PEMBANGUNAN SISTEM PERSELARIAN AGEN PELBAGAI UNTUK PERLOMBONGAN PETUA SEKUTUAN

(THE DEVELOPMENT OF A PARALLEL MULTI-AGENT SYSTEM FOR ASSOCIATION RULE MINING)

ABD MANAN BIN AHMAD

RESEARCH VOT NO:
71868

Jabatan Kejuruteraan Perisian
Fakulti Sains Komputer dan Sistem Maklumat
Universiti Teknologi Malaysia

2006

Lampiran 20

UTM/RMC/F/0024 (1998)

UNIVERSITI TEKNOLOGI MALAYSIA

BORANG PENGESAHAN
LAPORAN AKHIR PENYELIDIKAN

TAJUK PROJEK : **THE DEVELOPMENT OF A PARALLEL MULTI AGENT SYSTEM FOR ASSOCIATION RULE MINING**

Saya **ABD MANAN BIN AHMAD**
(HURUF BESAR)

Mengaku membenarkan **Laporan Akhir Penyelidikan** ini disimpan di Perpustakaan Universiti Teknologi Malaysia dengan syarat-syarat kegunaan seperti berikut :

1. Laporan Akhir Penyelidikan ini adalah hakmilik Universiti Teknologi Malaysia.
2. Perpustakaan Universiti Teknologi Malaysia dibenarkan membuat salinan untuk tujuan rujukan sahaja.
3. Perpustakaan dibenarkan membuat penjualan salinan Laporan Akhir Penyelidikan ini bagi kategori TIDAK TERHAD.
4. * Sila tandakan (/)

[ ]	SULIT	(Mengandungi maklumat yang berdarjah keselamatan atau Kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972).
[ ]	TERHAD	(Mengandungi maklumat TERHAD yang telah ditentukan oleh Organisasi/badan di mana penyelidikan dijalankan).
[X]	TIDAK TERHAD	

TANDATANGAN KETUA PENYELIDIK

Abd. Manan Ahmad
Fakulti Sains Komputer Dan Sistem Maklumat
Universiti Teknologi Malaysia

Nama & Cop Ketua Penyelidik

Tarikh : 15 Jan 07.

CATATAN : **Jika Laporan Akhir Penyelidikan ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa/ organisasi berkenaan dengan menyatakan sekali sebab dan tempoh laporan ini perlu dikelaskan sebagai SULIT dan TERHAD.*

ABSTRAK

Perkembangan maklumat yang berterusan di dalam Internet membuatkan alatan yang mudah dan tepat diperlukan oleh pengguna untuk membantu mereka mendapatkan maklumat-maklumat tersebut. Perlombongan data penggunaan Web telah menarik minat ramai penyelidik penggunaan Web untuk mengenalpasti kelakuan pengguna semasa melayari halaman Web dengan melombong log pelayan Web yang merekodkan semua aktiviti pengguna. Dengan mengaplikasikannya ke dalam enjin rekomendasi, personalisasi halaman Web dapat dilakukan berdasarkan kepada pengetahuan tentang kelakuan pengguna yang diperolehi. Walaupun begitu, kecekapan rekomendasi yang dijanakan masih menjadi isu para penyelidik. Kajian laporan akhir ini tertumpu kepada pembangunan model rekomendasi halaman Web berasaskan petua sekutuan (*association rule*) dan pengukuran nilai keserupaan, yang dinamakan sebagai ARsim. Pembangunan model ini menggunakan teknologi agen untuk membantu meningkatkan pemprosesan data-data dan penjanaan petua-petua rekomendasi. Persekitaran agen pula ialah Aglet 1.1b3. Selain itu, satu parameter tambahan digunakan untuk mengukur keserupaan antara URL iaitu masa yang diambil oleh pengguna untuk melihat suatu halaman Web. Untuk menjanakan senarai rekomendasi akhir, keserupaan antara URL-URL di dalam profail pengguna aktif diukur dengan merujuk kepada profail penggunaan Web yang berpadanan dan seterusnya *N* URL yang paling serupa dengan profail pengguna aktif tadi akan direkomenkan kepada pengguna. Tiga metrik yang lazim digunakan bagi mengukur kecekapan ARsim, iaitu *precision*, *coverage* dan *F1*. Keputusan perbandingan dengan dua teknik lain iaitu perlombongan petua sekutuan tradisional dan eVZpro menunjukkan

ARsim hanya merekomen URL-URL yang paling sesuai kepada pengguna dan seterusnya meningkatkan kecekapan enjin rekomendasi halaman Web.

ABSTRACT

The continuous growth of the information on the Internet makes it necessary for users to be provided with a convenient and yet accurate tools to capture the information needed. Web usage mining has gained more popularity among researchers in discovering the users browsing behavior by mining the web server log that records all the user's transactions activities. By applying it into the recommendation engine, Web personalization can be executed based on the discovered user's behavior. Nevertheless, the efficiency of the generated recommendations is still an issue for reseachers. This final report focusing on the development of a usage model for predictions based on association rule and similarity measures, named ARsim. Model development will used agent technology to improve data processing time and generate recommendations rules. Aglet 1.1b3 will be agent platform for this model development. Additional parameter was used to measure the similarities between URLs, which is the time user spend on a particular page. To generate the final recommendation, similarity between URLs contained in the active user profile was calculated upon the matched Web usage profiles and finally the top-N most similar URLs are then recommended to the user. Three evaluation metrics, which is commonly used by other researchers for evaluation of Web page recommendation model, was applied to evaluate the efficacy of ARsim, namely precision, coverage and F1. Comparison to two other different techniques, traditional association rule and eVZpro found that the integration of rules and similaty measures allow only the most appropriate URLs to be recommended and thus increase the efficiency of the Web page recommendation engine.

ISI KANDUNGAN

SENARAI RAJAH

SENARAI JADUAL

SENARAI ISTILAH

CLF	-	*Common Log Format*
Dalam Talian	-	*Online*
ECLF	-	*Extended Common Log Format*
HTML	-	*Hyper Text Markup language*
Luar Talian	-	*Offline*
Pengkadaran	-	*Rating*
Pengkadaran Tersirat	-	*Implicit Rating*
Pengkadaran Tersurat	-	*Explicit Rating*
Petua Sekutuan	-	*Association Rule*
Tapak Web	-	*Web Site*
URL	-	*Uniform Resource Locator*

BAB 1

PENGENALAN PROJEK

1.1 Pendahuluan

Perlombongan data adalah satu kaedah yang digunakan dengan meluas dalam mengenalpasti maklumat yang kritikal daripada gudang data untuk membantu pembuat keputusan perniagaan membuat keputusan. Jurang yang wujud antara sistem penyimpanan data yang ada di dalamnya membuatkan satu kaedah baru perlu dibangunkan bagi mengurangkan jurang ini. Justeru itu, prototaip sistem ini dicadangkan untuk membantu mengurangkan kerumitan pengguna dalam melaksanakan proses perlombongan data.

Agen untuk prototaip ini akan dibangunkan dengan menggunakan Aglet dan Java. Untuk perlombongan data pula, teknik yang digunakan ialah *Association Rule* kerana ia mudah untuk difahami dan sesuai untuk semua jenis data. Gudang data pula akan dibangunkan sendiri menggunakan Microsoft Access.

Metodologi yang akan digunakan adalah analisa dan rekabentuk berasaskan agen di mana kaedah UML*(Uniied Modeling Language)* akan digunakan. Metodologi ini

dipilih adalah berasaskan kepada kesesuaiannya dengan pembangunan prototaip yang berorietasikan objek dan menggunakan objek sebagai elemen utamanya. Setelah model rekomendasi halaman Web ini siap nanti, diharapkan tugas-tugas perlombongan data untuk membantu di dalam membuat cadangan halaman-halaman Web kepada pengguna akan menjadi lebih mudah.

1.2 Latar Belakang Masalah

Perlombongan Web (Srivastava, J., Cooley, R., Deshpande, M. and Tan, P. N., 2000, Cooley, R., Mobasher, B. and Srivastava, J., 1997, Kitsuregawa, M., Toyoda, M. and Pramudiono, I.) yang semakin popular dikalangan penyelidik Web adalah merupakan satu bidang yang masih memerlukan banyak perhatian dan menjanjikan masa depan yang lebih baik untuk meningkatkan prestasi Web berdasarkan keupayaannya untuk mengenalpasti dan mengelompokkan pengguna mengikut profil-profil tertentu (Smith, K. A. and Ng, A., 2003, Fu, Y., Sandhu, K. and Shi, M., 2000). Berdasarkan kepada profil-profil tersebut, pelayan Web boleh mencadangkan halaman-halaman Web lain yang dianggap berguna bagi pengguna tersebut. Selain itu, perlombongan Web juga berupaya untuk mengenalpasti kesilapan dan ketidak sesuaian dalam penyusunan kandungan halaman Web pentadbir Web (Spiliopoulou, M., 2000, Ishikawa, H., Ohta, M., Yokoyama, S., Nakayama, J., and Katayama, K., 2002). Ciri-ciri ini menunjukkan potensi perlombongan Web untuk meningkatkan prestasi system personalisai Web.

Laporan akhir ini mencadangkan satu model baru dinamakan ARsim untuk personalisasi Web menggunakan perlombongan Web, iaitu dengan mengintegrasikan perlombongan petua sekutuan (Association Rule Mining) dan pengkadaran. Pengkadaran di sini maksunya adalah nilai yang diberikan oleh pengguna ke atas suatu halaman Web yang telah dilihat. Tempoh masa yang digunakan oleh pengguna untuk melihat suatu halaman Web itu akan diambil dan dinormalkan untuk dijadikan pemberat atau perkadaran bagi suatu item. Setelah melakukan perlombongan petua sekutuan dan perkadaran, senarai item-item serupa bagi setiap URL unik akan dijanakan menggunakan

pengukur keserupaan. Pengukuran keserupaan antara item-item ini dilakukan ke atas peraturan-peraturan dan bukannya ke atas transaksi pengguna yang direkodkan ke dalam log pelayan Web. Model ini menggunakan pengukuran berasaskan item untuk mendapatkan keserupaan. Setelah itu, rekomendasi akan dibuat dengan mendapatkan item-item paling serupa bagi setiap URL yang telah diklik oleh pengguna dan disusun mengikut kedudukan ataupun rank. Perbandingan model yang dibangunkan akan dilakukan terhadap model-model yang telah dibangunkan oleh penyelidik lain sebelum ini, iaitu petua sekutuan tradisional (Mobasher, B. Dai, H., Luo, T., and Nakagawa, M., 2001) dan eVZpro(Demiriz, A., 2004).

Integrasi antara perlombongan petua sekutuan dan pengelompokan untuk mendapatkan profil pengguna telah dicadangkan oleh Mobasher et al. (Mobasher, B., Cooley, R. and Srivastava, J., 1999). Perlombongan petua sekutuan digunakan untuk mendapatkan hubungan antara URL yang diperolehi daripada corak capaian pengguna. Kelompok-kelompok pengguna kemudiannya mengumpulkan URL berkaitan berasaskan kepada kewujudannya di antara transaksi walaupun transaksi-transaksi tersebut tidak serupa. Pengelompokan data umum telah dicadangkan (Fu, Y., Sandhu, K. and Shi, M., 2000) dan telah berjaya mengurangkan dimensi data yang hendak dikelompokkan. Walaubagaimanapun, ia akan mencadangkan pautan-pautan yang tidak berkaitan kepada pengguna disebabkan oleh sifatnya yang umum. Ini adalah kerana ia akan menjanakan semua halaman yang terkandung di bawah halaman umum. Wang et al. (Wang, K, Xu, C. and Liu, B., 1999) telah mencadangkan teknik untuk mengkelompokkan transaksi yang mengandungi item-item yang serupa. Mereka menggunakan item besar untuk mengukur keserupaan dan bukannya keserupaan dari segi pasangan demi pasangan. Demiriz (2004) pula menggunakan petua sekutuan dan pengukuran keserupaan bagi mendapatkan rekomendasi. Kajian ini mendapati penggunaan petua sekutuan untuk sistem rekomendasi adalah lebih baik berbanding rangkaian kebersandaran. Kajian-kajian lain menggunakan petua sekutuan untuk melakukan pra-capaian terhadap halaman Web (Yang, Q., Zhang, H. H. and Li, T., 2001; Lan, B., Bressan, S. and Ooi, B. C., 2000; Lan, B., Bressan, S., Ooi, B. C. and Tan, K. L., 2000) menunjukkan kebaikan yang dapat diberikan oleh petua sekutuan dalam mengurangkan kelembapan rangkaian dan memahami corak capaian

pengguna. Dengan meramalkan halaman yang bakal dicapai oleh pengguna, pelayan akan secara automatik menghantar halaman tersebut terlebih dahulu ke dalam *cache* pelanggan sebelum halaman tersebut diminta dalam erti kata yang sebenar oleh pengguna. Walaubagaimanapun, saiz log transaksi yang sangat besar akan menghasilkan peraturan-peraturan yang tidak berguna. Untuk mengurangkan dimensi data yang hendak dilombong, Mobasher et. Al (Mobasher, B., Cooley, R., and Srivastava, J., 2000) telah mencadangkan untuk mengelompokkan sesi pengguna terlebih dahulu sebelum melaksanakan perlombongan data pada setiap kelompok. Kajian-kajian berkenaan telah menunjukkan keupayaan perlombongan petua sekutuan dalam menghasilkan keputusan yang baik untuk meramalkan halaman yang bakal dicapai ke tahap yang terbaik (Mobasher, B. Dai, H. Luo, T., and Nakagawa, M., 2001) dan beberapa pembaikan masih perlu dilakukan. Gabungan beberapa teknik atau penggunaan parameter baru ke dalam teknik perlombongan petua sekutuan yang sedia ada adalah perlu untuk menghasilkan rekomendasi yang lebih tepat.

1..3 Pernyataan Masalah

1. Bagaimanakah petua sekutuan dapat menghasilkan model profil pengguna yang baik untuk rekomendasi halaman-halaman Web kepada pengguna?
2. Bagaimanakah masa yang diambil oleh pengguna untuk melayari suatu halaman Web itu dapat membantu pernyataan masalah (1) menghasilkan rekomendasi yang lebih baik?
3. Apakah kaedah terbaik yang dapat meningkatkan kecekapan sistem rekomendasi dari segi ketepatan cadangan yang dihasilkan berasaskan kepada penyelesaian yang diperolehi daripada pernyataan masalah (1) dan (2)?

1.4 Matlamat Kajian

Untuk membangunkan model rekomendasi halaman Web yang dapat meningkatkan kecekapan dan keupayaan enjin rekomendasi untuk menghasilkan cadangan halaman-halaman Web kepada pengguna berasaskan kepada corak navigasi pengguna yang diekstrak daripada log capaian Web dengan menggunakan petua sekutuan dan pengukur keserupaan.

1.5 Objektif Kajian

1. Mengenalpasti bagaimanakah masa yang diambil oleh pengguna untuk melawat suatu halaman Web dapat digunakan di dalam perlombongan data penggunaan Web dan menghasilkan keputusan yang lebih baik.
2. Bangunkan model sistem rekomendasi halaman Web yang dinamakan ARsim untuk personalisasi Web dengan menggunakan petua sekutuan dan pengukur keserupaan.
3. Merekabentuk sistem dan agen yang bersesuaian dengan keperluan yang telah dispesifikasikan.
4. Menguji dan mengesahkan model yang dibangunkan.

1.6 Skop Kajian

1. Kecekapan model dalam merekomenkan halaman-halaman Web akan diuji menggunakan data yang sedia ada, iaitu log pelayan Web Fakulti Sains Komputer dan Sistem Maklumat, Universiti Teknologi Malaysia bertarikh 2 Julai 2003 hingga 17 Disember 2003, log pelayan Web Pusat Angkasa Keneddy Nasa bertarikh 1 Julai 1995 hingga 31 Ogos 1995 dan log pelayan Web Universiti Saskatchewan bertarikh 1 Jun 1995 hingga 31 Disember 1995.
2. Data yang digunakan adalah melibatkan dan terhad kepada log capaian Web,.

3. Perlombongan data akan dilakukan ke atas fail-fail HTML yang terdapat di dalam log pelayan Web yang diuji.
4. Dibangunkan menggunakan *Java Development Kit* 1.2.1 dan *Aglet* 1.1b3

BAB II

ANALISA MASALAH

2.1 Kajian Latarbelakang

Perlombongan data dan agen adalah dua teknologi yang semakin berkembang. Oleh itu, penggabungan kedua-dua teknologi ini pasti akan membawa kebaikan yang besar kepada organisasi yang amat bergantung kepada data-data dalam pengurusan perniagaannya. Ini akan membantu mengurangkan masalah penggunaan perisian-perisian perlombongan data yang mana selama ini agak sukar untuk dibangunkan oleh pengguna bukan teknikal.

Di dalam bab ini, setiap subjek yang terlibat dengan pembangunan prototaip sistem perlombongan data yang berasaskan agen akan disentuh dan diterangkan dengan lebih terperinci untuk pemahaman. Subjek-subjek yang akan disentuh adalah gudang data, teknologi agen, aplikasi agen di dalam perlombongan data, perbandingan perlombongan data berasaskan agen dan perlombongan data biasa Aglet (peralatan pembangunan agen).

2.2 Perlombongan Data

Perlombongan data adalah merupakan satu proses untuk mengenalpasti hubungkait, corak dan trend data yang baru dan bermakna dengan cara mengimbas sejumlah data yang besar yang disimpan di dalam gudang data. Ia melakukan carianterhadap maklumat yang dikehendaki dengan menggunakan teknologi mengenal corak seperti rangkaian neural dan statistikal. Nama lain bagi perlombongan data adalah Knowledge Discovery in Database (KDD).

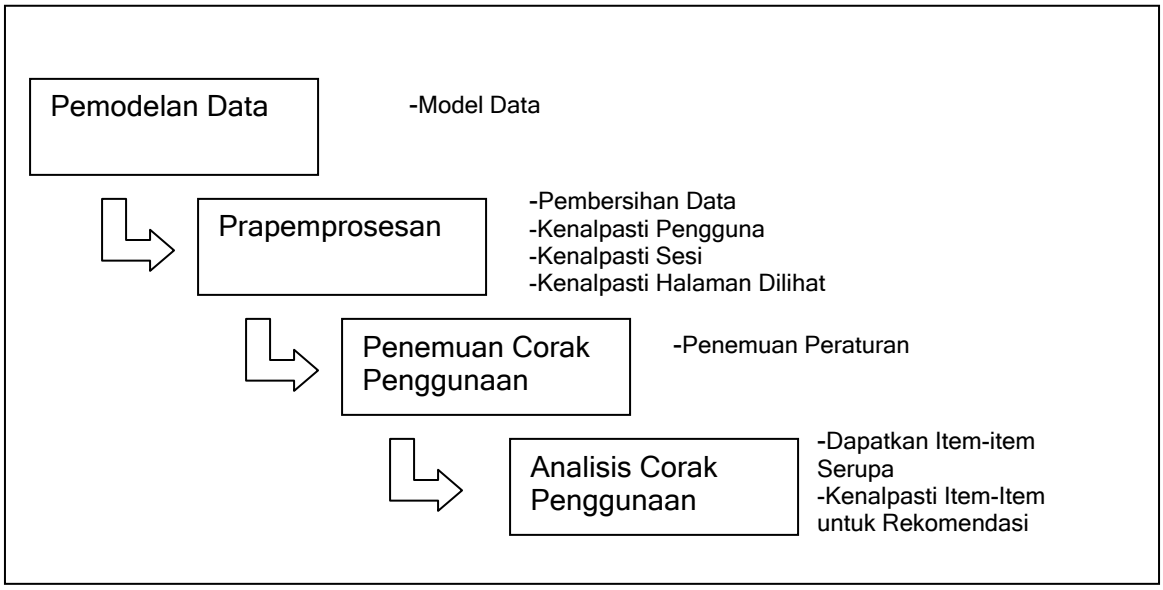
Terdapat dua jenis perlombongan data, iaitu perlombongan data deskriptif dan prediktif. Perlombongan data deskriptif adalah merupakan satu proses mencari corak-corak menarik pada data. Ia akan memberitahu sesuatu yang ada pada data yang tidak diketahui oleh pengguna. Perlombongan data jenis inilah yang selalunya akan memberikan pengguna maklumat-maklumat yang tidak dijangkakan. Perlombongan data prediktif pula adalah apabila pengguna memang telah tahu apa yang dikehendakinya dan dia hanya perlu memasukkan arahan untuk mendapatkan maklumat yang dikehendaki tersebut. Dalam kes ini, pengguna biasanya menggunakan pengalaman sebagai panduan di dalam perlombongan data yang hendak dilakukan. Sebagai contoh, siapakah pelanggan yang paling berpotensi untuk menjadi penyumbang terbesar dalam perniagaaan berasaskan kepada statistik sebelum ini dan sebagainya.

2.2.1 Proses-proses di dalam Perlombongan Data

Tujuan utama perlombongan data adalah untuk mencari corak yang menarik dan penting pada data-data. Walau bagaimanapun, ia adalah tidak mencukupi jika sekadar untuk mencari corak data sahaja. Menurut Michael J. A. Berry dan Gordon Linoff (Hickie et al):

Perlombongan penggunaan Web secara umumnya terdiri daripada empat fasa iaitu permodelan data, prapemprosesan, penemuan corak penggunaan dan analisis corak

penggunaan (Srivastava, J et al, 2000; Cooley, R et al, 1997). Rajah 2.1 menunjukkan proses-proses perlombongan data penggunaan Web yang umum. Permodelan data berperanan untuk memastikan hanya data-data yang sesuai sahaja digunakankan bagi tujuan perlombongan data penggunaan Web bagi menjamin pengetahuan yang diperolehi akan menjadi lebih bernilai. Prapemprosesan adalah fasa untuk menukarkan masukan-masukan log pelayan Web kepada bentuk data yang sesuai untuk dilombong. Antara proses yang terlibat adalah pembersihan data, mengenalpasti pengguna, sesi dah halaman dilihat. Penemuan corak penggunaan pula bertujuan untuk mengekstrak maklumat-maklumat penting daripada data-data yang telah disediakan tadi. Pada fasa ini, corak penggunaan Web dan kelakuan pengguna akan dapat dikenal pasti. Corak ini penting untuk digunakan pada fasa seterusnya, iaitu analisis corak penggunaan bagi membolehkan organisasi atau syarikat-syarikat e-perdagangan memberikan perkhidmatan atau promosi yang terbaik kepada pelanggan-pelanggan mereka.



Rajah 2.1: Perlombongan Data Penggunaan Web

Dalam proses penemuan corak penggunaan, pelbagai aplikasi dan algoritma perlombongan data boleh digunakan bergantung kepada tujuan ia dilaksanakan. Antara aplikasi yang biasa digunakan termasuklah penemuan petua sekutuan, corak berturutan, pengelompokkan pengguna serta halaman dan banyak lagi. Hasil daripada proses

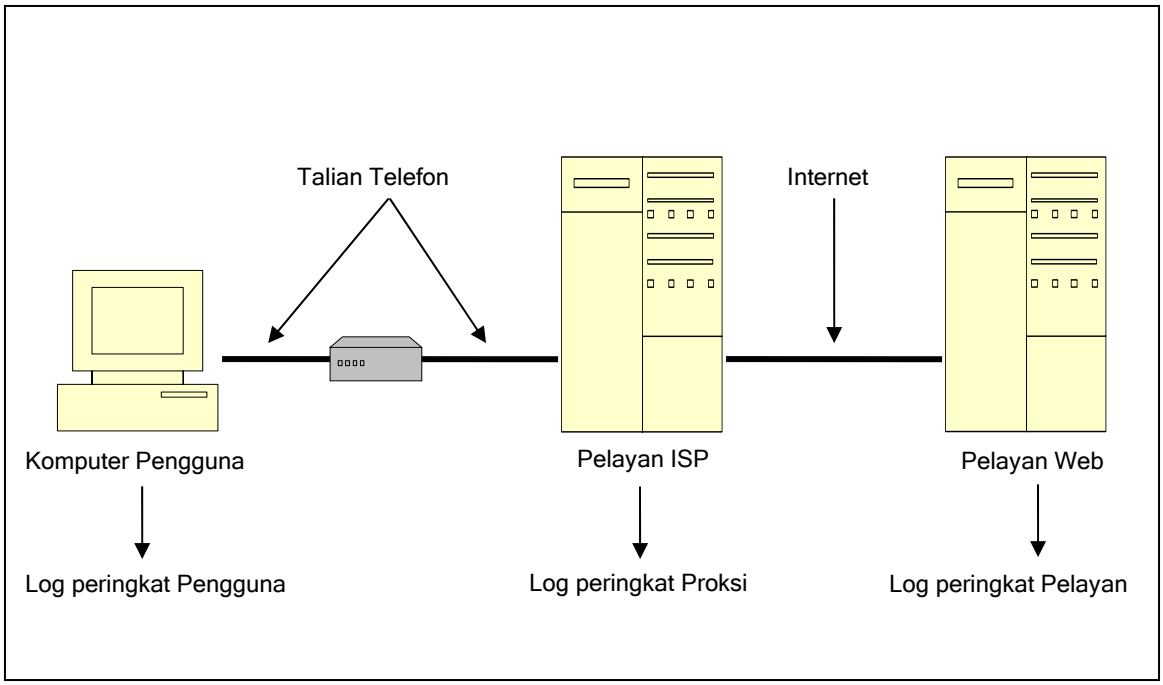
penemuan corak ini akan dianalisis untuk menghasilkan profil-profil penggunaan tapak Web yang mengandungi pengetahuan-pengetahuan penting bagi mengenalpasti halaman-halaman Web yang sesuai untuk direkomenkan kepada pengguna. Seterusnya, profil-profil penggunaan ini akan digunakan oleh bahagian dalam talian sistem untuk memaparkan halaman Web yang telah diubahsuai mengikut akriviti navigasi pengguna aktif.

2.2.1.1 Permodelan Data

Penggunaan data yang sesuai adalah sangat penting untuk personalisasi Web. Terdapat beberapa sumber data yang boleh digunakan untuk tujuan ini. Untuk itu, pemilihan yang tepat perlu dilakukan supaya pengetahuan yang diperolehi menjadi lebih bernilai.

a. Sumber Data Web

Salah satu perkara utama yang perlu dilakukan untuk mendapatkan pengetahuan daripada pangkalan data adalah membentuk set data yang sesuai untuk proses perlombongan data (Berry, M. J. A and Linoff, G., 1997). Data boleh dikumpul sama ada pada peringkat pelayan, pengguna dan juga proksi (Cooley, R. W., 2000). Sumber yang berlainan akan menghasilkan corak navigasi yang berlainan. Misalnya, pengumpulan data seorang pengguna dalam melayari halaman Web dari pelbagai tapak Web[9]. Data peringkat pelayan pula menggambarkan bagaimana satu tapak Web tunggal itu digunakan oleh ramai pengguna. Akhir sekali, data peringkat proksi pula merakamkan bagaimana pengguna-pengguna yang ramai melayari halaman-halaman Web dari pelbagai tapak Web. Rajah 2.2 menunjukkan peringkat-peringkat pengumpulan data.



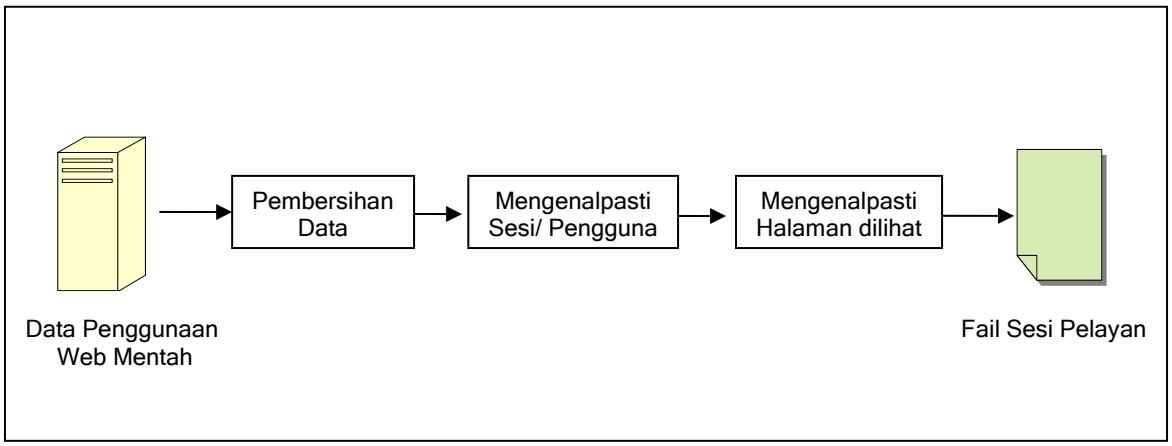
Rajah 2.2: Rajah capaian Web

Daripada Rajah 2.2 , dapat dilihat betapa proses merekod transaksi pengguna bukanlah mudah. Bagi log pelayan Web misalnya, tidak semua transaksi yang dilakukan oleh pengguna dapat direkodkan. Ini adalah disebabkan oleh *cache* halaman Web, yang mana pelayar Web pada computer pengguna akan mencari halaman Web yang diminta oleh pengguna di dalam *cache* terlebih dahulu. Sekiranya tidak dijumpai, barulah permintaan pengguna tadi akan dipanjangkan kepada pelayan Web.

2.2.1.2 PraPemprosesan Data

Fungsi fasa prapemprosesan adalah untuk menghasilkan model atau objek yang sesuai untuk perlaksanaan perlombongan data bagi suatu tapak Web. Kebiasaannya, proses penyediaan data adalah bahagian yang paling banyak menggunakan masa komputerisasi yang intensif dalam mana-mana sistem penemuan pengetahuan. Perlombongan data penggunaan Web tidak terkecuali daripada perkara ini. Rajah 2.3 menunjukkan proses-proses utama yang terlibat semasa fasa penyediaan data untuk perlombongan data penggunaan. Beberapa teknik dan heuristik boleh digunakan untuk

mendapatkan sesi pengguna yang boleh dipercayai. Fasa ini adalah sangat kritikal dalam menentukan kejayaan penemuan corak penggunaan semasa proses pengekstrakan pengetahuan pada fasa berikutnya.



Rajah 2.3: Fasa Prapemprosesan Data Penggunaan Web

a. Model Transaksi Penggunaan Web

Setelah melalui kesemua langkah-langkah yang terdapat di dalam fasa prapemprosesan data, secara umumnya sesi pelayan yang telah dihasilkan sudah sedia untuk fasa seterusnya iaitu penemuan corak penggunaan Web. Walaubagaimanapun, untuk laporan akhir ini, beberapa parameter yang terdapat pada sesi pelayan yang telah diperolehi akan dipilih untuk membangunkan model penggunaan Web bagi perlombongan data penggunaan Web. Ini adalah disebabkan pengkadaran akan digunakan bersama-sama dengan perlombongan data untuk mendapatkan model penggunaan Web yang akan dibincangkan seterusnya didalam bab ini.

Di dalam laporan akhir ini, masa yang digunakan untuk melihat suatu halaman Web akan digunakan sebagai pemberat bagi halaman tersebut. Pemilihan masa melihat halaman adalah disebabkan selalunya bagi algoritma-algoritma berasaskan keserupaan adan jarak, lebih terperinci vektor yang digunakan maka

keputusan yang akan dihasilkan adalah lebih tepat (Mobasher, B et al, 2001). Selain itu, ia boleh mengukur minat pengguna terhadap suatu halaman Web yang dilihatnya (Claypool, M, 2001). Disini pemberat bagi setiap halaman Web akan ditentukan oleh masa yang digunakan oleh pengguna untuk melihat halaman Web tersebut. Walaubagaimanapun, seperti yang dinyatakan sebelum ini, proses untuk mendapatkan masa yang sebenar adalah sukar.

Berpandukan kepada skop laporan akhir ini yang menggunakan log pelayan Web yang telah sedia ada sebagai data untuk pengujian, masa bagi melihat setiap halaman Web akan dikira dengan menolakkan masa halaman Web terkini mula diminta oleh pengguna dengan masa permintaan halaman Web yang seterusnya oleh pengguna yang sama. Andaikan masa permintaan halaman Web A yang diterima oleh pelayan adalah t_1 dan masa permintaan halaman Web B adalah t_2 . Maka, masa sebenar t_A untuk melihat halaman Web A adalah $t_A = t_2 - t_1$. Walaubagaimanapun, masa melihat bagi halaman terakhir dalam setiap sesi tidak boleh dikira menggunakan kawdah yang sama. Ini kerana tidak terdapat masa permintaan halaman Web seterusnya direkodkan. Untuk itu, *t* bagi halaman terakhir diperolehi dengan mendapatkan purata masa melihat halaman tersebut, yang bukan sebagai halaman terakhir, dari keseluruhan set data. Oleh kerana saiz fail dan keadaan rangkaian turut memainkan peranan dalam menentukan jumlah masa yang diambil untuk menerima permintaan daripada pelanggan dan memuat turun fail halaman daripada pelayan Web kepada pelayar Web, maka masa sebenar yang diperolehi tadi akan dibahagikan dengan saiz fail halaman web tersebut. Untuk itu, masa melihat bagi halaman Web A, $t\ \bar{A}$ di dalam laporan akhir ini dengan anggapan bahawa keadaan tradik rangkaian adalah sama bagi setiap halaman Web yang dilihat adalah :

$$t\ \bar{A} = \frac{t_A}{saiz\_fail_A} \qquad \textbf{(2.1)}$$

Akhir sekali, setelah melalui proses-prose penyediaan data, sebanyak *m* halaman Web, $P = \{ p_1, p_2, .., p_m \}$ dan n transaksi pengguna, $t_2 T = \{ t_1, t_{2,..,} t_m \}$ dijanakan. Setiap $t_i \in T$ adalah subset bagi *P*. Dengan mengambil w_i mewakili pemberat bagi setiap halaman Web $p_i \in P$, laporan akhir ini membangunkan formulasi model transaksi penggunaan Web yang akan digunakan untuk proses penemuan corak penggunaan Web sebagai:

$$t_i = \langle (p_1, w_1), (p_1, w_2), .., (p_m, w_m) \rangle \qquad \textbf{(2.2)}$$

2.2.1.3 Penemuan Corak

Setelah sesi pelayan berjaya dikenalpasti, beberapa teknik melombong corak boleh dilaksanakan seperti perlombongan petua sekutuan, pengkelompokan dan analisis jujukan (Cooley, R., Mobasher, B. and Srivastava, J., 1997). Perlombongan petua sekutuan (Agrawal, R. dan Srikant, R., 1994) menghasilkan peraturan-peraturan yang menjelaskan hubungan antara item-item, yang mana bagi kes perlombongan penggunaan Web, item adalah terdiri daripada URL. Ambang-ambang tertentu akan dikenakan untuk mendapatkan peraturan yang berguna. Pengelompokkan (Kohonen, T., 1982) akan mengumpulkan pengguna atau transaksi berdasarkan kepada keserupaan dari segi penggunaan manakala analisi jujukan (Srikant, R. and Agrawal, R., 1996) pula mendapatkan turutan item-item mengikut masa. Bagi kajian ini, teknik perlombongan data yang akan digunakan adalah perlombongan petua sekutuan. Teknik perlombongan sekutuan akan diterangkan dengan lebih lanjut di dalam bab 3.

2.2.1.4 Analisis Corak

Fasa analisis corak penggunaan adalah penting untuk menapis petua-petua yang tidak diperlukan daripada set keputusan yang dihasilkan oleh fasa penemuan corak penggunaan (Srivastava, J., Cooley, R., Deshpande, M. dan Tan, P. N., 2000). Terdapat banyak alatan yang digunakan untuk menganalisis corak-corak tersebut seperti SQL (Srivastava, J., Cooley, R., Deshpande, M. and Tan, P. N., 2000; Cooley, R., Mobasher, B. and Srivastava, J., 1997), memaparkan halaman indeks (Wang, S., Gao, W., Li, J. and Xie, H., 2000) mengubah secara dinamik organisasi hiperteks bagi halaman Web yang dilayari (Masseglia, F., Poncelet, P. and Teisseire, M., 1999) dan sistem rekomendasi (Konstan, J. A., Milner, B. N., Maltz, D., Herloker, J. L., Gordon, L. R. dan Riedl, J., 1997). Beberapa sistem rekomendasi yang telah dibangunkan (Mobasher, B., Cooley, R., and Srivastava, J., 2000; Ishikawa, H., Ohta, M., Yokoyama, S., Nakayama, J., and Katayama, K., 2002; Demiriz, A., 2004) menggunakan corak yang diperolehi untuk dianalisis bagi membolehkan rekomendasi tentang halaman-halaman Web yang releven diberikan kepada pengguna. Di dalam fasa ini juga enjin rekomendasi akan dibangunkan berasaskan kepada teknik dan model yang telah diperolehi semasa fasa prapemprosesan dan penemuan corak penggunaan Web.

a. Analisis Petua Penggunaan Web

Fasa analisis corak penggunaan Web adalah langkah untuk menukarkan petua-petua, corak dan statistik yang telah ditemui kepada pengetahuan yang berkaitan dengan tapak Web yang dianalisis. Antara masalah utama yang dihadapi semasa fasa analisis ini adalah untuk mendefinasikan dengan jelas apakah itu pengetahuan. Ini adalah kerana kepentingan suatu pengetahuan dan pemahaman ke atas pengetahuan-pengetahuan tersebut bergantung kepada orang yang melakukan analisis tersebut. Misalnya, adakalanya penggunaan pertanyaan yang mudah seperti pertanyaan SQL sudah memadai bagi penganalisis untuk mendapatkan pemahaman tentang data yang

dikaji, tetapi pada masa yang lain alatan yang lebih sofistikated dengan sistem visual yang baik diperlukan untuk analisis.

Perlombongan data dijalankan adalah untuk mendapatkan hubungan-hubungan atau corak-corak yang popular di dalam set data. Oleh itu, mencari hubungan yang paling popular atau menyusun corak-corak yang telah diperolehi mengikut populariti adalah penting supaya hanya corak yang menarik sahaja akan dipaparkan kepada pengguna.

i. Teknik untuk Analisis Corak Penggunaan Web

Terdapat banyak cara yang boleh digunakan untuk menganalisis corak yang telah dihasilkan oleh perlombongan data penggunaan Web. Cooley (2000) tealah menggunakan pengukuran interestingness untuk mengenalpasti corak yang ridak pernah diketahui oleh pengguna sebelum ini daripda hibungan yang telah ditemui. Teknik ini sesuai untuk tujuan kepintaran perniagaan, yang mana ia akan membantu dalam mengenalpasti bahaimana pengguna menggunakan tapak Web. Maklumat berkenaan perkara ini adalah sangat kritikal bagi syarikat e-perdagangan kerana ia akan menentukan sama ada kaedah pemasaran atau promosi yang digunakan sesuai atau tidak. Walaubagaimanapun, teknik ini tidak sesuai untuk diaplikasikan ke dalam laporan akhir ini kerana untuk menghasilkan sebuah sistem personalisasi Web yang baik, hubungan antara URL-URL yang telah diperolehi perlu dikekalkan untuk menjana senarai URL-URL yang akan direkomenkan kepada pengguna.

ii. Mengukur Keserupaan untuk Menjana Item-item Serupa

Model penggunaan Web seperti di dalam persamaan 3.5 sesuai untuk digunakan bagi mengukur keserupaan antara URL-URL yang terdapat di

dalam setiap petua. Pemberat, w_i bagi setiap item *i* adalah mewakili perngkadaran yang diberikan oleh pengguna. Merujuk kepada persamaan 2.2, pengkadaran adalah diperolehi daripada jumlah masa yang digunakan oleh pengguna untuk melihat 1 bait saiz halaman Web tersebut. Seterusnya, semasa proses penemuan corak penggunaan web, pemberat bagi setiap URL akan ditambah mengikut pengkadaran yang diberikan oleh pengguna setiap kali nilai kekerapan ataupun *support* URL tersebut bertambah. Oleh itu, selain daripada digunakan untuk mengukur keserupaan antara URL-URL, model ini juga sesuai digunakan untuk menjana halaman-halaman Web berdasarkan kepada *confidence*, σ petua tersebut untuk direkomenkan kepada pengguna.

Untuk mengukur keserupaan, beberapa kaedah untuk mengukur keserupaan boleh digunakan. Persamaan 2.1 dan persamaan 2.2 adalah dua metrik yang lazim digunakan untuk mengukur keserupaan. Untuk laporan akhir ini, pengukuran keserupaan menggunakan kosinus akan digunakan. Pemilihan ini dilakukan adalah kerana keserupaan berasaskan kosinus didapati lebih baik daripada kaedah-kaedah lain seperti pekali korelasi dan populariti (Demiriz, A, 2002). Selain itu, keserupaan berasaskan kosinus memberikan keputusan yang lebih baik sekiranya profil pengguna terdiri daripada nilai 0.0 hingga 1.0, yang mana nilai 1.0 menunjukkan bahawa kedua-dua item adalah sama.

Satu lagi teknik yang boleh digunakan adalah dengan menggunakan pengukur jarak *Euclidean*. Berbeza dengan kedua-dua teknik seperti di dalam persamaan 2.1 dan persamaan 2.2, pengukur jarak *Euclidean* berfungsi dengan mendapatkan jarak antara dua item, ataupun perbezaan antara dua item tersebut. Oleh itu, untuk mendapatkan keserupaan antara dua item tadi, 1 akan ditolak kepada nilai jarak *Euclidean* antara item-item tersebut. Prestasi ketiga-tiga teknik mengukur keserupaan ini akan dibandingkan melalui pengujian di dalam bab 4. Formulasi untuk mengira keserupaan menggunakan jarak *Euclidean* adalah seperti berikut:

$$similarity(i, j) = 1 - \sqrt{\sum_{n:r_n>0} (i_n - j_n)^2} \quad \textbf{(2.3)}$$

i dan *j* adalah dua item yang hendak diukur jaraknya, i_n dan j_n adalah mewakili nilai pemberat bagi item-item *i* dan *j* dari n petua yang ada. Bagi laporan akhir ini, nilai pemberat adalah diwakili oleh jumlah masa yang digunakan oleh pengguna untuk melihat setiap halaman atau item tersebut. Jarak yang diperolehi akan ditolak dengan satu untuk mendapatkan keserupaan. Sekiranya jarak adalah kecil, maka nilai keserupaan akan lebih tinggi, begitu juga sebaliknya.

Setelah mendapatkan keserupaan antara item-item, senarai *m* URL paling serupa akan disusun mengikut pangkat. Pangkat dikira dengan mendarabkan nilai keserupaan antara dua URL dengan *confidence* kedua-dua URL ini wujud dalam satu transaksi. Formulasi untuk mengira pangkat adalah seperti berikut:

$$pangkat(i, j) = similarity(i, j) \times confidence(i, j) \quad \textbf{(2.4)}$$

Pengiraan pangkat seperti di dalam persamaaan 2.4 adalah serupa dengan pengiraan *score* seperti yang digunakan oleh Demiriz (2002). Walaubagaimanapun, di dalam laporan akhir ini, *confidence* yang digunakan adalah kebarangkalian item *i* dan item *j* wujud di dalam satu transaksi, dan bukannya *confidence* bagi petua yang mana item *i* dan item *j* wujud. Ini adalah disebabkan fokus bagi laporan akhir ini lebih kepada menggunakan keserupaan antara URL untuk rekomendasi. Justeru, dengan melihat kepada kebarangkalian dua URL itu wujud bersama dalam suatu transaksi akan dapat menggambarkan hubungkait antara kedua-dua URL tersebut. Formulasi ini dipilih adalah kerana adakalanya pengguna secara tanpa disengajakan telah memberikan pengkadaran yang tinggi item *i* dan item *j* akibat daripada

gangguan luaran semasa melayari Web seperti menjawab panggilan telefon, membuat kopi dan masalah kelembapan rangkaian. Sekiranya menggunakan pengukuran keserupaan sahaja untuk menentukan pangkat, enjin rekomendasi (agen rekomendasi) sudah semestinya akan meramalkan bahawa terdapat hubungan yang sangat rapat antara item *i* dan item *j* dan seterusnya merekomenkan *i* kepada pengguna lain yang berminat terhadap item *j*. Dengan mengambil kira nilai *confidence* bagi item *i* dan item *j*, masalah ini akan dapat diatasi.

Setelah melalui proses pengukuran keserupaan dan penentuan pangkat, setiap URL unik, *u* yang terdapat di dalams set data, *D* akan mempunyai *m* URL lain, l yang hampir serupa dengannya, dan didefinasikan seperti berikut:

$$u_i = \left\langle \{l_1, l_2, .., l_m\}, \{w_1, w_2, .., w_m\} \right\rangle, u_i \in D, I_i \in D \qquad \textbf{(2.5)}$$

w_i adalah nilai pemberat bagi setiap URL l_i.

b. Agen Rekomendasi

Setelah memperolehi senarai item-item serupa bagi setiap URL unik seperti dalam persamaan 2.5, agen rekomendasi boleh memulakan proses meramal halaman-halaman Web yang berpadanan dengan corak aktiviti pengguna aktif. Agen rekomendasi ini berperanan sebagai sebagai enjin untuk menjana cadangan yang bersamaan dengan corak pemanduan pengguna. Secara keseluruhannya, untuk mendapatkan senarai *top-N* URL yang akan direkomen kepada setiap pengguna yang telah melihat satu set *P* URL, calon-calon URL yang akan direkomen, *C* hendaklah dijanakan terlebih dahulu dengan mendapatkan kesatuan daripada *m* URL yang paling serupa bagi setiap URL $p \in P$, dan menyingkirkan daripada kesatuan tersebut URL yang telah terdapat di

dalam *P*. Selepas itu, bagi setiap $c \in C$, keserupaan antara semua URL $p \in P$ dan *c*. Kemudian, jumlah keserupaan ini akan didarabkan dengan nilai *confidence* URL *c* tersebut wujud di dalam petua yang padan dengan aliran klik pengguna, *P*. Setelah itu, URL-URL di dalam *C* akan disusun secara menurun mengikut nilai keserupaan dan N URL yang pertama akan diambil untuk direkomen kepada pengguna. Untuk itu, formulasi yang akan digunakan untuk mendapatkan nilai rekomendasi bagi setiap URL yang akan direkomen kepada pengguna adalah seperti berikut:

$$rekomen(P,c) = \left(\sum_{p \in P} \left(S_p, c \right) \right) \times confidence(P,c) \qquad \textbf{(2.6)}$$

Sp,c adalah nilai keserupaan antara URL $p \in P$ dan URL $c \in C$.

2.3 Teknologi Agen

Agen perisian boleh ditakrifkan sebagai satu perisian yang melakukan tindakan bagi pengguna untuk memilih dan menyelesaikan tugas-tugas yang diperlukan dan pada masa yang sama ia boleh berinteraksi dan menggunakan program-program dan data-data lain. Agen berkemampuan untuk beroperasi sama ada apabila wujudnya rangkaian ataupun tidak. Ia bekerja tanpa henti dalam usaha untuk mencapai sasarannya.

Banyak definisi telah diberikan kepada agen. Definisi agen menurut The AIMA Agent (1995) adalah seperti berikut:

"*An agents is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors*"

Definisi agen menurut The Maes Agent (1995) pula menyatakan bahawa:

"*Autonomous agents are computatuional systems that inhabit some complex dynamic environment, sense and act autonomously in this environment, and by doing so realize a set of goals or tasks for which they are designed*"

Menurut definisi ini, persekitaran bagi agen adalah terhad kepada satu persekitaran yang kompleks dan dinamik. Agen akan bertindak secara bebas di dalam persekitaran ini dan ia sebenarnya akan menuju ke arah matlamat ia dibangunkan.

2.3.1 Ciri-ciri Agen

Selain daripada definisi-definisi yang telah diberikan sebelum ini, terdapat bebrapa lagi ciri-ciri atau sifat yang ada pada agen. Terdapat beberapa lagi ciri-ciri atau sifat yang ada pada agen. Terdapat dua jenis ciri-ciri atau sifat bagi agen, iaitu ciri-ciri mandatori atau wajib dan ciri-ciri pilihan.

Jadual 2.1: Ciri-ciri Mandatori Agen

Ciri-ciri	**Maksud**
Reactive	Tindakbalas terhadap perubahan yang berlaku di dalam persekitaran
Autonomous	Mempunyai kawalan terhadap tindakan
Goal-driven	Berorientasikan matlamat
Temporally continous	Melaksanakan proses secara berterusan

Jadual 2.2: Ciri-ciri Pilihan Agen

Ciri-ciri	**Maksud**
Communicative	Boleh berkomunikasi dengan agen lain
Mobile	Boleh bergerak dari hos ke hos
Learning	Mengubah tingkahlaku berasaskan pengalaman lalau
Believable	Mewujudkan rasa kebolehpercayaan kepada pengguna akhir

2.3.2 Kategori Agen

Setelah mengenal ciri-ciri yang mesti ada pada setiap agen dan ciri-ciri pilihan bagi agen, seterusnya agen ini akan dikategorikan kepada beberapa kategori seperti berikut untuk memudahkan pemahaman terhadap terminologi bagi agen:

- *Intelligent Agents*
  Kategori agen ayng paling luas bidang kajiannya dan yang paling diminati oleh kebanyakkan pembangun perisian
- *Learning Agents*
  Agen yang belajar daripada pengguna atau pemilik. Belajar bermaksud mengubah tingkahlaku berasaskan kepada pengalaman atau penilaian.
- *Mobile Agents*
  Agen yang aktif dan boleh bergerak di antara hos, komputer pengguna, rangkaian atau internet untuk melaksanakan tugasnya.
- *Believable Agents*
  Agen yang mempunyai animasi atau personaliti yang boleh meningkatkan kepercayaan pengguna kepada mereka.

2.3.3 Bidang Aplikasi Agen

Terdapat beberapa agen yang telah dibangunkan oleh para pengkaji agen. Senarai agen-agen tersebut berserta dengan fungsi dan bidangnya adalah seperti dalam Jadual 2.3:

Jadual 2.3: Senarai Agen

Bidang	**Nama Agen**	**Fungsi**
E-mail	Maxims (Maes[1994])	Mengawal apa yang dilakukan oelh pengguna terhadap e-mailnya setiap hari dan kemudiannya mengingati apa yang akan dilakukan oleh pengguna terhadap suatu situasi.
Web Browsing Assisting Agents	Letizia [1995]	Menggunakan pelayar web untuk mengawal aktiviti-aktiviti pengguna dan mengumpul maklumat tentang tingkahlaku pengguna. Dengan menggunakan heuristik agen ini akan mencadangkan perkara baru yang selari dengan minat pengguna kepada pengguna.
Agen Pengurusan Pentadbiran dan Aliran Kerja	FlowMark (IBM Corp.)	Menyediakan persekitaran untuk memanipulasikan objek berbentuk grafik yang mewakili langkah-langkah dalam proses perniagaan.
Pembangunan perisian	Neural Network Utility (NNU, daripada IBM)	Membantu pembangun perisian dalam mengecam corak dan belajar apa yang diperlukan untuk mencapai sesuatu maklumat.

Selain daripada agen-agen di atas, terdapat banyak lagi agen-agen lain merangkumi banyak lagi bidang seperti NewsAlert (Comshare Inc., agen pengesan dan berjaga-jaga) dan NotesAgents (Lotus Notes, collaboration agent).

2.4 Aplikasi Agen ke dalam Perlombongan Data

Perlombongan data merupakan suatu yang membolehkan aktiviti mencari, menapis dan menyusun maklumat dilakukan dengan mudah dan dalam bentuk yang sesuai. Salah satu cara yang boleh digunakan adalah mengaplikasikan agen ke dalam perlombongan data itu sendiri. Untuk perlombongan data, agen boleh digunakan untuk membuat penemuan terhadap hubungan-hubungan baru yang belum pernah ditemui sebelum ini. Selain itu, agen juga boleh digunakan untuk melaksanakan aktiviti-aktiviti lain yang termasuk di dalam persekitaran perlombongan data seperti antaramuka pengguna dan paparan hasil perlombongan data.

Untuk prototaip sistem dan model rekomendasi halaman Web yang hendak dibangunkan, agen akan digunakan untuk mencapai data daripada log pelayan dan memproses data-data yang telah dicapai tadi dengan menggunakan algoritma perlombongan data yang dipilih. Setelah keputusan diperolehi, maklumat tersebut akan dipaparkan kepada pengguna sebagai rekomendasi kepada halaman-halaman Web untuk pengguna. Teknik perlombongan data yang akan digunakan ialah petua sekutuan.

2.4.1 Agen Perlombongan Data Sedia Ada

Terdapat banyak agen yang telah dibangunkan untuk membantu dalam proses perlombongan data pada hari ini. Contoh-contoh agen yang telah dibangunkan adalah:

- *WebCrawler*
  Membantu dalam penemuan pengetahuan.

- *FAQFinder*
  Mengekstrak maklumat dari web
  .

- *ProdeBeacon*
  Dibangunkan oleh Prode Software Corp. yang mengandungi agen-agen yang boleh diprogramkan untuk memudahkan capaian pangkalan data di dalam gudang data. Agen ini membantu mengurangkan masa yang diambil untuk capaian data.

- *GENTIUM2*
  Dibangunkan oleh Planning Science International. Ia adalah merupakan suatu alatan DSS/EIS yang mengaplikasikan agen pintar untuk melaksanakan aktiviti perlombongan data. Tugas-tugas yang berulang dan pernyataan terhadap pangkalan data yang kompleks dilakukan secara automatik.

- *Convetics*
  Dibangunkan oleh HNC Software Inc. Ia menggunakan teknik rangkaian neural untuk mengecam imej teks bagi perkataan dan hubungan antara perkataan-perkataan tersebut untuk memahami maksud sesuatu dokumen.

- *IDM (Intelligent Data Miner)* [1]
  Merupakan satu prototaip sistem perlombongan data yang berasaskan agen perisian. IDM menyediakan beberapa jenis capaian data untuk membolehkan capaian dan analisa ke atas data yang disimpan di dalam gudang data. Ia terdiri daripada lima jenis agen iaitu agen antaramuka pengguna, agen perlombongan data, agen coordinator IDM, agen data-set dan agen visualisasi.

2.4.2 Kelebihan Penggunaan Agen di dalam Perlombongan Data

Terdapat beberapa kelebihan prototaip sistem yang akan dibangunkan. Kelebihan utama yang akan diperolehi adalah tugas-tugas perlombongan data akan dapat dilakukan dengan lebih mudah. Maksudnya di sini, pengguna tidak perlu pengetahuan yang tinggi untuk melaksanakan proses perlombongan data. Pengguna hanya perlu memasukkan

input yang mudah dan selebihnya untuk menukarkan input ke dalam bahasa pertanyaan seperti SQL akan dilakukan oleh agen.

Selain daripada itu, penggunaan agen di dalam proses perlombongan data turut membolehkan pemprosesan data dapat ditingkatkan. Terdapat dua jenis pengoptimuman yang boleh dilakukan di dalam laporan akhir ini iaitu agen pengoptimuman prestasi dan agen pengukuran prestasi. Pengukuran prestasi melibatkan rekabentuk teknik-teknik yang memperoleh maklumbalas samada daripada pengguna ataupun persekitaran untuk menentukan prestasi yang dalam konteks ini ia melibatkan model rekomendasi halaman-halaman Web kepada pengguna. Bagi agen pengoptimuman prestasi pula bertanggungjawab dalam mengoptimumkan parameter-parameter yang digunakan di dalam menentukan penjanaan petua-petua yang dihasilkan menerusi teknik petua sekutuan. Parameter-parameter ini berkemungkinan akan berubah mengikut prestasi agen pengoptimuman tersebut.

2.5 Alatan Pembangunan Agen

Sebelum memilih peralatan yang sesuai untuk membangunkan agen, beberapa perkara perlu diambil kira terlebih dahulu. Perkara-perkara tersebut terdiri daripada:

- Protokol komunikasi
  Penggunaan protokol komunikasi yang piawai akan memudahkan agen-agen bagi sistem yang berlainan saling berkomunikasi.
- Bahasa komunikasi agen *(Agent Communication Language, ACL)*
  Penggunaan bahasa yang piawai seperti KQML juga akan memudahkan agen-agen saling bekerjasama.
- Penjilidan bahasa
  Lebih banyak bahasa yang boleh dijilidkan bersama oleh suatu peralatan itu, maka ia akan memudahkan lagi pengintegrasian suatu sistem ke dalam sistem agen.
- *Middleware*

Penggunaan *Middleware* yang piawai akan mengurangkan kerja-kerja yang diperlukan untuk mengagihkan sistem agen. Contoh *Middleware* adalah CORBA *(Common Object Request Broker Architechture).*

2.5.1 Peralatan Pembangunan Agen di Pasaran

Terdapat beberapa peralatan pembangunan agen yang ada di pasaran:

- ***Agent Services Layer (ASL)***
  ASL dibangunkan oleh Unit Sistem Pintar bagi Broadcom Eireann Ltd. Ia menyokong pembangunan agen menggunakan bahasa-bahasa pengaturcaraan seperti C/C++, CLISP, Java dan Prolog. ASL dibangunkan dengan spesifikasi OMG CORBA 2.0 dan komunikasi antara agen adalah menggunakan KQML.

- ***Open Agent Architechture (OAA)***
  OAA dikeluarkan oleh SRI International menggunakan beberapa bahasa untuk membangunkan agen seperti C, Java, Prolog dan LISP. Komunikasi agen menggunakan ICL *(InterAgent communication Langguage).* OOA menyokong antaramuka pengguna pelbagai modal *(Multi Modal User Interface)* yang mana ia membenarkan pengguna berinteraksi dengan agen dengan cara melukis, menulis atau bercakap.

- ***Java-based Agent Franework for Multi-Agent Systems (JAFMAS)***
  Ciri utama JAFMAS adalah ia menggunakan metodologi generic untuk membangunkan sistem pelbagai agen berasaskan percakapan. Terdapat lima langkah dalam metodologi ini iaitu kenalpasti agen di dalam sistem, kenalpati perbualan agen dan akhir sekali perlaksanaan sistem pelbagai agen. Protokol komunikasi yang disokong oleh JAFMAS adalah TCP/IP, UDP/IP, KQML atau mana-mana protokol komunikasi yang berasaskan percakapan.

- ***JATLite***

  JATLite terdiri daripada set pakej Java yang memudahkan lagi pembangunan sistem pelbagai agen menggunakan Java. Ia membenarkan penghantaran dan penerimaan fail dan mesej dan berinteraksi dengan program lain pada komputer-komputer di mana agen tersebut berada. Ia menggunakan bahasa KQML untuk agen berinteraksi. Protokol yang digunakan adalah TCP/IP, FTP dan SMTP.

- ***Aglet Workbench***

  Aglet adalah objek-objek Java yang boleh berhijrah dari satu hos ke satu hos yang lain. Ia dihoskan oleh pelayan Aglet yang mana konsepnya adalah sama seperti mana applet dihoskan oleh pelayan web. Protokol yang digunakan adalah ATP *(Agent Transfer Protocol).*

Jadual 2.4 menunjukan dengan lebih jelas ciri-ciri setiap alatan pembangun agen yang telah dinyatakan tadi.

Jadual 2.4: Ringkasan ciri-ciri Alatan Pembangun Agen

Nama	ASL Broadcom Eireann RFesearch	OOA SRI International	JAFMAS University of Cincinnati	JATLite Stanford University	Aglet WorkBench IBM Corporation
Platform	JDK 1.1 (Java) Sun Solaris (C++)	Sun OS, Solaris, MS Windows	JDK 1.1	JDK 1.1	JDK 1.2
Perlaksanaan	Java/C++	Pelbagai	Java	Java	Java
Protokol Komunikasi	IIOP/POOP	TCP/IP	TCP/IP, UDP/IP	TCP/IP, SMTP, FTP	ATP
Bahasa Pembangunan	Java,C/C++, JESS,CLIPS,	C, Java, Prolog, Lisp,	Java	Java	Java

Agen	Prolog	Microsoft's Visual Basic dan Borland's Delphi			
Bahasa Komunikasi Agen	KQML	ICL	Sebarang ACL berasaskan percakapan	KQML	KQML
Perlesenan	Percuma untuk kegunaan bukan komersil	Percuma untuk kegunaan bukan komersil	Percuma	Percuma	Percuma

2.6 Aglet Workbench sebagai Pembangun Agen

Aglet adalah agen bergerak Java yang menyokong perlaksanaan automatik *(Autonomous)*. Ia mempunyai API, iaitu alatan pembangunan agen di mana dalam erti kata lain ia bermaksud set bagi kelas-kelas dan antaramuka-antaramuka Java yang membenarkan pembangunmembangunkan agen bergerak Java. Sekumpulan penyelidik di Makmal Kajian Tokyo IBM membangunkan API Aglet di Jepun sebagai tindakbalas terhadap permintaan untuk menghasilkan platform yang piawai bagi agen-agen bergerak di dalam persekitaran yang berbeza seperti internet. Dengan menggunakan API ini, keupayaan Java yang sesuai dengan sebarang platform dapat diaplikasikan kepada agen bergerak.

2.6.1 Kit Pembangunan Perisian Aglet IBM (ASDK)

Kit Pembangunan Perisian Aglet (ASDK) adalah pelaksanaan terhadap API Aglet dan ia boleh dipindah turunkan daripada halaman web Makmal Kajian Tokyo IBM (www.trl.ibm.co.jp/aglets). ASDK mengandungi pakej API Aglet, dokumentasi, agen-agen contoh dan pelayan aglet Tahiti. Tahiti adalah aplikasi Java yang membenarkan pengguna untuk menghantar dan menerima aglet daripada komputer lain yang turut mempunyai pelayan Tahiti.

Contoh sistem agen bergerak yang pertama yang dibangunkan menggunakan aglet adalah TabiCan. Ia adalah merupakan satu perkhidmatan komersil yang merupakan ruang perniagaan elektronik bagi penempahan tiket penerbangan dan pakej pelancongan. TabiCan direkabentuk untuk menjadi hos kepada beribu-ribu agen. Di bahagian pelayan, ia mempunyai agen kedai yang menanti permintaan daripada agen-agen pengguna. Pengguna yang mengunjungi halaman web ini boleh menyerahkan tugas kepada agen untuk mendapatkan tawaran yang terbaik dan meninggalkannya bekerja sendiri selama 24 jam. Walaubagaimanapun, perkhidmatan ini hanya boleh didapati dalam bahasa Jepun.

BAB III

METODOLOGI

3.1 Pengenalan

Untuk pembangunan model rekomendasi menggunakan petua sekutuan ini, metodologi pembangunan agen perisian akan digunakan. Metodologi ini dipilih adalah atas dasar kesesuaiannya dengan pembangunan model rekomendasi ini.

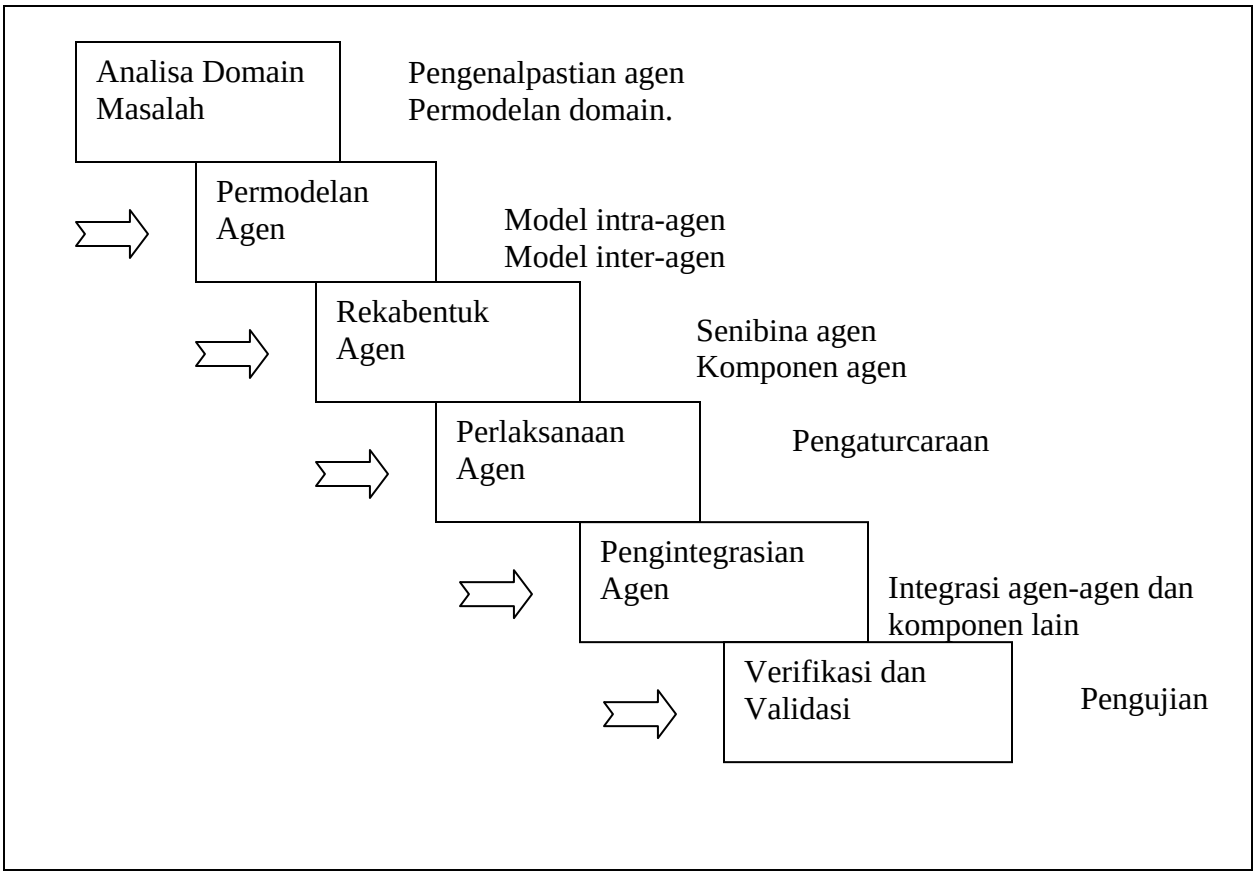
3.2 Metodologi Perisian Berorientasikan Agen

Peningkatan penggunaan agen perisian di dalam pembangunan aplikasi-aplikasi perisian tidak dapat dinafikan pada hari ini. Sejak ia mula diperkenalkan kepada dunia kepintaran buatan, banyak aplikasi yang telah menggnakan agen perisian sebagai elemen utamanya seperti penapis mel elektronik, pencari maklumat di web dan sebagainya. Satu pendekatan sistematik yang berasaskan kepada kejuruteraan perisian adalah diperlikan di dalam membangunkan sistem berorientasikan agen ini. Pendekatan yang digunakan mestilah mempunyai kaedah dan alatan yang merangkumi keseluruhan kitar hayat pembangunan sistem ini. Untuk itu, UML yang telah diterima dengan meluas di dalam

kejuruteraan perisian yang berorientasikan objek akan ditambahkan dengan beberapa keperluan agen yang spesifik untuk menghasilkan AUML *(Agent Unified Modeling Language).*

Untuk membangunkan perisian berorientasikan agen ini, satu model proses seperti pada rajah 3.1 (Park, Sooyong et al, 2000) akan digunakan. Model proses ini adalah adaptasi daripada model air terjun tradisional yang mana ia mempunyai enam aktiviti utama seperti berikut:

- Analisa domain – pemahaman dan permodelan domain masalah dan pengenalpastian agen.
- Permodelan agen – permodelan inter-agen dan intra-agen.
- Rekabentuk agen – senibina agen dan komponen.
- Perlaksanaan agen – melaksanakan agen dengan menggunakan alatan pembangunan agen dan bahasa komunikasi agen.
- Pengintegrasian agen – pengintegrasian agen-agen dan komponen-komponen lain.
- Verifikasi dan validasi – menguji fungsi-fungsi agen.



Rajah 3.1: Model Proses Pembangunan Perisian Berorientasikan Agen.

3.2.1 Analisa Domain Masalah

Pada fasa yang pertama ini, pendekatan berasaskan UML akan digunakan sepenuhnya. Di sini, rajah-rajah yang sering digunakan di dalam analisa dan rekabentuk berorientasikan objek akan digunakan sepenuhnya. Melalui fasa ini, bukan sahaja aspek-aspek statik sistem akan diperolehi, malah aspek dinamik sistem juga akan dapat dikenalpasti. Hasil daripada fasa ini nanti adalah senarai objek-objek yang terlibat dan dari sini agen akan dikenalpasti. Rajah-rajah yang akan digunakan pada fasa ini untuk memodelkan domain masalah adalah seperti berikut:

- Rajah Kes Guna
  Melalui permodelan kes guna, aktor luaran yang terlibat dalam pelaksanaan sistem perlombongan data yang akan dimodelkan bersama-sama dengan fungsi yang diperlukandaripada sistem (kes guna). Ia akan menggambarkan hubungan antara kedua-duanya. Rajah ini digunakan untuk memodelkan aspek statik sistem.

- Rajah Jujukan
  Rajah ini menunjukkan jujukan peristiwa yang erlaku, menunjukkan hubngan antara objek dan menekankan kepada turutan penghantaran mesej antara objek. Ia memodelkan aspek dinamik sistem.

- Rajah Kelas
  Rajah ini menunjukkan set kelas-kelas, antaramuka, hubungan dan kerjasama antara setiap kelas serta antaramuka sistem.

- Rajah Aktiviti
  Rajah ini menunjukkan aliran daripada satu aktiviti kepada aktiviti yang lain. Ia digunakan untuk menggambarkan aspek dinamik sistem.

3.2.2 Permodelan Agen

Setelah memperolehi aspek-aspek statik dan dinamik sistem, objek yang dirasakan sesuai untuk dijadikan agen akan dikenalpasti dan dimodelkan. Pada fasa ini, proses memodelkan agen terbahagi kepada dua iaitu permodelan intra-agen dan permodelan inter-agen.

Di dalam permodelan intra-agen, penumpuan diberikan kepada cirri-ciri yang perlu ada pada agen itu sendiri. Ia terbahagi kepada dua bahagian iaitu:

- Model gol
  Menyatakan objektif yang perlu dicapai oleh agen.
- Model keupayaan
  Menyimpan set operasi yang boleh dilakukan oleh agen.

Permodelan inter-agen ini pula akan melakukan proses pengintegrasian agen-agen yang telah dikenalpasti. Ia akan memodelkan komunikasi antara agen dan *mobility*. Model-model yang terlibat adalah seperti berikut:

- Model agen bergerak
  Menunjukkan bagaimana agen bergerak atau berhijrah untuk melaksanakan tugas-tugas tertentu.

- Model komunikasi agen
  Menunjukkan pertukaran mesej antara agen berasaskan kepada rajah jujukan.

3.2.3 Rekabentuk Agen

Pada fasa ini, rekabentuk agen akan dibangunkan berasaskan kepada model-model yang telah dibangunkan sebelum ini. Fasa ini menumpu kepada pembangunan senibina agen dan komponen-komponennya. Senibina agen ini akan melaksanakan teknik petua sekutuan yang digunakan untuk menghasilkan rekomendasi halaman-halaman Web untuk pengguna.

3.2.4 Perlaksanaan Agen

Fasa ini melibatkan proses pembangunan agen menggunakan kod aturcara menggunakan alatan pembangunan agen seperti Aglets 1.1b3 dan Java Development Kit v1.2.1. Data-data bagi membangunkan model rekomendasi ini adalah dengan menggunakan data-data dari log pelayan.

3.2.5 Integrasi Agen

Pada fasa kelima ini, agen-agen yang telah dibangunkan dan diuji secara berasingan sebelum ini akan diintegrasikan ke dalam satu pakej. Kesemua fungsi-fungsi agen dalam melaksanakan fasa-fasa dalam perlombongan data menggunakan petua sekutuan akan dintergrasikan. Pengintegrasian agen-agen ini dan juga komponennya akan membentuk sistem yang lengkap dan seterusnya berkeupayaan untuk menghasilkan model rekomendasi halaman-halaman Web kepada pengguna.

3.2.6 Verifikasi dan Validasi

Fasa yang terakhir adalah fasa verifikasi dan validasi. Pada fasa ini, sistem yang telah diintegrasikan tadi akan melalui pengujian secara keseluruhan, mengikut kes ujian yang telah ditetapkan. Sekiranya berlaku sebarang ralat, pembetulan akan dilakukan sehingga tidak terdapat ralat lagi pada sistem. Selain itu juga, model rekomendasi halaman Web yang dihasilkan juga akan diuji dan disahkan.

3.3 Pembangunan Sistem sebagai satu Prototaip

Pembangunan sistem adalah dalam bentuk prototaip. Prototaip terbahagi kepada dua iaitu prototaip keperluan dan prototaip evolusi. Prototaip keperluan adalah digunakan

untuk mengenalpasti keperluan-keperluan bagi sistem yang hendak dibangunkan. Model bagi setiap prototaip yang dibangunkan akan dibuang sebaik sahaja model tersebut diakui memenuhi keperluan dan ia akan dijadikan asas bagi pembangunan model prototaip yang seterusnya. Prototaip keperluan sesuai digunakan andai pengguna dan pembangun masih tidak jelas tentang spesifikasi keperluan bagi sistem.

Prototaip evolusi digunakan sekiranya pembangun berniat untuk menjadikan prototaip yang dibangunkan menjadi sistem sebenar akhirnya. Model bagi setiap prototaip akan diubahsuai sehingga ia memenuhi kehendak pengguna yang sebenarnya. Prototaip ini sesuai digunakan andai pembangunan sistem dilaksanakan secara dinamik mengikut perkembangan perniagaan atau pembangun yang terlibat tidak mempunyai pengetahuan yang mendalam tentang penggunaan peralatan yang terkini.

Prototaip evolusi dipilih dalam pembangunan sistem ini kerana masa yang diperuntukkan adalah terhad dan tidak sesuai untuk melaksanakan prototaip keperluan yang memerlukan peminaan model yang baru pada setiap kali selepas pengesahan suatu model. Selain itu, kemampuan yang agak terbatas tentang pengunaan peralatan juga menjadi suatu faktor. Satu lagi sebab adalah kerana pembangunan sistem ini tidak memerlukan kos yang tinggi dan ini sesuai dengan prototaip evolusi yang boleh dikatakan murah berbanding prototaip keperluan.

3.3.1 Kaedah UML (Unified Modeling Language)

UML ialah singkatan daripada *Unified Modeling Language* untuk pembangunan berorientasikan objek. UML ditulis oleh Grandy Booch, Jim Rumbaugh and Ivar Jacobson. UML menggabungkan kaedah serta teknik daripada metodologi yang lain.

UML digunakan untuk mentakrifkan pemetaan yang kurang jelas daripada fasa analisa keperluan kepada fasa analisa terperinci kepada fasa rekabentuk dan diikuti oleh

fasa implementasi. Objektif penggunaan kaedah UML ialah mentarifkan proses pembangunan pengulangan semula dan peningkatan.

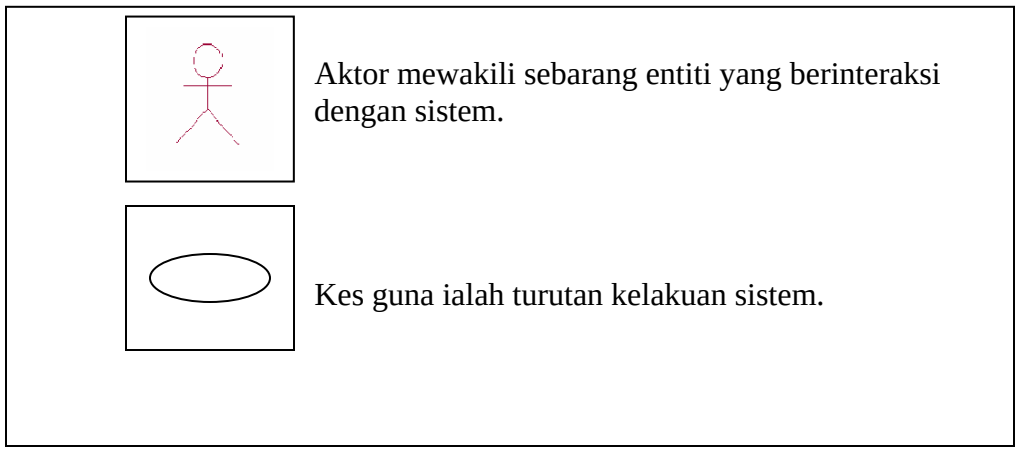
3.3.2 Teknik Permodelan Kes Guna

Permodelan kes guna ialah model sistem yang digambarkan o;eh kes guna dan aktor. Permodelan kes guna yang sama digunakan di dalam fasa analisa keperluan, analisa terperinci dan rekabentuk. Ia terdiri daripada empat jenis rajah iaitu rajah kes guna, rajah kelas, rajah jujukan dan rajah kerjasama *(collaboration)*.

Tujuan permodelan kes guna adalah untuk menjamin pemahaman pengguna dan pakar domain terhadap sistem. Ia juga digunakan untuk mengenalpasti entiti yang berinteraksi dengan sistem dan fungsi sistem serta antaramuka yang diperlukan oleh sistem.

3.3.2.1 Rajah Kes Guna

Objek yang digunakan di dalam permodelan kes guna adalah aktor dan kes guna seperti dalam Rajah 3.2



Rajah 3.2: Perwakilan dalam Kes Guna

Aktor bukanlah sebahagian daripada sistem, ia mewakili entiti di dalam sistem. Kes guna merupakan model dialog di antara actor dan sistem. Ia dimulakan oleh aktor yang melibatkan sesetengah fungsi di dalam sistem.

3.3.2.2 Rajah Kelas

Proses pengkelasan adalah proses di mana kita akan mengenalpasti kelas-kelas yang terdapat dalam sistem. Proses mengenalpasti kelas adalah peringkat yang amat sukar. Rajah kelas digunakan untuk menggambarkan setiap kelas dan hubungan yang wujud antara kelas-kelas tersebut.

3.3.2.3 Rajah Jujukan

Rajah jujukan menunjukkan hubungan dan interaksi antara objek-objek mengikut turutan masa. Rajah jujukan mengandungi objek, garis hayat objek *(object life line)* dan mesej antara objek-objek. Objek diwakilkan dalam bentuk kekotak segiempat dengan nama yang digariskan. Garis hayat objek pula diwakilkan garisan menegak putus-putus manakala mesej pula diwakilkan oleh anak panah melintang daripada garisan menegak yang mana setiap anak panah ini akan dilabelkan dengan mesej.

3.3.2.4 Rajah Kerjasama *(colloboration)*

Rajah kerjasama adalah sebagai alternative bagi menggambarkan aliran mesej-mesej antara objek-objek. Rajah ini mempunyai objek-objek, hubungan antara objek dan mesej antara objek.

Rajah kerjasama terdiri daripada objek yang diwakilkan dalam bentuk segiempat dengan nama yang digariskan. Hubungan antara objek diwakilkan melalui garisan yang menghubungkan setiap objek manakala aliran mesej-mesej adalah dalam bentuk anak panah yang berlabel. Label itulah yang menunjukkan nama mesej yang dihantar.

3.4 Justifikasi Pemilihan Metodologi, Kaedah / Teknik

AUML dipilih sebagai metodologi pembangunan sistem adalah disebabkan:

- Konsep UML yang digunakan memudahkan penggunaan prototaip kerana kebolehannya untuk di guna semula.
- Objek-objek merupakan perwakilan daripada dunia sebenar. Ini akan memudahkan pemahaman terhadap prototaip sistem.
- Prototaip sistem yang hendak dibangunkan adalah berorientasikan objek. Maka, UML adalah pilihan analisa dan rekabentuk yang sesuai untuk digunakan.
- Penggunaan agen di dalam sistem memerlukan satu metodologi yang menyokong pembangunan agen.

3.5 Senibina Multi Agen untuk Perlombongan Penggunaan Web

Agen perisian menurut (Lange, D. B., and Oshima, M, 1998) adalah kod aturcara yang tidak bergantungan dan mempunyai keupayaan untuk bergerak diantara hos, pintar dan berautonomi yang akan digunakan untuk melaksanakan setiap tugas yang diberikan. Senibina bagi sistem yang dicadangkan diterangkan pada gambarajah 3.3. Agen akan digunakan untuk melaksanakan setiap tugas yang diperlukan seperti mendapatkan capaian bagi log data, menyaring data, perlombongan petua sekutuan dan akhir sekali, untuk memaparkan jangkaan hasil keputusan kepada pengguna.

Senibina multi-agen yang dicadangkan mempunyai beberapa komponen utama seperti berikut:

1. Agen

   Agen merupakan entiti yang berautonomi dengan kemampuan untuk berkomunikasi dan menaakul (Lemaitre, C., and Excelente, C. B, 1998), sebahagian perisian yang dicipta oleh dan berkelakuan bagi pihak pengguna (Lange, D. B., and Oshima, M, 1998) dan agen juga merupakan pembantu elektronik kepada pengguna manusia yang mana ia boleh bertindak bagi pihak pengguna tetapi dengan memenuhi darjah autonomi untuk menyelesaikan beberapa permasalahan bersama-sama dengan agen yang lain (Bose, Ranjit and Sugumaran, Vijayan,1999). Kesemua definisi tersebut menepati dan bersesuaian dengan sebahagian perisian yang dinamakan agen. Lapan agen akan dibangunkan untuk sistem ini iaitu *UserAgent, Pre-Mining Agent, Data Distributor Agent, Data Mining Agent, Merge Agent, Output Agent, Recommendation Agent dan Monitor Agent.* Gambarajah 1 menunjukkan senibina yang dicadangkan untuk sistem multi agen ini.

2. Workspaces

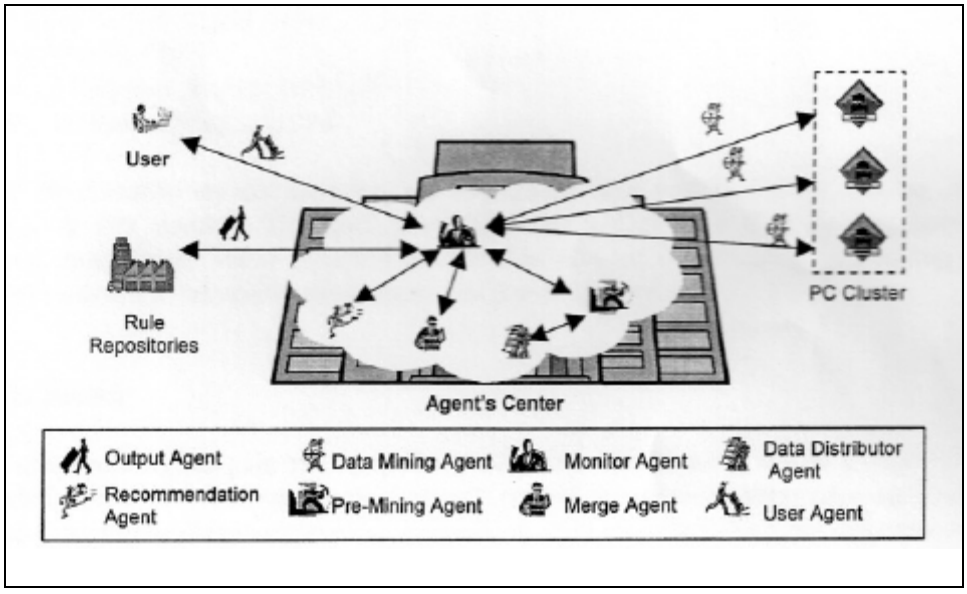
   Tempat dimana pengguna melaksanakan perisian sistem. Ia juga berfungsi sebagai pusat pelaksanaan perlombongan. Di sini *Merge Agent* akan diimplementasikan bertujuan untuk menyatukan hasil-hasil yang diperolehi.

3. Repositories

   Rule Repositories akan dibangunkan untuk menyimpan peraturan-peraturan yang telah dijana. Fail teks yang mudah akan digunakan berbanding dengan sistem pangkalan data dimana ia menggunakan kurang ruang dan pada masa yang sama, didapati bahawa prototaip ini tidak memerlukan ruang sistem penyimpanan data yang komprehensif.

4. Clusters of PC

   Beberapa buah komputer peribadi akan digunakan untuk melaksanakan proses perlombongan selari berbanding dengan penggunaan beberapa nod pemprosesan yang mana ia lebih ekonomi begitu juga dengan proses penyelenggaraan prestasi. Pengguna ruang kerja mungkin juga akan digunakan seperti unit pemprosesan untuk proses perlombongan



Rajah 3.3: Senibina Sistem Multi-Agen

Agent's Center merupakan pengguna ruang kerja. Setiap perincian agen digambarkan seperti berikut:

i. *Monitor Agent* bertanggungjawab untuk mengawal keseluruhan proses yang terlibat. Ia akan berkomunikasi dengan hampir setiap agen yang wujud di dalam sistem tersebut kecuali dengan *Data Distributor Agent* dan bertanggungjawab dalam menentukan agen yang mana akan diberikan tugasan bagi setiap tugas. Boleh dikatakan bahawa *Monitor Agent* merupakan nadi pengerak kepada sistem tersebut.

ii. *User Agent* bertindak sebagai antaramuka antara sistem dan juga pengguna. Ia boleh memahami akan nilai masukan yang diberikan oleh pengguna, permintaan akan nilai keluaran daripada monitor agent dan akhir sekali memaparkan hasi keputusan kembali kepada pengguna.

iii. *Pre-Mining Agent* bertanggungjawab dalam menyediakan data seperti yang digambarkan pada seksyen 2.2.1.1 dan 2.2.1.2 (rujuk pada bab 2). Ia boleh bergerak ke lokasi data dalam keadaan yang diingini oleh pengguna.

iv. *Data Distributor Agent* pula bertanggungjawab dalam membahagi-bahagikan data kepada beberapa bahagian, bergantung kepada bilangan pemproses yang ada. Peranan agen ini adalah penting kerana ia dapat meningkatkan tahap kelajuan proses perlombongan data antara agen pelbagai menerusi kaedah perselarian atau pemprosesan data teragih. Selain itu juga, Data Distributor Agent juga berperanan dalam mengkoordinasikan aktiviti agen-agen yang bersedia dalam melaksanakan tugas-tugas melombong data.

v. *Data Mining Agent* berfungsi untuk mengesktrak pengetahuan daripada data yang telah disediakan. Ia boleh bergerak ke pemproses yang bersedia melaksanakan tugas bersama-sama dengan data untuk dilaksanakan dalam proses perlombongan data. Data Mining Agent juga bertanggungjawab dalam mencari dan mendapatkan corak pengetahuan daripada data-data yang ingin dilombong.

vi. *Merge Agent* bertanggungjawab dalam menyatukan peraturan-peraturan yang telah di jana oleh setiap agen perlombongan data.

vii. *Output Agent* bertanggungjawab dalam memegang semua hasil keputusan daripada proses perlombongan dan menyimpan keputusan tersebut ke tempat simpanan peraturan.

viii. *Recommendation Agent* berperanan untuk menjana cadangan petua-petua yang bersamaan dengan corak pemanduan pengguna. Dalam mendapatkan hasil janaan cadangan yang baik, parameter-parameter yang digunakan perlulah dalam nilai yang terbaik bergantung kepada keadaan semasa. Diantara parameter yang diambil kira semasa proses penjanaan halaman ialah berdasarkan kepada parameter precision, coverage dan F1. Oleh yang demikian, recommendation

agent juga berperanan dalam mengoptimumkan prestasi nilai parameter-parameter tersebut bagi menjamin hasil janaan adalah yang terbaik.

3.5.1 Perlombongan Penggunaan Web menggunakan Petua Sekutuan

Menurut (Agrawal, R.., and Srikant, R, 1994), *Association Rule Mining* (ARM) atau Pelombongan Petua Sekutuan biasanya di gunakan untuk mendapatkan set item-item kerap bagi mengenalpasti hubungan diantara item-item tersebut. Untuk mendapatkan set item-item kerap ini, susunan halaman-halaman Web di dalam suatu sesi pelayan adalah tidak penting. Sebaliknya, kekerapan setiap halaman akan dikira pada setiap sesi yang mana halaman itu wujud. Seterusnya, sekumpulan item-item yang kerap wujud bersama di dalam suatu transaksi akan dicari untuk mendapatkan jumlah kekerapannya yang dinamakan *support*. Satu ambang yang dinamakan *minimum support* akan ditentukan pada peringkat awal proses pelombongan untuk membataskan bilangan petua yang akan dijanakan dan juga masa pemprosesan. Andaikan *D* adalah set data yang hendak dilombong dan i adalah item yang terkandung di dalamnya, *support*, σ bagi item *i* didefinasikan sebagai:

$$\sigma(i_j) = \frac{jumlah(i_j \in D)}{jumlah(D)} \qquad \textbf{(3.1)}$$

jumlah $(i_j \in D)$ adalah jumlah kekerapan item i_j wujud di dalam set data *D*. Manakala jumlah $jumlah(D)$ adalah jumlah bilangan keseluruhan item yang terdapat di dalam set data *D*.

Salah satu ciri utama yang ada pada petua sekutuan adalah sekiranya support suatu set item itu tidak melepasi minimum support yang telah ditetapkan, maka semua set item yang lebih terperinci daripada set item tadi juga akan disingkirkan daripada proses penemuan petua. Kriteria ini sangat penting untuk mengurangkan dimensi data yang

hendak dilombong pada setiap peringkat penemuan petua. Secara asasnya, terdapat tiga langkah terlibat didalam pelombongan petua sekutuan:

i. Janakan calon-calon set item *k* (setiap calon di dalam set item terdiri daripada sekumpulan k item) dari set item-item kerap $k-1$, F_{k-1}. Calon-calon dijanakan dengan menyatukan setiap item-item di dalam set item F_{k-1}, dan menghasilkan calon-calon set item *k*, C_k yang mana $C_k = F_{k-1} \cap F_{k-1}$. Contohnya, jika $F_1 = \{AB, BC, CD\}$, maka ia akan menjanakan $C_2 = \{ABC, ACD, BCD\}$.

ii. Singkirkan calon-calon yang mempunyai sekurang-kurangnya satu item tidak kerap. Contohnya, item ACD akan disingkirkan kerana item AC adalah bukan item kerap.

iii. Imbas keseluruhan set data untuk mendapatkan *support* bagi setiap item calon.

Petua sekutuan adalah pernyataan $A \geq B$, di mana A dan B adalah set-set butiran. Objektif peraturan ini adalah untuk menghitung kemungkinan akan nilai-nilai transaksi yang terkandung di dalam butiran B, di beri kandungan butiran A yang mana dikenali sebagai *confidence* bagi petua sekutuan. *Confidence* diberi sebagai $support(A \cup B) / support(A)$, dimana *support* adalah jumlah keadaan butiran dalam pangkalan data.

Bagi keadaan perlombongan penggunaan web, butiran akan diwakili oleh URL untuk membentuk set-set butiran, $U = \{u_1, u_2, u_3, ..., u_n\}$. Bagi setiap transaksi t_i, terdiri daripada set-set butiran, $t_i = \{u_1, u_2, ..., u_m\} \in T$, dimana ia menunjukkan halaman yang telah dicapai secara berturutan oleh pengguna. Mana-mana butiran dalam setiap urusan mempunyai *support* < *minimum support* akan di kurangkan. Proses ini dikenali sebagai *support filtering*. Keluaran daripada proses ini ialah menghasilkan sesuatu set bagi butiran yang besar, $L = \{l_1, l_2, ..., l_k\}$. *Support*, σ bagi set-set butiran $l_i \in L$, akan memberikan set *T* yang ditakrifkan sebagai

$$\sigma(l_i) = \frac{|\{t \in T : l_i \subseteq t\}|}{|T|} \quad \textbf{(3.2)}$$

dimana *confidence,* α bagi set-set butiran ditakrifkan sebagai

$$\alpha(l_i \Rightarrow l_j) = \frac{\sigma(l_i \cup l_j)}{\sigma(l_i)} \quad \textbf{(3.3)}$$

Tahap kepuasan nilai ambang minimum untuk *confidence* bagi petua sekutuan akan dijana daripada kekerapan set butiran, dimana kekerapan set butiran adalah set-set butiran yang memenuhi kepuasan bagi nilai ambang minimum *support*. Di dalam penyelidikan ini, setiap peraturan *r* akan dipertunjukkan didalam bentuk $A \Rightarrow B, (\sigma_x, \sigma_y)$, dimana A dan B adalah set butiran, σ_x adalah *support* $A \cup B$, dan σ_y adalah *confidence* bagi *r*. *Confidence* akan digunakan untuk menghasilkan cadangan penyelesaian semasa fasa analisis corak. Pelombongan petua sekutuan akan dilaksanakan secara selari untuk mengurangkan penggunaan masa.

Terdapat banyak algoritma perlombongan petua sekutuan telah dibangunkan. Antarnya ialah *Apriori* (Agrawal, R.., and Srikant, R., 1994), *FP-Growth* (Han, J., Pei. Dan Yin, Y, 2000) dan *Charm (*Zaki, M. J, 2000). Apriori berfungsi dengan menjana set item-item kerap terlebih dahulu dan kemudiannya menghasilkan petua sekutuan dari set item kerap tadi. Untuk itu, *Apriori* boleh menjana kedua-dua set item kerap dan petua sekutuan. *FP-Growth* pula digunakan untuk menjanakan set item-item kerap bagi petua sekutuan. Set item-item kerap dijanakan dengan membangunkan struktur pepohon corak kerap yang menyimpan semua maklumat-maklumat berkaitan dengan item-item tadi. *FP-Growth* tidak mempunyai fungsi untuk menjana petua sekutuan. *Charm* pula akan menjana set item-item kerap yang tidak bertindih. Maksudnya, algoritma ini akan mencari subset daripada satu set item kerap yang mana subset ini sudah memadai untuk mewakilkan set item kerap tadi tanpa perlu menghilangkan maklumat yang tersimpan di dalam set item kerap tersebut. Dengan itu, bilangan petua yang akan dijanakan daripada set item kerap akan dapat dikurangkan. Walaubagaimanapun, sama seperti *FP-Growth, Charm* hanya menjana set item-item kerap dan bukannya petua sekutuan. Zheng et al

(2001), telah membuat kajian untuk membandingkan beberapa algoritma perlombongan petua sekutuan. Hasil daripada kajian tersebut, di dapati Apriori masih masih merupakan algoritma yang terbaik untuk perlombongan petua sekutuan. Walaupun ia lambat dari segi masa untuk mendapatkan set item-item kerap, namun perbezaan masa tersebut tidak terlalu besar dan masih boleh diterima. Kelebihan Apriori adalah ia mampu untuk menjana set item-item kerap dan petua sekutuan berbanding algoritma yang lain. Faktor ini telah mendorong penggunaan Apriori di dalam laporan akhir ini. Selain itu, kajian-kajian lain juga menggunakan algoritma Apriori untuk mengenalpasti kelakuan pengguna (Mobasher, B et al, 2001, Lan, B., et al, 2000).

Pemilihan perlombongan petua sekutuan sebagai teknik untuk melombong data penggunaan Web di dalam laporan akhir ini adalah berdasarkan kepada beberapa faktor. Pertama, perlombongan petua sekutuan mempunyai keupayaan untuk membuat ramalan lebih baik berbanding teknik lain. Demiriz (2002), telah membuktikan bahawa petua sekutuan lebih baik dalam membuat ramalan berbanding teknik-teknik pembelajaran mesin dan rangkaian kebersandaran. Selain itu, beberapa kajian lain juga telah menunjukkan bahawa petua sekutuan dapat memberikan keputusan yang lebih baik berbanding pengelompokan (Mobasher, B., Cooley, R., and Srivatava, J, 2000) dan juga sesuai dijadikan sebagai asas untuk membuat rekomendasi (Sarwar, B., Karypis et al, 2000, Fu, X et al, 2000). Kedua, laporan akhir ini ingin mengenal pasti corak kelakuan pengguna semasa melayari halaman Web dan seterusnya menggunakan corak ini untuk meramal capaian halaman Web pengguna tersebut. Dengan petua sekutuan, hubungan antara halaman Web dari aspek penggunaannya akan dapat dikenalpasti dan ini membolehkan sistem rekomendasi meramalkan apakah halaman yang mungkin akan dicapai oleh pengguna berasaskan kepada halaman-halaman yang telah dicapai sebelumnya.

Daripada kajian-kajian yang telah dilakukan , belum ada yang menggunakan masa melihat setiap halaman Web untuk membantu mengukur keserupaan antara halaman-halaman Web dan seterusnya menjana rekomendasi halaman-halaman Web. Penggunaan masa melihat halaman Web hanya terhad untuk fasa pemprosesan sahaja. Didalam

laporan akhir ini, pada setiap pengulangan untuk mengimbas set data bagi mendapatkan *support* bagi setiap item calon, sperti yang dirunjukkan dalam langkah ketiga, pemberat bagi setiap item, *w* juga akan ditambah nilainya. Merujuk kepada model transaksi penggunaan Web di dalam persamaan 3.2, setiap halaman Web bagi setiap transaksi akan disertakan sekali pemberat bagi halaman tersebut. Jumlah pemberat bagi setiap halaman akan diperolehi dengan menambahkan nilai pemberat-pemberat halaman tersebut yang terdapat di dalam set data.

Setelah mendapatkan set item-item kerap, petua sekutuan yang memenuhi satu lagi ambang iaitu *confidence* akan dijanakan daripada set item-item kerap tadi. *Confidence* di kira untuk mendapatkan kebarangkalian, misalnya item A dan item B wujud bersama di dalam satu transaksi. Pengiraan *confidence*, α adalah seperti berikut:

$$\alpha\left(i_j, i_m\right) = \frac{\sigma\left(i_j \cup i_m\right)}{\sigma\left(i_j\right)} \qquad \textbf{(3.4)}$$

Di dalam laporan akhir ini, set petua sekutuan ataupun model penggunaan Web, R yang telah dijanakan secara konsepnya didefinasikan sebagai:

$$r_j = \left\langle \left(p_1, p_2, ..p_n\right), \left(w_1, w_2, ..w_n\right), \sigma, \alpha \right\rangle \in R \qquad \textbf{(3.5)}$$

Sekiranya ada antara petua-petua yang dijanakan mempunyai nilai *support* yang kurang daripada *minimum support*, maka petua itu akan disingkirkan daripada *R*. Proses ini dinamakan sebagai penapisan *support*. Algoritma perlombongan petua sekutuan yang digunakan di dalam laporan akhir ini adalah algoritma *Apriori* (Agrawal, R.., and Srikant, R, 1994). Model dalam persamaan 3.5 akan digunakan untuk proses analisis corak penggunaan Web.

3.5.1.1 Algoritma *Apriori*

Algoritma *Apriori* adalah algoritma yang menggunakan teknik gelintaran lebar-dahulu (Zheng, Z et al, 2001). Algoritma ini mempunyai dua langkah: pertama adalah untuk mencari set item-item kerap. Set ini mestilah mempunyai *support* sekurang-kurangnya menyamai nilai *minimum support*. Langkah kedua adalah menjana petua-petua sekutuan daripada set item-item kerap tadi.

```
APRIORI (data, minsup, minconf)
 1. Data = data;
 2. L1 = {set 1-item kerap};
 3. for (k=2; Lk−1 ≠ null; k++)
 4. Ck = jana_calon (Lk−1);
 5. foreach transaksi t ∈ D
 6.       Ct =subset (Ck, t)
 7.       foreach calon c ∈ Ct
 8.           c.Kira++;
 9.           c.Masa += c.Masa;
10.       end
11. Lk = {c ∈ Ck | c.Kira ≥ min sup}
12. P = jana_petua( Lk , minconf)
13. end
14. return P;
```

Rajah 3.4: Algoritma Apriori

Rajah 3.4 menunjukkan algoritma Apriori dengan jelas (Agrawal, R.., and Srikant, R, 1994). Pada baris dua, set1-item kerap L_1 diperolehi. Set ini akan digunakan untuk menjana item-item kerap yang seterusnya. Baris tiga menyatakan syarat untuk pengulangan proses penjanaan item-item kerap. Penjanaan akan berulang sehingga set $(k-1)$ - item kerap adalah kosong. Pada setiap pengulangan, calon item-item kerap C_k akan dijana menggunakan prosidur jana_calon, seperti yang ditunjukkan pada baris empat. Setelah memperolehi senarai calon-calon C_k, baris lima akan mengimbas setiap

transaksi di dalam set data. Sekiranya suatu calon itu merupakan sebset bagi suatu transaksi (baris enam), maka nilai *support* bagi calon tersebut akan ditambahkan seperti yang ditunjukkan di baris tujuh dan baris lapan. Baris sembilan adalah merupakan pengubahsuaian yang telah dilakukan pada algoritma *Apriori* asal. Oleh kerana laporan akhir ini akan menggunakan masa sebagai pemberat untuk mengukur keserupaan antara halaman Web, maka jumlah masa yang diambil oleh pengguna setiap kali satu atau set halaman Web itu dilihat akan dikira. Setelah semua transaksi diimbas dan nilai support serta jumlah masa melihat bagi semua calon, *c* diperolehi, proses penyingkiran calon yang tidak kerap akan dilakukan (baris 11). Calon-calon kerap ini akan disimpan di dalam set item kerap L_k dan digunakan untuk pengulangan seterusnya. Setelah senarai item-item kerap diperolehi, proses terakhir adalah untuk menjanakan petua-petua sekutuan seperti yang ditunjukkan pada baris 12. Proses penjanaan petua-petua ini akan dilakukan oleh prosidur jana_petua.

```
jana_calon (L_{k-1})
1.   insert into C_k
2.   select p.item_1, p.item_2, ..., p,item k-1, q.item k-1
3.   from L k-1, p, L k-1 q
4.   where p.item_1 = q.item_1,...,p.item k-2 = q.item k-2, p.item k-1 < q.item k-1
5.   foreach calon c ∈ C_k
6.        foreach (k-1)-subset s bagi c
7.             if (S ∉ L_{k-1})
8.                  delete c dari C_k
9.   end
10.  return C_k
```

Rajah 3.5: Prosedur jana_calon

Rajah 3.5 pula menunjukkan prosedur jana_calon yang digunakan untuk menjana calon-calon item kerap. Prosedur ini terbahagi kepada dua fasa, iaitu penggabungan dan penyingkiran item. Penggabungan ditunjukkan pada baris satu hingga baris empat. Fasa ini melibatkan penggabungan antara L $_{k-1}$ dengan L $_{k-1}$. Sebagai contoh, sekiranya $L_2 = \{\{1,2\},\{1,3\},\{2,3\},\{2,5\},\{3,4\},\{3,5\}\}$, selepas penggabungan ia akan menghasilkan

$C_3 = \{\{1,2,3\},\{2,3,5\},\{3,4,5\}\}$. Semasa fasa penyingkiran pula, sekiranya (k-1)-subset calon $c \in C_k$ yang dijanakan tidak terdapat di dalam L_{k-1}, maka calon *c* tadi akan disingkirkan daripada C_k. Daripada contoh yang diberikan, set item {3,4,5} akan disingkirkan kerana set item {4,5} tidak wujud di dalam L_2. Proses ini ditunjukkan pada baris lima hingga baris lapan prosedur jana_calon. Hasilnya, set 3-item kerap, $L_3 = \{\{1,2,3\},\{2,3,5\}\}$.

```
Jana_petua ( L_k , minconf)
1.  foreach a_k ∈ L_k
2.      A = {set (k-1)-item a_{k-1} | a_{k-1} ⊂ a_k };
3.      foreach a_{k-1} ∈ A
4.          conf = support (a_k) /support (a_{k-1});
5.          if (conf ≥ minconf)
6.              Petua.append (a_{k-1} ⇒ (a_k − a_{k-1}), a.masa_{k-1} ⇒ (a.masa_k − a.masa_{k-1})
7.                  support = support (a_k), confidence = conf);
8.          end
9.      end
10. return Petua;
```

Rajah 3.6: Prosedur jana_petua

Setelah mendapatkan senarai item-item kerap, penjanaan petua-petua sekutuan akan dilakukan. Rajah 3.6 menunjukkan langkah-langkah yang terlibat semasa penjanaan petua sekutuan. Setiap item di dalam set (K)-item kerap akan menjana beberapa petua, bergantung kepada *confidence* yang dihasilkan. Daripada contoh sebelm ini, telah diperolehi set 3-item kerap adalah $L_3 = \{\{1,2,3\},\{2,3,5\}\}$. Proses dimulakan dengan melihat kepada item pertama di dalam set (baris satu). Set (k-1)-item kerap akan dijanakan daripada item pertama set 3-item kerap, seperti yang ditunjukkan di dalam baris dua. Berdasarkan contoh, A akan mengandungi set item $\{\{1,2\},\{1,3\},\{2,3\}\}$.

Seterusnya baris tiga dan empat menunjukkan pengiraan *confidence* bagi setiap item di dlam L_3 berpandukan kepada item-item di dalam A. Sekiranya *confidence* melepasi nilai minimum *confidence*, maka item kerap tadi akan ditukar ke dalam bentuk petua dan dimasukkan ke dalam senarai petua-petua sekutuan (baris enam dan baris tujuh). Misalnya, jika *support* ({1,2,3})/*support* (2,3) lebih besar daripada *minimum confidence*, maka petua "$\{2,3\} \Rightarrow \{1\}$, {2.masa, 3.masa} $\Rightarrow$ {1.masa} *support* = σ, *confidence* = α" akan dijanakan. Laporan akhir ini turut menyenaraikan masa ke dalam petua yang dihasilkan kerana ia akan digunakan sebagai parameter untuk mengukur keserupaan.

BAB IV

PENGUJIAN DAN KEPUTUSAN

4.1 Pengenalan

Pada fasa implementasi, prototaip sistem yang telah direkabentuk akan dibangunkan dengan menggunakan alatan perisian seperti ASDK (Aglet Software Development Kit 1.1b3), JDK 1.2.2 (Java Development Kit 1.2.1), Textpad 4 dan sebagainya.

Perlaksanaan pembangunan prototaip ini dilakukan mengikut secara bermodul mengikut senario yang telah dikenalpasti.Setelah selesai pembangunan modul-modul, pengintegrasian dan penggabungan modul akan dilakukan untuk menjadi sebuah sistem yang lengkap. Modul-modul ini berasaskan kepada fasa-fasa dalam perlombongan data menggunakan teknik petua sekutuan. Kaedah ini digunakan agar pengesanan ralat dapat dilakukan dengan lebih mudah dan sistematik dan seterusnya pengujian terhadap hasil model rekomendasi dapat dilaksanakan dan proses itu digambarkan dalam perenggan yang seterusnya.

Selanjutnya, untuk menghasilkan sebuah sistem rekomendasi bagi personalisasi Web yang baik, beberapa langkah yang berkesan perlulah diambil. Pertama, kaedah

penyediaan data yang digunakan mestilah mampu menyediakan data yang terbaik untuk dilombong. Kedua, kaedah mendapatkan senarai URL untuk direkomen kepada pengguna. Di dalam bab ini, kecekapan rekomendasi akan diukur dengan melihat kepada sejauh mana ketepatan (*precision*) ramalan yang diperolehi dan pada masa yang sama merangkumi semua URL (*coverage*) yang berkemungkinan akan dipilih oleh pengguna. Terlebih dahulu, kebetulan pengekodan algoritma *Apriori* yang dibangunkan perlu dibuktikan terlebih dahulu. Setelah terbukti bahawa kod tersebut betul, pengujian untuk mendapatkan nilai ambang yang terbaik akan dijalankan dengan menggunakan prosidur pengujian *ten-fold cross validation*. Seterusnya, keberkesanan dalam menentukan kaedah pengkadaran yang terbaik seperti yang diterangkan di dalam Bab 3 akan diuji. Kemudian, pemilihan pengukur keserupaan yang sesuai seperti yang disenaraikan di dalam Bab 4, akan ditunjukkan sebelum perbandingan antara model yang dibangunkan dengan model ataupun teknik-teknik lain dilaksanakan.

4.2 Persekitaran Pembangunan

Selain daripada itu, terdapat beberapa lagi perisian yang diperlukan untuk membolehkan pengkodan aturcara dilakukan. Antaranya, terdapat dua perisian yang perlu dimuat turun daripada internet iaitu:

i. JDK 1.2.1 – http://www.javasoft.com
ii. Aglets 1.1b3 – http://www.tri.ibm.co.jp/aglets

4.3 Pakej Aglets 1.1b3 dan JDK 1.2.1

Untuk melaksanakan prototaip sistem ini, pakej Aglets 1.1b3 dan JDK 1.2.1 digunakan. Perisian Aglets 1.1b3 digunakan untuk menjadikan modul yang akan dibangunkan sebagai agen dan menyediakan platform untuk perlaksanaan sistem berasaskan agen manakala JDK 1.2.1 pula dibangunkan untuk membangunkan kod

aturcara sistem. Beberapa perkara perlu dilakukan sebelum aplikasi Java dapat dilarikan di dalam persekitaran Aglets seperti berikut:

1. Salin fail JDK 1.2.1 dan Aglets 1.1b3 ke dalam cakera keras. Sebagai contoh, Sekiranya cakera keras anda berlabel C, maka salin fail tadi ke dalam C.
2. Ubah fail *autoexec.bat* seperti di bawah untuk mensetkan direktori bagi JDK 1.2.1 dan Aglets 1.1b3.

```
set AGLET_HOME =C:\Aglets1.1b3
set JDK_HOME =C:\JDK1.2.1
set PATH=%PATH%;C:\%JDK_HOME%\bin;%AGLET_HOME%\bin
set CLASSPATH=%CLASSPATH%;C:\5JDK_HOME%\lib\tools.jar;.;
                %AGLET_HOME%\lib\aglets1_2.jar
set AGLET_PATH=%AGLET_HOME%\public
set AGLET_EXPORT_PATH+%AGLET_PATH%
```

3. Salin fail *agletsd1_2.bat* ke dalam C:\Aglets1.1b3\bin.
4. Salin fail *agletsd1_2.jar* ke dalam C:\Aglets1.1b3\lib.
5. Masukkan pernyataan di bawah pada fail *java.policy* pada direktori C:\jdk1.2.2\jre\lib\security.

```
grant codeBase “file://C:/Aglets1.1b3/lib/-“ {
        permission java.security.AllPermision;
};
```

6. Masukkan pernyataan berikut pada fail *java.security* pada directori C:\jdk1.2.2\jre\lib\security.

```
policy.url.3=file:${user.home}/.aglets/security/aglets.policy
```

7. Taip pernyataan berikut dan namakannya sebagai fail *aglets.props* ke dalam cakera keras berlabel C

```
aglets.owner.name=<user_name>
aglets.owner.password=<user_ password >
aglets.keystore.password=<user_login_ password >
```

8. Taipkan pernyataan di bawah pada *MS-DOS Prompt* untuk melarikan Aglet.

```
C:WINDOWS>aglets1_2 –f aglets.props
```

9. Taip pernyataan di bawah pada *MS-DOS Prompt* untuk mencipta key di dalam *keystore*.

```
C:WINDOWS>keytool –genkey –alias user_name
```

4.4 Pelayan Tahiti

Pakej alatan pembangunan Aglet turut mengandungi pelayan Tahiti. Tahiti adalah aplikasi Java yang membenarkan pengguna untuk menerima, mengurus dan menghantar aglet ke komputer-komputer lain yang turut memasukkan pernyataan di bawah pada *MS-DOS Prompt*.

```
C:WINDOWS>agletsd1_2
```

Pengguna juga boleh melarikan lebih daripada satu pelayan Tahiti pada satu-satu masa.Untuk melaksanakannya, pengguna perlu menyatakan nombor port agi pelayan Tahiti. Secara lalainya, nombor port adalah 4434.

```
C:WINDOWS>agletsd1_2 –port 9000
```

4.5 Set Data

Data daripada tiga tapak Web telah diperolehi untuk pengujian di dalam tesis ini. Maklumat-maklumat yang berkaitan dengan data-data ini adalah seperti di dalam Jadual 4.1. Ketiga-tiga set ini hanya mempunyai log pelayan Web tanpa maklumat tambahan seperti struktur tapak Web dan kandungan setiap halaman Webnya. Walaubagaimanapun, set-set data ini sesuai untuk pengujian model yang dibangunkan kerana tujuan ia dibangunkan adalah untuk melihat bagaimana data penggunaan Web dapat membantu dalam personalisasi halaman Web.

Tapak Web Fakulti Sains Komputer dan Sistem Maklumat (FSKSM), Universiti Teknologi Malaysia bertarikh 2 Julai 2003 hingga 17 Disember 2003 menyediakan maklumat tentang fakulti tersebut. Ia mempunyai pautan kepada setiap jabatan yang terdapat di fakulti seperti Jabatan Kejuruteraan Perisian, Jabatan Sistem Maklumat dan Jabatan Sistem dan Komunikasi Komputer. Selain itu, pautan kepada halaman kemudahan, staf dan pusat-pusat kecemerlangan juga turut dimuatkan. Pelajar-pelajar boleh mengakses tapak Web setiap pensyarah untuk mendapatkan nota-nota kuliah dan bahan-bahan untuk ulangkaji. Data ini akan digunakan untuk kebanyakkan pengujian yang terdapat di dalam bab ini.

Tapak Web Pusat Angkasa Kennedy NASA15 bertarikh 1 Julai 1995 hingga 31 Ogos 1995 pula menyediakan maklumat tentang aktiviti, berita dan perkembangan terkini tentang NASA. Walaubagaimanapun, tiada capaian dilakukan pada 1 Ogos 1995 jam 14:52:01 hingga 3 Ogos 1995 jam 04:36:13 kerana pelayan ditutup akibat daripada taufan Erin.

Tapak Web Universiti Saskatchewan16 bertarikh 1 Jun 1995 hingga 31 Disember 1995 menyediakan maklumat tentang Universiti Saskatchewan yang terletak di Saskatoon, Saskatchewan, Kanada. Ia mempunyai pautan kepada perpustakaan, fakulti-fakulti, jabatan-jabatan dan segala maklumat yang berkaitan dengan universiti ini. Ia juga membolehkan pelajar menyemak hal-hal berkaitan kursus yang diambil, mengakses

bahan-bahan rujukan atau kuliah dan sebagainya. Ia merupakan sebuah halaman Web rasmi yang biasa dibangunkan oleh setiap universiti untuk menyebarkan informasi dan aktiviti-aktiviti terkini universiti tersebut.

Jadual 4.1: Set data yang digunakan untuk pengujian

Nama Tapak	FSKSM	NASA	SASKATCHEWAN
Bilangan Masukan	1,046,442	3,461,553	405,014
HTML unik	3,656	2,419	2,736
Bilangan Sesi Pelayan	36,250	281,188	94,420
Bilangan Masukan (selepas ditapis)	292,937	939,659	334,381
HTML unik (selepas permbersihan)	140	308	553
Bilangan Sesi Pelayan (selepas permbersihan)	4,640	42,922	12,481
Panjang transaksi maksimum (selepas pembersihan)	36	26	33
Purata panjang transaksi (selepas pembersihan)	3	3	3
Sisihan piawai (selepas pembersihan)	169.16	1,540.92	322.48
Tahap kejarangan data (selepas pembersihan)	0.9523	0.9803	0.9860

Tahap kejarangan data diukur untuk melihat kesan kepadatan data terhadap rekomendasi yang dijanakan oleh enjin rekomendasi. Adakah data yang lebih padat akan menjana rekomendasi yang lebih baik? Formulasi untuk mengira tahap kejarangan data adalah seperti berikut (Demiriz, A, 2002):

$$1-\frac{\sum_{i=1}^{k} support(L)}{KxT} \qquad \textbf{(4.1)}$$

Support (L) adalah jumlah kekerapan setiap item di dalam set data, *K* adalah bilangan URL unik dan *T* adalah bilangan transaksi ataupun sesi pelayan.

4.6 Metodologi dan Metrik untuk Pengujian

Metodologi pengujian yang digunakan di dalam tesis ini adalah sama dengan kaedah pengujian yang biasa digunakan bagi mengukur kecekapan rekomendasi yang dihasilkan oleh sebuah sistem rekomendasi (Gunduz, S., dan Ozsu, M. T. A, 2003, Demiriz, A, 2004, Mobasher, B., Dai, H., Luo, T., and Nakagawa, M, 2001). Setelah melalui fasa prapemprosesan, halaman-halaman Web yang mempunyai kekerapan kurang daripada 0.1% atau melebihi 80% daripada jumlah sesi pelayan yang diperolehi akan disingkirkan bagi mengurangkan kebisingan dan mengelakkan data yang mempunyai kekerapan yang sangat tinggi daripada mendominasi pembinaan petua-petua. Ini adalah kerana halaman yang mempunyai kekerapan kurang daripada 0.1% akan menghasilkan nilai *minimum support* yang tinggi. Akibatnya, lebih banyak halamana lain akan disingkirkan semasa proses penjaanaan petua kerana tidak dapat melepasi nilai ambang yang ditetapkan. Begitu juga dengan halaman yang terlalu kerap sehingga melebihi 80% daripada jumlah keseluruhan halaman. Ia akan mengakibatkan banyak halaman-halaman yang kurang kerap disingkirkan semasa penjanaan petua dan seterusnya mendominasi petua-petua sekutuan yang dijanakan. 70% daripada baki sesi yang tinggal akan digunakan untuk latihan dan mendapatkan petua-petua manakala 30% lagi digunakan untuk pengujian.

Rekomendasi akan dilakukan berasaskan kepada halaman-halaman yang telah diklik oleh pengguna aktif. Sebelum itu, dimaklumkan bahawa petua sekutuan adalah merupakan asas bagi model yang dibangunkan. Proses melombong petua sekutuan adalah seperti yang dijelaskan di dalam Bab 3. Seterusnya, keserupaan suatu URL akan diukur dengan mendapatkan petua-petua yang melibatkan URL tersebut. Mengukur keserupaan berasaskan kosinus akan digunakan untuk pengujian di dalam laporan akhir ini, yang

mana proses untuk mendapatkan item-item yang paling serupa adalah seperti yang telah juga dibincangkan di dalam Bab 2.

Setiap transaksi, t di dalam set data pengujian akan dibahagikan kepada dua bahagian. *Y* halaman yang pertama di dalam *t* akan diambil sebagai sesi pengguna aktif, u_t manakala baki $t-n$ halaman lagi akan digunakan untuk menguji ketepatan rekomen yang dijanakan, dinamakan sebagai set penilaian, E_t. *w* mewakili saiz tetingkap, yang mana saiz tetingkap sesi pengguna aktif adalah sebhagian daripada aliran klik pengguna yang diperlukan oleh enjin rekomendasi yntuk menghasilkan set rekomendasi. Daripada Jadual 4.1, didapati purata panjang transaksi bagi set data FSKSM adalah tiga, begitu juga dengan data NASA dan Saskatchewan. Oleh itu, saiz tetingkap yang akan digunakan adalah dua untuk pengujian ini. Set calon rekomendasi ini diwakili oleh $c(u_t, \tau)$. Walaubagaimanapun, hanya *N* calon paling serupa akan direkomenkan kepada pengguna, di wakili oleh $R(c(u_t, \tau), N)$.

Pengukuran keupayaan enjin rekomendasi untuk melaksanakan personalisasi Web yang baik oleh setiap teknik yang akan diuji diukur menggunakan tiga metrik pengujian yang biasa digunakan di dalam perlombongan penggunaan Web iaitu *precision* (Mobasher, B., Dai, H., Luo, T., and Nakagawa, M., 2001), *coverage* (Mobasher, B., Dai, H., Luo, T., and Nakagawa, M., 2001) dan *F1* (Lewis, D., dan Gale, W. A, 1994). Langkah terakhir, bagi setiap nilai ambang τ yang dikenakan, purata nilai ukuran yang diperolehi bagi keseluruhan transaksi yang terdapat di dalam set penilaian akan dijadikan sebagai nilai ukuran pengujian yang sebenar bagi setiap metrik. Nilai ambang *minimum support* dan *minimum confidence* akan ditentukan di dalam pengujian penentuan nilai ambang di dalam tesis ini. Nilai-nilai ambang yang terbaik akan dipilih dan digunakan untuk pengujian-pengujian lain di dalam laporan akhir ini.

4.6.1 Precision

Precision mengukur darjah ketepatan ramalan yang dilakukan oleh enjin rekomendasi. Ia diperolehi dengan mendapatkan bilangan item-item yang bersilang antara item yang terdapat di dalam set yang direkomen kepada pengguna dengan set penilaian. Selepas itu, bilangan item direkomenkan. Formulasi untuk mendapatkan *precision* adalah seperti berikut (Mobasher, B., Dai, H., Luo, T., and Nakagawa, M., 2001):

$$pre(R(c(u_t,\tau),N)) = \frac{|\, R(c(u_t,\tau),N) \cap E_t \,|}{|\, R(c(u_t,\tau),N) \,|} \qquad \textbf{(4.2)}$$

$R(c(u_t,\tau),N)$ adalah *N* item calon yang paling setupa dengan sesi pengguna aktif, u_t. E_t pula mewakili set penilaian untuk menguji rekomendasi yang dijanakan.

4.6.2 Coverage

Coverage mengukur keupayaan sesuatu enjin rekomendasi merekomen halaman-halaman Web yang mungkin akan dilawati oleh pengguna. Ia diperolehi dengan mendapatkan bilangan persilangan antara item-item yang terdapat di dalam set rekomendasi dan set penilaian. Selepas itu, bilangan persilangan ini akan dibahagikan dengan bilangan item di dalam set penilaian. Formulasi untuk mendapatkan *coverage* adalah seperti berikut (Mobasher, B., Dai, H., Luo, T., and Nakagawa, M., 2001):

$$\mathrm{cov}(R(c(u_t,\tau),N)) = \frac{|\, R(c(u_t,\tau),N) \cap E_t \,|}{|\, E_t \,|} \qquad \textbf{(4.3)}$$

Sama seperti *precision*, $R(c(u_t,\tau),N)$ adalah *N* item calon yang paling serupa dengan sesi pengguna aktif, u_t. E_t pula mewakili set penilaian untuk menguji rekomendasi yang dijanakan.

4.6.3 F1

Metrik *precision* dan *coverage* selalunya akan saling berkonflik antara satu sama lain, yang mana sekiranya nilai *precision* naik, maka nilai *coverage* akan turun. Untuk itu, metrik *F1* yang menggabungkan kepentingan kedua-dua nilai *precision* dan *coverage* akan digunakan di dalam laporan akhir ini. Penggunaan metrik ini memudahkan pengukuran terhadap kecekapan teknik atau model sistem rekomendasi yang dibangunkan dengan mengambil kira kedua-dua nilai metrik *precision* dan *coverage.* Formulasi untuk mendapatkan *F1* adalah seperti berikut (Lewis, D., dan Gale, W. A, 1994):

$$F1(R(c(u_t,\tau),N)) = \frac{2\times pre(R(c(u_t,\tau),N))\times \mathrm{cov}(R(c(u_t,\tau),N))}{pre(R(c(u_t,\tau),N)) + \mathrm{cov}(R(c(u_t,\tau),N))} \quad \textbf{(4.4)}$$

4.7 Skema Input dan Output Pengujian

Input dan output bagi setiap pengujian yang akan dilakukan adalah sama dari segi skema. Cuma, perbezaan hanya akan berlaku pada nilai pemberat kerana pengujian akan turut di lakukan untuk melihat kesan pemberat yang berlainan dalam menghasilkan rekomendasi yang baik. Rajah 4.1 dan Rajah 4.2 menunjukkan semua skema input dan output bagi pengujian laporan akhir ini. Walaubagaimanapun, skema ini tidak akan digunakan untuk pengujian kebetulan kod algoritma yang dibangunkan.

Petua	URL-URL serupa	Aliran Klik Pengguna
URL-URL Pemberat *Support* *Confidence*	URL-URL Pangkat	URL-URL

Rajah 4.1: Skema Input

Rekomendasi
URL-URL Pangkat

Rajah 4.2: Skema Output

4.8 Pengujian Kebetulan Algoritma

Di dalam Bab 3 laporan akhir ini, huraian lengkap tentang algoritma *Apriori* yang digunakan untuk melombong petua sekutuan telah dilakukan. Sedikit pengubahsuaian telah dibuat untuk menyesuaikan penggunaan algoritma tersebut di dalam laporan akhir ini. Kebetulan ataupun kesahihan algoritma yang telah dibangunkan hendaklah diuji terlebih dahulu untuk memastikan petua yang diperolehi adalah betul dan boleh dipercayai. Untuk tujuan ini, perbandingan antara kod *Apriori* yang telah dibangunkan dengan kod yang telah dibangunkan oleh (Borgelt, C., dan Kruse, R, 2002) perlu dilakukan. *Apriori v4.24* ini boleh dimuat turun secara percima di Internet. Versi-versi terdahulu kod ini telah diaplikasikan ke dalam alatan perlombongan data *Clementine* yang dikeluarkan oleh *SPSS*. Kod yang dibangunkan ini menggunakan teknik gelintaran kelebaran dahulu dan menyusun set-set item ke dalam bentuk pepohon awalan. Teknik ini dapat mempercepatkan masa pengiraan kekerapan setiap set item. Kaedah pengujian kod algoritma ini berlainan daripada pengujian-pengujian lain di dalam laporan akhir ini

kerana tujuan utamanya adalah untuk menentukan sama ada kod yang telah dibangunkan (*ARsim*) akan menghasilkan keputusan ataupun petua sekutuan yang sama dengan kod bandingan *Apriori v4.24*.

4.8.1 Kaedah

Kedua-dua algoritma *Apriori* yang telah dibangunkan dan diperolehi, iaitu *ARsim* dan *Apriori v4.24* ini akan dilaksanakan ke atas data FSKSM yang digunakan untuk pengujian di dalam laporan akhir ini. Input bagi pengujian ini adalah log capaian Web yang telah dibersihkan dan output adalah petua-petua sekutuan yang dijana daripada log tersebut.

4.8.2 Keputusan

Nilai-nilai ambang, iaitu *minimum support* dan *minimum confidence* telah dipilih secara rawak. Nilai-nilai ini dipilih secara rawak kerana pengujian ini hanya perlu memastikan bahawa bilangan petua dan petua yang dijanakan oleh kod *ARsim* adalah sama dengan *Apriori v4.24.* Jadual 5.2 menunjukkan bilangan petua yang diperolehi menggunakan nilai-nilai ambang yang dipilih.

4.8.3 Perbincangan

Keputusan yang diperolehi adalah konsisten bagi setiap nilai ambang yang digunakan. Seperti yang ditunjukkan didalam Jadual 4.2, bilangan petua mengikut bilangan item turut dikaji dan didapati semuanya memberikan bilangan petua yang sama. Untuk itu, keputusan ini telah membuktikan bagawa kod algoritma *ARsim* adalah betul dan boleh digunakan untuk pengujian-pengujian seterusnya di dalam bab ini. Perbezaan antara kod algoritma *ARsim* dengan *Apriori v4.24* adalah dari segi masa yang digunakan

untuk menjana petua-petua. *Apriori v4.24* adalah lebih laju kerana ia menggunakan pepohon. Walaubagaimanapun, kod ini tidak dapat menjana petua sekiranya nilai ambang yang digunakan terlalu rendah. Ini adalah kerana ia akan memerlukan ruang memori yang sangat besar untuk menyimpan pepohon semasa proses penjanaan petua.

Jadual 4.2: Bilangan petua sekutuan untuk setiap set data

Support%	Confidence%	Bilangan Petua
1	10	Jumlah : 974 1-item : 127 2-item : 441 3-item : 306 4-item : 100
1	50	Jumlah : 373 1-item : 127 2-item : 117 3-item : 89 4-item : 40
2	10	Jumlah : 183 1-item : 64 2-item : 84 3-item : 25 4-item : 10
2	40	Jumlah : 156 1-item : 64 2-item : 60 3-item : 22 4-item : 10

4.9 Pemilihan Nilai Ambang

Pelombongan petua sekutuan akan menghasilkan jumlah petua yang banyak sekiranya tidak dikawal. Nilai ambang diperlukan untuk membataskan jumlah petua yang akan dijanakan. Pemilihan nilai ambang yang sesuai dalam mana-mana perlombongan petua sekutuan adalah penting untuk memastikan petua-petua yang dijanakan tidak terlalu banyak ataupun tidak terlalu sedikit. Oleh itu, pengujian untuk menentukan nilai ambang yang sesuai bagi tesis ini perlu dilakukan. Nilai-nilai ambang yang didapati sesuai dan terbaik akan digunakan untuk pengujian-pengujian seterusnya di dalam laporan akhir ini.

4.9.1 Kaedah

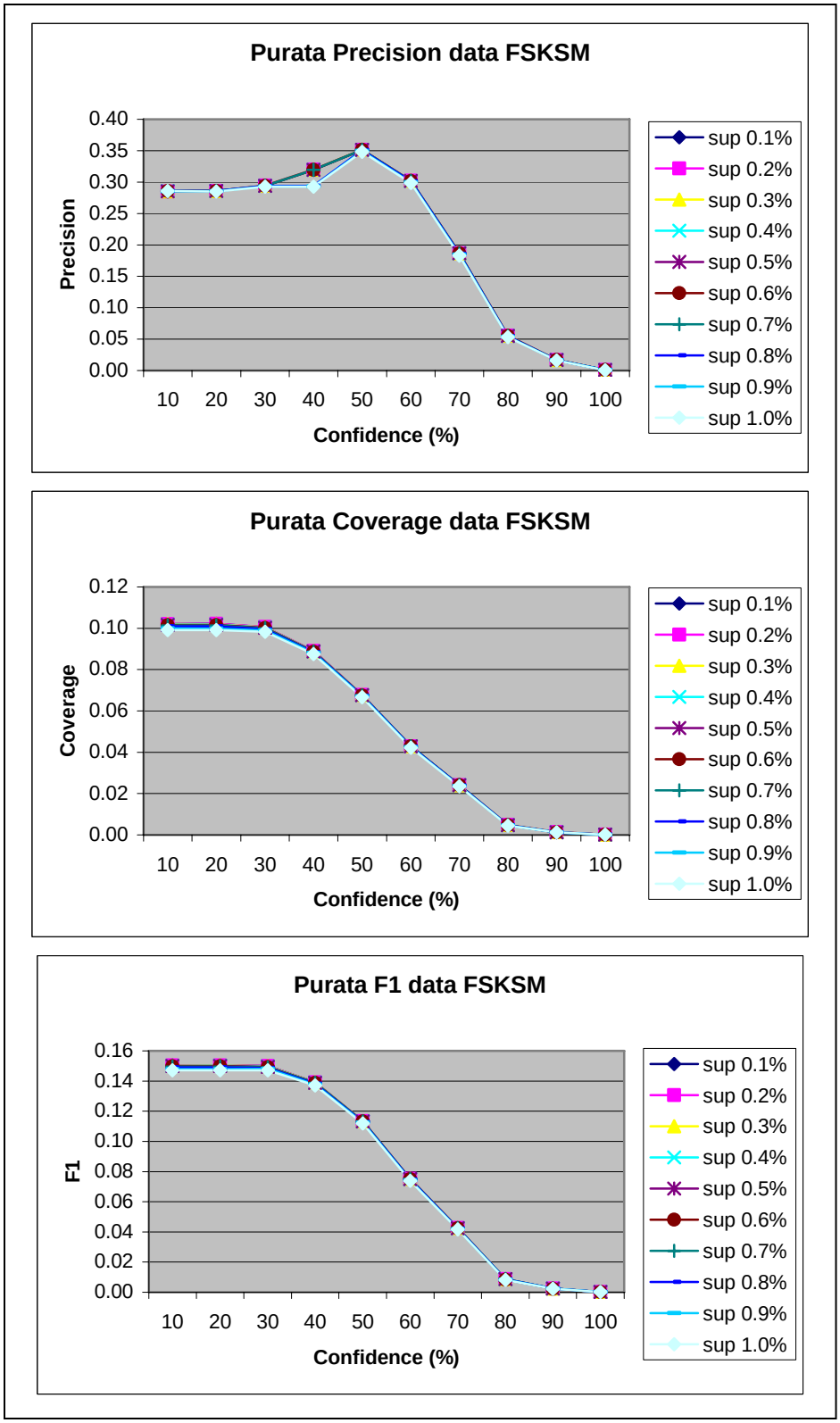
Untuk mendapatkan nilai ambang yang terbaik, pengujian ini akan dilakukan menggunakan *ten-fold cross validation*. Set data asal FSKSM akan dibahagikan kepada dua, iaitu 70% untuk set data latihan manakala baki 30% lagi untuk set data pengujian. Pembahagian data-data ini akan dilakukan secara rawak. Proses pembahagian data secara rawak ini akan diulang sebanyak sepuluh kali. Maksudnya disini, set data FSKSM akan mempunyai sepuluh set data latihan dan sepuluh set data pengujian. Seterusnya, set-set data lain iaitu NASA dan Saskatchewan juga akan melalui proses yang sama. Keseluruhannya, 30 set data akan diuji untuk menentukan nilai ambang yang sesuai. Teknik yang digunakan untuk pengujian *Apriori* sahaja untuk menjana petua-petua sekutuan. Input dan output bagi pengujian ini adalah seperti yang ditunjukkan di dalam Rajah 4.1 dan Rajah 4.2.

4.9.2 Keputusan

Nilai-nilai ambang yang diuji untuk *minimum support* adalah dari 0.1% hingga 1.0% manakala untuk *minimum confidence* pula adalah dari 10% hingga 100%. Jadual 4.3, Jadual 4.4, Jadual 4.5 menunjukkan nilai-nilai purata keputusan bagi setiap set data dengan jelas. Graf purata keputusan untuk *precision, coverage* dan *F1* bagi setiap set data pula ditunjukkan di dalam Rajah 4.3, Rajah 4.4 dan Rajah 4.5. Nilai ukuran tertinggi bagi tiga-tiga metrik *precision, coverage* dan *F1* yang dicatatkan oleh enjin rekomendasi ditebalkan hurufnya seperti yang ditunjukkan di dalam ketiga-tiga jadual. Jadual 4.6 pula menunjukkan purata bilang petua yang dihasilkan oleh setiap nilai ambang bagi setiap set data.

Jadual 4.3: Purata keputusan set data FSKSM

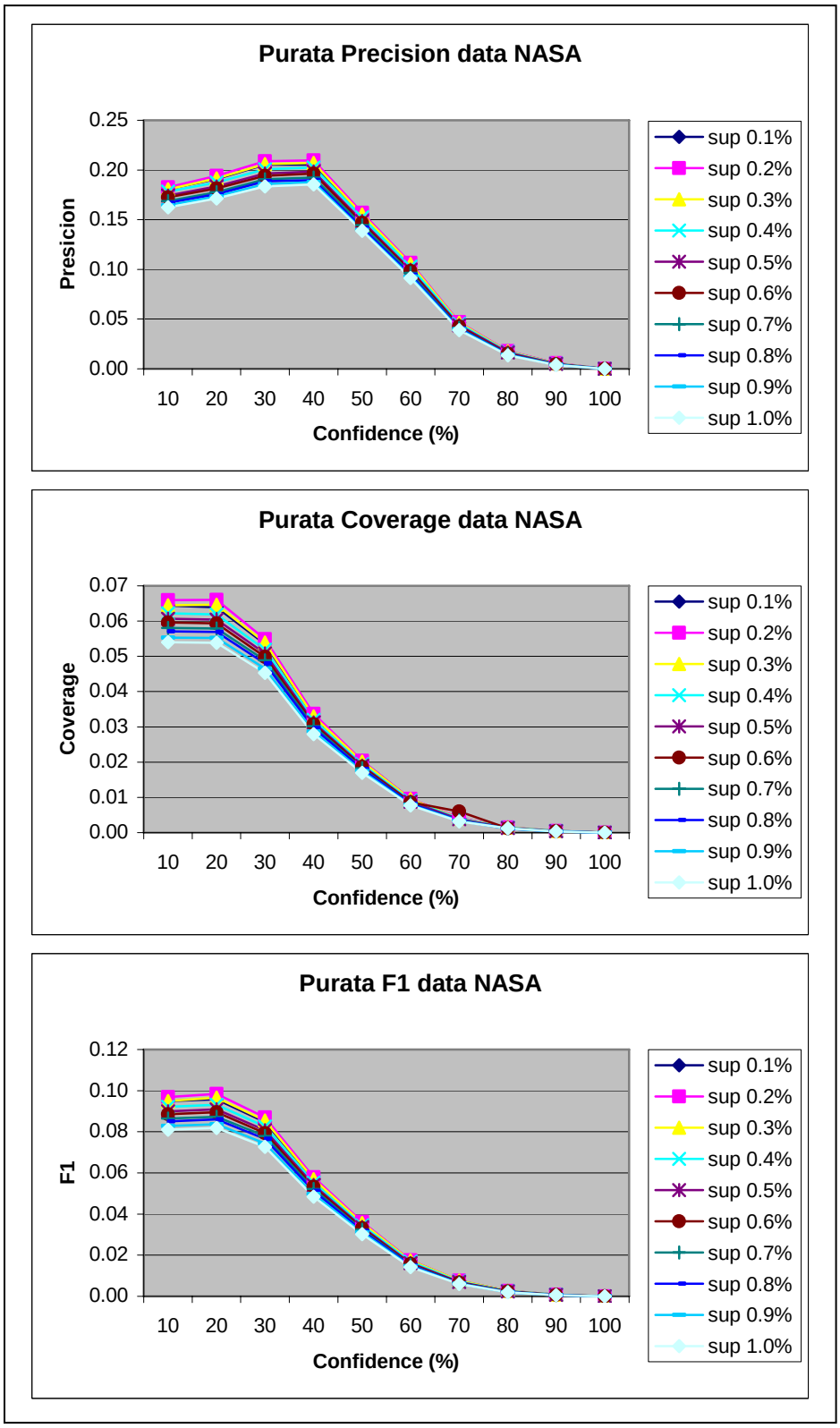
NILAI AMBANG	Conf. (%) / Sup. (%)	10	20	30	40	50	60	70	80	90	100
Precision (%)	0.1	28.50	28.57	29.44	31.99	35.15	30.22	18.67	5.55	1.70	0.09
	0.2	28.50	28.57	29.44	31.98	35.15	30.22	18.67	5.55	1.70	0.09
	0.3	28.54	28.59	29.44	31.99	35.16	30.23	18.68	5.55	1.70	0.09
	0.4	28.58	28.58	29.41	31.95	35.12	30.19	18.65	5.51	1.67	0.09
	0.5	28.53	28.57	29.40	31.94	35.11	30.18	18.64	5.50	1.67	0.09
	0.6	28.56	28.59	29.41	31.94	35.12	30.18	18.65	5.51	1.67	0.09
	0.7	28.57	28.59	29.40	31.94	35.12	30.17	18.62	5.48	1.64	0.09
	0.8	28.58	28.59	29.38	29.38	35.10	30.13	18.57	5.46	1.64	0.09
	0.9	28.51	28.52	29.32	29.32	34.94	29.96	18.41	5.39	1.62	0.08
	1.0	28.59	28.49	29.28	29.28	34.80	29.83	18.28	5.34	1.60	0.06
Coverage (%)	0.1	10.20	10.21	10.05	8.90	6.77	4.29	2.41	0.48	0.14	0.01
	0.2	10.20	10.21	10.06	8.90	6.77	4.29	2.41	0.48	0.14	0.01
	0.3	10.20	10.19	10.05	8.89	6.77	4.29	2.41	0.48	0.14	0.01
	0.4	10.18	10.18	10.03	8.88	6.76	4.28	2.40	0.47	0.13	0.01
	0.5	10.16	10.16	10.02	8.87	6.76	4.28	2.40	0.47	0.13	0.01
	0.6	10.15	10.15	10.01	8.86	6.75	4.28	2.40	0.47	0.13	0.01
	0.7	10.13	10.13	9.99	8.86	6.75	4.28	2.40	0.47	0.13	0.01
	0.8	10.09	10.09	9.95	8.83	6.73	4.27	2.39	0.47	0.13	0.01
	0.9	9.99	9.98	9.90	8.80	6.70	4.24	2.37	0.46	0.13	0.00
	1.0	9.91	9.91	9.83	8.75	6.67	4.21	2.35	0.44	0.12	0.01
F1 (%)	0.1	15.03	15.04	14.99	13.92	11.35	7.52	4.26	0.88	0.25	0.01
	0.2	15.03	15.04	14.99	13.92	11.35	7.52	4.26	0.88	0.25	0.01
	0.3	15.02	15.03	14.98	13.91	11.34	7.51	4.26	0.88	0.25	0.01
	0.4	15.00	15.01	14.96	13.90	11.33	7.50	4.25	0.87	0.24	0.01
	0.5	14.99	14.99	14.94	13.88	11.33	7.50	4.25	0.87	0.24	0.01
	0.6	14.98	14.98	14.93	13.88	11.32	7.50	4.25	0.83	0.23	0.01
	0.7	14.96	14.96	14.90	13.87	11.32	7.49	4.25	0.87	2.40	0.01
	0.8	14.91	14.91	14.86	13.83	11.29	7.47	4.23	0.86	0.24	0.01
	0.9	14.79	14.79	14.80	13.77	11.24	7.42	4.19	0.84	0.24	0.01
	1.0	14.71	14.71	14.72	13.71	11.19	7.38	4.16	0.82	0.23	0.01



Rajah 4.3: Graf purata keputusan data FSKSM

Jadual 4.4: Purata keputusan set data NASA

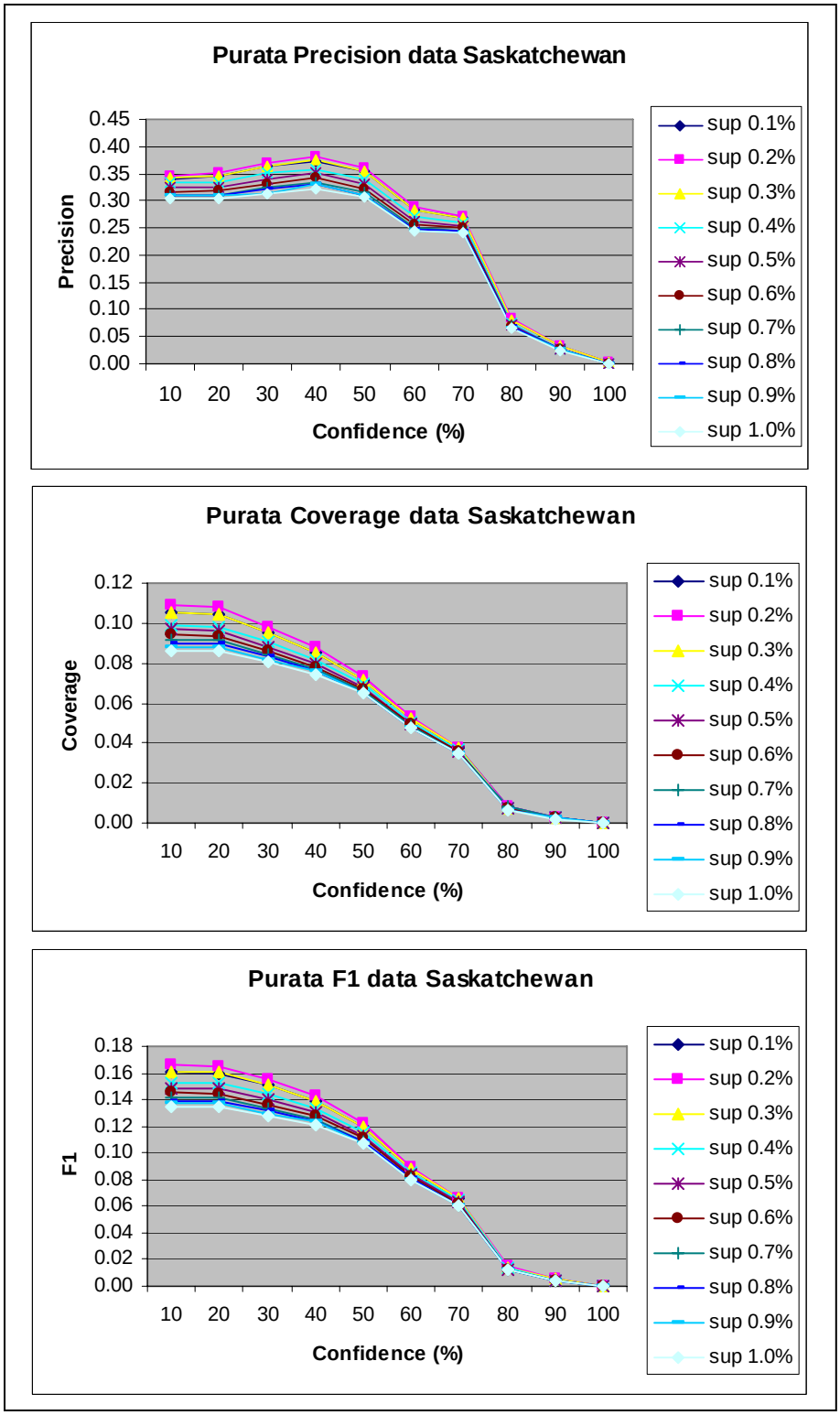
NILAI AMBANG	Conf. (%) / Sup. (%)	10	20	30	40	50	60	70	80	90	100
Presicion (%)	0.1	18.09	19.10	20.54	20.54	15.58	10.58	4.65	1.76	0.56	0.00
	0.2	18.25	19.40	20.88	20.96	15.71	10.65	4.70	1.79	0.57	0.00
	0.3	18.05	19.14	20.58	20.72	15.53	10.55	4.64	1.76	0.55	0.00
	0.4	17.82	18.76	20.08	20.24	15.23	10.29	4.52	1.69	0.52	0.00
	0.5	17.43	18.34	19.63	19.84	14.85	9.97	4.38	1.63	0.49	0.00
	0.6	17.24	18.14	19.42	19.64	14.70	9.85	4.27	1.54	0.45	0.00
	0.7	16.87	17.76	19.07	19.25	14.46	9.66	4.13	1.48	0.42	0.00
	0.8	16.70	17.58	18.87	18.98	14.21	9.46	4.05	1.45	0.41	0.00
	0.9	16.45	17.33	18.57	18.77	13.98	9.27	3.97	1.40	0.40	0.00
	1.0	16.24	17.13	18.36	18.56	13.87	9.13	3.90	1.36	0.38	0.00
Coverage (%)	0.1	6.44	6.39	5.35	3.30	2.03	0.95	0.41	0.15	0.05	0.00
	0.2	6.59	6.60	5.49	3.37	2.04	0.96	0.42	0.15	0.05	0.00
	0.3	6.45	6.46	5.39	3.30	2.01	0.95	0.41	0.15	0.05	0.00
	0.4	6.22	6.19	5.19	3.20	1.94	0.91	0.40	0.14	0.05	0.00
	0.5	6.07	6.04	5.08	3.13	1.90	0.88	0.38	0.13	0.04	0.00
	0.6	5.96	5.94	4.99	3.08	1.87	0.86	0.60	0.12	0.04	0.00
	0.7	5.81	5.79	4.89	3.02	1.83	0.83	0.35	0.12	0.03	0.00
	0.8	5.71	5.69	4.80	2.96	1.80	0.82	0.34	0.11	0.03	0.00
	0.9	5.53	5.52	4.65	2.87	1.75	0.79	0.33	0.11	0.03	0.00
	1.0	5.40	5.38	4.53	2.79	1.70	0.77	0.32	0.12	0.03	0.00
F1 (%)	0.1	9.50	9.58	8.48	5.69	3.59	1.74	0.75	0.27	0.09	0.00
	0.2	9.69	9.84	8.70	5.79	3.62	1.76	0.77	0.27	0.09	0.00
	0.3	9.50	9.66	8.55	5.70	3.57	1.74	0.76	0.27	0.08	0.00
	0.4	9.22	9.31	8.25	5.52	3.45	1.68	0.73	0.25	0.08	0.00
	0.5	9.00	9.09	8.07	5.41	3.37	1.62	0.70	0.24	0.07	0.00
	0.6	8.86	8.95	7.94	5.33	3.31	1.58	0.67	0.23	0.07	0.00
	0.7	8.64	8.73	7.79	5.22	3.25	1.53	0.64	0.22	0.06	0.00
	0.8	8.51	8.60	7.65	5.12	3.19	1.50	0.63	0.21	0.06	0.00
	0.9	8.28	8.37	7.44	4.97	3.11	1.46	0.61	0.20	0.06	0.00
	1.0	8.10	8.19	7.27	4.85	3.02	1.42	0.60	0.20	0.05	0.00



Rajah 4.4: Graf purata keputusan data NASA

Jadual 4.5: Purata keputusan set data Saskatchewan

NILAI AMBANG	Conf. (%) / Sup. (%)	10	20	30	40	50	60	70	80	90	100
Presicion (%)	0.1	33.99	34.44	36.26	37.25	35.41	28.33	26.80	7.98	3.27	0.30
	0.2	34.67	35.13	36.95	38.27	36.20	28.96	27.22	8.23	3.42	0.40
	0.3	34.31	34.63	36.36	37.48	35.41	28.36	26.80	7.95	3.26	0.31
	0.4	33.25	33.48	35.07	35.90	33.95	27.05	25.92	7.39	2.83	0.08
	0.5	32.41	32.61	34.12	35.06	33.06	26.22	25.43	7.18	2.74	0.07
	0.6	31.60	31.77	33.21	34.24	32.24	25.69	25.01	6.98	2.69	0.07
	0.7	30.86	31.03	32.43	33.39	31.49	25.04	24.52	6.88	2.68	0.07
	0.8	30.91	30.96	32.20	33.03	31.07	24.64	24.34	6.77	2.63	0.07
	0.9	31.09	31.14	31.16	32.97	30.92	24.50	24.22	6.65	2.58	0.06
	1.0	30.47	30.50	31.38	32.15	30.71	24.31	24.03	6.52	2.45	0.06
Coverage (%)	0.1	10.50	10.40	9.52	8.52	7.19	5.24	3.73	0.78	0.28	0.02
	0.2	10.88	10.77	9.84	8.76	7.34	5.33	3.79	0.81	0.30	0.03
	0.3	10.54	10.44	9.56	8.54	7.19	5.24	3.74	0.77	0.28	0.02
	0.4	9.93	9.84	9.04	8.18	6.94	5.08	3.65	0.73	0.25	0.00
	0.5	9.67	9.59	8.80	7.98	6.81	4.98	3.60	0.71	0.24	0.01
	0.6	9.43	9.35	8.61	7.83	6.68	4.91	3.54	0.69	0.24	0.01
	0.7	9.15	9.12	8.42	7.69	6.58	4.84	3.49	0.80	0.28	0.01
	0.8	8.99	8.98	8.31	7.63	6.53	4.80	3.48	0.66	0.23	0.01
	0.9	8.78	8.77	8.13	7.56	6.51	4.79	3.47	0.65	0.23	0.00
	1.0	8.62	8.61	8.02	7.46	6.47	4.76	3.46	0.64	0.22	0.00
F1 (%)	0.1	16.04	15.97	15.08	13.87	11.96	8.84	6.55	1.42	0.52	0.03
	0.2	16.56	16.49	15.54	14.25	12.21	9.00	6.65	1.47	0.54	0.05
	0.3	16.12	16.04	15.13	13.92	11.96	8.84	6.56	1.41	0.52	0.03
	0.4	15.29	15.21	14.37	13.32	11.52	8.55	6.39	1.33	0.46	0.00
	0.5	14.89	14.82	13.99	13.00	11.29	8.37	6.30	1.29	0.45	0.01
	0.6	14.52	14.45	13.67	12.74	11.07	8.24	6.21	1.25	0.44	0.01
	0.7	14.12	14.09	13.37	12.51	10.88	8.11	6.11	1.24	0.44	0.01
	0.8	13.93	13.92	13.20	12.39	10.79	8.04	6.09	1.20	0.43	0.01
	0.9	13.70	13.69	12.98	12.30	10.75	8.01	6.07	1.19	0.42	0.01
	1.0	13.44	13.43	12.77	12.12	10.69	7.97	6.04	1.17	0.40	0.01



Rajah 4.5: Graf purata keputusan data Saskatchewan

Jadual 4.6: Purata jumlah purata bagi setiap set data

NILAI AMBANG	Conf. (%) / Sup. (%)	10	20	30	40	50	60	70	80	90	100
Presicion (%)	0.1	55927	51824	47174	42478	37466	31199	24596	17836	10645	6312
	0.2	50650	47318	43224	38965	34339	28524	22401	15931	9434	5101
	0.3	42808	40003	36507	32834	28912	24046	18605	12867	7317	3388
	0.4	36652	34240	31150	27948	24468	20249	15508	10554	5854	2520
	0.5	30909	28903	26176	23396	20385	16677	12521	8362	4548	1828
	0.6	25787	24125	21734	19303	16695	13529	9922	6406	3329	1121
	0.7	21532	20186	18097	15996	13715	10949	7924	5018	2551	793
	0.8	17870	16772	14981	13139	11168	8778	6206	3853	1981	602
	0.9	13784	12890	11458	9902	8302	6414	4433	2645	1340	419
	1.0	11352	10576	9382	8007	6619	5071	3434	2041	1022	325
Coverage (%)	0.1	59152	55091	50078	45129	39989	33556	26469	19320	11814	7253
	0.2	52574	49236	44746	40177	35397	29503	23065	16346	9669	5108
	0.3	44791	41985	38104	34029	29791	24695	19005	12985	7248	3205
	0.4	38549	36101	32661	29085	25308	20837	15819	10451	5526	2076
	0.5	32877	30780	27672	24563	21265	17375	12945	8335	4264	1326
	0.6	27973	26192	23429	20758	17879	14526	10657	6695	3323	978
	0.7	23592	22103	19685	17383	14903	12027	8677	5372	2631	697
	0.8	19997	18756	16635	14593	12465	9979	7089	4307	2171	558
	0.9	15703	14723	13017	11355	9617	7557	5274	3132	1550	398
	1.0	13052	12228	10797	9350	7842	6099	4159	2457	1187	308
F1 (%)	0.1	57540	53458	48626	43804	38728	32378	25533	18578	11230	6783
	0.2	51612	48277	43985	39571	34868	29014	22733	16139	9552	5105
	0.3	43800	40994	37306	33432	29352	24371	18805	12926	7283	3297
	0.4	37601	35171	31906	28517	24888	20543	15664	10503	5690	2298
	0.5	31893	29842	26924	23980	20825	17026	12733	8349	4406	1577
	0.6	26880	25159	22582	20031	17287	14028	10290	6551	3326	1050
	0.7	22562	21145	18891	16690	14309	11488	8301	5195	2591	745
	0.8	18934	17764	15808	13866	11817	9379	6648	4080	2076	580
	0.9	14744	13807	12238	10629	8960	6982	4854	2889	1445	409
	1.0	12202	11402	10090	8679	7231	5585	3797	2249	1105	317

4.9.3 Perbincangan

Melalui analisis yang dibuat ke atas keputusan-keputusan pengujian, didapati ketiga-tiga set data memberikan keputusan yang konsisten. Nilai *minimum support* yang paling baik untuk *precision* adalah 0.2%, manakala untuk *minimum confidence* pula 40%. Bagi *coverage* pula, nilai *minimum support* terbaik adalah 0.2% dan 20% untuk *minimum*

confidence. Bagi pengukuran *F1*, nilai *minimum support* terbaik yang didapati adalah 0.2% dan *minimum confidence* 20%. Secara keseluruhannya, nilai *minimum support* terbaik adalah konsisten, iaitu 0.2% tetapi tidak bagi *minimum confidence*. Keputusan terbaik mengikut *minimum confidence* adalah dari 10% hingga 50%.

Jika dilihat pada purata petua seperti di dalam Jadual 5.6, didapati jumlah petua yang paling banyak bagi setiap set data adalah pada *minimum support* 0.1%. Ini menunjukkan bahawa bilangan petua yang terlalu banyak tidak semestinya akan menghasilkan rekomendasi yang baik. Ini kerana apabila bingan petua bertambah, maka bilangan item utnuk rekomendasi juga akan bertambah. Daripada bilangan item-item yang banyak, hanya lima item terbaik dari segi nilai *confidence* akan direkomen kepada pengguna untuk laporan akhir ini. Oleh itu, banyak item-item akan disingkirkan semasa proses penjanaan senarai rekomendasi. Ini sudah pasti akan membawa kepada keburukan kerana tidak semua item yang mempunyai nilai *confidence* yang tinggi itu memenuhi kehendak pengguna. Pemilihan nilai ambang yang sesuai adalah penting supaya enjin rekomendasi mempunaui bilangan petua yang cukup dan bersesuaian, tidak terlalu sedikit dan tidak terlalu banyak untuk membuat rekomendasi. Bagi ketiga-tiga set data, nilai *minimum support* 0.2% memberikan keputusan yang terbaik. Walaupun bilangan petua yang dijanakan juga agak tinggi, namun ia berupaya untuk menjana rekomendasi yang baik. Berdasarkan keputusan-keputusan seperti yang ditunjukkan di dalam Jadual 4.4, Jadual 4.5 dan Jadual 4.6, nilai *minimum support* 0.2% dan 0.3% akan digunakan untuk pengujian-pengujian seterusnya di dalam laporan akhir ini. Bagi *minimum confidence* pula, nilai-nilai 10%, 20%, 30%, 40% dan 50% akan digunakan. Nilai-nilai ini dipilih kerana keputusan yang baik diperolehi bermula dari nilai ambang 10% hingga 50%. Bagi nilai-nilai ambang seterusnya, keputusan didapati semakin menurun.

4.10 Pengujian Prapemprosesan

Untuk menguji keberkesanan enjin rekomendasi yang telah dibincangkan di dalam Bab 2, data mentah yang diperolehi daripada log pelayan Web mestilah

dibersihkan dan ditukar kepada bentuk sesi pelayan. Sesi pelayan yang dihasilkan mestilah tepat supaya hanya halaman-halaman Web yang sepatutnya akan direkomen kepada pengguna oleh enjin rekomendasi. Oleh kerana set-set data yang digunakan untuk pengujian tidak disertakan dengan ID bagi setiap sesi, maka pengujian untuk membuktikan bahawa sesi yang dikenalpasti melalui heuristik adalah sama dengan sesi yang sebenar tidak dapat dilakukan. Oleh yang demikian, pengujian akan dilakukan dengan mendapatkan kecekapan rekomendasi yang dilakukan oleh setiap teknik yang digunakan. Kesan ambangan terhadap kecekapan enjin rekomendasi akan ditunjukkan di dalam setiap pengujian yang dilaksanakan.

4.10.1 Menguji Kaedah Pengiraan Masa Bagi Setiap Halaman

Log pelayan Web yang biasa tidak dapat merekodkan masa yang digunakan oleh seseorang pengguna untuk melihat sesuatu halaman Web. Biasanya, masa melihat halaman Web akan dikira dengan menolakkan masa ia mula diminta dengan masa halaman Web seterusnya diminta oleh pengguna yang sama, yang mana kedua-dua masa ini direkodkan ke dalam log pelayan Web. Seperti yang telah dihuraikan di dalam Bab 2, disebabkan kesukaran untuk mendapatkan masa yang tepat akibat kelembapan rangkaian dan juga saiz fail yang boleh mempengaruhi masa yang direkodkan oleh log, persamaan 2.1 cuba mengurangkan ralat tersebut dengan mendapatkan purata masa yang digunakan oleh setiap pengguna untuk melihat satu bait fail halaman-halaman Web tersebut.

Hipotesis awal pengujian ini adalah dengan menggunakan masa yang digunakan oleh pengguna untuk melihat satu bait saiz fail halaman sebagai pengkadaran akan dapat meningkatkan keberkesanan sistem rekomendasi kerana ia telah menafikan kesan saiz fail dalam mempengaruhi jumlah masa yang digunakan oleh pengguna untuk melihat suatu halaman Web.

4.10.1.1 Kaedah

Input bagi pengujian ini adalah model transaksi penggunaan Web yang merupakan output daripada proses mengenalpasti sesi pelayan, petua-petua dan aliran klik pengguna. Skema input dan output adalah seperti di dalam Rajah 4.1 dan Rajah 4.2. Pemberat bagi setiap halaman dilihat adalah berasaskan kepada jumlah masa yang digunakan oleh pengguna untuk melihat satu bait saiz fail halaman Web tersebut. Setelah melalui proses perlombongan petua sekutuan dan mengukur keserupaan anara URL, enjin rekomendasi akan merekomen URL-URL lain yang paling serupa dengan halaman dilihat. Peratus ketepatan output daripada kaedah ini akan dibandingkan dengan output yang dihasilkan oleh model transaksi penggunaan Web yang menggunakan jumlah masa penuh melihat suatu halaman Web sebagai pemberat. Web FSKSM dengan direktori asas http://www.fsksm.utm.my/ akan digunakan untuk pengujian ini.

4.10.1.2 Keputusan

Kedua-dua jenis pemberat ini menghasilkan bilangan petua yang sama bagi setiap nilai ambang yang digunakan. Bilangan petua yang dihasilkan adalah seperti ditunjukkan di dalam Jadual 4.7. Contoh petua yang dihasilkan adalah seperti di dalam Rajah 4.3. Nilai ukuran tertinggi bagi tiga-tiga metrik *precision, coverage* dan *F1* yang dicatatkan oleh enjin rekomendasi ditebalkan hurufnya seperti yang ditunjukkan di dalam Jadual 4.8. Jadual 4.9 pula menunjukkan contoh rekomendasi yang dijanakan oleh enjin rekomendasi berdasarkan kepada pemberat yang diberikan kepada setiap URL, nilai *support* 0.1% dan *confidence* 10%. Aliran klik pengguna akrif yang diambil untuk mendapatkan rekomendasi adalah /~aryati dan /~aryati/ajar.htm. Lajur terakhir di dalam rajah menunjukkan aliran klik seterusnya yang dilakukan oleh pengguna. Rekomendasi yang dijanakan oleh pengkadaran menggunakan satu bait saiz fail menghasilkan ukuran *precision* 60% dan *coverage* 100%. Rekomendasi bagi pengkadaran yang menggunakan saiz fail penih pula menghasilkan ukuran *precision* 40% dan *coverage* 67%.

Jadual 4.7: Bilangan petua mengikut nilai ambang untuk menguji kaedah pengiraan masa

Support (%)	Confidence (%)	Bilangan Petua
0.2	10	72188
	20	60241
	30	55169
	40	49593
	50	43828
0.3	10	52341
	20	50372
	30	46026
	40	41223
	50	36229

If /~aryati, /~aryati/ajar.htm **Then** /~aryati/sadm/silibus.htm, 0.5 0.004 0.32, Support 25%, confidence 60%

Rajah 4.6: Contoh petua sekutuan

Jadual 4.8: Keputusan nilai *precision, coverag*e dan *F1* mengikut jenis pemberat untuk data FSKSM.

Sup (%)	Conf (%)	Precision (%)		Coverage (%)		F1 (%)	
		1 Bait	Penuh	1 Bait	Penuh	1 Bait	Penuh
0.2	10	30.1	29.5	27.4	25.3	28.8	27.5
0.2	20	30.0	29.1	27.4	25.1	28.6	26.9
0.2	30	32.3	31.3	28.0	25.9	30.4	28.4
0.2	40	32.5	32.8	25.0	22.6	28.2	26.7
0.2	50	28.0	31.4	16.7	16.1	20.9	21.2
0.3	10	30.4	30.0	27.5	25.4	28.9	27.4
0.3	20	30.1	29.4	27.3	25.2	28.6	27.1
0.3	30	32.7	31.6	**29.4**	26.4	**31.0**	28.8
0.3	40	33.3	32.1	25.2	21.6	28.7	25.8
0.3	50	**33.6**	30.4	17.5	16.3	23.0	21.3

Jadual 4.9: Senarai rekomendasi yang dijanakan mengikut jenis pemberat

Jenis Pemberat	Rekomendasi Top-5	Nilai Pemberat	Set Penilaian
1 Bait Saiz Fail	/~aryati/sadm/silibus.htm	110.73	
	/~nazir	101.112	
	/~foad	87.6027	
	/~nazir/myteaching.htm	80.1107	
	/~aryati/dss/sil200304.htm	78.581	**/~**aryati/sadm/silibus.htm
Saiz Fail Penuh	/~index.htm	124.2242	/~nazir
	/~nazir	123.1124	/~nazir/myteaching.htm
	/~aslinda	122.2919	
	/~aryati/sadm/silibus.htm	115.3879	
	/~aryati/dss/sil200304.htm	101.5292	

4.10.1.3 Perbincangan

Bilangan petua yang dijanakan didapati adalah semakin berkurang apabila nilai ambang bertambah. Semakin banyak petua akan meningkatkan keupayaan enjin rekomendasi untuk merekomen semua URL yang berkemungkinan akan diklik oleh pengguna tetapi ketepatan enjin rekomendasi adalah rendah kerana tidak semua URL yang direkomen kepada pengguna adalah memenuhi kehendak pengguna trersebut. Bilangan petua yang kurang pula akan meningkatkan ketepatan rekomendasi yang dijanakan kerana hanya URL yang paling sesuai sahaja akan direkomen kepada pengguna tetapi ia tidak dapat memenuhi kehendak pengguna secara keseluruhannya kerana hanya memepunyai bilangan URL yang terhad untuk rekomendasi. Jelas di sini menunjukkan betapa pentingnya untuk memilih nilai-nilai ambangan yang bersesuaian bagi membolehkan enjin rekomendasi berfungsi secara optimum.

Jadual 4.8 jelas menunjukkan kelebihan menggunakan masa bagi melihat satu bait fail sebagai pemberat berbanding menggunakan masa melihat keseluruhan saiz fail dari mula hingga akhir, kecuali pada beberapa ambangan ayang mana ukurannya adalah sama. Rujuk Jadual 4.9, URL yang tidak direkomen oleh pemberat kedua (saiz fail penuh) iaitu /~nazir/myteaching.htm sebenarnya tersenarai sebagai calon awal untuk rekomendasi dengan nilai pemberatnya 98.2584. Tetapi nilai ini bukan merupakan antara lima nilai yang tertinggi. Melalui kajian yang lebih terperinci, didapati saiz fail

/~nazir/myteaching.htm adalah 308 bait. Jika dilihat pada fail yang tidak termasuk di dalam set penilaian, misalnya fail /~aslinda, didapati saiz failnya adalah 310 bait. Walaupun bezanya sedikit, namun jumlah pemberat yang dikira adalah melalui proses perlombongan petua sekutuan terlehih dahulu, yang mana kekerapan sesuatu item itu wujud bersamaan juga akan mepengaruhi nilai pemberat. Jadi disini, saiz fail yang lebih daripada fail /~aslinda telah mengakibatkan masa yang lebih sedikit digunakan untuk memuat turun fail tersebut ke pelayar Web pengguna. Akibatnya, pengguna akan mengambil masa yang lebih lama sebelum boleh membuat keputusan tentang halaman mana yang hendak diklik seterunsya. Masa ini dianggap sebagai masa melihat halaman Web tersebut oleh log pelayan Web walaupun sebenarnya bukan. Maka, hipotesis awal pengujian ini telah dibuktikan benar. Walaubagaimanapun, terdapat sedikit kekeliruan tentang kecekapan sebenar rekomendasi yang dijanakan. Ini kerana pertanyaan pasti akan timbul tentang apakah hubungan ataupun perkaitan antara dua halaman yang berbeza iaitu /~aryati/sadm/silibus.htm dan /~nazir/myteaching.htm. Untuk menyelesaikan kekeluruan ini, kedua-dua Web halaman tersebut telah dicapai dan dilihat. Didapati, /~aryati/sadm/silibus.htm memaparkan tentang silibus matapalejaran Kaedah Analisa dan Rekabentuk Sistem. Di dalam halaman ini juga ada disenaraikan nama-nama pensyarah lain yang turut mengajar subjek yang sama dan salah satu daripadanya adalah En. Mohd Nazir Ahmad iaitu tuan punya halaman Web /~nazir/myteaching.htm. Halaman /~nazir/myteaching.htm. pula memaparkan tentang sebari subjek yang diajar oleh tuan punya tapak Web. Antara subjek yang turut tersenarai di dalam halaman in aadalah Kaedah Analisa dan Rekabentuk Sistem. Oleh itu, pengguna mungkin berminat untuk mengetahui tentang siapakah pensyarah yang kan mengajar subjek Kaedah Analisa dan Rekabentuk Sistem dan seterusnya ingin melihat apakah topik-topik yang terangkum di dalam subjek tersebut. Pautan kepada nota-nota yang berkaitan juga terdapat di dalam kedua-dua halaman Web ini. Maka, dapat disimpulkan bahawa rekomendasi yang dijanakan adalah tepat kerana ia menghubungkan dua halaman Web yang berbeza tetapi saling berkaitan deari segi kandungan.

4.11 Perbandingan dengan Teknik-teknik Lain

Untuk mengenalpasti pembaikan dan kelebihan teknik yang telah dibangunkan, iaitu gabungan petua sekutuan dan pengukur keserupaan (ARsim), perbandingan dengan teknik-teknik lain yang sedia ada perlu dilakukan. Di dalam tesis ini, teknik rekomendasi menggunakan petua sekutuan tradisional (AR) akan dibangunakn untuk melihat kesan pengukur keserupaan dalam meningkatkan kecekapan petua sekutuan. Kaedah rekomendasi petua-petua sekutuan dan aliran klik pengguna untuk menjana rekomendasi. Petua yang mempunyai *confidence* tertinggi akan diambil untuk rekomendasi.

Terdapat dua hipotesis untuk pengujian ini. Pertama, dengan mengenalpasti keserupaan antara URL-URL yang menjadi calon awal untuk rekomendasi oleh petua sekutuan, URL akhir yang terpilih untuk dicadangkan kepada pengguna akan menjadi lebih tepat dan sesuai untuk personalisasi. Hipotesis kedua pula berkaitan dengan satu lagi teknik yang akan dibandingkan, iaitu eVZpro. eVZpro mendapatkan senarai rekomendasi dengan mengira keserupaan antara aliran klik penggunadengan petua-petua yang dijanakan oleh sekutuan dan bukannya melalui pemadanan seperti yang biasa digunakan oleh pengkaji-pengkaji lain. Kelebihan utama teknik ini adalah ia pasti tidak akan menghasilkan rekomendasi kosong. Selain itu, eVZpro juga dikatakan mampu untuk mengatasi keupayaan rangkaian kebersandaran dalam membuat peramalan. Untuk itu, hipotesis kedua untuk pengujian ini adalah dengan mendapatkan keserupaan dan *confidence* antara URL dengan aliran klik pengguna akan memberikan keputusan rekomendasi yang lebih baik daripada sekadar mendapatkan keserupaan antara aliran klik pengguna dengan petua.

4.11.1 Kaedah

Input bagi pengujian ini adalah model transksi penggunaan Web yang merupakan output daripada proses mengenalpasti sesi pelayan, petua-petua dan aliran klik pengguna. Skema input dan output adalah seperti di dalam Rajah 4.1 dan Rajah 4.2. Perlombongan

AR akan dilakukan ke atas transaksi-transaksi tersebut untuk mendapatkan petua-petua yang diperlukan. Seterusnya, petua-petua ini akan digunakan untuk menjana rekomendasi menggunakan ketiga-tiga teknik yang hendak dibandingkan. Ketiga-tiga log Web data FSKSM, NASA dan Saskatchewan akan digunakan di dalam pengujian ini. Tujuannya adalah untuk melihat konsistensi model yang telah dibangunkan dalam menghasilkan rekomendasi yang baik.

4.11.2 Keputusan

Jadual 4.10, Jadual 4.11 dan Jadual 4.12 menunjukkan keputusan yang diperolehi oleh ketiga-tiga teknik yang diuji. Bilangan petua bagi setiap teknik adalah sama mengikut set data yang digunakan. Jadual 5.18 menunjukkan dengan lebih terperinci bilangan petua yang digunakan untuk menjana senarai rekomendasi.

Jadual 4.10 : Keputusan perbandingan antara teknik untuk data FSKSM

Sup (%)	Conf (%)	Precision (%)			Coverage (%)			F1 (%)		
		Arsim	Ar	eVZpro	Arsim	Ar	eVZpro	Arsim	Ar	eVZpro
0.2	10	28.4	29.3	19.0	25.1	10.7	8.6	26.7	15.7	11.8
0.2	20	29.3	30.5	19.2	25.6	11.0	8.8	27.3	16.2	12.1
0.2	30	31.9	31.4	20.3	27.1	11.0	9.6	29.3	16.2	13.0
0.2	40	31.1	33.7	19.5	23.1	9.9	10.7	26.5	15.3	13.6
0.2	50	30.1	37.8	19.1	16.7	7.4	9.4	21.5	12.4	12.6
0.3	10	29.4	29.1	19.1	25.5	10.7	8.1	27.3	15.6	11.4
0.3	20	29.7	30.4	19.6	26.1	11.0	8.6	27.8	16.1	12.0
0.3	30	33.1	31.3	20.8	28.8	10.9	10.1	30.8	16.1	13.6
0.3	40	31.6	33.6	20.4	24.7	9.9	10.5	27.7	15.3	13.9
0.3	50	30.2	37.8	19.4	17.6	7.4	9.9	22.2	12.4	13.1

Jadual 4.11 : Keputusan perbandingan antara teknik untuk data NASA

Sup (%)	Conf (%)	Precision (%)			Coverage (%)			F1 (%)		
		Arsim	Ar	eVZpro	Arsim	Ar	eVZpro	Arsim	Ar	eVZpro
0.2	10	23.2	22.1	19.9	18.7	8.1	15.8	20.7	11.9	17.6
0.2	20	23.6	22.4	20.8	19.5	7.8	16.1	21.4	11.5	18.1
0.2	30	22.8	24.0	21.5	13.0	6.5	15.9	16.5	10.3	18.3
0.2	40	18.5	24.3	20.5	6.7	4.1	15.6	9.8	6.9	17.6
0.2	50	11.4	18.5	19.5	3.5	2.5	15.0	5.4	4.4	17.0
0.3	10	22.7	21.8	20.9	18.6	7.9	15.6	20.4	11.6	17.9
0.3	20	23.4	22.0	21.7	19.2	7.6	15.6	21.1	11.3	18.1
0.3	30	22.9	23.6	23.5	13.2	6.3	15.8	16.7	10.0	18.9
0.3	40	18.3	23.9	23.6	6.4	3.9	15.8	9.5	6.8	18.9
0.3	50	11.2	18.2	23.6	3.5	2.4	15.6	5.4	4.3	18.8

Jadual 4.12 : Keputusan perbandingan antara teknik untuk data Saskatchewan

Sup (%)	Conf (%)	Precision (%)			Coverage (%)			F1 (%)		
		Arsim	Ar	eVZpro	Arsim	Ar	eVZpro	Arsim	Ar	eVZpro
0.2	10	38.8	34.9	16.5	29.7	10.2	18.5	33.6	15.8	17.4
0.2	20	39.1	35.2	16.8	30.1	10.7	19.2	34.0	16.4	17.9
0.2	30	41.6	36.9	16.5	26.5	10.0	17.7	32.4	15.7	17.1
0.2	40	41.1	39.2	15.6	22.8	8.8	16.7	39.3	14.4	16.1
0.2	50	38.4	36.1	16.8	18.5	7.3	16.7	25.0	12.1	16.8
0.3	10	37.8	33.7	18.6	28.1	9.9	18.0	32.2	15.3	18.3
0.3	20	37.9	34.4	18.7	28.7	10.3	18.3	32.6	15.8	18.5
0.3	30	40.5	36.0	17.8	25.6	9.6	17.3	31.4	15.2	17.6
0.3	40	39.9	38.0	18.0	22.1	8.6	16.4	28.4	14.0	17.2
0.3	50	37.3	34.9	17.0	18.0	7.0	15.3	24.2	11.7	16.1

Jadual 4.13 : Bilangan petua bagi setiap data mengikut nilai *minimum support* dan *minimum confidence*

Sup (%)	Conf (%)	FSKSM	NASA	Saskatchewan
0.2	10	70990	28006	112876
0.2	20	65861	25313	116142
0.2	30	59169	21943	113005
0.2	40	52103	18481	107547
0.2	50	47928	15279	98874
0.3	10	61259	15560	88875
0.3	20	55778	14372	87307
0.3	30	51526	12679	84852
0.3	40	36322	10825	80651
0.3	50	40411	9269	73990

4.11.3 Perbincangan

Terlebih dahulu, dapat diperhatikan bahawa bilangan petua yang dihasilkan oleh data FSKSM dan Saskatchewan adalah lebih berbanding data NASA. Jika dirujuk kembali Jadual 4.1, tahap kejarangan data Saskatchewan adalah lebih tinggi daripada dua data yang lain. Walaubagaimanapun, ia menghasilkan petua yang paling banyak. Bilangan URL yang banyak mengakibatkan lebih banyak petua dapat dihasilkan kerana kombinasi dan hubungan antara item-item yang lebih banyak dapat dijanakan. Walaupun mempunyai tahap kejarangan data yang rendah berbanding data Saskatchewan dan bilangan URL yang tinggi berbanding data FSKSM, tetapi sisihan piawaian data NASA yang agak tinggi memungkinkan banyak URL disingkirkan semasa proses penjanaan petua akibat daripada kegagalan untuk melepasi had *minimum support*. Ini mengakibatkan bilangan petua yang sedikit dihasilkan. Bagi data FSKSM pula, walaupun mempunyai sisihan piawai yang rendah, namun bilangan URL dan transaksi yang juga rendah berbanding data Saskatchewan memungkinkan bilangan petua yang lebih rendah dihasilkan. Jika dilihat pada Jadual 4.10, Jadual 4.11 dan Jadual 4.12, keputusan yang dihasilkan oleh data NASA adalah lebih rendah daripada data FSKSM dan Saskatchewan. Bilangan petua yang kurang akan mengakibatkan bilangan URL yang akan direkomenkan turut berkurangan dan secara langsung mempengaruhi kecekapan enjin rekomendasi.

Tenik AR menghasilkan keputusan yang terbaik dari segi ketepatan rekomendasi bagi kedua-dua log data FSKSM dan NASA. Walaubagaimanapun, bagi data Saskatchewan, ARsim memberikan ketepatan yang terbaik. Bilangan petua yang banyak membolehkan lebih banyak pilihan rekomendasi dijanakan oleh enjin rekomendasi. Di dalam hal ini, penggunaan pengukur keserupaan di dalam ARsim nampaknya mempunyai kelebihan berbanding AR tradisional daalam menjana rekomendasi. Berpandukan kepada keputusan pengujian, ketepatan rekomendasi ARsim bagi data Saskatchewan dan FSKSM adalah lebih tinggi berbanding data NASA. Dari segi keupayaan untuk memberikan rekomendasi yang merangkumi semua URL yang sesuai dan releven kepada pengguna, AR adalah sangat rendah. Jadual 4.10, Jadual 4.11 dan Jadual 4.12

menunjukkan prestasi enjin-enjin rekomendasi yang menggunakan tiga teknik yang berlainan jelas membuktikan perkara ini. Didapati, rekomendasi yang berasaskan AR adalakalanya tidak dapat merekomenkan URL itu wujud dalam suatu petua. Adakalanya, URL-URL berkenaan tidak wujud langsung dalam mana-mana petua disebabkan ia tidak dapat melepasi nilai ambang yang telah ditetapkan. Oleh itu, URL-URL yang berpotensi tidak dapat direkomenkan kepada pengguna dan ini secara tidak langsung mempengaruhi nilai ukuran *coverage* bagi enjin rekomendasi tersebut. Penggunaan pengukur keserupaan dapat mengimbangi kelemahan ini. Ia adalah kerana walaupun *confidence* suatu URL itu wujud dalam suatu petua tidak melepasi ambangan yang telah ditetapkan, namun keserupaan antara URL tersebut dengan URL-URL lain yang terkandung di dalam petua tadi akan turut diambil kira.

Teknik eVZpro juga menghasilkan ketepatan yang hampir setara dengan teknik-teknik yang lain. Jika dilihat kepada data NASA di dalam Jadual 4.11, persaingan antara ketiga-tiga teknik ini adalah sengit dan hampir serupa. Jika diperhatikan, kelebihan utama eVZpro adalah keputusan yang dihasilkan adalah konsisten bagi setiap nilai ambang yang diberi. Berbanding dengan teknik-teknik lain, terutamanya ARsim yang menghasilkan keputusan yang semakin menurun mengikut pertambahan nilai ambang, eVZpro mampu mengekalkan peratus ketepatannya. Selain itu, eVZpro juga tidak akan menghasilkan rekomendasi kosong, iaitu rekomendasi pasti akan dijanakan bagi setiap input ataupun aliran klik pengguna yang diberikan. Berlainan dengan teknik-teknik lain yang mana adakalanya tidak ada rekomendasi akan dijanakan sekiranya enjin rekomendasi tidak menemui corak atau petua yang bertepatan dengan aliran klik pengguna, terutamanya apabila nilai ambang adalah tinggi. Walaupun eVZpro menghasilkan keputusan yang kosisten, namun ia tidak dapat menghasilkan rekomendasi yang terbaik jika dibandingkan dengan ARsim.

ARsim didapati berjaya memberikan kecekapan rekomendasi yang terbaik berbanding dua lagi teknik yang lain. Walaupun ia tidak mampu untuk menghasilkan ketepatan yang terbaik untuk semua data, namun dari segi *coverage* dan keseluruhan ia adalah yang terbaik. Seperti yang telah dinyatakan di dalam Bab 1 laporan akhir ini,

matlamat loporan akhir ini adalah untuk menghasilkan sebuah model enjin rekomendasi yang cekap, yang mana kecekapan suatu enjin rekomendasi bukanlah diukur berasaskan kepada ketepatan rekomendasi yang dijanakan semata-mata, sebaliknya rekomendasi yang dijanakan juga mestilah mampu untuk merekomenkan semua URL yang mungkin akan dipilih oleh pengguna. Daripada Jadual 4.10, Jadual 4.11 dan Jadual 4.12, prestasi ARsim tidaklah begitu jauh dari segi ukuran *precision* berbanding teknik-teknik lain, malah mampu menjadi yang terbaik untuk data Saskatchewan. Cuma, antara kelemahan utama ARsim adalah keputusan yang dihasilkan akan semakin merosot apabila nilai ambang bertambah. Dari segi ukuran *coverage* pula, didapati ARsim adalah yang terbaik. Keputusan ukuran *F1* telah membuktikan kedua-dua hipotesis yang telah ditetapkan untuk pengujian ini. Didapati, dengan mendapatkan item-item yang paling serupa dengan item input dan kemudiannya mendarabkan nilai-nilai keserupaan ini dengan *confidence* item tersebut wujud dalam suatu petua akan meningkatkan kecekapan enjin rekomdasi. Ini telah menunjukkan bahawa hipotesis pertama adalah benar. Mendapatkan keserupaan antara aliran klik pengguna dengan petua untuk menjana rekomendasi sememangnya dapat menghasilkan keputusan yang baik dan konsisten. Walaubagaimanapun, dengan menggabungkan keserupaan antara URL dan kebarangkalian URL tersebut untuk diklik oleh pengguna akan menghasilkan keputusan yang lebih baik. Ini telah membenarkan hipotesis kedua bagi pengujian ini. Walaupun dari segi konsistensi eVZpro adalah terbaik berbanding ARsim, namun pemilihan nilai ambang yang sesuai akan dapat mengatasi masalah ini.

BAB 5

KESIMPULAN

Perlombongan data penggunaan Web telah menjadi satu lagi bidang kajian yang sangat penting dalam membekalkan pelayan Web dengan kepintaran untuk lebih memahami pengguna. Perkembangan bidang ini boleh dilihat dengan bertambahnya bilangan kajian dan pengkaji-pengkaji yang mengaplikasikan pelbagai teknik perlombongan data ke atas pelayan Web. Antara yang paling penting adalah personalisasi Web yang mampu memberikan pengguna apa yang mereka mahu atau perlukan tanpa perlu mereka nyatakan dengan jelas. Walaupun data demografik boleh diaplikasikan ke dalam perlombongan data penggunaan Web, namun laporan akhir ini telah membuktikan bahawa data penggunaan Web yang lazim juga mampu untuk menjana rekomendasi yang baik. Beberapa isu penting juga telah dibangkitkan antaranya privasi pengguna dan kesukaran untuk mendapatkan data yang benar-benar menggambarkan kelakuan pengguna semasa melayari halaman web.

Keputusan-keputusan di dalam Bab 4 telah menunjukkan betapa pentingnya untuk melaksanakan proses penyediaan data yang baik bagi mendapatkan peraturan-peraturan yang relevan dan seterusnya menjanakan rekomendasi yang tepat. Selain itu, penggunaan teknologi agen dalam perlombongan data ke atas pelayan Web juga menyumbang impak

yang besar dalam menjana model rekomendasi halaman Web untuk pengguna. Ciri-ciri yang dimiliki oleh agen membantu mempercepatkan masa pemprosesan dalam menghasilkan penjanaan rekomendasi yang baik. Kelancaran pemprosesan yang melaksanakan banyak modul-modul dalam perlombongan data dapat ditingkatkan dengan menggunakan teknologi agen.

Di dalam laporan akhir ini juga, penggunaan masa melihat halaman Web sebagai pengkadaran bagi setiap halaman Web telah berjaya meningkatkan keupayaan enjin rekomendasi. Pengunaan masa untuk menentukan kepentingan suatu halaman Web meskipun kurang popular namun dengan melakukan sedikit pengubahsuaian ke atas masa tersebut telah berjaya meningkatkan kecekapan model yang telah dibangunkan, iaitu ARsim, seperti yang dibuktikan di dalam laporan akhir ini. Ini adalah kerana masa melihat suatu halaman Web itu adalah mewakili minat seseorang pengguna terhadap halaman tersebut. Jika ia direkod dengan betul, pentadbir Web ataupun organisasi pasti akan lebih memahami apa yang dikehendaki oleh pengguna mereka dan seterusnya beberapa perubahan boleh dilakukan ke atas tapak Web mereka demi memenuhi kehendak pengguna, yang juga pelanggan mereka.

Penggunaan pengukur keserupaan bersama-sama dengan petua sekutuan telah berjaya meningkatkan lagi kecekapan peramalan untuk tujuan rekomendasi. Ia berjaya hampir menyeimbangkan keputusan *precision* dan *coverage* yang saling berkonflik sekiranya nilai ambang, iaitu *minimum support* dan *minimum confidence* yang digunakan adalah sesuai. Walaubagaimanapun, kajian lebih mendalam diperlukan untuk meningkatkan ketepatan rekomendasi yang dijanakan kerana gabungan kedua-dua pengukur keserupaan dan petua sekutuan ini adakalanya lebih rendah jika dibandingkan dengan petua sekutuan tradisional.

Menerusi laporan akhir ini juga, teknologi agen yang diaplikasikan bersama-sama dengan teknik petua sekutuan juga turut menyumbang kepada peningkatan prestasi agen rekomendasi (enjin rekomendasi). Teknik perselarian agen yang digunakan membantu proses rekomendasi halaman-halaman Web personalisasi dengan lebih cepat jika diambil

kira dari segi masa pemprosesan dalam pelbagai peringkat. Selain itu, pengintegrasian teknologi agen bersama-sama dengan teknik petua sekutuan juga membolehkan petua yang dihasilkan lebih bernilai bagi membolehkan satu proses rekomendasi terhadap halaman-halaman Web personalisasi. Ini kerana kebolehupayaan agen-agen yang saling berkomunikasi dan bekerjasama dalam menyelesaikan sesuatu tugasan membolehkan proses eksplorasi terhadap pengetahuan dapat dilaksanakan dengan berkesan.

Dengan mengaplikasikan teknologi agen di dalam perlombongan data bagi model rekomendasi halaman-halaman Web ini juga dapat membantu mengoptimumkan prestasi model tersebut. Model rekomendasi halaman Web yang baik terdiri daripada hasil janaan petua-petua dengan menggunakan teknik petua sekutuan. Dengan menggunakan teknologi agen, parameter-parameter bagi janaan petua-petua tersebut dapat ditingkatkan dan seterusnya memastikan hasil rekomendasi halaman-halaman Web adalah yang terbaik dan mengikut kelakuan pengguna.

Banyak lagi isu yang boleh dikaji di dalam personalisasi Web amnya dan perlombongan data penggunaan Web khasnya. Daripada laporan akhir ini sahaja ada beberapa perkara yang memerlukan kajian lebih mendalam. Dari segi penyediaan data, dengan mengambil kira misalnya semantik Web untuk proses mengenalpasti sesi pelayan dan pengguna, mendapatkan masa melihat halaman Web yang sebenar, heuristik untuk mengenalpasti laluan yang diambil pengguna dalam melayari tapak Web dan juga mengambil kira semua jenis fail yang dicapai pengguna seperti fail imej dan bunyi adalah penting untuk menangkap kelakuan pengguna. Untuk penemuan corak pula mungkin boleh dibaiki dengan menggabungkan beberapa teknik perlombongan data yang sesuai seperti pengelompokkan dan petua sekutuan, manakala analisis corak pula mungkin boleh diluaskan dengan tidak hanya menumpu kepada personalisasi Web, tetapi juga ramalan perniagaan dan sebagainya. Selain itu, perlombongan data ke atas pelayan Web yang pelbagai juga telah mula mendapat perhatian para pengkaji. Dengan melombong data ke atas pelayan Web yang pelbagai, rekomendasi ataupun personalisasi Web tidak akan hanya terhad kepada tapak Web individu, malah ia akan turut merangkumi semua tapak Web yang terlibat yang mana ini sudah semestinya memberikan pilihan yang lebih

banyak dan baik kepada pengguna. Selain itu, konsep disebalik model yang telah dibangunkan ini juga mungkin boleh diluaskan lagi penggunaannya. Ia boleh digunakan untuk melombong data perubatan bagi mengesan penyakit ataupun data bioteknologi bagi mengesan identiti gen, sel dan sebagainya. Walaubagaimanapun, beberapa perubahan perlu dilakukan ke atas model seperti menggunakan atribut yang sesuai sebagai penggukur keserupaan.

Secara keseluruhannya, walaupun banyak isu yang timbul dan kesukaran untuk mendapatkan data yang tepat, perlombongan data penggunaan Web pasti akan terus berkembang ia akan menjadi salah sebuah alat yang penting untuk bukan sahaja personalisasi Web, tetapi juga untuk membantu organisasi ataupun syarikat e-dagang meningkatkan perkhidmatan dan produk mereka kepada pelanggan. Ini adalah berpunca daripada kebolehannya untuk memahami kelakuan penggna dan seterusnya mengekstrak pengetahuan penting daripada timbunan data Web yang menjangkau sehingga beratus gigabait saiznya.

RUJUKAN

Agrawal, R.., and Srikant, R., "Fast Algorithm for Mining Association Rules", in Proceedings of 20th International Conference on Very Large Databases, 487-499, 1994.

Berry, M. J. A. and Linoff, G. Data Mining Techniques: For Marketing, Sales and Customer Support, Canada: *Wiley Computer Publishing*; 1997.

Bose, Ranjit and Sugumaran, Vijayan (1999). "Application of Intelligent Agent Technology for Managerial Data Analysis and Mining", Database for Advances in Information Systems, vol. 30, no. 1, pp. 77-94.

Borgelt, C., dan Kruse, R. Induction of Association Rules: Apriori Implementation. 15th Conference on Computational Statictics, Germany. 2002.

Cooley, R. W. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. *Tesis Doktor Falsafah. University of Minnesota*; 2000.

Cooley, R., Mobasher, B., and Srivastava, J., "Data Preparation for Minig World Wide Web Browsing Patterns", Journal of Knowledge and Information Systems, 1, 1-27, 1999.

Cooley, R., Mobasher, B., and Srivastava, J., “Web Mining: Information and Pattern Discovery on the World Wide Web”, in International Conference on Tools with Artificial Intelligence, 558-567, 1997.

Demiriz, A. Enhancing Product Recommender System on Sparse Binary Data. Journal of Intelligent System in Accounting, Finance and Management, 2002, 11(3):127-134.

Difnum, V., Weigand, H., and Xu, L. “Agent Societies : Towards frameworks-Based Design”, in Woodridge, M. J., Wei, G., and Ciancarini, P., editors, Agent Oriented Software Engineering. Springer-Verlag Berlin Heidelberg, pp. 33-49, 2002.

Eirinaki, M., and Vazirgiannis, M., “Web Mining for Web Personalization”, in ACM Transactions on Internet Technology, 3(1), 1-27, 2003.

Fong, J., Hughes, J. G., and Zhu, J., “ Online Web Mining Transactions Association Rules using Frame Metadata Model”, 121-129, 2000.

Fu, X., Budzik, J., and Hammond, K. J., “Mining Navigation History for Recommendation”, in Proceedings of ACM 2000, 106-112, 2000.

Fu, Y., Sandhu, K., and Shi, M., “Clustering of Web Users Based on Access Patterns”, Lecture Notes in Artificial Intelligence, 1836, Springer-Verlag, 21-38, 2000.

Gery, M., and Haddad, H. Evaluation of Web Usage Mining Approaches for User’s Next Request Prediction. Proceedings of Fifth ACM International Workshop on Web Information and Data Management. New Orleans, Louisiana, USA. 2003. 74-81.

Gunduz, S., dan Ozsu, M. T. A. User Interest Model for Web Page Navigation. Proceedings of International Workshop on Data Mining for Actionable Knowledge. Seoul, Korea, 2003.

Han, J., Pei. Dan Yin, Y. Mining Frequent Patterns without Candidate Generation. SIGMOD 2000. 2000. 1-20.

Hong, T. P., Lin, K. Y., and Wang, S. L., "Mining Linguistic Browsing Patterns in the World Wide Web", in Soft Computing, 6,, 329-336, 2002.

Ishikawa, H., Ohta, M., Yokoyama, S., Nakayama, J., and Katayama, K. Web Usage Mining Approaches to Page Recommendation and Restructuring. *International Journal of Intelligent Systems in Accounting, Finance and Management*, 2002, 11(3): 137-148.

Kitsuregawa, M., Toyoda, M., and Pramudiono, I., "Web Community Mining and Web Log Mining: Commodity Cluster based Execution", in Proceedings of Thirteenth Australasian Database Conference 2002, 3-10, 2002.

Kohonen, T., "Self-organized Formation of Topology Correct Features Maps", in Biological Cybernatics, 43, 59-69, 1982.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J. Honkela, J., Paatero, V. and Saarela, A., "Self Oragnization of a Massive Document Collection", IEEE Transactions on Neural Networks, 11(3), 574-585, 2000.

Konstan, J. A., Milner, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., and Reidl, J. Applying Collaborative Filtering to Usenet News. *Communications of the ACM*, 1997, 40(3):77-87.

Lan, B., Bressan, S., and Ooim B. C., “Making Web server Pushier”, Lecture Notes in Artificial Intelligence, 1836, Spring-Verlag, 112-125, 2000.

Lan, B., Bressan, S., Ooi, B. C., and Tan, K. L., “Rule-Assisted Prefetching in Web-Server Caching”, in Proceedings of CIKMM 2000, 504-511, 2000.

Lange, D. B., and Oshima, M. (1998). “Programming and Deploying Java Mobile Agents with Aglets”, Canada, Addison-Wesley.

Lee, C. H., and Yang, H. C., “A Web Text Mining Approach Based on Self-Organizing Map”, WIDM 99, 59-62, 1999.

Lemaitre, C., and Excelente, C. B., “Multi-Agent Organization Approach”, in Proceedings of the Second Iberoamerican Workshop on Distributed Artificial Intelligence and Multi-Agent Systems, 1998.

Lewis, D., dan Gale, W. A. A Sequential Algorithm for Training Text Classifiers. Procedding of the 1th Annual ACM-SIGIR Conference, London. 1994.

Liu, C., Zhong, N., and Ohsuga, S. “A Multi-Agent Based Architecture for Distributed KDD Process”, in Ras, Z. W., and Ohsuga, S., editors, ISMIS. Springer-Verlag Berlin Heidelberg, pp. 591-600, 2000.

Masseglia, F., Poncelet, P., and Teisseire, M., “Using Data Mining Techniques on Web Access Logs to Dynamically Improve Hypertext Structure”, in ACM SigWeb Letters, 8(3), 1-9, 1999.

Mobasher, B., Dai, H., Luo, T., and Nakagawa, M., “Effective Personalization Based on Association Rule Discovery from Web Usage Data”, in Proceedings of WIDM 2001, 9-15, 2001.

Mobasher, B., Dai, H., Luo, T., and Nakagawa, M., Using Sequential and Non-Sequential Patterns for Predictive Web Usage Mining Tasks. Proceeding of the 2002 International Conference on Data Mining. 2002. 669-672.

Mobasher, B., Cooley, R., and Srivatava, J., "Automatic Personalization based on Web Usage Mining", in Communications of the ACM, 43(8), 142-151, 2000.

Mobasher, B., Cooley, R., and Srivatava, J., "Creating Adaptive Web Sites through Usage-Based Clustering of URLs", in Knowledge and Data Engineering Workshop, 1999.

Park. Sooyong et al. "Agent-Oriented Software Modeling with UML Approach", IEICE Trans. Inf. & Syst., vol. E83-D, no. 8, pp. 1631-1641.

Sarwar, B., Karypis, G., Konstan, J., dan Rield, J. Analysis of Recommendation Algorithm for E-Commerce. 2nd ACM E-Commerce Conference. Minneapolis. 2000. 158-167.

Silva, L., M., batista, V., Martins, P., and Soare, G., (1999). "Using Mobile Agents for Parallel Processing", Proceedings of the International Symposium on Distributed Objects and Applications, pp. 34-42.

Smith, K. A., and NG, A. "Web Page Clustering using a Self-Organization Map of User Navigation Patterns", in Journal of Decision Support Systems, 35, 245-256, 2003.

Spiliopoulou, M., "Web Usage Mining for Web Site Evaluation", in Communications of the ACM, 43(8), 127-134, 2000.

Srikant, R. and Agrawal, R. Mining Sequential Patterns: Generalizations and Performance Improvements. *Fifth International Conference on Extending Database Technology*. 1996.

Srivastava, J., Cooley, R., Deshpande, M., and Tan, P. N., “Web Usage Mining Discovery and Applications of Usage Patterns from Web Data”, SIGKDD Explorations, 1(2), 12-23, 2000.

Wang, K, Xu, C., and Liu, B., “Clustering Transactions Using Large Items”, in Proceedings of CIKM 99. 483-490, 1999.

Wang, S., Gao, W., Li, J., and Xie, H., “Web Clustering and Association Rule Discovery for Web Broadcast”, Lecture Notes in computer Science, 1846, Springer-Verlag, 227-231, 2000.

Yang, Q., Zhang, H. H., and Li, T., “Mining Web Logs for Prediction Models in WWW Caching and Prefetching”, in Proceedings of KDD 01, 473-478, 2001.

Zaki, M. J. Generating Non-Redundant Association Rules. Proceedings of the Sixth ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston. 2000. 32-43.

Zheng, Z., Kohazi, R., dan Mason, L. Real World Performance of Association Rule Algorithms. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2001. 404-406.

SUMBANGAN AKADEMIK

1. Abd Manan Ahmad and Mohd Hanafi. An Application of Multi-Agent System for Data Mining using Market Based Analysis Technique. International Conference on Artificial Intelligence in Engineering and Technology (ICAIET-2002), Kota Kinabalu, Sabah, Malaysia. 17th – 18th, June 2002.

2. Abd Manan Ahmad, Mohd Hanafi Hijazi and Sazali Abd Manaf. An Application of Mobile Agent for Association Rule Mining. Proceeding of International Arab Conference on Information Technology, University of Qatar, Dec 16th – 19th 2002.

3. Abd Manan Ahmad, Sazali Abd Manaf and Mohd Hanafi Hijazi. Agent Based Association Rule Mining. Conference on Agent Technology and Application (ATA'02), Faculty Science Computer and Information System, Universiti Teknologi Malaysia, 17th August 2002.

4. Abd Manan Ahmad and M. Hanafi. Association Rule for More Efficient Page Recommendation in Web usage Mining. International Conference on Artificial Intelligence in Engineering and Technology (ICAIET-2004), Kota Kinabalu, Sabah, Malaysia. Aug-2004.