Teknik Pengukuhan Perangkak Tumpuan melalui Modul Pengesan Bahasa bagi Capaian Web Bahasa Melayu

Masnizah Mohd
masnizah.mohd@ukm.edu.my
Fakulti Teknologi dan Sains Maklumat
Universiti Kebangsaan Malaysia

Wan Fariza Paizi@Fauzi
fariza.fauzi@ukm.edu.my
Fakulti Teknologi dan Sains Maklumat
Universiti Kebangsaan Malaysia

Amri Jasin
amrijasin.ukm@gmail.com
Fakulti Teknologi dan Sains Maklumat
Universiti Kebangsaan Malaysia

ABSTRAK

Perangkak ialah antara komponen utama dalam seni bina sistem capaian maklumat atau enjin gelintar. Ia berfungsi mengumpul laman web yang relevan bertujuan untuk diuruskan melalui pengindeksan maklumat pautan dan kandungan. Perangkak tumpuan adalah aplikasi perangkak yang direka khas untuk memilih dan mengumpul laman web yang mempunyai kaitan tentang domain atau pertanyaan khusus di Internet. Perangkak yang baik mampu memberikan keputusan maklumat yang tepat, pantas, luas dan relevan kepada pengguna semasa proses pencarian maklumat menggunakan enjin gelintar. Kesukaran malah ketidakupayaan mengesan pautan serta kandungan berbahasa Melayu merupakan antara isu utama. Kesannya ialah ada di antara kandungan laman web Bahasa Melayu tidak dapat diindeks seterusnya diproses untuk capaian maklumat. Malah kekurangan perangkak yang khusus bagi carian laman web Bahasa Melayu sebagai bahasa carian utama menjadi pendorong utama penyelidikan ini. Maka objektif utama kajian ini adalah untuk mengenalpasti strategi merangkak yang baik untuk perangkak tertumpu memilih pautan yang relevan dan berkualiti berdasarkan pertanyaan Bahasa Melayu. Perangkak tumpuan yang digunakan dalam penyelidikan ini telah melalui pengubahsuaian hasil daripada gabungan beberapa teknik pengukuhan merangkak. Hasil pengujian yang berulang menunjukkan kehadiran modul pengukuhan perangkak tumpuan telah memberi keputusan yang baik iaitu berupaya mengesan laman web bahasa Melayu yang tepat. Penyelidikan ini juga menjadi titik tolak kepada perkembangan pencarian maklumat berdasarkan pertanyaan Bahasa Melayu di Internet, di samping dapat memartabatkan Bahasa Melayu di dunia siber.

Kata Kunci: perangkak; capaian maklumat; Bahasa Melayu; enjin gelintar; web

eISSN: 2550-2131
ISSN: 1675-8021

Focused Crawler Enhancement Technique with Language Detection Module for Malay Web Retrieval

ABSTRACT

Crawler is one of the major components in the architecture of information retrieval systems or search engines. The function is to gather relevant websites aimed to be managed through indexing of links and content. A focused crawler application is designed to select and collect web pages that are relevant to domains or specific topics in the Internet. A good crawler can provide accurate, extensive and relevant information to the user during the process of information seeking using search engines. The inability to detect links and content of Malay language is one of the main issues. Therefore, some of the content of the Malay website cannot be indexed and processed for information retrieval. The lack of research in focused crawler especially for Malay website has motivated this research. The main objective of this study is to identify good crawling strategies for focused crawler in detecting relevant and quality links for Malay website. The focused crawler employed in this research has undergone some modifications resulting from a combination of some crawling strengthening techniques. Findings indicate that the presence of a focused crawler enhancement module provides good results because it can detect Malay language webs accurately. This research is also a turning point for the development of information retrieval for Malay websites as well as enhancing the prominence of Malay language in cyberspace.

Keywords: crawler; information retrieval; Malay language; search engine; web

PENGENALAN

Perangkak tumpuan adalah aplikasi yang direka khas bertujuan memilih dan mengumpul laman web yang mempunyai kaitan tentang domain atau pertanyaan tertentu di Internet. Ia ditugaskan untuk menyelusur bentuk kerangka web dan mendapatkan sebahagian dari kandungan sesebuah laman web untuk dinilai persamaan dengan pertanyaan yang diberikan. Proses ini bermula dari satu set laman web yang dirujuk sebagai set benih (Pant et al., 2004). Proses pemilihan dan penilaian kesamaan bagi setiap prospek yang dilawati memberikan kesan penjimatan dalam penggunaan rangkaian jalur lebar dan kapasiti storan hasil carian. Ini dibuktikan apabila perangkak hanya mengikuti laluan yang mempunyai nilai kesamaan yang tinggi dan berhenti merangkak di laluan yang tidak lagi berkaitan dengan pertanyaan yang ditugaskan.

Model perangkak tumpuan boleh disesuaikan bagi pengumpulan kumpulan laman web yang mengandungi bahasa yang sama. Berdasarkan kepada beberapa kajian lepas, laman web yang menggunakan satu bahasa tertentu mempunyai banyak pautan ke laman web lain dalam bahasa yang sama (Masomeh, 2010). Kewujudan kelompok tumpuan koleksi laman web yang menggunakan bahasa yang sama juga dipercayai mempunyai sumber maklumat berkaitan dengan sesebuah pertanyaan bahasa tersebut. Perangkak tumpuan bercirikan bahasa tertentu adalah sejenis perangkak yang menggabungkan elemen konsep dan juga linguistik. Kajian lepas menyelami konteks bahasa bersama perangkak tumpuan bagi domain bahasa Thailand dan Jepun (Somboonviwat, Kitsuregawa & Tamura, 2015; Somboonviwat, Tamura & Kitsuregawa, 2016; Tamura, Somboonviwat & Masaru, 2016). Pendekatan yang diambil adalah berasaskan penempatan sesuatu bahasa yang wujud di laman web. Mereka mengumpul laman web yang tergolong dalam bahasa tertentu yang mempunyai identiti linguistik yang sama, menggunapakai ciri domain serta struktur yang terdapat di laman web

untuk mengenalpasti kandungan bahasa utama dan memilih domain yang besar untuk menghasilkan korpus khusus pelbagai bahasa daripada sumber web sedia ada.

Walau bagaimanapun, kajian menyeluruh dalam konteks perangkak tumpuan berdasarkan laman web Bahasa Melayu masih kurang diterokai. Kajian ini memberi penekanan kepada pengukuhan perangkak tumpuan khusus untuk bahasa tertentu berserta kaedah pengujian bagi mengoptimumkan proses merangkak dan capaian laman web di Internet.

Bermula di penghujung tahun 90-an, penyelidikan perlombongan web hanya tertumpu kepada pengoptimuman teknik merangkak dan bahan kajian adalah berupa laman web Bahasa Inggeris (Masomeh et al., 2010; Somboonviwat, Kitsuregawa & Tamura, 2015; Somboonviwat, Tamura & Kitsuregawa, 2016; Tamura, Somboonviwat & Masaru, 2016). Pada dasarnya, laman web bukan sahaja dipersembahkan di dalam satu bahasa, namun terdapat juga dalam bahasa lain. Kebarangkalian bagi pautan di dalam laman web Bahasa Melayu merujuk ke laman web Bahasa Inggeris adalah tinggi. Manakala kebarangkalian bagi pautan dalam laman web Bahasa Inggeris merujuk ke laman web Bahasa Melayu adalah rendah. Namun jika sebuah perangkak tumpuan dilengkapi dengan kebolehan untuk mengenalpasti kandungan bahasa laman web, sudah tentunya hasil pengumpulan laman web berdasarkan pertanyaan dari Bahasa Melayu adalah baik dan peratusan laman web yang relevan juga tinggi berdasarkan pertanyaan Bahasa Melayu. Justeru, kajian ini bertujuan mengaplikasi kaedah pengesanan kandungan bahasa laman web dalam aplikasi perangkak tumpuan bagi mencapai laman web dalam kelompok bahasa yang sama, dan menilai prestasi ketepatan kaedah pengesanan kandungan bahasa laman web dalam aplikasi perangkak tumpuan.

Makalah ini dimulakan dengan pengenalan mengenai konsep perangkak tumpuan diikuti dengan kajian lepas yang memperincikan proses, kaedah, struktur serta komponen perangkak tumpuan. Metodologi kajian dibincangkan seterusnya, khasnya implementasi modul pengesan bahasa Melayu dalam mengukuhkan perangkak tumpuan bagi capaian web bahasa Melayu. Hasil eksperimen dianalisis dari dua aspek iaitu kejituan dan kepantasan masa yang diambil oleh perangkak tumpuan dengan membandingkan teknik pengukuhan merangkak yang dicadangkan (dengan modul pengesan bahasa Melayu) dengan kaedah asas yang digunakan (tanpa modul pengesan bahasa Melayu). Makalah ini dikukuhkan dengan perbincangan dan diakhiri dengan kesimpulan keseluruhan kajian ini.

KAJIAN LEPAS

KAEDAH DAN STRUKTUR PERANGKAK

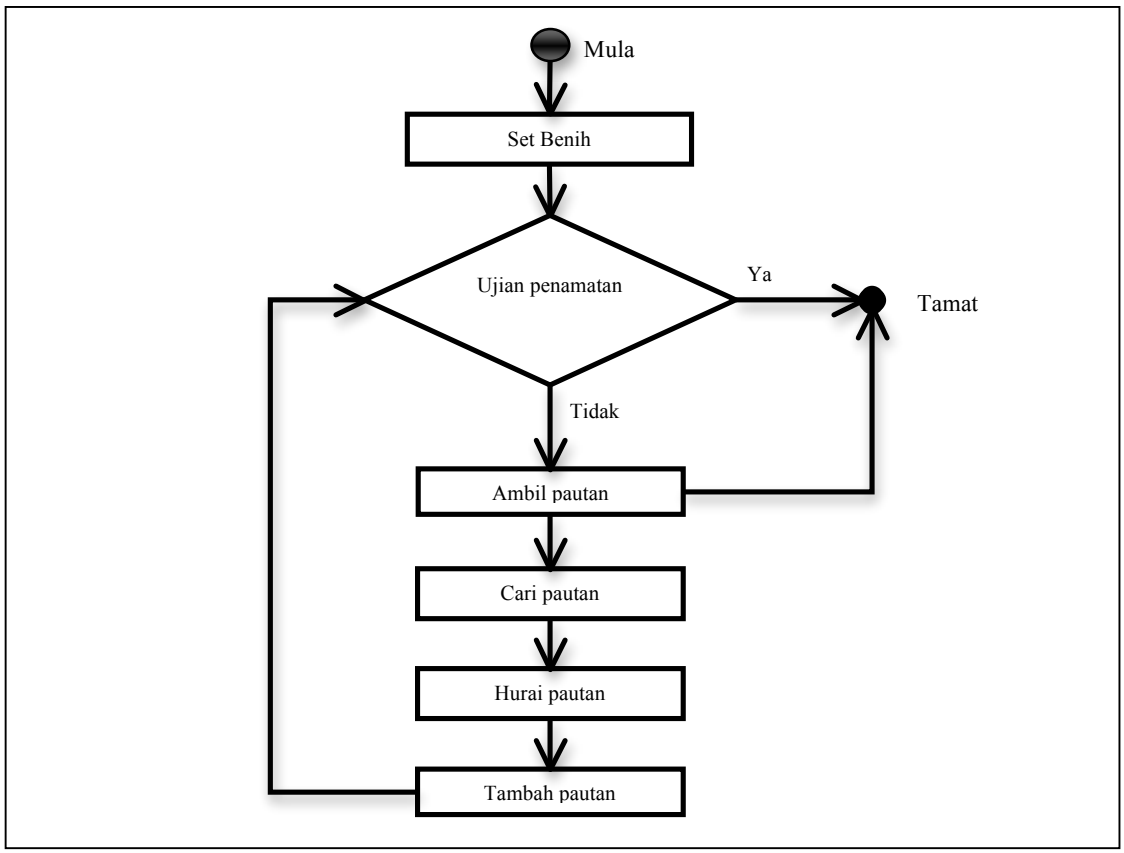
Penyelidikan telah dijalankan dalam meneroka dan menghasilkan strategi baru untuk perangkak tumpuan dalam bidang capaian maklumat khususnya. Seiring dengan itu, beberapa kaedah telah dicadangkan dalam hasil kajian (Masomeh et al., 2010; Somboonviwat, Kitsuregawa & Tamura, 2015; Somboonviwat, Tamura & Kitsuregawa, 2016; Tamura, Somboonviwat & Masaru, 2016) khusus bagi menghasilkan perangkak tumpuan yang lebih cekap dan pantas. Beberapa kajian penting mengklasifikasikan kaedah merangkak kepada beberapa kaedah seperti pembelajaran mesin atau berasaskan suap balik, berasaskan warisan, dan strategi merangkak berpandukan maklumat tambahan yang pelbagai. Malahan terdapat sebahagian daripada penyelidikan yang diteliti telah mempelbagaikan teknik merangkak melalui gabungan pengelasan dan pengugusan (Fatma Howedi & Masnizah Mohd, 2014).

Kaedah berasaskan suap balik mengambilkira pengetahuan dari hasil merangkak berulang yang dijadikan sebagai landasan untuk melakukan proses merangkak seterusnya (Angkawattanawit & Rungsawang, 2002). Kaedah ini juga dikenali sebagai kaedah

pembelajaran. Manakala kaedah berasaskan warisan berfokus terhadap proses menganalisis dan menganggar kesesuaian laman web di kejiranan laman web induk sebelum mencapai laman web seterusnya (Hersovici et al., 1998).

Perangkak juga sesuai digunakan bersama sumber maklumat tambahan seperti kata kunci, leksikon dan set pertanyaan (Nazlia Omar, Masnizah Mohd & Yusman Jamat, 2013; Hamood Alshalabi, Sabrina Tiun, Nazlia Omar & Mohammed Albared, 2013; Hamed Zakeri Rad, Sabrina Tiun & Saidah Saad, 2018), atau pelengkap pengetahuan seperti tesaurus dan kepintaran ontologi (Badia, Muezzinoglu & Nas-raoui, 2006; Arifah Che Alhadi, Shahrul Azman Mohd Noah & Lailatul Qadri Zakaria, 2012; Shahrul Azman Mohd Noah, Nazlena & Mohd Sabri, 2018).

Merangkak boleh dilihat sebagai satu masaalah carian graf. Web boleh diumpamakan sebagai sebuah graf yang besar dengan laman web sebagai nodus dan pautannya sebagai tepian. Asas proses merangkak bermula daripada beberapa set benih dan kemudian mengikuti setiap tepian untuk mencapai nodus yang lain. Namun bagi sebuah perangkak tumpuan, objektif utama adalah untuk menyusuri setiap nodus di bahagian-bahagian graf yang berkaitan dengan pertanyaan yang ditugaskan. Rajah 1 menggambarkan aliran perangkak asas berjujukan. Perangkak ini mengekalkan senarai laman web yang belum dimuatturun menggunakan pemboleh ubah jajaran yang dinamakan sebagai barisan pautan piawai. Senarai ini kebiasaannya dimulakan dengan set benih yang disediakan oleh pengguna atau aplikasi lain. Setiap kitaran capaian kandungan laman web melibatkan proses pemilihan pautan. Setelah selesai memuatturun serpihan kandungan laman web tersebut, pautan yang terkandung di dalam laman web tersebut diekstrak dan pautan baharu ditambah ke dalam barisan pautan piawai sebagai pautan yang belum diteroka. Proses ini boleh ditamatkan apabila sejumlah laman web telah diperolehi oleh perangkak atau senarai laman web yang belum diteroka dalam barisan pautan piawai telah pun kosong.



RAJAH 1. Struktur dan proses am perangkak

Perangkak hanya menggunakan satu proses tunggal untuk menyelusur seluruh kandungan web secara berurutan. Kaedah ini mengambil masa yang tinggi walaupun kapasiti pemprosesan komputer atau antara muka rangkaian masih ada lagi ruang belum digunapakai. Teknologi pelbagai jalur sudah tidak asing lagi dalam dunia pengkomputeran. Rutin ulangan sebuah proses dalam mana-mana aplikasi boleh dibahagikan kepada beberapa jalur, di mana setiap jalur boleh melaksanakan rutin tersebut secara selari. Jika teknologi ini dibentuk di dalam sebuah perangkak, teknologi ini sudah tentu memberikan ia satu kelebihan untuk memantapkan lagi proses merangkak yang berulang-ulang.

PENGUMPULAN DAN PENGHURAIAN LAMAN WEB

Dalam usaha untuk mengumpul kandungan laman web, perangkak memerlukan sebuah pelanggan aplikasi protokol yang berupaya meminta maklumat sesebuah laman web dari pelayan web dan membaca jawapan balas permintaan tersebut. Protokol Pengecualian Robot menyediakan mekanisma untuk pentadbir pelayan web berkomunikasi dengan pelanggan mengenai polisi capaian fail mereka. Melalui medium ini, perangkak dengan lebih khusus dan pantas mengenalpasti fail yang tidak boleh dicapai olehnya. Perjanjian ini dilakukan melalui fail bernama *robots.txt*. Fail ini disimpan di bawah direktori akar pelayan web seperti *http://www.bnm.gov.my/robots.txt*. Setiap perangkak yang berhasrat untuk memuatturun kandungan laman web daripada sebuah pelayan web mestilah memuatturun fail robots.txt dan memastikan pautan yang tersenarai di dalamnya tidak diambil.

Apabila kandungan laman web telah dimuatturun, maklumat di dalamnya diekstrak dan dijadikan penentu hala tuju perangkak. Proses penghuraian laman web lebih mudah digambarkan sebagai proses pengekstrakan pautan dan pengemaskinian kandungan laman web. Penghuraian ini juga melibatkan langkah-langkah seperti mengubah kandungan pautan kepada bentuk berstruktur seperti di bawah:

1. Menukar protokol dan nama domain kepada huruf kecil kepada http://www.bnm.gov.my.
2. Mengeluarkan rujukan yang menjadi sebahagian daripada pautan. Oleh itu, http://www.bnm.gov.my/faq.html#what diperbaiki menjadi http://www.bnm.gov.my/faq.html.
3. Melakukan pengekodan pautan untuk beberapa karakter yang biasa digunakan seperti "~". Ini menyekat perangkak daripada menganggap http://www.bnm.gov.my/~indeks/ sebagai pautan yang berbeza daripada http://www.bnm.gov.my/%7eindeks/.
4. Menambah ketiadaan karakter "/" di penghujung pautan. http://www.bnm.gov.my dan http://www.bnm.gov.my/ mesti dipetakan kepada bentuk struktur yang sama.
5. Menggunakan kebolehan heuristik untuk mengenalpasti kedudukan fail utama laman web. Nama-nama fail seperti index.html atau index.htm kebiasaannya tidak dihiraukan atas andaian bahawa mereka adalah fail utama.
6. Mengeluarkan karakter ".." dari direktori induknya. Oleh itu, http://www.bnm.gov.my/../seeds.dat diperbaiki kepada http://www.bnm.gov.my/seed.dat.
7. Membiarkan nombor port di pautan melainkan nombor port 80. Sebagai alternatif, membiarkan nombor port di pautan dan menambah port 80 apabila tiada nombor port ditetapkan.

Adalah penting bagi sebuah perangkak untuk memastikan penghuraian berstruktur dipatuhi setiap masa supaya perangkak boleh mengumpul maklumat struktur dan kandungan laman web semasa proses merangkak.

PERANGKAK TUMPUAN BAHASA TERTENTU

Perangkak tumpuan bercirikan bahasa tertentu adalah sejenis *perangkak* tertumpu yang menggabungkan elemen konsep dan juga linguistik (Somboonviwat, Kitsuregawa & Tamura, 2015). Pendekatan yang diambil adalah berasaskan penempatan sesuatu bahasa yang wujud dalam web. Mereka mengumpul laman web yang tergolong dalam bahasa tertentu yang mempunyai identiti linguistik yang sama seperti penggunaan perkataan, kata akar dan bahasa kebangsaan.

Hasil kajian Somboonviwat, Kitsuregawa & Tamura (2015), merupakan salah satu usaha mewujudkan arkib web bagi negara dengan identiti linguistik. Penggunaan kaedah linguistik, maklumat meta-tag dan algoritma N-gram diimplementasi bagi mencari laman yang relevan dan bersesuaian (Cavnar & Trenkle, 1994). Kajian terhadap kesesuaian ciri linguistik dalam aktiviti merangkak tidak lagi asing dalam bidang ini.

Somboonviwat, Kitsuregawa & Tamura (2015) mendapati bahasa laman web ditentukan dari pengekodan skema aksara. Berlainan pula bagi Botha & Barnards (2005) di mana mereka mengutarakan idea mereka melalui pemilihan domain yang besar untuk menghasilkan korpus khusus pelbagai bahasa daripada sumber web sedia ada dengan menggunapakai ciri yang terdapat dalam laman web untuk mengenalpasti kandungan bahasa utama dalam sesebuah laman web.

METODOLOGI

Pada asasnya, hasil dapatan perangkak tumpuan amat dipengaruhi oleh senarai pautan permulaan atau set benih. Tiga kumpulan set benih diwujudkan hasil daripada proses pemilihan secara rawak dan saringan pautan. Kategori set benih pertama adalah pautan dari sektor perbankan. Manakala kategori set benih yang kedua adalah pautan dari sektor kesihatan dan yang terakhir adalah dari sektor pendidikan.

Shervin Daneshpajouh, Mojtaba Mohammadi Nasiri, & Mohammad Ghodsi (2003), yang menjalankan kaedah penetapan set benih bagi perangkak, telah mengesyorkan penggunaan algoritma pencarian pertanyaan berasaskan pautan semasa menjana set benih. Pemilihan dan penyusunan set benih yang baik memastikan hasil capaian laman web yang baik walaupun proses merangkak dimulakan dengan senarai pautan yang terhad. Usul tersebut digunakan dalam penghasilan set benih eksperimen melalui pemilihan domain yang mempunyai kaitan dan prospek kepada pautan laman web Bahasa Melayu. Lima langkah utama dipatuhi semasa proses penghasilan set benih kumpulan eksperimen.

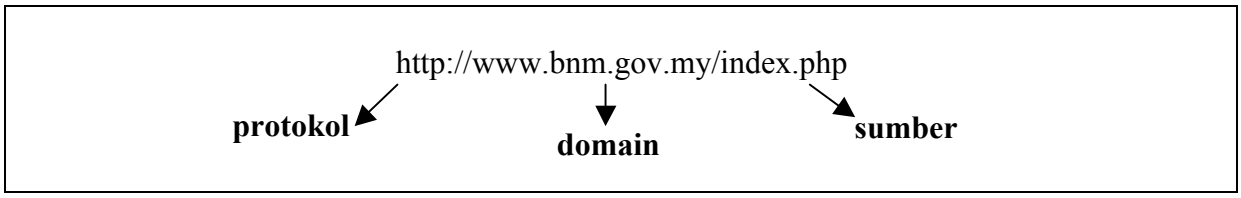
1. Memuatturun fail dari laman web DMOZ.
2. Mengekstrak pautan yang terkandung dalam format XML dari fail yang dimuatturun.
3. Memilih pautan yang tersenarai di bawah direktori: "*Top/Business/Financial_Services/Banking_Services/Banks_and_Institutions/Regional/Asia/Malaysia*".
4. Memilih pautan yang tersenarai di bawah direktori: "*Top/Regional/Asia/Malaysia/Health/Hospitals*".
5. Memilih pautan yang tersenarai di bawah direktori "*Top/World/Bahasa_Melayu/Rujukan/Pendidikan/Institusi_Pengajian_Tinggi*".

Perangkak tumpuan berasaskan bahasa tertentu perlu memenuhi dua keperluan utama. Aplikasi perangkak perlu mempunyai kerangka yang baik dalam mengesan pautan laman web yang berkaitan dengan pertanyaan yang diberikan. Keperluan kedua pula adalah mengaplikasikan kaedah yang sesuai untuk mengenalpasti kumpulan laman web yang

mempunyai kandungan bahasa yang sama. Setiap pautan yang diekstrak dari laman web induk diklasifikasi sebagai laman web Bahasa Melayu apabila satu nilai skor yang dikira melebihi nilai piawai yang ditetapkan dalam menentukan bahasa kandungan laman web.

SARINGAN DOMAIN TINGKAT-ATAS

Pelayar web dan perangkak adalah dua contoh pelanggan web yang berlainan. Namun kedua-duanya mengumpul laman web melalui kaedah yang sama. Setiap laman web mempunyai pautan yang unik. Pautan merupakan simbol yang merangkumi tiga bahagian iaitu protokol, domain (nama hos), dan nama sumber. Rajah 2 memberi visual yang lebih jelas mengenai ciri sebuah pautan.



RAJAH 2. Tiga bahagian pautan

Semasa langkah pengekstrakan pautan dalam setiap proses rungkaian laman web, tumpuan diberi kepada laman web yang mempunyai domain berakhiran dengan simbol ".my" atau simbol sisipan "/my/" atau "/ms/". Jika salah satu prinsip di atas dipatuhi, perangkak tumpuan membenarkan laman web ini diproses untuk kitaran seterusnya. Sebaliknya, aplikasi perangkak tidak mendaftar laman-laman web tersebut ke dalam barisan pautan piawai.

Proses saringan ini bertujuan supaya aplikasi perangkak lebih cenderung kepada pengumpulan laman web Bahasa Melayu melalui maklumat domain dan teks pautan. Sebagai contoh, "http://www.bnm.gov.my/". Beberapa pautan dikenalpasti hadir di dalam kandungan laman web induk ini. Penghakiman tambahan dilakukan ke atas domain pautan yang diekstrak melalui laman web induk tersebut. Jika domain pautan berakhir dengan simbol ".my", aplikasi perangkak menyimpan pautan ini ke dalam barisan pautan piawai. Jika domain pautan tidak mengandungi simbol ".my", aplikasi perangkak seterusnya memeriksa keseluruhan teks pautan sama ada mengandungi simbol carian "/my/" atau "/ms/". Jika kedua-dua logik aturcara tersebut tidak dipatuhi, aplikasi perangkak menggugurkan pautan tersebut. Sebahagian daripada keputusan saringan bagi laman web "http://www.bnm.gov.my/" dipadan dalam Jadual 1.

JADUAL 1. Contoh Saringan Domain dan Teks Pautan

Pautan
http://power.akpk.org.my/
http://www.ssm.com.my/ms/
http://www.tourismmalaysia.gov.my/en/my/

Maklumat kod bahasa dalam laman web diguna untuk mengenalpasti bahasa utama atau sebahagian kandungan laman web. Kebiasaannya, maklumat ini diguna oleh aplikasi enjin gelintar dan pelayar untuk melaraskan paparan skrin pengguna. Melalui kajian terdahulu, penyelidik seperti Tamura dan Kitsuregawa (2016) juga menggunakan maklumat tag laman web untuk mengumpul laman web yang mempunyai kod bahasa utama yang sama. Setiap pengaturcara web perlu mengisytiharkan bahasa utama laman web dengan penggunaan ciri bahasa di dalam tag laman web. Bahasa Melayu diwakili oleh simbol "ms" berdasarkan

kod bahasa yang terkandung di dalam ISO639-1. Ciri tag laman web ini dijadikan sebagai maklumat tumpuan bagi aplikasi perangkak untuk mengesan laman web Bahasa Melayu.

JADUAL 2. Contoh Penggunaan Tag Laman Web

Skim	Penggunaan Tag
HTML	<html lang=”ms”> ... </html>
XHTML	<html xmlns=http://www.w3.org/1999/xhtml lang=”ms” xml:lang=”ms”> ... </html>

MODUL PENGESANAN BAHASA MELAYU

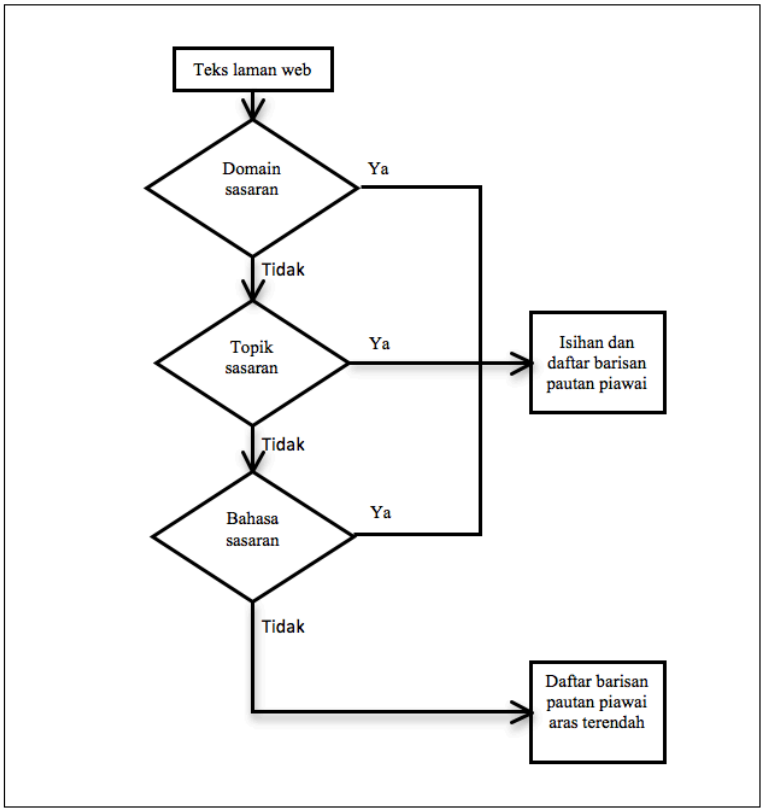
Penentuan bahasa dalam modul pengesan Bahasa Melayu ini dibuat melalui teknik pengekodan aksara tag laman web. Selain daripada penyaringan domain dan penggunaan tag laman web bagi mengesan laman web Bahasa Melayu, kajian ini mencadangkan penggunaan indeks kamus Bahasa Melayu bagi memperkukuhkan prestasi teknik pengecaman seperti dipaparkan di Rajah 3. Pendekatan tag diberi keutamaan dalam mengesan bahasa laman web memandangkan teknik tersebut mudah diaplikasikan ke dalam aplikasi perangkak. Pengecaman bahasa melalui kandungan laman web adalah teknik kedua yang digunapakai dalam kajian ini. Teknik ini berkesan dalam situasi apabila ciri tag tidak wujud dalam dokumen web. Botha dan Barnards (2005) membuktikan bahawa kandungan laman web boleh digunakan dalam membuat perbandingan dan penilaian terhadap bahasa linguistik laman web. Gabungan elemen tag dan kandungan digunakan dalam kajian ini bagi mengenalpasti laman web Bahasa Melayu.

Laman web yang dimuatturun dianggap relevan jika menggunakan perkataan dalam bahasa sasaran. Kesemua kaedah dibangunkan untuk bekerja secara automatik bagi menentukan bahasa laman web dalam aplikasi perangkak tumpuan bahasa tertentu. Menurut Shanjian dan Katsuhiko (2001), modul pengesanan bahasa terbuka boleh digunakan untuk menganalisis skim pengekodan aksara di dalam sesebuah laman web. Malahan, alat pengesan bahasa melalui kaedah ini adalah lebih dipercayai daripada merujuk kepada tag laman web. Namun tidak semua bahasa disokong oleh modul ini.

Oleh itu, penentuan bahasa laman web dibuat melalui peratusan jumlah perkataan Bahasa Melayu yang terkandung dalam laman web berbanding keseluruhan perkataan. Pemarkahan oleh modul pengesan bahasa ini digunakan oleh perangkak untuk dijadikan sebagai mekanisme pembias sekunder bagi mengarah aktiviti merangkak ke kelompok Bahasa Melayu selepas pemarkahan pertanyaan yang relevan. Algoritma pengesan linguistik yang dibina adalah berdasarkan persamaan 1

$$\text{malay}(q, p) = \frac{|| q ||}{|| p ||} \quad (1)$$

di mana, q adalah jumlah kekerapan perkataan Bahasa Melayu yang hadir dalam laman web. Manakala p pula adalah jumlah perkataan keseluruhan laman web. Perkataan Bahasa Melayu dikenalpasti hadir melalui proses perbandingan dengan sebuah indeks perkataan Bahasa Melayu yang diperolehi melalui laman web “*http://ms.wiktionary.org/wiki/Kategori:Perkataan_Melayu*”. Sebanyak 638 patah perkataan diperoleh melalui pemprosesan teks telah didaftar dalam indeks sebagai rujukan kata akar Bahasa Melayu.



RAJAH 3. Proses Pengesanan Bahasa Melayu oleh Perangkak

ANALISIS

Dua matriks penilaian dipertimbangkan termasuklah kejituan dan kepantasan merangkak bagi mengukur prestasi aplikasi perangkak tumpuan Bahasa Melayu. Tiga pertanyaan dipilih sebagai pertanyaan pengujian aplikasi perangkak tumpuan. Pertanyaan tersebut adalah "Akaun Semasa", "Jantung" dan "Kolej Kediaman". Set pertanyaan ini dipilih berdasarkan kepada tiga kategori set benih iaitu dari sektor perbankan, kesihatan dan pendidikan. Bagi setiap pengujian pertanyaan, tiga parameter jumlah sasaran laman web yang perlu dikumpul perlu dijalankan. Jumlah sasaran pengumpulan laman web adalah 5000 laman, 10,000 laman dan 50,000 laman. Bagi pengujian yang mensasarkan pengumpulan 5000 laman web, setiap pertanyaan merekodkan perbezaan masa yang sedikit panjang bagi ketiga-tiga pertanyaan apabila menggunakan teknik pengukuhan bahasa.

KEJITUAN

Kejituan merujuk kepada ketepatan laman web bahasa Melayu yang dimuatturun oleh aplikasi perangkak tumpuan. Jadual 3, 4 dan 5 menunjukan peratusan kejituan yang lebih tinggi direkodkan apabila aktiviti merangkak diaplikasi bersama teknik pengukuhan bahasa berbanding dengan ketiadaannya dalam pengujian pengumpulan sasaran laman web bahasa Melayu.

Analisis kejituan dari segi kandungan mendapati pertanyaan "akaun semasa" telah menyenaraikan laman web yang mempunyai kata kunci perbankan seperti akaun simpanan, akaun deposit, akaun tetap termasuk terma akaun semasa. Malah frasa yang menjurus kepada bank tertentu turut dikesan seperti "akaun semasa hong leong" dan "akaun semasa bank rakyat" turut disimpan oleh perangkak dalam senarai laman web modul pengesan bahasa. Set benih yang disenaraikan dalam barisan pautan piawai menyenaraikan kata kunci berkaitan

sektor kewangan. Ini bertujuan membentuk konsep kewangan seperti senarai institusi kewangan seperti Bank Rakyat, Hong Leong, Affinbank dan CIMB; aspek proses seperti *transaksi*, *deposit* dan *simpanan*; dan aspek manusia seperti *pegawai bank*, *pengurus*, *pencarum*, *penyimpan* dan *pengguna*. Malah mekanisme ini didapati menyumbang kepada teknik pengukuhan perangkak tumpuan melalui modul pengesan bahasa bagi capaian web Bahasa Melayu. Perangkak tumpuan tanpa modul pengesan bahasa Melayu didapati telah menyenaraikan laman web yang kurang tepat seperti produk bank (insuran, kontrak, cek) dan perkhidmatan bank yang umum (pinjaman, perniagaan, bisnes, perdagangan).

Mekanisme yang sama telah diterapkan semasa membuat pertanyaan seperti "jantung" telah menyenaraikan laman web yang mempunyai kata kunci kesihatan seperti *rawatan jantung*, *serangan jantung* dan *penyakit jantung*. Namun peratusan kejituan pertanyaan "jantung" didapati agak rendah berbanding pertanyaan yang lain seperti "akaun semasa" dan "kolej kediaman" pada kelompok 5000, 10 000 dan 50 000 laman web kerana istilah saintifik seperti *koronari* dan *arteri*, didapati mendominasi berbanding istilah Bahasa Melayu.

Perangkak tumpuan dengan modul pengesan bahasa Melayu membentuk konsep pendidikan seperti senarai institusi pengajian tinggi seperti UKM, USM dan lain-lain; aspek proses seperti *penginapan*, *sewa*, *permohonan* dan *keluar*; dan aspek manusia seperti *mahasiswa*, *mahasiswi,* pelajar, *pengetua*, dan *felo*. Konsep ini diterap dalam set benih yang disenaraikan dalam barisan pautan piawai.

JADUAL 3. Kejituan teknik pengukuhan modul perangkak pada kelompok 5,000 laman web

Pertanyaan	**Kejituan**	
	Dengan modul pengesan bahasa Melayu	**Tanpa modul pengesan bahasa Melayu**
"Akaun Semasa"	3.72%	1.56%
"Jantung"	0.4%	0.2%
"Kolej Kediaman"	3.68%	0.4%

JADUAL 4. Kejituan teknik pengukuhan modul perangkak pada kelompok 10,000 laman web

Pertanyaan	**Kejituan**	
	Dengan modul pengesan bahasa Melayu	**Tanpa modul pengesan bahasa Melayu**
"Akaun Semasa"	4.43%	4.33%
"Jantung"	6.92%	0.82%
"Kolej Kediaman"	**81.2%**	33.9%

JADUAL 5. Kejituan teknik pengukuhan modul perangkak pada kelompok 50,000 laman web

Pertanyaan	**Kejituan**	
	Dengan modul pengesan bahasa Melayu	**Tanpa modul pengesan bahasa Melayu**
"Akaun Semasa"	2.21%	2.16%
"Jantung"	1.88%	1.27%
"Kolej Kediaman"	1.44%	1.04%

Perbezaan kejituan yang signifikan ialah pada pertanyaan "Kolej Kediaman" dengan kelompok 10,000 laman web. Keputusan ujian jelas menunjukkan bahawa terdapat peningkatan dari segi peratusan kejituan laman web Bahasa Melayu yang dimuatturun apabila teknik pengukuhan merangkak dibangunkan dalam aplikasi perangkak tumpuan Bahasa Melayu.

KEPANTASAN

Kepantasan merujuk kepada masa yang diambil untuk memuatturun dengan lengkap laman web bahasa Melayu oleh aplikasi perangkak tumpuan. Hasil pengujian kelompok sasaran pengumpulan laman web menunjukkan aktiviti merangkak yang dilengkapi modul pengesan bahasa berjaya menamatkan pencarian dalam jangkamasa yang lebih singkat berbanding dengan pencarian tanpa modul pengesan bahasa.

Pengumpulan laman web melalui struktur pautan web bahasa Melayu seperti ".my" dan set benih yang mengimplementasi konsep, leksikon, kata kunci dan frasa dalam barisan pautan piawai menjadikan proses pengesanan laman web Bahasa Melayu lebih cekap. Pertanyaan "akaun semasa" menyebabkan perangkak tertumpu terus mendaftarkan laman web yang mempunyai perkataan berkaitan institusi kewangan seperti Bank Rakyat, Hong Leong, Affinbank dan CIMB; aspek proses seperti *transaksi*, *deposit* dan *simpanan*; dan aspek manusia seperti *pegawai bank*, *pengurus*, *pencarum*, *penyimpan* dan *pengguna*, dalam barisan pautan piawai. Padanan pertanyaan dengan kandungan laman web lebih pantas apabila perangkak tumpuan dipandu dengan sumber pengetahuan berasaskan leksikon oleh set benih yang disenaraikan dalam barisan pautan piawai. Mekanisme yang sama telah diterapkan semasa membuat pertanyaan seperti "jantung" atau "kolej kediaman" yang telah menyenaraikan laman web yang mempunyai kata kunci kesihatan dan kata kunci pendidikan pada kelompok 5000, 10 000 dan 50 000 laman web. Pendekatan berasaskan pembelajaran ini akhirnya menjadikan prestasi merangkak meningkat bukan hanya dari aspek kejituan malah kepantasan dalam mengesan laman web Bahasa Melayu. Jelas teknik pengukuhan merangkak yang dicadangkan dengan modul pengesan bahasa Melayu membantu kecekapan perangkak tumpuan.

JADUAL 6. Kepantasan teknik pengukuhan modul perangkak pada kelompok 5,000 laman web

Pertanyaan	**Kepantasan (saat)**	
	Dengan modul pengesan bahasa Melayu	**Tanpa modul pengesan bahasa Melayu**
"Akaun Semasa"	528	566
"Jantung"	557	822
"Kolej Kediaman"	**506**	3791

JADUAL 7. Kepantasan teknik pengukuhan modul perangkak pada kelompok 10,000 laman web

Pertanyaan	**Kepantasan (saat)**	
	Dengan modul pengesan bahasa Melayu	**Tanpa modul pengesan bahasa Melayu**
"Akaun Semasa"	902	1013
"Jantung"	1407	2366
"Kolej Kediaman"	1061	1335

JADUAL 8. Kepantasan teknik pengukuhan modul perangkak pada kelompok 50,000 laman web

Pertanyaan	**Kepantasan (saat)**	
	Dengan modul pengesan bahasa Melayu	**Tanpa modul pengesan bahasa Melayu**
"Akaun Semasa"	2255	2532
"Jantung"	1265	1675
"Kolej Kediaman"	1022	1622

Perbezaan kepantasan yang signifikan ialah pada pertanyaan "Kolej Kediaman" dengan kelompok 5000 laman web dengan peratus perbezaan kepantasan sebanyak 86.65%. Begitu juga, tempoh yang lebih singkat diperlukan oleh perangkak tumpuan untuk

mengumpulkan sejumlah laman web bersama modul pengukuhan tambahan berbanding penggunaan struktur perangkak asas. Ini menjelaskan bahawa teknik pengukuhan merangkak yang dicadangkan (dengan modul pengesan bahasa Melayu) adalah lebih efisien berbanding kaedah asas yang digunakan (tanpa modul pengesan bahasa Melayu).

PERBINCANGAN

Kajian ini bertujuan memperbaiki kejituan perangkak tumpuan bagi mengumpul laman web Bahasa Melayu berdasarkan pertanyaan. Beberapa kaedah pengesanan kandungan bahasa laman web dianalisis bagi meningkatkan kecekapan sebuah perangkak tumpuan. Hasilnya sebuah modul pengesanan bahasa dibangunkan di dalam aplikasi perangkak tumpuan untuk bekerja secara automatik bagi mengenalpasti laman web Bahasa Melayu.

Hasil analisis menunjukkan peningkatan kejituan dan kepantasan pada aplikasi perangkak tumpuan dengan adanya modul pengesan bahasa. Melalui pemerhatian, dengan wujudnya pautan laman web Bahasa Melayu yang terkandung dalam laman web induk Bahasa Melayu dan analisis kandungan melalui implementasi konsep, leksikon serta kata kunci menjadi penyumbang kepada keputusan ini. Aplikasi perangkak berupaya menyelusur bentuk graf web yang terarah kepada pertanyaan dan bahasa laman web di kumpulan yang sama. Ini turut didiperhatikan dalam kajian oleh (Kitsuregawa & Tamura, 2015; Somboonviwat, Tamura & Kitsuregawa, 2016; Tamura, Somboonviwat & Masaru, 2016) dan mendapati perangkak tumpuan dengan modul pengesan bahasa tertentu memberi keputusan serta hasil yang tepat dan masa yang diambil lebih singkat (Shan-Bin Chan & Hayato Yamana, 2010). Ini kerana konsep linguistik seperti konsep bahasa, kata kunci yang menjurus sera relevan kepada pertanyaan, menyokong perangkak tumpuan dalam mengesan laman web bahasa Melayu.

Namun beberapa isu dihadapi, semasa atau sebelum proses merangkak dijalankan. Antaranya adalah berkenaan isu persendirian. Sebuah perangkak yang baik perlu memelihara hak persendirian pelayan web. Pastinya, pelayan web tidak mahu diganggu beberapa kali untuk dilawati bagi mendapatkan maklumat yang sama setiap kali pengujian dijalankan. Aplikasi perangkak hendaklah diuji di ruang kerja sendiri dengan teliti sebelum dilancarkan ke dalam dunia web sebenar. Disebabkan oleh isu persendirian ini, aplikasi perangkak tumpuan yang dibangunkan terlebih dahulu merisik fail robot.txt dari setiap pelayan web sebelum memuat turun kandungan yang dibenarkan oleh pengurus laman web tersebut.

Web pada dasarnya adalah sangat dinamik, amat mustahil untuk memastikan bahawa semua hasil ujian yang dibandingkan di bawah syarat yang sama dapat mengulang sumber penilaian yang sama. Di samping itu, terdapat beberapa laman web yang gagal diteroka disebabkan oleh pengaturcara yang tidak mematuhi bahasa laman web dengan tepat. Sebagai contoh, simbol <table> tidak diakhiri atau ditutup dengan simbol </table>. Walaupun kesilapan kecil seperti ini masih boleh diterima oleh aplikasi pelayar namun bagi aplikasi perangkak adalah penting baginya kesempurnaan dalam menulis laman web supaya dapat mengenalpasti pautan dan kandungan teks sesebuah laman web. Kajian oleh Shan-Bin Chan dan Hayato Yamana (2010) dan Shervin Daneshpajouh, Mojtaba Mohammadi Nasiri dan Mohammad Ghodsi (2003) turut menambahbaik kejituan perangkak melalui kaedah pemprosesan pautan, kandungan dan tag pada laman web yang dikesan.

Proses merangkak sememangnya memerlukan masa, rangkaian jalur lebar dan ruang simpan data yang besar. Spesifikasi komputer yang digunakan untuk menjalankan pengujian terhadap perangkak perlu dilengkapi dengan unit pemprosesan pusat serta storan ingatan utama yang besar. Di samping itu, kapasiti cakera keras juga perlu diberi perhatian disebabkan oleh saiz laman web yang dimuatturun adalah besar. Keperluan pemprosesan dan storan hendaklah dipastikan sebelum sebarang pengujian dibuat bagi mengelakkan kejadian

seperti kehabisan masa sewaktu operasi muatturun. Bagi setiap kejadian kehabisan masa yang berlaku menurunkan kadar kejituan sesebuah kitaran pengujian. Kadangkala aplikasi Java diganggu oleh ralat memori apabila sumber komputer atau memori tidak mencukupi bagi melaksanakan sebarang operasi oleh jalur tertentu.

KESIMPULAN

Penyelidikan ini di harap dapat membantu perkembangan ilmu dalam teknik pengumpulan maklumat melalui rangkaian web khusus bagi dokumen Bahasa Melayu. Dengan adanya indeks khas bagi laman web Bahasa Melayu, dapat memberikan pengalaman carian baru samada dari aplikasi enjin carian atau aplikasi carian maklumat yang lain. Melalui indeks dokumen Bahasa Melayu yang khusus, dapat memberi keputusan carian yang pantas. Pelbagai komponen atau modul dalam aplikasi perangkak tumpuan yang dibangunkan ini boleh diperincikan dan diperbaiki bagi menghasilkan sebuah aplikasi yang lebih mantap dan lebih dipercayai.

Kajian ini menggunakan sepenuhnya senarai pautan yang dipilih melalui direktori *DMOZ* sebagai set benih bagi permulaan proses merangkak. Senarai pautan yang merujuk ke domain ".my" mempunyai peratusan yang amat rendah berbanding senarai keseluruhan pautan di dalam data direktori *DMOZ*. Kajian lanjutan boleh disarankan agar melaksanakan pengujian kaedah terhadap set benih yang diperoleh dari direktori web seperti Yahoo! Directory dan Google Directory. Pengukuhan keberkesanan modul pengesan Bahasa Melayu juga boleh dilakukan ke atas kategori yang lain seperti sukan mahupun berita.

Pendekatan yang dipilih adalah dengan menggabungkan ciri konteks laman web dan kandungannya bagi memperkukuhkan proses merangkak lebih pantas dan sentiasa mengulang prestasi yang baik. Keputusan pengujian menunjukkan bagaimana nilai kejituan dari hasil merangkak meningkat melalui pelaksanaan kaedah pengesanan kandungan bahasa laman web yang diperkenalkan. Keputusan ujian yang positif dalam menghimpunkan laman web Bahasa Melayu dengan menggunakan aplikasi perangkak tumpuan ini dapat diaplikasikan serta diubahsuai bagi mencari maklumat dalam sebarang ruang atau medium penyimpanan. Jelas pendekatan teknik pengukuhan perangkak tumpuan melalui modul pengesan bahasa bagi capaian web Bahasa Melayu mampu memberikan keputusan maklumat yang tepat, luas dan relevan kepada pengguna semasa proses pencarian maklumat.

RUJUKAN

Acharjya, D.P. & Mitra, Anirban. (2017). *Bio-Inspired Computing for Information Retrieval Applications*. United States: GI Global.

Almpanidis, G, Kotropoulos, C & Pitas, I. (2007). Combining Text and Link Analysis for Focused Crawling: An Application for Vertical Search Engines. *Information System*. *Vol. 32*(6), 886-908.

Angkawattanawit, N & Rungsawang, A. (2002). Learnable Crawling: An Efficient Approach to Topic-Specific Web Resource Discovery. *Proceedings of the 2nd International Symposium on Communications and Information Technology (ISCIT)*. Bangkok, Thailand. March.

Arifah Che Alhadi, Shahrul Azman Noah & Lailatul Qadri Zakaria. (2012). Pendekatan Ontologi dalam Capaian dan Perwakilan Semantik Dokumen Web. *Jurnal Teknologi*. *Vol. 46*(1), 103-120.

Badia, A, Muezzinoglu, T. & Nas-raoui, O. (2006). Focused Crawling: Experiences in a Real World Project. Proceedings of the 15 International Conference on World Wide Web. Edinburgh, Scotland, May. 1043-1044.

Bhatia, M. P. S. & Gupta, D. (2008). Discussion on Web Crawlers of Search Engine. *Proceedings of 2nd National Conference on Challenges & Opportunities in Information Technology* (COIT-2008). RIMT-IET, India. March.

Botha, G. & Barnards, E. (2005). Two Approaches to Gathering Text Corpora from the World Wide Web. Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa. Stellenbosch, South Africa, December. 194-199.

Cavnar, W. B. & Trenkle, M. (1994). N-Gram Based Text Categorization. *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*. Las Vegas, US, September. 161-175.

Chakrabarti, S, Punera, K. & Subramanyam, M. (2002). Accelerated Focused Crawling Through Online Relevance Feedback. *Proceedings of the 11th international conference on World Wide Web*. Honolulu, Hawaii, US. May.

Cho, J, Garcia-Molina, H. & Page, L. (1998). Efficient Crawling Through URL Ordering. *Proceedings of the 7th International World-Wide Web Conference*. Brisbane, Australia, April. 161-172.

De Bra, P, Houben, G.-J, Kornatzky, Y. & R. Post. (1994). Information Retrieval in Distributed Hypertexts. *Proceedings of RIAO"94, Intelligent Multimedia, Information Retrieval Systems and Management*. New York, US. October.

Ehrig, M. & Maedche, A. (2003). Ontology-Focused Crawling of Web Documents. *Proceedings of the 2003 ACM Symposium on Applied Computing (SAC)*. Melbourne, FL, USA. March.

Fatimah Ahmad. (1995). A Malay Language Document Retrieval System: An Experimental Approach and Analysis. Ph.D thesis, Universiti Kebangsaan Malaysia, Bangi, Malaysia.

Fatma Howedi & Masnizah Mohd. (2014). Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data.*Computer Engineering and Intelligent Systems. Vol. 5*(4), 48-56.

Hamed Zakeri Rad, Sabrina Tiun & Saidah Saad. (2018). Lexical Scoring System of Lexical Chain for Quranic Document Retrieval. *GEMA Online® Journal of Language Studies. Vol. 18*(2), 59-79.

Hamood Alshalabi, Sabrina Tiun, Nazlia Omar & Mohammed Albared. (2013). Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization. *Procedia Technology. Vol. 11*, 748-754.

Hersovici, M , Jacovi, M., Maarek, Y.S., Pelleg, D., Shtalheim, M. & Ur, S. (1998). The Shark-Search Algorithm: An Application Tailored Web Site Mapping. *Computer Networks and ISDN Systems*. *Vol. 30*(1-7), 317-326.

Katharine Jarmul & Richard Lawson. (2017). *Python Web Scraping*. Birmingham: Packt Publishing Ltd.

Kumar, V., Grama, A., Gupta, A. & Karypis, G. (2008). *Introduction to Parallel Computing: Design and Analysis of Algorithms*. Michigan: Benjamin/Cummings Pub. Co.

Masomeh, A., Yari, A., & Mohammad, J. V. (2010). Language Specific Crawling based on Web Pages Features. *In Proceedings of the 2010 International Conference on Multimedia Computing and Information Technology (MCIT)*. Sharjah, United Arab Emirates. March.

Menczer, F., Pant, G., Srinivasan, P. & Ruiz, M. (2001). Evaluating Topic -Driven Web Crawlers. *In Proceedings of the 24th Annual International ACM/SIGIR Conference*. Melbourne, Australia. August.

Nazlia Omar, Masnizah Mohd & Yusman Jamat. (2013). Kebarangkalian Secara Automatik Dalam Alkhwarizmi Bayes untuk Aplikasi Agen Bualan. *Asia-Pacific Journal of Information Technology and Multimedia. Vol. 2*(1), 27-37

Neel, S. & Jeonghee, Y. (2001). Mining the Web for Relations. *Computer Networks Vol. 33*(1-6), 699-711.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1998). The Pagerank Citation Ranking: Bringing Order to The Web. *Proceedings of the 7th International World Wide Web Conference.* Brisbane, Australia, April.

Pant, G. & Menczer, F. (2003) Topical crawling for business intelligence. *Proc. 7th European Conference on Research and Advanced Technology for Digital Libraries.* Bath, UK. September.

Pant, G, Srinivasan, P. & Menczer, F. (2004). Crawling the Web. *In Web Dynamics: Adapting to Change in Content, Size, Topology and Use*. Trondheim, Norway. August.

Pant, G. (2003). Deriving Link-Context from HTML Tag Tree. *8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*. San Diego, California, US. June.

RaviKumar, S., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A. & Upfal, E. (2000). Stochastic Models for the Web Graph. *Proceedings of 41st Annual Symposium on Foundations of Computer Science*. Redondo Beach, CA, US. November.

Savoy, J. (1993). Stemming of French Words on Grammatical Categories. *Journal of American Society for Information Science*. *Vol. 44*(1), 1-9.

Shahrul Azman Mohd Noah, Nazlena Mohamad Ali & Mohd Sabri Hasan. (2018). Penentuan Fitur bagi Pengekstrakan Tajuk Berita Akhbar Bahasa Melayu. *GEMA Online® Journal of Language Studies. Vol. 18*(2), 154-167.

Shan-Bin Chan & Hayato Yamana. (2010). The Method of Improving the Specific Language Focused Crawler. *Proceedings of the 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing*. Beijing, China. August.

Shanjian Li & Katsuhiko Momoi. (2001). A Composite Approach To Language/Encoding Detection. *Proceedings of the 19th International Unicode Conference*. San Jose, California, US. September.

Shervin Daneshpajouh, Mojtaba Mohammadi Nasiri & Mohammad Ghodsi. (2003). A Fast Community Based Algorithm For Generating Web Crawler Seeds Set. *Proceedings of the Fourth International Conference on Web Information Systems and Technologies*. Madeira, Portugal. May.

Somboonviwat, K., Kitsuregawa, M. & Tamura, T. (2015). Simulation Study of Language Specific Web Crawling. *ICDE 21st International Conference on Data Engineering (ICDE'05)*. Tokyo, Japan. April.

Somboonviwat, K., Tamura, T. & Kitsuregawa, M. (2016). Finding Thai Web Pages in Foreign Web Spaces. *International Conference on Data Engineering (ICDE) Workshop*. Helsinki, Finland. May.

Tamura, T., Somboonviwat, K. & Masaru, K. (2016). A Method for Language Specific Web Crawling and Its Evaluation. *IEICE Transactions on Iinformation and Systems Vol. 38*(2), 10-20.

PENULIS

Masnizah Mohd mendapat PhD daripada University of Strathclyde, United Kingdom. Merupakan Profesor Madya di Fakulti Teknologi dan Sains Maklumat UKM. Bidang kepakaran beliau ialah capaian maklumat dan pemprosesan bahasa tabii.

Wan Fariza Paizi@Fauzi merupakan Pensyarah Kanan di Fakulti Teknologi dan Sains Maklumat UKM. Mendapat PhD daripada Universiti Monash pada tahun 2012. Bidang kepakaran beliau ialah pemprosesan bahasa tabii dan teknologi semantik.

Amri Jasin memperoleh Sarjana Teknologi Maklumat (Sains Maklumat) daripada Fakulti Teknologi dan Sains Maklumat, UKM pada tahun 2014. Bidang penyelidikan beliau adalah perangkak web dan enjin gelintar.

eISSN: 2550-2131
ISSN: 1675-8021