JURNAL TEKNOLOGI TECHNOSCIENTIA ISSN: 1979-8415
Vol. 3 No. 1 Agustus 2010

SISTEM TRAFIK NAVIGASI WEB MENGGUNAKAN METODE SEQUENTIAL PATTERN

Uning Lestari[1]

[1]Jurusan Teknik Informatika, Institut Sains & Teknologi AKPRIND Yogyakarta

Masuk: 19 Mei 2010, revisi masuk : 21 Juni 2010, diterima: 11 Juli 2010

ABSTRACT

Analysis performed on web server logs have traditionally been widely used to see the web visitor activities such as determining the amount of access to web pages, anyone accessing the web user and the time to visit the web and visited URLs and more statistical analysis. But it is still very little information about the relationship between existing data on the web, the sequence of requests made by a user or group of users who have access to a web site. To overcome these problems required a web mining techniques. Web mining is a technique to automatically discover and extract information on the web. This study used different techniques to sequential pattern Apriori algorithm to identify patterns of web site navigation. The the applications that have been made will show recapitulation of information about the number of visits per day, a search of browser used, details of visits that include IP addresses, proxy server, the location of the sequence pattern visitor and web traffic. Web navigation patterns using Apriori is able to produce the sequence pattern of answering pages accessed by visitors. This pattern always varies depending on the minimum support are included. The higher MinSup, so little sequence variation pattern of visits and reverse the lower the value of the Minsup more and more variations of the sequence pattern of visits by users.

Keywords: Sequential Pattern, Apriori Algorithm, web navigation

INTISARI

Analisa yang dilakukan pada web server log secara tradisional telah banyak digunakan untuk melihat aktifitas pengujung web seperti menentukan jumlah akses terhadap halaman web, siapa saja user yang mengakses web dan waktu mengunjungi web dan URL yang dikunjungi dan masih banyak lagi analisa secara statistik. Tetapi hal tersebut masih sangat sedikit memberikan informasi tentang hubungan antara data yang ada dalam web, urutan kunjungan yang dilakukan oleh seorang user atau siapa saja kelompok user yang melakukan akses ke suatu situs web. Untuk mengatasi permasalahan tersebut dibutuhkan teknik *web mining.* Web mining merupakan teknik untuk secara otomatis menemukan dan mengekstrak informasi dalam *web*. Dalam penelitian ini digunakan teknik *sequential pattern* dengan algoritma Apriori untuk mengetahui pola navigasi dari sebuah web site. Dari aplikasi yang telah dibuat akan mengasilkan informasi tentang rekapitulasi jumlah kunjungan per hari, pencarian browser yang digunakan, detail kunjungan yang meliputi alamat IP, proxy server, lokasi keberadaan pegunjung dan pola urutan kunjungan web. Pola navigasi web dengan metode Apriori mampu menghasilkan pola urutan halaman wab yang diakses oleh pengunjung. Pola ini selalu bervariasi tergantung dari *minimum support* yang dimasukkan. Semakin tinggi *Min-Sup*nya maka akan semakin sedikit variasi pola urutan kunjungan dan sebaliknya semakin rendah nilai *Minsup*-nya semakin banyak variasi pola urutan kunjungan user.

Kata kunci: Sequential Pattern, Algoritma Apriori, avigasi web

PENDAHULUAN

Teknologi *web* dalam sepuluh tahun terakhir merupakan teknologi yang paling mempengaruhi kehidupan masyarakat dunia. Pertumbuhan yang eksplosif dari *World Wide Web* telah memungkinkan tersedianya sumber informasi *on-line* yang sangat besar. Pada prinsipnya, ad-

[1] uning@akprind.ac.id

ministrasi web tidak hanya sebatas pada pengelolaan data dilakukan secara dinamis. Pengelolaan data pengunjung merupakan hal penting yang perlu diperhatikan.

Web Mining didefinisikan sebagai kajian ini tentang teknik-teknik untuk secara otomatis menemukan dan mengekstrak informasi dalam *web*. *Web pada server log* merupakan sebuah data yang sangat berharga bagi sebuah organisasi yang melakukan aktivitasnya di dalam web. Karena jumlah data yang sangat besar maka sebuah organisasi perlu melakukan analisa terhadap data tersebut sehingga informasi tersembunyi dapat diperoleh. Analisa ini diharapkan akan memberikan informasi bagaimana melakukan restrukturisasi web site agar dapat meningkatkan efektifitas, memberikan layanan komunikasi yang lebih baik, dan juga dapat untuk meningkatkan target tertentu terhadap sebuatu kelompok pemakai tertentu. Pemantauan sebuah web menjadi hal penting dan wajib diperhatikan oleh setiap *web developer*. Pekerjaan pemantauan web pada kenyataannya sangat kompleks dan tidak bisa dilakukan secara manual, dibutuhkan berbagai *tools* (alat/aplikasi/program) pembantu yang dapat melakukan tugas ini secara otomatis.

Penelitian bidang pemanfaatan informasi dari log data saat ini umumnya menggunakan dua pendekatan, yaitu pertama memetakan data navigasi ke bentuk table-tabel relasi kemudian menggunakan teknik baku dalam data mining seperti associaton rule dan sequential pattern (Gunawan, 2000) dan kedua, pendekatan langsung yang dapat diterapkan pada rekaman *log-data* dengan memodelkan rekaman navigasi user sebagai *hypertext probabilistic grammar*, dimana grammar dengan nilai probabilitas besar akan membangkitkan string yang mewakili jejak akses yang paling banyak diakses user. Cooley.et.al (1997) telah membuat arsitektur umum dari *Web Usage Mining* yang disebut WEBMINER. WEBMINER mendukung proses-proses pembersihan data (*data cleaning*) yang secara otomatis menemukan pola dan telah memperkenalkan query yang mendukung proses analisa terhadap pola yang dihasilkan. Algoritma yang digunakan menggunakan data mining yaitu *Association Rule* dan *Sequential Pattern*

Data mentah dapat dikumpulkan secara dinamis dari sebuah halaman web dan disimpan dalam sebuah basis data. Dari data ini, dapat di*generate* secara *real time*, di-*update* untuk tiap pengunjung dan tidak hanya dibuat pada akhir bulan atau akhir tahun. Hasil ini dapat digunakan untuk menunjukkan peningkatan trafik/lalu-lintas atau juga rasio kunjungan pengunjung. Perhitungan trafik web menyajikan berbagai detail mengenai situs yang dipantau, mulai dari informasi pengunjung terakhir hingga ke ringkasan kunjungan per bulan sejak web dipublikasikan. Adapun detail yang dimaksud antara lain alamat IP (*Internet Protocol*), *proxy server*, *browser* yang digunakan, lokasi keberadaan pengunjung yang mengakses halaman web serta pola urutan kunjungan web.

Dari latar belakang masalah ini yang sudah dijelaskan terlihat bahwa banyak masalah dalam *web* khususnya dalam pencarian pola navigasi yang berhubungan dengan user. Masalah tersebut dikarenakan minat user bermacam-macam dan selalu berubah tiap waktu sehingga menyulitkan pengelola dalam memantau trafik web dan mengenali pola pengunjung. Dalam penelitian ini dibuat sistem aplikasi yang dapat menganalisis trafik/lalu lintas kunjungan web untuk mencari pola navigasi yang berguna untuk administrator web tersebut.

Proses Sequential Pattern, Pada proses ini merupakan pencarian pola kunjungan dengan metode Sequential Pattern ini, istilah *session* sering disebut dengan *sequence.* Pencarian pola dengan metode Apriori dimulai dengan menentukan minimum supportnya (minsup), kemudian menentukan *large itemset*-nya yaitu itemset yang memiliki support diatas minimum supportnya.

Menurut Agrawal., Srikant (1995) sebuah *sequence* didfinisikan sebagai urutan dari *itemset.* Sebuah *itemset* dinotasikan dalam $(i_1, i_2, \ldots i_j)$ dimana i_j adalah sebuah item. Sebuah *sequence* mungkin berada dalam *sequence* yang lainnya. Sebagai contoh missal ter-

dapat *sequence* { (3) (4 5) (8)} berada pada *sequence* { (7) (3 8) ((9) (4 5 6) (8)} karena (3) $\subseteq$ (3 8), (4 5) $\subseteq$ (4 5 6), (8) $\subseteq$ (8).

Perlu diperhatikan bahwa { (3) (5)} $\not\subset$ { (3 5)} yang pertama berarti bahwa item 3 dibeli setelah item 5 sedang yang kedua berarti item 3 dan item 5 dibeli secara bersama-sama. Dalam sekumpulan *sequence*, sebuah *sequences*, adalah maksimal jika *s* tidak berada di dalam s*equence* yang lain. *Support* dari *sequence* menurut Agrawal (1995) didefinisikan sebagai fraction dari total pelanggan yang memberikan *Support* terhadap *sequence.*

Permasalahan utama ini dalam melakukan *mining sequential pattern* adalah menemukan maksimal *sequence* diantara seluruh *sequence* yang memiliki *mini-mum support* tertentu. Untuk setiap maksimal *sequence* yang memenuhi akan mewakili sebuah *sequential pattern.* Proses untuk menemukan pola navigasi pengunjung web menggunakan algoritma Apriori All terdiri 5 tahap yaitu: ProsesTahap *Sort* yaitu dengan melakukan *sorting* terhadap basis data dengan kunci primer adalah identitas dari pelanggan/ user sedang kunci sekundernya adalah waktu kunjungan. Proses Tahap Pencarian Large Itemset (l-itemset) *Large Itemset* merupakan *itemset* yang memiliki *support* lebih besar dari *minimum support*nya. Kumpulan *itemset* yang ditemukan akan dipetakan ke dalam urutan integer. Proses Tahap *Transformasi,* dalam sebuah transformasi *customer sequence*, setiap transaksi diganti dengan bagian dari seluruh *l-itemset* yang terdapat dalam transaksi. Jika pada sebuah transaksi tidak terdapat *l-iemset* maka *sequence* dihapus dalam basis data trans-formasi.Jika pada sebuah *customer sequence* tidak terdapat dalam *l-itemset*, maka juga akan dihapus dalam basis data transformasi. ProsesTahapan *sequence,* tahapan ini digunakan algoritma untuk pencarian urutan kunjungan*.* Algoritma yang digunakan adalah algoritma *AprioriAll* (Agrawal, 1995). Proses Tahap *Maximal Sequence,* ini merupakan tahapan untuk menemukan *Maximal Sequence* diantara *l-itemset*. Untuk algoritma AprioriAll tahapan ini digabungkan pada tahapan *sequence,* hal ini dilakukan dengan mengurangi waktu yang diperlukan untuk menghitung non maksimal *Sequence*.

Metode penelitian, data penelitian diambil dari web server log salah satu web site pendidikan yaitu web site IST AKPRIND dengan alamat situs : www.akprind.ac.id. Adapun batasan-batasan dari sistem ini adalah : Sistem dapat mencari pola urutan kunjungan yang dilakukan oleh user; Data dari web server log yang dipilih hanya dilihat dari halaman yang dikunjungi dan tidak memperhatikan lama waktu kunjungan; Rekapitulasi jumlah kunjungan per hari; Pencarian browser yang digunakan; Detail kunjungan meliputi alamat IP, proxy server, lokasi keberadaan pegunjung; Pola urutan kunjungan web.

Pre Processing, proses persiapan yang dilakukan dengan serangkaian proses-proses. Diawali dengan membuka file *web server log* yang asli. Banyaknya file web server log dapat lebih dari satu. File *web server log* sementara disimpan dalam file yang diberi nama *logtemp*, yang gunanya menampung seluruh data *web server log*. Kemudian lakukan persiapan suatu file untuk menampung hasil proses persiapan dan juga file yang menampung data yang tidak berhasil dibersihkan (*cleaning*). File yang berisi data yang valid disimpan dalam tabel LogImage. Selanjutnya File *web server log* dibaca perbarisnya kemudian dilakukan pemisahan sekaligus juga melakukan proses cleaning. Kemudian hasilnya disimpan dikedua file yang telah dipersiapkan. Proses pemisahan dilakukan dengan cara membaca satu demi satu karakter pada setiap baris dari *web server log* dan juga suatu *string* pembatas.

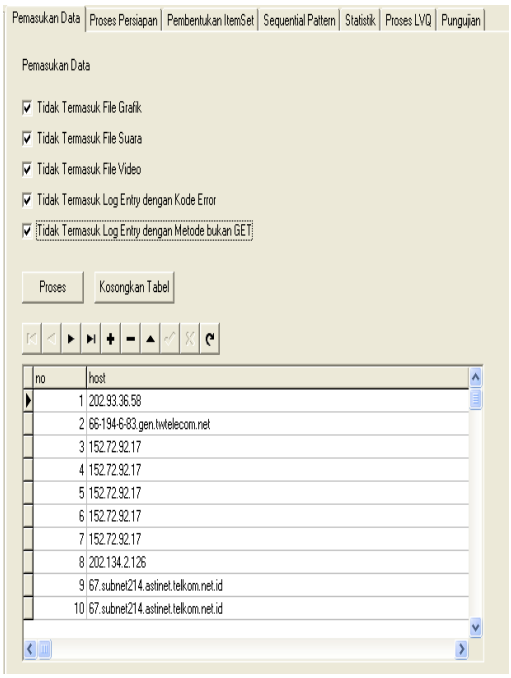
Pemisahan ini perlu dilakukan hati-hati mengingat karakter pembatas (*delimeter*) dari setiap field dari *web server log*. Bersamaan dengan melakukan proses pemisahan, dapat diketahui jenis dari *web server log* karena yang membedakan antara ECLF dan CLF hanyalah apakah mempunyai *agent* dan *referrer* atau tidak. Kalau jenis CLF tidak memiliki *agent* dan *referrer* maka keduanya akan ditandai dengan karakter '-" se-

dangkan jika ECLF tidak mempunyai *agent* dan *referrer*.

Selanjutnya proses pemisahan tersebut dibarengi dengan proses *cleaning*. Harus diuji dari setiap *request*nya yaitu dengan memperhatikan ekstensi (suffiknya). Sebagai contoh bila tidak diinginkan file grafik maka file-file dengan ekstensi gif, jpeg, jpg, bmp dan lainnya akan dihilangkan. Selain memperhatikan request juga diperhatikan apakah merupakan baris yang tidak error. Baris yang error ditandai dengan kode status dan requestnya bukan GET. Melakukan transfer dari file teks LogTemp ke file database.

PEMBAHASAN

Item set adalah kumpulan dari item-item , yaitu nama URL hasil request user. Pembentukan itemset menunjukkan pembentukan session. Satu *session* adalah urutan URL yang diakses dalam satu kunjungan tunggal.

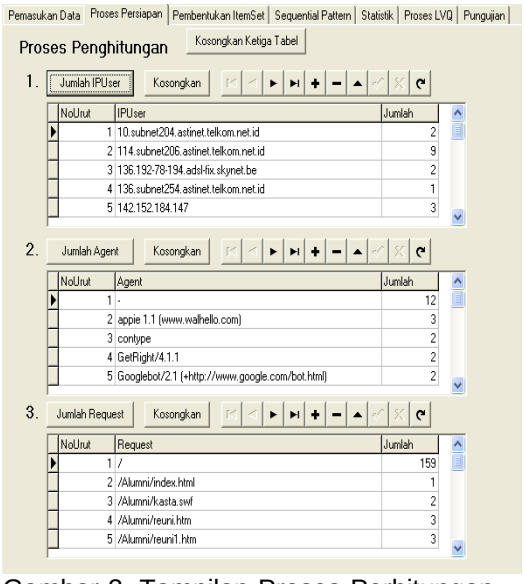


Gambar 1. Tampilan Pemasukan data

Sebelum dilakukan proses pembentukan itemset terlebih dahulu database harus diurutkan berdasarkan IP User, agent serta Tanggal. Hasil pengurutan ini yang akan dijadikan dasar pembentukan itemset. Cara pembentukan itemset yaitu dengan melakukan pengelompokan berdasarkan IP user, agent dan tanggal akses. Jika IP User berbeda dianggap sebagai session yang berbeda. Sering satu buah IP digunakan bersama-sama oleh banyak user, maka akan di cek jika agent nya brbeda, maka dianggap *session* yang berbeda. Tetapi jika agentnya sama maka dianggap *session* yang sama

Proses Pemasukan Data, sistem pencarian pola kunjungan web memerlukan *pre processing* sebelum melakukan proses berikutnya. Pre processing tersebut meliputi proses pengambilan data dari log server data dan pembersihan data (*cleaning*) sehingga data siap untuk digunakan.. Tampilan proses ini dapat dilihat pada Gambar 1.

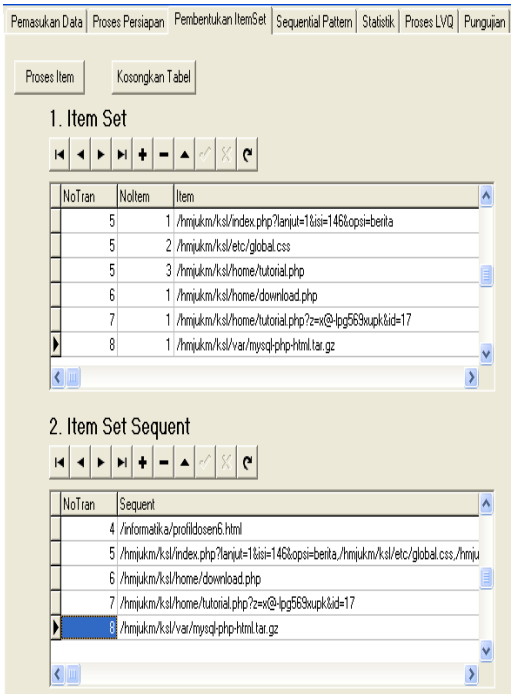
Proses Perhitungan Data, proses persiapan data ini dengan mencari jumlah seluruh jumlah IP User, jumlah agent dan seluruh jumlah request dari data yang sudah dinputkan dan dibersihkan. Tampilan proses dapat dilihat pada Gambar 2.



Gambar 2. Tampilan Proses Perhitungan Data

Proses Pembentukan Itemset, pembentukan itemset ini merupakan data inti dari proses pencarian urutan kunjungan user web. Ada 2 proses yaitu mencari item set seluruh akses pengunjung dan mencari sequence dari itemset. Dari sequence ini nantinya digunakan untuk menghitung support

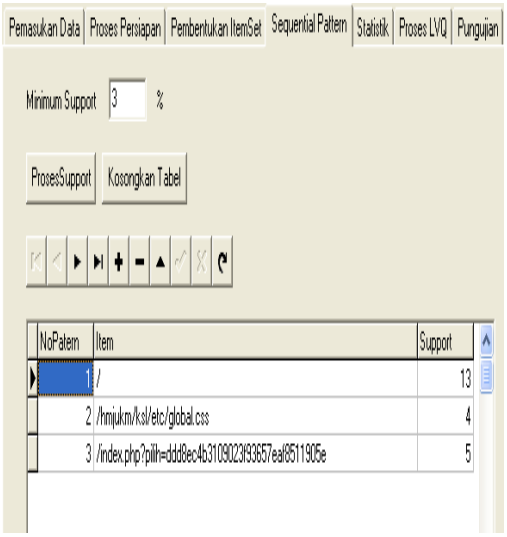
setiap sequence. Tampilan proses ini terlihat pada Gambar 3.



Gambar 3. Proses Pembentukan Item Set

Proses Sequential Pattern, setelah sequence didapatkan, dilanjutkan dengan menghitung support setiap sequence yang muncul. Proses ini dapat dilihat pada Gambar 4.

Dari hasil kerja metode LVQ ini, dapat dilihat bahwa tidak seperti kebanyakan algoritma pembelajaran terawasi dimana neuron bekerja dengan cara memproses penjumlahan terbobot (input dikalikan dengan bobot), pada metode LVQ ini neuron-neuron bekerja dengan mencari jarak normal Euclidian antara input dengan bobot-bobot yang bersangkutan.



Gambar 4. Tampilan Proses Sequential Pattern

Tabel 1 Tabel Hasil Sequential Pattern

MinSup (%)	Jumlah Pola Sekuential
1	14
2	4
3	3
4	2
10	1

NoPatern	Item	Support (%)
5	/hmjukm/ksl/home/tutorial.php	2
6	/hmjukm/ksl/index.php	1
7	/hmjukm/ksl/var/Modul_data_base_tg.pdf	1
8	/index.php	1
9	/index.php?pilih=09e0d14a6bd0dd3cb98867c690df1842	1
10	/index.php?pilih=22316aad943f045e9e7061a387fc0be9	1
11	/index.php?pilih=ddd8ec4b3109023f93657eaf8511905e	4
12	/pmb/index.php?pilih=ba4006333d90893962cebf16bf4821f6	1
13	/profil.php?pilih=86fb81d511935f66d22a98f0e33e5faa	1
14	/seminar/Pengumuman_Abtrak_SNAST03.pdf	2

Gambar 5. Hasil Sequence dengan minsup 1%

NoPatern	Item	Support (%)
1	/	12
2	/hmjukm/ksl/etc/global.css	4
3	/index.php?pilih=ddd8ec4b3109023f93657eaf8511905e	4
4	/seminar/Pengumuman_Abtrak_SNAST03.pdf	2

Gambar 6 . Hasil Sequence dengan minsup 2%

Sebagian contoh hasil sequential pattern dapat dilihat pada Gambar 5. untuk minimum support 1% . Arti dari baris pertama adalah 2% user melakukan kunjungan dengan urutan akses /hmjukm/ksl/home/tutorial.php. Pada Gambar 6. baris pertama memperlihatkan 12% user melakukan kunjungan dengan akses indeks http: //www. Akprind.ac.id

Pada percobaan pelatihan ini diambil data dari log server IST AKPRIND dengan format ECLF. Data percobaan sistem diambil selama 12 hari dari tanggal 1–12 September 2007. Setelah dilakukan pengolahan terhadap data tersebut didapat hasil sebagai berikut: Jumlah seluruh transaksi selama 11 hari adalah 4314 transaksi kunjungan web, Jumlah seluruh IP User ada 793 IP User, Jumlah seluruh Agent ada 227 Agent, Jumlah seluruh request ada 577 request, dan Jumlah item set sequence ada 1846 sequence. Hasil pencarian pola dengan Sequential Pattern dengan contoh nilai minimum support yang berbeda dapat dilihat pada Tabel 1

KESIMPULAN

Pola Kunjungan web dengan metode Apriori mampu menghasilkan pola urutan halaman wab yang diakses oleh pengunjung. Pola ini selalu bervariasi tergantung dari *minimum support* yang dimasukkan. Semakin tinggi *MinSup*-nya maka akan semakin sedikit variasi pola urutan kunjungan dan sebaliknya semakin rendah nilai *Minsup*nya semakin banyak variasi pola urutan kunjungan user.

DAFTAR PUSTAKA

Agrawal, R, Srikant, R., 1995, *Mining Sequential Patterns*, Proc. of the Int'l Conference on Data Engineering (ICDE), Taipei, Taiwan (ICDE 1995).

Borges, J.L.C, 2000, *A Data Mining Model to Capture User Web Navigation Pattern*, PhD Thesis, Departement of Computer Science, Universitas College London

Cooley, R., Mobaster, B and Srivastava,J., 1997, *Web Mining Information and Pattern Discovery on the World Wide Web*, Departement of Computer Science and Engineering.

Gunawan, R., 2002, *Pencarian Pola Navigasi dari Data Web dengan Teknik Association Rule dan Sequential Pattern*, Thesis Magister Teknik, Teknik Elektro, Universitas Gadjah Mada, Yogyakarta.

Rafiudin, R., 2004, *Panduan Menjadi Seorang Webmaster*, Andi, Yogyakarta