PEMODELAN REKOMENDASI HALAMAN WEB BERASASKAN TEKNIK PERLOMBONGAN DATA

MOHD HANAFI BIN AHMAD HIJAZI

Tesis ini dikemukakan
sebagai memenuhi syarat penganugerahan
ijazah Sarjana Sains (Sains Komputer)

Fakulti Sains Komputer dan Sistem Maklumat
Universiti Teknologi Malaysia

MAC 2005

Buat isteriku Mariati binti Jaafar, ayah Hj. Ahmad Hijazi bin Hj. Ismail dan ibu Hjh. Salmah binti Ag. Damit

PENGHARGAAN

Dengan Nama Allah yang Maha Pengasih lagi Maha Penyayang

Terlebih dahulu, segala pujian dan syukur diucapkan kepada Yang Maha Esa atas bantuan-Nya untuk saya menyiapkan tesis ini. Seterusnya, saya mengucapkan penghargaan kepada penyelia tesis saya iaitu Prof. Madya Abdul Manan bin Ahmad kerana memberikan panduan, kritikan dan idea yang bernas sepanjang saya membangunkan tesis ini.

Terima kasih juga diucapkan kepada Universiti Teknologi Malaysia kerana memberi saya peluang untuk menyelesaikan tesis ini dengan kemudahan-kemudahan yang disediakan. Tidak dilupakan majikan saya Universiti Malaysia Sabah yang sanggup membiayai yuran sepanjang pengajian saya di UTM.

Tidak dilupakan rakan-rakan seperjuangan yang sentiasa bersedia untuk mendengar dan memberikan pendapat yang baik untuk membantu meringankan masalah-masalah yang dihadapi. Terima kasih juga kerana sudi memberikan pandangan-pandangan yang boleh membantu saya meningkatkan kualiti tesis yang dihasilkan. Akhir sekali, terima kasih juga kepada keluarga saya yang tidak jemu-jemu memberikan dorongan untuk saya menyiapkan tesis ini.

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 transaksi 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 tesis ini tertumpu kepada pembangunan model rekomendasi halaman Web berasaskan petua sekutuan (*association rule*) dan pengukuran nilai keserupaan, yang dinamakan sebagai ARsim. 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 oleh para penyelidik model rekomendasi halaman Web untuk pengujian 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 users' 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 researchers. This thesis is focusing on the development of a usage model for predictions based on association rule and similarity measures, named ARsim. Additional parameter was used to measure the similarities between URLs, which is the time user spent 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 similarity measures allow only the most appropriate URLs to be recommended and thus increase the efficiency of the Web page recommendation engine.

KANDUNGAN

SENARAI JADUAL

SENARAI RAJAH

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 Mining*
Tapak Web	-	*Web Site*
URL	-	*Uniform Resource Locator*

BAB 1

PENGENALAN

1.1 Pendahuluan

Pemuatan maklumat yang tidak terbatas ke dalam Web membuatkan bilangan pengguna Internet semakin bertambah setiap hari. Keinginan melayari Internet sama ada untuk mendapatkan maklumat akademik, pekerjaan, perkhidmatan dan sebagainya mengakibatkan pelayan-pelayan Web terpaksa memenuhi permintaan yang tinggi terhadap halaman Web yang dihoskan dalam tempoh masa yang singkat. Akibat daripada permintaan yang menggalakkan, popularity yang tinggi dan kemudahan untuk menggunakannya mendorong banyak pihak memuatkan maklumat ke dalam Web, termasuk kerajaan, syarikat-syarikat swasta dan individu.

Longgokkan maklumat yang merangkumi maklumat-maklumat penting dan yang kurang penting mengakibatkan kesukaran kepada pengguna Internet dalam mencari maklumat yang dikehendaki, mendapatkan pengetahuan-pengetahuan yang baru daripada maklumat-maklumat yang terdapat di dalam Web dan mendapatkan maklumat dalam bentuk yang dikehendaki. Pemberi maklumat seperti organisasi swasta dan kerajaan juga akan menghadapi kesukaran untuk mempelajari dan memahami tingkahlaku pelanggan-pelanggan mereka [1] dengan adanya maklumat

yang terlalu banyak ini. Terdapat beberapa kaedah yang telah dikaji untuk membantu pengguna dalam melayari Internet, antaranya memperkenalkan strategi *cache*[1] untuk pengguna dan pelayan [2], personalisasi Web [3, 4, 5] dan perlombongan Web [6, 7, 8]. Kajian-kajian yang dijalankan ini adalah bertujuan untuk meningkatkan prestasi Web supaya ia boleh dikembangkan lagi pada masa hadapan.

Personalisasi Web yang boleh didefinisikan sebagai sebarang tindakan yang boleh mengubah kandungan atau struktur suatu halaman Web berasaskan kepada ciri-ciri dan kehendak pelanggan [5] adalah merupakan kaedah yang baik untuk membantu pengguna melayari Internet. Dengan mengubah penampilan halaman Web mengikut citarasa pengguna individu, organisasi boleh menawarkan produk dan perkhidmatan pada harga yang sesuai, konteks yang sesuai dan pada masa yang tepat [9]. Terdapat tiga kaedah untuk sistem personalisasi iaitu sistem peraturan keputusan manual, agen penapisan berasaskan kandungan dan sistem penapisan kolaboratif [4]. Kaedah rekomendasi yang biasa digunakan adalah penapisan kolaboratif. Teknik ini melibatkan pemadanan pengkadaran yang diberikan oleh pengguna ke atas objek (contohnya filem, lagu dan sebagainya) dengan pengguna-pengguna lain yang hampir serupa dengan pengguna tersebut untuk mencadangkan objek-objek lain yang belum dilihat atau diketahui oleh pengguna tersebut. Pengkadaran terbahagi kepada dua iaitu pengkadaran tersurat dan pengkadaran tersirat [10]. Pengkadaran tersurat adalah pengkadaran yang biasanya akan diberikan oleh pengguna melalui borang atau tetingkap yang terdapat pada halaman Web yang dilayari. Ia mudah difahami dan jelas. Pengkadaran tersirat pula adalah pengkadaran yang diberikan oleh pengguna tanpa mereka sedari. Misalnya aliran klik halaman-halaman Web yang dilayari, masa yang digunakan untuk melayari suatu halaman Web, aktiviti tetikus dan papan kekunci [10]. Peruntukkan masa ke atas suatu halaman Web adalah antara parameter yang sesuai untuk menentukan pengkadaran pengguna kerana ia mewakili minat pengguna terhadap halaman tersebut [11, 12]. Kaedah pengkelasan yang biasa digunakan untuk sistem personalisasi ini adalah *k-Nearest-Neighbor* (*k*NN). Kaedah ini akan mencari rekod-rekod pengguna lepas yang hampir serupa dengan rekod pengguna semasa untuk mendapatkan *k* pengguna yang paling serupa untuk

[1] *cache* adalah memori kecil yang menyimpan data terkini yang dicapai dan bertujuan untuk mempercepatkan proses capaian terhadap data yang sama pada capaian seterusnya.

mencadangkan halaman-halaman Web kepada pengguna semasa. Walaubagaimanapun, sistem penapisan kolaboratif mempunyai kelemahan yang mana ia memerlukan masa yang banyak kerana proses pemadanan dilakukan secara dalam talian [13, 14].

Perlombongan Web [6, 7, 8] 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 mengkelompokkan pengguna mengikut profil-profil tertentu [15, 16]. Bardasarkan 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 oleh pentadbir Web [17, 18]. Ciri-ciri ini menunjukkan potensi perlombongan Web untuk meningkatkan prestasi sistem personalisasi Web.

Tesis ini mencadangkan satu model baru yang dinamakan ARsim untuk personalisasi Web menggunakan perlombongan Web, iaitu dengan mengintegrasikan perlombongan petua sekutuan (*Association Rule Mining*) dan pengkadaran. Pengkadaran di sini maksudnya 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 pengkadar bagi suatu item. Setelah melakukan perlombongan petua sekutuan dan pengkadaran, 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 [3] dan eVZpro [19].

1.2 Latarbelakang Masalah

Sejak mula dibangunkan lebih kurang 36 tahun yang lalu, Internet yang bermula sebagai salah satu kajian yang dijalankan oleh Agensi Projek Kajian Termaju Pertahanan Amerika Syarikat (DARPA) untuk mengkaji teknik-teknik dan teknologi bagi membolehkan variasi rangkaian paket saling dihubungkan antara satu sama lain, mengalami evolusi yang pantas dan ketara. Pada ketika itu, ia dikenali sebagai ARPAnet dan berfungsi untuk membolehkan perkongsian maklumat antara jabatan-jabatan dalam DARPA. Pada tahun 1970, lima nod pertama telah dibangunkan yang melibatkan UCLA, Stanford, UC Santa Barbara, University of Utah dan BBN (Bolt, Beranek dan Newman). Empat tahun kemudian, iaitu pada tahun 1974, protokol Internet yang sangat penting iaitu TCP/IP (Transmission Control Protocol/ Internet Protocol) telah diperkenalkan oleh Vinton Cerf dan Robert Kahn. Protokol ini masih digunakan hingga ke hari ini sebagai protokol komunikasi Internet yang berjaya menghubungkan seluruh dunia dalam satu medium yang tunggal. Sehingga 1 September 2002, bilangan hos IP telah mencecah lebih 200 juta berbanding 1 juta pada tahun 1992. Pada hari ini juga bilangan pengguna Internet telah melebihi 840 juta orang, lebih dua kali ganda berbanding tahun 2000 yang hanya merekodkan lebih 407 juta orang [20].

Web telah menjadi antara tempat yang paling popular bagi membeli, menukar dan mencari barangan dan perkhidmatan. Berjuta-juta orang melayari Web setiap hari untuk mendapatkan keperluan hidup dan juga untuk menawarkan barangan atau perkhidmatan. Perkembangan maklumat di dalam Internet mengakibatkan alatan yang dapat membantu pengguna dalam mendapatkan maklumat yang dikehendaki adalah perlu dan kritikal. Selain itu, keselesaan pengguna juga adalah antara ciri

penting yang perlu diambil kira. Sistem yang boleh melombong pengetahuan dan mengekstrak corak navigasi pengguna adalah merupakan penyelesaian kepada permasalahan di atas dan diharap dapat membantu pengguna dari aspek capaian maklumat [7]. Beberapa alatan komersial untuk menganalisis penggunaan Web oleh pengguna telah dibangunkan seperti WebTrends [21] dan MarketWave [22]. Alatan-alatan ini dapat membantu organisasi dalam melihat statistik penggunaan Web dari segi purata panjang transaksi setiap pengguna, halaman yang kerap dilihat dan ia juga menyediakan kemudahan untuk menjana laporan secara mingguan, bulanan dan tahunan. Walaubagaimanapun, alatan-alatan ini tidak mempunyai keupayaan untuk merekomen halaman Web yang sesuai kepada pengguna.

Sistem personalisasi Web yang boleh bertindak mengubah sendiri kandungan atau struktur Web mengikut ciri atau pilihan pengguna [5] telah mula mendapat perhatian daripada organisasi-organisasi e-perdagangan dalam usaha untuk memberikan layanan atau perkhidmatan yang terbaik kepada pelanggan-pelanggan mereka. Tindakan yang boleh dilakukan adalah merangkumi pemaparan halaman Web yang lebih memudahkan sehinggalah kepada penyesuaian halaman Web mengikut kehendak pengguna dan membekalkan maklumat yang telah disesuaikan [3]. Personalisasi boleh dilakukan dengan menyerlahkan pautan-pautan, memasukkan secara dinamik pautan baru yang dijangkakan boleh menarik minat pengguna, penghasilan halaman indeks baru dan juga dengan menggunakan sistem rekomendasi. Sistem rekomendasi boleh dibahagikan kepada tiga kategori iaitu sistem peraturan keputusan manual, agen penapisan berasaskan kandungan dan sistem penapisan kolaboratif (CF) [4], yang mana CF adalah sistem yang paling dominan dalam membekalkan sistem e-perdagangan dengan kepintaran untuk mendapatkan profil-profil pengguna dan seterusnya merekomen halaman yang berkaitan dengan minat pengguna. Walaubagaimanapun, terdapat beberapa kekangan dalam sistem CF terutamanya dari aspek kecekapan dan *scalability* [14]. Beberapa pembaikan telah dilakukan oleh para penyelidik untuk mengatasi masalah ini seperti menggunakan pengukuran keserupaan berasaskan item [23, 24] dan pengurangan dimensi [25]. Satu teknik baru yang dapat mengeksploitasi teknik-teknik sedia ada adalah diperlukan untuk menghasilkan sebuah sistem rekomendasi yang lebih baik.

Melombong data penggunaan Web [6] telah mendapat perhatian yang tinggi dikalangan penyelidik dalam mengenalpasti *kelakuan pengguna dalam melayari Internet*, yang mana selepas ini dan pada bab-bab seterusnya akan dirujuk sebagai *navigasi pengguna*. Satu-satunya cara untuk mencapai tujuan ini adalah dengan melombong log pelayan Web yang menyimpan rekod-rekod bagi setiap aktiviti transaksi pengguna. Terdapat banyak teknik melombong data yang digunakan termasuk pengelompokan [15, 16, 26, 27, 28, 29], perlombongan petua sekutuan [2, 3, 8, 30, 31, 32, 33], dan corak berturutan [29, 33, 34].

Permodelan profil pengguna adalah penting bagi organisasi yang menggunakan Web sebagai medium perniagaannya. Sebagai contoh, mereka boleh tahu siapakah pelanggan mereka, apakah yang dibeli, bagaimana mereka membeli dan bila mereka membeli. Semua maklumat ini boleh diperolehi dengan merujuk kepada aliran klik yang dilakukan oleh pengguna[2]. Rekod-rekod ini akan disimpan di dalam log pelayan Web. Data-data di dalam log tersebut mempunyai nilai yang sangat besar sekiranya organisasi tahu mengeksploitasikannya. Selain daripada yang telah dinyatakan di atas, data-data ini juga boleh digunakan untuk membantu pengguna dalam melayari Internet dan mengenalpasti sama ada struktur tapak Web adalah sesuai atau tidak.

Integrasi antara perlombongan petua sekutuan dan pengelompokan untuk mendapatkan profil pengguna telah dicadangkan oleh Mobasher et al. [26]. 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 [16] 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

[2] http://www.dmreview.com

halaman umum. Wang et al. [27] 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 [19] 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 [2, 32, 33] 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 [4] 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 oleh pengguna. Walaubagaimanapun, keputusan yang dihasilkan masih belum mencapai ke tahap yang terbaik [3] 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.

Beberapa penyelidik telah menggunakan masa yang diperuntukkan oleh pengguna untuk melayari sesuatu halaman Web sebagai pemberat untuk mengelompokkan pengguna [10, 12, 35]. Kajian-kajian ini telah menunjukkan keputusan yang baik dalam penghasilan kelompok profil pengguna. Ini membuktikan bahawa masa pelayaran halaman Web yang direkodkan ke dalam log pelayan Web adalah berpotensi untuk menghasilkan profil pengguna yang baik dan seterusnya membantu meningkatkan ketepatan rekomendasi. Walaubagaimanapun, setakat ini masa tersebut hanya digunakan untuk menyediakan data sebelum pengelompokkan

dilakukan. Penggunaan masa ini boleh diluaskan lagi ke peringkat-peringkat lain dalam perlombongan data seperti penemuan pengetahuan dan analisis pengetahuan.

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 sebagai ARsim untuk personalisasi Web dengan menggunakan perlombongan petua sekutuan dan pengukur keserupaan.
3. 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 Kennedy 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.

1.7 Struktur Tesis

Tesis ini disusun seperti berikut: Bab 2 akan mengkaji latarbelakang bagi perlombongan penggunaan Web. Metodologi yang digunakan untuk membangunkan model dibincangkan di dalam Bab 3. Penerangan tentang fasa pemodelan data dan prapemprosesan juga dilakukan di dalam bab ini. Bab 4 pula menerangkan tentang bagaimana model corak navigasi pengguna dapat diekstrak daripada data log Web dan seterusnya teknik-teknik yang boleh digunakan untuk menganalisis corak penggunaan Web bagi rekomendasi halaman Web. Pengujian dan keputusan akan ditunjukkan di dalam Bab 5. Bab 6 menyimpulkan keseluruhan kajian.

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 pengukur 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-syarikat e-dagang meningkatkan perkhidmatan dan produk mereka kepada pelanggan. Ini adalah berpunca daripada kebolehannya untuk memahami kelakuan pengguna dan seterusnya mengekstrak pengetahuan penting daripada timbunan data Web yang menjangkau sehingga beratus gigabait saiznya.

RUJUKAN

1. Kosala, R. and Blockeel, H. Web Mining Research: A Survey. *SIGKDD Explorations*, 2000, 2(1): 1-15.

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

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

4. Mobasher, B., Cooley, R., and Srivastava, J. Automatic Personalization based on Web Usage Mining. *Communications of the ACM,* 2000, 43(8): 142-151.

5. Eirinaki, M. and Vazirgiannis, M. Web Mining for Web Personalization. *ACM Transactions on Internet Technology*, 2003, 3(1): 1-27.

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

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

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

9. Tam, K. Y., and Ho, S. Y. Web Personalization: Is it Effective?. *IT Pro*. 2003. 53-57.

10. Claypool, M., Le, P., Waseda, M., and Brown, D. Implicit Interest Indicators. *Proceedings of ACM Intelligents User Interfaces Conference*, Santa Fe, New Mexico:ACM, 2001. 33-40.

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

12. Shahabi, C., Zarkesh, A. M., Adibi, J., and Shah, V. Knowledge Discovery from Users Web-Page Navigation. *7th International Conference on Research Issues in Data Engineering*. 1997. 20-29.

13. Munindar P. Singh. *The Practical Handbook of Internet Computing*. CRC Press; 2004.

14. Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. Analysis of Recommendation Algorithms for E-Commerce. *2nd ACM E-Commerce Conference*. Minneapolis. 2000. 158-167.

15. Smith, K. A. and Ng, A. Web Page Clustering using a Self-Organization Map of User Navigation Patterns. *Journal of Decision Support Systems*, 2003, 35: 245-256.

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

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

18. 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.

19. Demiriz, A. Enhancing Product Recommender Systems on Sparse Binary Data. *Journal of Data Mining and Knowledge Discovery*, 2004, 9(2): 147-170.

20. Internet Society, http://www.isoc.org.

21. http://www.webtrends.com

22. http://www.marketwave.com

23. Karypis, G. Evaluation of Item-Based Top-N Recommendation Algorithms. *Technical Report CS-TR-00-46*, Computer Science Department, University of Minnesota. 2000.

24. Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. Item-Based Collaborative Filtering Recommendation Algorithms. *10th International World Wide Web Conference*. Hong Kong, 2001. 285-295.

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

26. Mobasher, B., Cooley, R. and Srivastava, J. Creating Adaptive Web Sites through Usage-Based Clustering of URLs. *Proceedings of Knowledge and Data Engineering Workshop*. Washington DC, 1999. 19-25.

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

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

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

30. 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, 2000. 227-231.

31. Fong, J., Hughes, J. G. and Zhu, J. (2000). Online Web Mining Transactions Association Rules using Frame Metadata Model. *Proceedings of the First International Conference on Web Information Systems Engineering*. Hong Kong, 2000. 2121-2130.

32. Lan, B., Bressan, S. and Ooi, B. C. Making Web Servers Pushier. *Lecture Notes in Artificial Intelligence*. 1836. Springer-Verlag, 2000. 112-125.

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

34. Masseglia, F., Poncelet, P. and Teisseire, M. Using Data Mining Techniques on Web Access Logs to Dynamically Improve Hypertext Structure. *ACM SigWeb Letters*, 1999, 8(3): 1-19.

35. Banerjee, A., and Ghosh, J. Clickstream Clustering using Weighted Longest Common Subsequences. *1st SIAM Conference on Data Mining*. Chicago, 2001.

36. Perkowitz, M., and Etzioni, O. Adaptive Web Sites: Automatically Synthesizing Web Pages. *Fifteenth National Conference on Artificial Intelligence*. Madison, Wisconsin, United States. 1998. 727-732.

37. Shahabi, C., and Banaei-Kashani, F. Efficient and Anonymous Web-Usage Mining for Web Personalization. *INFORMS Journal on Computing*, 2003, 15(2): 123-147.

38. Mulvenna, M. D., Anand, S. S., and Buchner, A. G. Personalization on the Net using Web Mining. *Communications of the ACM*, 2000, 43(8): 123-125.

39. Schafer, B. J., Konstan, J. and Riedl, J. Recommender Systems in E-Commerce. *1st ACM Conference on E-Commerce*. Denver, Colorado. 1999. 158-166.

40. Ardissono, L., Goy, A., Petrone, G., and Segnan, M. Personalization in Business-to-Customer Interaction. *Communications of the ACM*, 2002, 45(5): 52-53.

41. Breese, J. S., Heckerman, D., and Kadie, C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*. 1998. 43-52.

42. Yu, P. S. Data Mining and Personalization Technologies. *Proceedings of 6th International Conference on Database Systems for Advanced Applications*. Hsinchu, Taiwan. 1999. 6-16.

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

44. 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.

45. Chase, W., dan Bown, F. *General Statistics Fourth Edition*. John Wiley and Sons Inc.; 2000.

46. Belew, R. K. *Finding Out About: A Cognitive Perspective on Search Engine Technology and WWW*. Cambridge University Press; 2001.

47. Mobasher, B., Dai, H., Luo, T., Nakagawa, M., Sun, Y., dan Wiltshire, J. Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. *Data Mining and Knowledge Discovery*. 2002, 6: 61-82.

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

49. Pabarskaite, Z. Decision Trees for Web Log Mining. *Intelligent Data Analysis*. 2003, 7: 141-154.

50. http://www.w3.org

51. Cooley, R., Mobasher, B. dan Srivastava, J. Data Preparation for Mining World Wide Web Browsing Patterns. *Journal of Knowledge and Information Systems,* 1999, 1: 1-27.

52. Baglioni, M., Ferraara, U., Romei, A., Ruggieri, S. dan Turini, F. Preprocessing and Mining Web Log Data for Web Personalization. Cappelli, A. and Turini, F. (eds.): AI*IA 2003, *Lecture Notes on Artificial Intelligent*, Springer-Verlag Berlin Heidelberg, 2003. 237-249.

53. Berendt, B., Mobasher, B., Spiliopoulou, M., and Wiltshire, J. Measuring the Accuracy of Sessionizers for Usage Analysis. *Proceedings of the Web Mining Workshop at the First SIAM International Conference on Data Mining*. Chicago, 2001. 7-14.

54. Mobasher, B., Dai, H., Luo, T., dan Nakagawa, M. Improving the Effectiveness of Collaborative Filtering on Anonymous Web Usage Data. *Proceedings of the IJCAI 2001 Workshop on Intelligent Techniques for Web Personalization (ITWP01)*. Seattle. 2001.

55. Agrawal, R. dan Srikant, R. Fast Algorithm for Mining Association Rules. *Proceedings of 20th International Conference on Very Large Data Bases*. 1994. 487-499.

56. Kohonen, T. Self-Organized Formation of Topology Correct Feature Maps. *Biological Cybernatics,*1982, 43: 59-69.

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

58. 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.

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

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

61. Zaki, M. J. Generating Non-Redundant Association Rules. *Proceedings of the Sixth ACM-SIGKDD International Conference on Knowledge Discovery and Data Mining*. Boston. 2000. 34-43.

62. 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. 401-406.

63. Fu, X., Budzik, J. dan Hammond, K. J. Mining Navigation History for Recommendation. *Proceedings of ACM 2000*. 2000. 106-112.

64. Lewis, D., dan Gale, W. A. A Sequential Algorithm for Training Text Classifiers. *Proceedings of the 17th Annual ACM-SIGIR Conference.* London. 1994.

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

66. Kamus Komputer, Dewan Bahasa dan Pustaka, Kuala Lumpur, 1995.

67. Kamus Komputer Bahasa Inggeris – Bahasa Melayu, Penerbit Fajar Bakti Sdn. Bhd., Selangor, 2002.

68. Glosari Teknologi Maklumat, Dewan Bahasa dan Pustaka, Kuala Lumpur, 1996.

69. Istilah Matematik Bahas Inggeris – Bahasa Melayu, Dewan Bahasa dan Pustaka, Selangor, 1992.

70. E-Kamus KaryaNet, http://www.karyanet.com.my.

71. Mohd. Hanafi Ahmad Hijazi dan Abdul Manan Ahmad. Using Time Spent on A Web Page for Web Personalization. *Proceedings of the Joint International Conference on Informatics and Research Women in ICT*, Kuala Lumpur, Malaysia. 2004. 79-86.

72. Mohd. Hanafi Ahmad Hijazi dan Abdul Manan Ahmad. Association Rule for More Efficient Page Recommendation Using Web Usage Mining. *Proceeding of the 2nd International Conference on Artificial Intelligence in Engineering and Technology*, Kota Kinabalu, Sabah, Malaysia. 2004. 582-587.

73. Abdul Manan Ahmad dan Mohd. Hanafi Ahmad Hijazi. Web Page Recommendation Model for Web Personalization. *Proceedings of 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science*. 3214. Springer-Verlag Heidelberg. 2004. 587-593.