ANALISIS DAN IMPLEMENTASI WEB USAGE MINING DENGAN MENGGUNAKAN ALGORITMA C5.0 (STUDY KASUS PADA WEB STUDENT IT TELKOM)

Danmaseka Maryowati[1], Eko Darwiyanto, ST., MT.[2], Erda Guslinar, ST.[3]

[1,2,3]Program Studi Teknik Informatika Institut Teknologi Telkom, Bandung

[1,2,3]Fakultas Informatika – Institut Teknologi Telkom
Jl. Telekomunikasi, Dayeuh Kolot Bandung 40257 Indonesia
danmaseka@gmail.co.id[1], ekd@ittelkom.ac.id[2], egp@ittelkom.ac.id[3]

Abstrak

***World Wide Web* atau WWW merupakan salah satu fenomena teknologi yang berkembang sangat pesat saat ini. WWW menyediakan berbagai layanan informasi mengenai berita, iklan, pendidikan, e-commerce dan sebagainya. Informasi yang tersedia dalam WWW tersebut memiliki ukuran yang sangat besar dan terdistribusi secara global di seluruh dunia.Web juga mengandung kekayaan informasi dilihat dari struktur dan penggunaannya (*web usage*). Web merupakan kumpulan data dan informasi yang sangat berpotensi untuk dilakukan penggalian (*mining*) agar menghasilkan pengetahuan (*knowledge*) yang dapat berguna bagi masyarakat maupun pihak-pihak tertentu.**

Algoritma C5.0 merupakan algoritma untuk mengklasifikasikan dengan menghasilkan *decision tree*. Pemilihan atribut yang akan diproses menggunakan ukuran *information gain*. Atribut dengan nilai *information gain* tertinggi akan terpilih sebagai *parent* bagi *node* selanjutnya. Algoritma ini membentuk pohon keputusan dengan cara pembagian dan menguasai sampel secara rekursif dari atas ke bawah. Untuk memudahkan pengguna informasi dalam menafsirkan terhadap hasil klasifikasi C5.0 disajikan dalam dua bentuk, menggunakan pohon keputusan dan sekumpulan aturan IF-T HEN yang lebih mudah untuk dimengerti.

Berdasarkan hasil analisa yang telah dilakukan dapat diketahui bahwa akurasi untuk *tree* yang digenerate dari data training menghasilkan akurasi kurang baik terhadap data testing, oleh karena itu dilakukan proses *pruning*. Rule yang dihasilkan setelah proses pruning memiliki akurasi lebih baik terhadap data testing dan memiliki simplisitas aturan yang rendah, sehingga dihasilkan aturan yang lebih sederhana dibandingkan *tree* sebelumnya. Pola akses dari user yang telah terklasifikasi kurang memberikan perbedaan yang signifikan hal ini disebabkan user pengakses web tersebut memiliki kepentingan terhadap informasi yang hampir sama. Dari hasil klasifikasi ini dapat diperoleh feedback terhadap admin web untuk peningkatan performansi web dalam hal navigasi.
Kata kunci : web usage,* C5.0*, decision tree, information gain, pruning

Abstract

World Wide Web or WWW is one of the technological phenomenon that is growing very rapidly at this time. WWW provides a range of information services about news, advertising, education, e-commerce and etc. The information available on the WWW has a very large size and distributed globally in the world. Web also contains a wealth of information seen from the structure and *web usage*. Web is a collection of data and information that is potentially to mining in order to generate knowledge that can be useful for the community as well as certain parties.

C5.0 algorithm is an algorithm to generate *decision tree* to classify. Selection of attributes that will be processed using a measure of *information gain*. Size *information gain* is used to select the test attribute at each node in the tree. Attributes with highest *information gain* value will be selected as the parent for the next node. These algorithms form the *decision tree* by way of division and mastering the sample recursively from top to bottom. The users get interpreting information on the *C5.0* classification results are presented in two forms, using a *decision tree* and a set of IF-T HEN rules that are easier to understand.

Based on the results of the analysis has been done can be seen that accuracy for the tree that are generated from training data produces poor accuracy on testing data, therefore a process of *pruning*. Rule generated after *pruning* process has better accuracy on testing data and have the low simplicity of rules, so that the resulting rules are much simpler than the previous tree. Access patterns of users who have been classified not give a significant difference it is because the user accessed the web has an interest in almost the same information. From the results of this classification can be obtained feedback on the web admin for increased performance in terms of web navigation.

Keywords: *web usage, C5.0, decision tree, information gain, pruning*

1. Pendahuluan

Aplikasi web banyak digunakan oleh perusahaan-perusahaan, sekolah-sekolah, perguruan tinggi, dan lembaga atau organisasi lainnya dalam kegiatan penjualan, promosi, belajar dan kegiatan lainnya dimana dibutuhkan pengiriman, penyebaran dan penerimaan informasi sehingga memberikan kemudahan bagi pengguna (*user*) yang membutuhkan. Kenyataannya ada beberapa kelompok user yang masih sulit dalam memperoleh informasi yang tepat pada suatu web, salah satunya dikarenakan navigasi yang rumit dalam mencapai suatu informasi yang terdapat dalam suatu web.

Dengan adanya masalah tersebut maka perlu diimplementasikan suatu teknik untuk mengenali perilaku user *(usage)* dan struktur web melalui informasi tertentu sehingga akan dihasilkan website yang lebih informatif yaitu memberikan nilai guna secara tepat bagi user dengan menempatkan informasi yang sangat diperlukan oleh user sehingga informasi dapat tepat sampai kepada user tanpa dibingungkan oleh hyperlink yang rumit.

Dalam Tugas Akhir ini dibangun suatu sistem Web Usage Mining untuk web student IT Telkom dengan memanfaatkan algoritma C5.0 dalam pengklasifikasian data penggunaan (usage), sehingga dari hasil pola akses user dapat memberikan feedback kepada developer, sebagai informasi dalam meningkatkan performansi terutama dalam hal navigasi web.

2. Landasan Teori

2.1. Web Mining

Web mining merupakan aplikasi teknik data mining untuk mengekstrak pengetahuan (*knowledge*) dari data web [26]. *Web mining* dapat dibagi dalam tiga kategori berdasarkan jenis data yang diekstrak, yaitu:

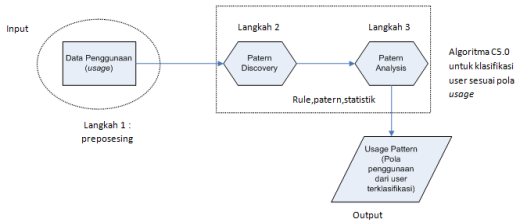
a. *Web content mining* (WCM) Merupakan discovery informasi terhadap content web, yang terdiri dari text, image, audio, video, metadata, dan hyperlinks.

b. *Web structure mining* (WSM) Merupakan *discovery* model yang berkaitan dengan struktur hubungan web yang meliputi *intra-page structure* dan *inter-page structure*.

c. *Web Usage Mining* (WUM) Merupakan proses untuk mengaplikasikan teknik *data mining* dalam melakukan *discovery* terhadap pola penggunaan *(usage pattern)* dari data web. WUM mengenerate data dari *session* dan *behavior user* dalam berinteraksi dengan data web.

2.2. Web Usage Mining

Web usage mining adalah teknik *data mining* dalam melakukan *discovery* atau pencarian terhadap data akses dari data suatu web. Data yang menjadi inputan dari WUM adalah data dari web server seperti *access logs, browser logs, user profiles, registration data, user session* atau *transaction, cookies, user queries, mouse clicks*, dan data lainnya sebagai hasil transaksi dengan web. Sedangkan output dari WUM adalah *usage pattern* atau pola penggunaan *user* terhadap web, sehingga dengan pola tersebut dapat berfungsi untuk pengembang dari web tersebut. Arsitektur web usage mining secara umum adalah sebagai berikut:



2.3. Algoritma C5.0

Algoritma C5.0 merupakan merupakan penyempurnaan dari algoritme terdahulu yang dibentuk oleh Ross Quinlan pada tahun 1987, algoritma ini dikembangkan dari algoritma sebelumnya yaitu algoritma ID3 dan C4.5. Dalam algoritma C5.0, pemilihan atribut yang akan diproses menggunakan ukuran information gain. Ukuran information gain digunakan untuk memilih atribut uji pada setiap node di dalam *tree*. Ukuran ini digunakan untuk memilih atau membentuk node pada pohon. Atribut dengan nilai information gain tertinggi akan terpilih sebagai parent bagi node selanjutnya. Algoritma ini membentuk pohon keputusan dengan cara pembagian dan menguasai sampel secara rekursif dari atas ke bawah. Algoritma ini dimulai dengan semua data yang dijadikan akar dari pohon keputusan sedangkan atribut yang dipilih akan menjadi pembagi bagi sampel tersebut.

Pseucode algoritma C5.0 adalah sebagai berikut :

1. Pada tahap awal, *tree* digambarkan sebagai node tunggal yang merepresentasikan training set.
2. Jika sampel seluruhnya berisi kelas yang sama, maka node tersebut menjadi *leaf* dan dilabeli dengan kelas tersebut.
3. Jika tidak, algoritma dengan menggunakan ukuran berbasis entropi (information gain) akan memilih variabel prediktor yang akan memisahkan *record* ke dalam kelas-kelas individual. Variabel tersebut menjadi variabel tes atau keputusan pada node tersebut.
4. Cabang dikembangkan untuk tiap nilai yang diketahui dari variabel tes, dan sampel dipartisi berdasarkan cabang tersebut.
5. Algoritma menggunakan proses yang sama secara rekursif membentuk *decision tree*.
6. Partisi rekursif berakhir hanya ketika satu dari kondisi-kondisi berikut terpenuhi:

a. Seluruh *record* pada node tertentu memiliki kelas yang sama.
b. Tidak ada atribut yang tersisa pada record yang dapat dipartisi lebih lanjut. Dalam kasus ini suara mayoritas digunakan. Node tersebut menjadi leaf node dan dilabeli dengan kelas yang menjadi mayoritas dalam record yang ada. Tidak ada record untuk cabang variabel tes. Dalam kasus ini, leaf terbentuk dengan mayoritas kelas sebagai label record tersebut.

3. Perancangan Sistem

Sistem yang dibangun pada Tugas Akhir ini adalah sistem untuk menentukan pola penggunaan user terhadap suatu web (*Web usage mining*) dengan menggunakan algoritma C5.0 dalam proses klasifikasi user berdasarkan pola aksesnya. Studi kasus yang digunakan dalam menganalisa sistem ini menggunakan web student IT Telkom. Data yang digunakan adalah data yang berasal dari access log web tersebut. Proses dalam membangun sistem pengklasifikasian Web Usage Mining dengan algoritma C5.0 adalah

3.1. Preprocessing

Tahapan dalam melakukan preprocessing data untuk keperluan pengolahan data dalam menemukan pola akses pada WUM adalah

a. *Load Data*: pengambilan data history dari access log.
b. *Read Data & choose data* : Pembacaan File dan pemilihan data yang akan dipakai dengan melakukan filter data, menghilangkan noise dan data yang tidak relevan. Pada proses ini dilakukan data cleaning, transformation dan reduction.
c. *Statistika:* Menghitung kemunculan konten untuk menentukan statistika awal dari distribusi data sebelum masuk ke dalam proses klasifikasi.
d. *Session identification*: mengklasifikasikan data akses berdasarkan *user session*nya, untuk diketahui *path* akses dari masing-masing user.

3.2. Pattern Discovery

Pada tahap *pattern discovery* dilakukan proses identifikasi path akses dan menjadikannya sebagai inputan untuk klasifikasi dengan memberikan label kelas terhadap setiap data *record* yang akan dijadikan dataset. Proses pada tahap *pattern discovery* adalah sebagai berikut :

a. Path Identification:Mengurutkan konten yang diakses oleh user dalam 1 session.
b. Pemberian label *class* untuk masing-masing *record* dataset.
c. Pengklasifikasian dalam menentukan kategori user dan prioritasnya dengan menggunakan algoritma C5.0.

3.3. Pattern Analysis

Pada tahap ini dilakukan pengujian terhadap pattern klasifikasi yang sudah ditemukan pada proses sebelumnya, dengan menampilkan semua *rule* yang dihasilkan dari klasifikasi. Prosesnya adalah sebagai berikut :

a. Testing rule yang dihasilkan dari proses sebelumnya dengan menggunakan data testing.
b. Pruning rule yang dihasilkan pada tahap pattern discovery apabila akurasi terhadap data testing kurang baik.
c. Result & feedback: Analisa untuk dari hasil rule sebagai feedback untuk pengembang web dalam membuat suatu keputusan salahsatunya untuk performansi navigasi website

4. Pengujian dan Analisa

4.1. Tujuan Pengujian

Pengujian dilakukakan untuk menghitung akurasi dari *rule* yang telah dihasilkan dalam proses klasifikasi, dimana tingkat akurasi berbanding terbalik dengan nilai error yang dihasilkan. Pengujian yang dilakukan terhadap data training maupun testing adalah menggunakan pendekatan akurasi prediktif dan simplisitas aturan.

4.2. Analisa akurasi prediktif

Akurasi data testing akan dihitung berdasarkan jumlah nilai yang sama antara kelas dalam test set dengan kelas yang dihasilkan oleh pohon dengan menggunakan keseluruhan aturan yang terbentuk, perhitungan akurasi adalah sebagai berikut [28]:

$$\text{Akurasi} = \frac{\text{kasus}_{\text{benar}}}{\text{kasus}_{\text{benar}} + \text{kasus}_{\text{salah}}} \qquad (4,1)$$

Atau

$$Akurasi = \frac{f11 + f00}{f01 + f11 + f10 + f00} \qquad (4,2)$$

Setiap entri fij menyatakan banyaknya *record* dari kelas i yang diprediksi menjadi kelas j.

4.2.1. Pengujian sebelum pruning

Perhitungan akurasi untuk data training sebanyak 60% dari data acces log menghasilkan data sebanyak 2541 record dan data testing sebanyak 10% dari data acces log menghasilkan data sebanyak 420 record.

Pengujian akurasi sebelum pruning menggunakan threshold information gain yaitu nilai yang diberikan untuk membatasi *expand*/pertumbuhan *node* pada saat pembentukan tree. Pengujian akurasinya dengan

memberikan *threshold information gain* sebagai berikut :

Threshold	Jumlah *rule* (aturan)	Akurasi train (%)	Akurasi testing (%)
0%	998	86,15	67,38
10%	955	85,95	67,38
20%	717	84,14	68,57

Dari hasil percobaan dengan mengubah nilai *threshold* dari information gain hanya untuk menyederhanakan *tree* dan aturan *(rules)* tetapi kurang bisa memperbaiki akurasi data training maupun data testing. Dari hasil pengujian nilai *threshold* menghasilkan *rules* yang baik dalam meng-*cover* kelas yang memiliki kelas mayor dan kurang baik untuk data dari kelas minor, hal ini dikarenakan ketika kondisi *threshold* terpenuhi maka dilakukan penghentian *expand* node dengan memberikan label *(leaf)* kelas mayoritas dari subset pada node tersebut. Kondisi yang baik adalah ketika nilai *threshold* sebesar 20% karena akurasi testing mencapai 68,57% dan dapat mengcover data dari kelas user baru lebih banyak dari pada menggunakan nilai *threshold* lain.

4.2.2. Pengujian setelah pruning

Pada proses pruning menggunakan *Rule Post Pruning*, dimana pemangkasan dilakukan pada sekumpulan *rule* yang saling independent, dari setiap kumpulan *rule* tersebut dipotong prekondisi secara bertahap untuk mendapatkan akurasi data validasi yang sesuai dengan *threshold* yang telah ditentukan. *Threshold* validasi digunakan untuk menentukan persentasi akurasi suatu *rule* terhadap data validasi. Pada percobaan pruning dilakukan uji menggunakan *threshold validasi* : 0%, 10%, 20%, 30%, 40%, 50%, 70%, 80% dan 90%.

Threshold Validasi *Rule*	Jumlah *Rule*	Akurasi Data Validasi (%)	Akurasi Data Testing (%)
0%	**176**	**90,39**	**84,76**
10%	**174**	**90,39**	**84,76**
20%	165	86,14	79,29
30%	153	85,28	78,57
40%	144	85,12	78,57
50%	140	84,8	76,9
60%	121	83,15	75,48
70%	111	83,15	75,24
80%	94	82,76	75,48
90%	80	82,76	75,48

Dari hasil pengujian *threshold* validasi seperti pada tabel 4.16 tersebut dapat dilihat bahwa semakin tinggi *threshold* yang diberikan untuk memvalidasi suatu *rule*, maka semakin sedikit *rule* yang dihasilkan, hal ini karena aturan-aturan yang tidak memberikan validasi lebih baik dari nilai *threshold* dipangkas sehingga diperoleh jumlah *rule* yang lebih sedikit. Threshold validasi yang paling optimal adalah pada saat bernilai 0% dan 10%, karena menghasilkan akurasi untuk data validasi dan data testing yang paling baik. Pada posisi tersebut tercapai rule yang lebih baik dalam mengklasifikasikan data testing sehingga memberikan nilai akurasi yang baik.

4.3. Analisa Simplisitas Aturan

Simplisitas aturan merupakan nilai rata-rata jumlah antisenden dari setiap rule yang dihasilkan. Formula untuk menghitung simplisitas aturan adalah

$$simplisitas = \frac{\sum_{i=1}^{n} jml\_antisenden}{n} \quad (4,3)$$

Tree yang dihasilkan sebelum proses pruning memiliki 717 aturan dengan nilai simplisitas aturan sebanyak 4 dan nilai akurasi terhadap data testing sebesar 68,57%. *Tree* yang dihasilkan setelah proses pruning secara optimal memiliki 174 aturan dengan nilai simplisitas aturan sebanyak 3 dan nilai akurasi terhadap data testing sebesar 84,76%.

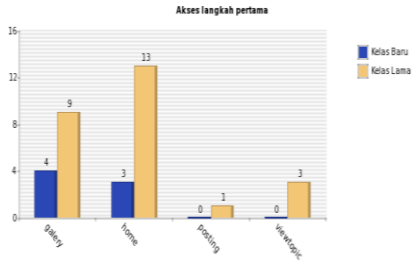
Dari analisa simplisitas aturan diperoleh hasil:

a. *Tree* sebelum dilakukan proses *pruning* memiliki aturan yang lebih kompleks jika dibandingkan dengan rule yang dihasilkan setelah proses *pruning*.
b. Penurunan jumlah rata-rata antisendent diakibatkan proses pemangkasan cabang, sehingga semakin banyak cabang yang terpotong menghasilkan rule yang lebih sederhana.

4.4. Analisa Pola Akses User

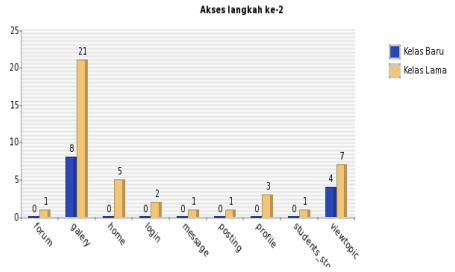
Pola akses yang diperoleh dari hasil klasifikasi dengan algoritma C5.0 maka dapat diperoleh:

a. Pola akses langkah pertama



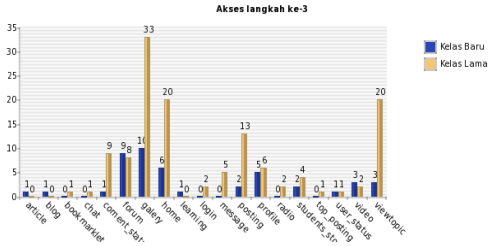
Hasil analisa dari gambar tersebut, dieroleh informasi bahwa sebagian besar user baru dan user lama memulai akses dari halaman home.

b. Pola akses langkah ke-2



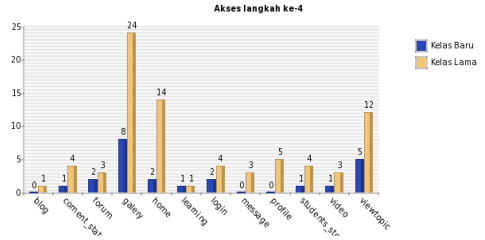
Hasil analisa dari gambar tersebut, dieroleh informasi bahwa sebagian besar kelas user baru pada langkah ke-2 melakukan akses terhadap halaman galery dan viewtopic begitu juga dengan user kelas lama.

c. Pola akses langkah ke-3



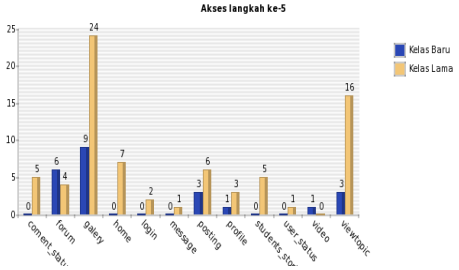
Hasil analisa dari gambar tersebut, dieroleh informasi bahwa sebagian besar user kelas baru dan user kelas lama pada langkah ke-3 mengakses halaman viewtopic dan galery.

d. Pola akses langkah ke-4



Hasil analisa dari gambar tersebut, dieroleh informasi bahwa sebagian besar kelas user baru pada langkah ke-4 mengakses halaman galery dan home, sedangkan pada user lama mengakses halaman galery dan viewtopic.

e. Pola akses langkah ke-5



Hasil analisa dari gambar tersebut, dieroleh informasi bahwa sebagian besar kelas user baru dan kelas user lama pada langkah ke-5 mengakses halaman galery dan viewtopic.

4.5. Analisa Umpan Balik (Feedback)

4.5.1. Analisa Langkah Akses User

Hasil analisa dari langkah pertama gambar 4.5 sampai langkah ke-5 gambar 4.9 dapat diperoleh informasi untuk umpan balik (feedback) kepada admin web bahwa :

a. Langkah ke-3 merupakan atribut yang sangat berpengaruh pada proses klasifikasi, karena memiliki ke aneka ragaman nilai yang berbeda antara user baru dan user lama.

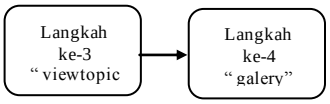
b. Halaman galery, home, posting dan viewtopic merupakan halaman yang paling strategis diantara halaman lain, karena halaman-halaman tersebut sebagian besar di akses pada langkah pertama, sehingga link informasi yang penting dapat disimpan pada halaman ini.

c. Halaman galery merupakan halaman yang paling banyak diakses oleh user, sehingga kemungkinan user lebih menyenangi halaman yang bergambar.

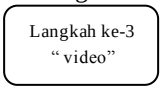
4.5.2. Analisa Pola Urutan Akses User

4.5.2.1. Pola Urutan Akses Kelas User Baru

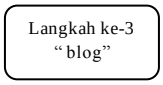
Dibawah ini adalah 5 pola akses pada kelas user baru yang memiliki akurasi rata-rata tertinggi:

a. 

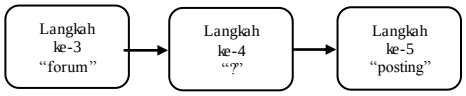
Pola akses user baru dengan urutan langkah ke-3 viewtopic dan ke-4 galery memiliki support sebesar 87,5% terhadap akurasi data training dan data testing.

b. 

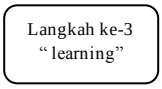
Pola akses user baru dengan urutan langkah ke-3 video memiliki support sebesar 83,34% terhadap akurasi data training dan data testing.

c. 

Pola akses user baru dengan urutan langkah ke-3 blog memiliki support sebesar 72,73% terhadap akurasi data training dan data testing.

d. 

Pola akses user baru dengan urutan langkah ke-3 forum, langkah ke-4 halaman apapun tetapi langkah ke-5 adalah posting memiliki support sebesar 69,23% terhadap akurasi data training dan data testing.

e. 

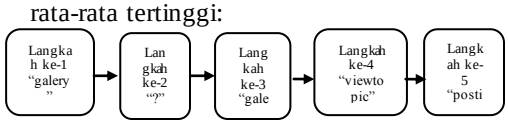
Pola akses user baru dengan urutan langkah ke-3 learning memiliki support sebesar 68,25% terhadap akurasi data training dan data testing.

Dari usage pattern kelas user baru tersebut dapat diperoleh informasi kepada admin web bahwa:

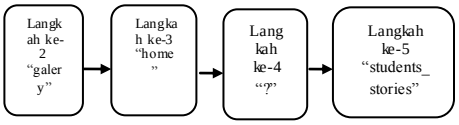
a. Langkah ke-3 merupakan atribut yang bernilai penting bagi proses klasifikasi pada user kelas baru.
b. Untuk langkah ke-3 sebagian besar kelas user baru mengakses halaman blog dan learning.
c. Apabila langkah ke-3 adalah halaman viewtopic maka akses akan dilanjutkan ke halaman galery. Sehingga kemungkinan topik yang menarik menurut user kelas baru adalah yang memiliki gambar dan *entertain*, karena selain galery user baru juga mengakses video.

4.5.2.2. Pola Urutan Akses Kelas User Lama

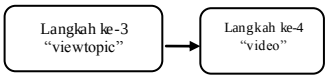
Dibawah ini adalah 5 pola akses pada kelas user lama yang memiliki akurasi rata-rata tertinggi:

a. 

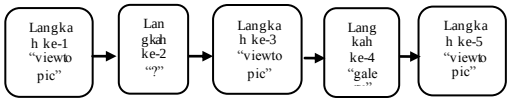
Pola akses user baru dengan urutan langkah ke-1 galery, langkah ke-2 halaman manapun tetapi langkah ke-3 galery, ke-4 viewtopic dan langkah ke-5 posting memiliki support sebesar 89,5% terhadap akurasi data training dan data testing.

b. 

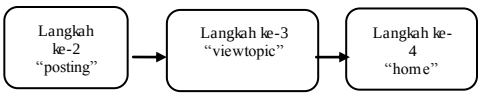
Pola akses user baru dengan urutan langkah ke-2 galery, langkah ke-3 home, langkah ke-4 halaman manapun tetapi langkah ke-5 students_stories memiliki support sebesar 89,3% terhadap akurasi data training dan data testing.

c. 

Pola akses user baru dengan urutan langkah ke-3 viewtopic, langkah ke-4 video memiliki support sebesar 88,9% terhadap akurasi data training dan data testing.

d. 

Pola akses user baru dengan urutan langkah ke-1 viewtopic, langkah ke-2 halaman manapun tetapi langkah ke-3 viewtopic, langkah ke-4 galery dan langkah ke-5 viewtopic memiliki support sebesar 88,9% terhadap akurasi data training dan data testing.

e. 

Pola akses user baru dengan urutan langkah ke-2 posting, langkah ke-3 viewtopic, langkah ke-4 home memiliki support sebesar 87,5% terhadap akurasi data training dan data testing.

Dari usage pattern kelas lama tersebut dapat diperoleh informasi kepada admin web bahwa:

a. Sebagian besar user akan mengulang atau kembali ke halaman yang pernah diakses pada langkah sebelumnya, sehingga untuk menghindari akses yang berulang-ulang maka perlu disediakan url yang menarik mengenai suatu informasi yang berkaitan dengan halaman yang sedang dibuka.
b. Halaman yang paling banyak diakses pada kelas user lama adalah halaman viewtopic dan galery. Sehingga untuk

menyampaikan informasi yang tidak terakses oleh kelas user lama bisa ditambahkan link pada halaman tersebut.

5. Kesimpulan

a. Informasi perilaku user dalam memanfaatkan web, memberikan kontribusi dalam memperbaiki navigasi dan penempatan informasi berdasarkan *behaviour* dari masing-masing user yang telah terklasifikasi.
b. Hasil klasifikasi berdasarkan pola navigasi user dapat diketahui bahwa kedua kategori user yang diklasifikasikan memiliki kecenderungan langkah akses yang hampir sama.
c. Klasifikasi dengan menggunakan algoritma C5.0 pada kasus *WUM* ini dihasilkan cabang yang besar, maka dilakukan pruning untuk memperbaiki akurasi data testing.
d. Pruning ketika *tree* dibangun diperoleh ***threshold information gain*** yang paling optimal sebesar 20% dan setelah *tree* terbentuk dilakukan pemangkasan *rule* dengan nilai ***threshold* validasi** yang paling optimal adalah 10% sehingga dihasilkan akurasi untuk data testing 84,76%.

6. Daftar Pustaka

[1] Abdurrahman, Bambang Riyanto T., Rila Mandala,Rajesri Govindaraju.2008. *Pemanfaatan Algoritma Ant Colony Untuk Web Usage Mining*:Bandung.

[2] Abraham,Ajith.*Business Intelligence from Web Usage Mining*. Oklahoma State University:USA

[3] Al Fatta,hanif. *WEB MINING: PENCARIAN POLA DAN INFORMASI PADA WORLD WIDE WEB*.

[4] Algoritme C5.0 untuk klasifikasi model decision tree. http://zulfani.000a.biz/isi.php?id=112 diakses pada 25 Oktober 2010.

[5] Anon.2003. *Clementine Algorithms Guide*.Chicago:SPSS Inc.

[6] Bettina Berendt, Bamshad Mobasher, Myra Spiliopoulou.2002.*Web Usage Mining for E-Bisnis Applications*.

[7] Chaofeng,Li. *Research and Development of Data Preprocessing in Web Usage Mining*. School of Management, South-Central University for Nationalities:China.

[8] Dixit,Dipa. Kiruthika,M. 2010.PREPROCESSING OF WEB LOGS. CRIT:Vashi

[9] Dixit,Dipa; Gadge,Jayant. 2010.*Automatic Recommendation for Online Users Using Web Usage Mining*. Thadomal Shahani Engineering College:Bandra.

[10] Gambetta,Windi.2008.*Pohon Keputusan(Decision Tree)*.Departemen Teknik Informatika,ITB:Bandung.

[11] Han J dan Kamber. 2001. *Data Mining: Concepts and Techniques*. Simon Fraser University. USA: Morgan Kaufman Publisher.

[12] Hengshan Wang, Cheng Yang, Hua Zeng.2006. *Design and Implementation of a Web Usage Mining Model Based On Fpgrowth and Prefixspan*. Business School, University of Shanghai for Science and Technology Shanghai:Cina.

[13] Holden,Nicholas.*Web Page Classification with an Ant Colony Algorithm*. University of Kent Canterbur: Kent UK.

[14] Iko Pramudiono. *"Parallel Platform for Large Scale Web Usage Mining"*, Tesis Ph.D, Universitas Tokyo, 2004.