognition*, Proceedings of the IEEE, Vo;. 77 No. 2, pp 257 – 286
51. Dempster A, dkk, 1977, *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society, Series B, 39(1):1-38.
52. Finding Number of Clusters in Text Databases, http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set, diakses terakhir tanggal 20 Juli 2012.
53. L. Haibin, K. Vlado, *Combined mining of Web server logs and web contents for classfying user navigation patterns and predicting users' future request*, Data & Knowledge Engineering Journal, Vol. 61, pp. 304-330, 2007.

54. Yuhefizar, B. Santosa, I Ketut E.P, Y. K. Suprpto, *Combination of Hierarchical and Non-Hierarchical Cluster Method for Segmentation of Web Visitors, Telkomnika Journal, Vol. 11 No. 1, pp. 207-214, 2013.*
55. Pani S.K., Panigrahy L., Sankar V.H., Ratha B.K., Mandal A.K., Padhi S.K., *Web Usage Mining: A Survey on Pattern Extraction from Web Logs, International Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1. Pp 15 – 23, 2011.*
56. Mamoun A, Awad, I. Khalil, *Prediction of User's Web-Browsing Behavior: Application of Markov Model, IEEE Transactions on System, Man, And Cybernetics, Vol. 42, No. 04, pp 1131-1142, 2012*
57. J. Srivastava, R. Cooley, M. Deshpande, P-N Tan, *Web Usage Mining: Discovery and Application of Usage Patterns from Web Data, SIGKDD Exploration, Vo., 1, Issue 2, pp. 1-12, 2000.*
58. N. K. Tyagi, A.K. Solanki, S. Tyagi, *An Algorithmic Approach to Data Preprocessing in Web Usage Mining. International Journal of Information Technology and Knowledge Management, Volume 2, No. 2, pp. 279-283, 2010.*
59. R. Cooley, B. Mobasher, J. Srivastava, *Data preparation for mining World Wide Web browsing patterns, Knowledge and Information Systems, Vol. 1, No. 1, pp. 5-32, 1999.*
60. Youquan H., *Decentralized Association Rule Mining on Web Using Rough Set Theory, Journal of Communication and Computer, Volume 2, No.7, ISSN1548-7709, 2005.*
61. G. Castellano, A. M. Fanelli, M. A. Torsello, *Log Data Preparation for Mining Web Usage Patterns, IADIS International Conference Applied Computing, pp 371-378, 2007.*
62. Chen H. W., Zong X, Wei L.C., Haw Y.J., *World Wide Web Usage Mining Systems and Technologies, Journal Systemic, Cybernetic and Informatics, Volume ,1 No. 4, pp 53 – 59, 2004.*

63. Suneetha K.R., Krishnamoorthi D.R., *Identifying User Behavior by Analyzing Web Server Access Log File*, *IJCSNS International Journal of Computer Science and Network Security*, Vol .9, No.4, 2009.
64. Wahab M. H. A., Mohd M. N. H., *Data Preprocessing on Web Server Logs for Generalized Association Rules Mining Algorithm*, *World Academy of Science, Engineering and Technology*.

Halaman ini sengaja dikosongkan

Lampiran : Identitas halaman web berdasarkan nama variabel

<i>Var</i>	Halaman Web
<i>p1</i>	GET /semuaberita.php
<i>p2</i>	GET /personal/homebase.php
<i>p3</i>	GET /en/academic.php
<i>p4</i>	GET /en/index.php
<i>p5</i>	GET /en/about.html
<i>p6</i>	GET /en/services.php
<i>p7</i>	GET /index.php
<i>p8</i>	GET /en/admission.php
<i>p9</i>	GET /semuaaktivitas.php
<i>p10</i>	GET /tour_cmhs.php
<i>p11</i>	GET /en/news.php
<i>p12</i>	GET /sk_rektor_its/tahun_sk.php
<i>p13</i>	GET /en/research.php
<i>p14</i>	GET /en/library.php
<i>p15</i>	GET /sk_rektor_its/tahun_peraturan.php
<i>p16</i>	GET /rektor.php
<i>p17</i>	GET /sk_rektor_its/tahun_edaran.php
<i>p18</i>	GET /en/admission.html
<i>p19</i>	GET /en/map.php
<i>p20</i>	GET /tour_mhs.php
<i>p21</i>	GET /tour_ortu.php
<i>p22</i>	GET /tour_umum.php
<i>p23</i>	GET /personal/index.php
<i>p24</i>	GET /pengumuman/SK.html
<i>p25</i>	GET /tour_staf.php
<i>p26</i>	GET /en/library.html
<i>p27</i>	GET /en/sports.php
<i>p28</i>	GET /en/lc.php
<i>p29</i>	GET /personal/priv/index.php
<i>p30</i>	GET /personal/pengajaran.php

<i>Var</i>	Halaman Web
<i>p31</i>	GET /personal/publikasi.php
<i>p32</i>	GET /en/contact.html
<i>p33</i>	GET /personal/priv/profil.php
<i>p34</i>	GET /en/vision.html
<i>p35</i>	GET /en/history.html
<i>p36</i>	GET /personal/riset.php
<i>p37</i>	GET /en/objectives.html
<i>p38</i>	GET /personal/rekap_pub.php
<i>p39</i>	GET /sk_rektor_its/login.php
<i>p40</i>	GET /en/administrasi.php
<i>p41</i>	GET /personal/priv/publikasi.add.php
<i>p42</i>	GET /en/baak.html
<i>p43</i>	GET /personal/priv/publikasi.edit.php
<i>p44</i>	GET /en/scs.php
<i>p45</i>	GET /en/siskal.php
<i>p46</i>	GET /en/despro.php
<i>p47</i>	GET /en/geomatika.php
<i>p48</i>	GET /en/si.php
<i>p49</i>	GET /pengumuman/PPArs1.html
<i>p50</i>	GET /personal/credits.php
<i>p51</i>	GET /en/tekpai.php
<i>p52</i>	GET /p_links.html
<i>p53</i>	GET /en/arsitektur.php
<i>p54</i>	GET /en/elektro.php
<i>p55</i>	GET /en/hotspot.php
<i>p56</i>	GET /en/scs/guesthouse.html
<i>p57</i>	GET /personal/publikasi_penelitian.php
<i>p58</i>	GET /snmptn_link.html
<i>p59</i>	GET /en/cultural.php
<i>p60</i>	GET /en/facilities.html
<i>p61</i>	GET /en/kelautan.php
<i>p62</i>	GET /en/staff.html
<i>p63</i>	GET /en/tekkim.php

CURRICULUM VITAE



IDENTITAS DIRI

Nama	: Yuhfizar
NIP	: 19760113 200604 1 00 2
Tempat dan Tanggal Lahir	: Lubuk Jantan, 13 Januari 19i76
Jenis Kelamin	: Laki-laki
Status Perkawinan	: Kawin
Agama	: Islam
Golongan / Pangkat	: IIIId / Penata Tk. I
Jabatan Akademik	: Lektor Kepala
Perguruan Tinggi	: Politeknik Negeri Padang
Alamat	: Kampus Limau Manis – Padang
Telp./Faks.	: 0751-72590/ 0751-72576
Alamat Rumah	: Komp. Griya Kharisma Permai II Blok C9, Kel. Koto Lalang, Kec. Lubuk Kilangan - Padang
Handphone	: 08126777956
Alamat e-mail	: ephi.lintau@gmail.com

RIWAYAT PENDIDIKAN

1. SD Negeri 07, Desa Nusa Indah, Lintau, Batusangkar, 1989
2. MTSN Pekanbaru, 1993
3. MAN 1 Pekanbaru, Jurusan Biologi. 1996
4. D.1 Andalas Institusi Manajemen, Padang 1997
5. D.3 AMIK Jayanusa, Padang, 2001
6. S.1 STMIK Jayanusa, Padang, 2004
7. S.2 Magister Ilmu Komputer, UPI YPTK Padang, 2008
8. S.3, Mahasiswa Jurusan Teknik Elektro ITS sejak tahun 2010

RIWAYAT KEPANGKATAN/GOLONGAN

1 April 2006	: Calon Pegawai Negeri Sipil / III-a
1 April 2007	: Penata Muda / III-a
1 April 2011	: Penata Muda Tingkat 1 / III-b
1 April 2013	: Penata / III-c
1 April 2015	: Penata Tingkat I / III-d

RIWAYAT JABATAN FUNGSIONAL

1 Februari 2008	Asisten Ahli
1 Februari 2011	Lektor
1 Agustus 2014	Lektor Kepala

PUBLIKASI ILMIAH

1. Jurnal Internasional :
 - a. **Yuhefizar**, Budi Santosa, I Ketut Eddy P, Yoyon K. Suprpto, “*Two Level Clustering Approach For Data Quality Improvement in Web Usage Mining*”, Journal of Theoretical and Applied Information Technology, E-ISSN 1817-3195 / ISSN 1992-8645, Vol. 62 No. 2, Hal. 404 – 409, 2014. (**Terindeks Scopus**).
 - b. **Yuhefizar**, Budi Santosa, I Ketut Eddy P, Yoyon K. Suprpto, “*Combination Hierarchical and Non-Hierarchical Cluster Method For Segmentation of Web Visitors*”. TELKOMNIKA Indonesian Journal of Electrical Engineering. Vol :11 No. 1, 2013. (**Terindeks Scopus dan Terakreditasi A Dikti**).
2. Jurnal Nasional
 - a. **Yuhefizar**, Yoyon K. Suprpto, Mochamad Hariadi, I Ketut Eddy P., “*Preprocessing Data Web Log Untuk Klasterisasi Pengguna Web Menggunakan Algoritma K-Means*”, Jurnal JAVA EE ITS. Vol 8 No. 1, 2010.
 - b. **Yuhefizar**, Mochamad Hariadi, Yoyon K. Suprpto, “*Peringkat Website Perguruan Tinggi Berbasis Analisa Hyperlink Menggunakan Factor Analysis*”, Jurnal Ilmiah Kursor, Vol. 6 No. 1 Januari 2011, Universitas Trunojoyo Madura, (**Terakreditasi B Dikti**).
3. Seminar
 - a. **Yuhefizar**, Yoyon K. Suprpto, Mochamad Hariadi, I Ketut Eddy P., “*Pre-Processing Data Web Log Menggunakan Pendekatan Query*”, Prosiding Seminar Nasional Sistem Informasi Indonesia, Jurusan Sistem Informasi Institut Teknologi Sepuluh Nopember, Surabaya, 3 Desember 2011.
 - b. **Yuhefizar**, Budi Santosa, I Ketut Eddy P. Yoyon K. Suprpto, “*Klasterisasi Pengunjung Web Berdasarkan Data Web Log Menggunakan Metoda K-Means*”, Prosiding Seminar Nasional Matematika dan Pendidikan Matematika Universitas Andalas, Padang, 31 Oktober 2012