Jurnal Antarabangsa (Teknologi Maklumat) 3(2002): 1-10

Dapatan Semula Maklumat Bahasa Melayu Menggunakan Pengindeksan Semantik Terpendam

MUHAMAD TAUFIK ABDULLAH, FATIMAH AHMAD,
RAMLAN MAHMOD & TENGKU MOHAMMED TENGKU SEMBOK

ABSTRAK

Pengindeksan semantik terpendam ialah satu varian daripada kaedah ruang vektor iaitu satu anggaran pangkat-rendah kepada perwakilan ruang vektor untuk pangkalan data digunakan. Idea utama dalam model pengindeksan semantik terpendam adalah untuk memetakan setiap vektor dokumen dan pertanyaan ke dalam satu ruang berdimensi lebih rendah yang berkaitan dengan konsep-konsep. Ini dilaksanakan dengan memetakan vektor-vektor istilah indeks ke dalam ruang berdimensi lebih rendah tersebut. Dakwaannya, capaian di dalam ruang yang dikecilkan mungkin lebih baik daripada capaian di dalam ruang istilah-istilah indeks. Dalam makalah ini, sebagai tambahan kepada kaedah ruang vektor, kaedah pengindeksan semantik terpendam digunakan untuk membina sistem dapatan semula maklumat bahasa Melayu.

Kata Kunci: dapatan semula maklumat bahasa Melayu, pengindeksan semantik terpendam, dapatan semula maklumat, kaedah ruang vektor.

ABSTRACT

Latent semantic indexing is a variant of the vector space method in which a low-rank approximation to the vector space representation of the database is employed. The main idea in latent semantic indexing model is to map each document and query vector into a lower dimensional space which is associated with concepts. This is accomplished by mapping the index term vectors into this lower dimensional space. The claim is that retrieval in the reduced space may be superior to retrieval in the space of index terms. In this paper, in addition to vector space method, latent semantic indexing method is applied to develop Malay language information retrieval system.

Keywords: Malay language information retrieval, latent semantic indexing, information retrieval, vector space method.

PENGENALAN

Dapatan semula maklumat (IR) melibatkan perwakilan, simpanan, organisasi dan capaian item maklumat. IR mempunyai paradigma seperti berikut: pengguna ingin mendapatkan dokumen tentang tajuk tertentu; pengguna memberikan penerangan teks dalam bentuk bebas tentang tajuk untuk menjadi pertanyaan; daripada pertanyaan tersebut enjin IR mendapatkan istilah-istilah indeks; semua istilah indeks dipadankan dengan istilah indeks yang diperolehi dari dokumen yang telah diproses terdahulu; dokumen yang mempunyai padanan terbaik diberikan kepada pengguna dalam susunan berpangkat [Baeza-Yates 1999; Grefenstette 1998].

Penilaian IR ke atas kejayaan ialah kepersisan dan panggilan semula. Kepersisan ialah berapa banyak dokumen dalam senarai berpangkat yang berkaitan dengan pertanyaan. Manakala panggilan semula ialah berapa banyak dokumen yang berkaitan yang mungkin boleh dijumpai dalam koleksi dokumen yang berada dalam senarai capaian.

Antara kaedah popular dapatan semula maklumat yang dibangunkan ialah kaedah yang berasaskan ruang vektor. Data dimodelkan sebagai satu matriks dan pertanyaan pengguna terhadap pangkalan data diwakilkan sebagai satu vektor. Dokumen-dokumen yang berkenaan dalam pangkalan data ditentukan melalui operasi vektor. Setiap dokumen diwakilkan dengan satu vektor, dengan setiap komponen menunjukkan kepentingan untuk suatu istilah dalam mewakili semantik atau makna bagi dokumen. Vektor untuk semua dokumen dalam pangkalan data disimpan sebagai lajur bagi satu matriks.

Kaedah pengindeksan semantik terpendam (LSI) ialah satu variasi daripada model ruang vektor dengan anggaran pangkat rendah kepada perwakilan ruang vektor untuk pangkalan data digunakan. Matriks asal digantikan dengan matriks lain yang hampir sama dengan matriks asal tetapi ruang lajurnya hanya subruang daripada ruang lajur matriks asal.

PERWAKILAN ISTILAH DAN DOKUMEN

Dalam model ruang vektor dan LSI, istilah dan dokumen diwakilkan oleh satu matriks $m \times n$, A. Setiap istilah unik m dalam koleksi dokumen diberikan satu baris dalam matriks, manakala setiap dokumen n dalam koleksi diberikan satu lajur dalam matriks. Elemen bukan sifar a_{ij}, iaitu $A = [a_{ij}]$ menunjukkan istilah i wujud dalam dokumen j dan memberikan bilangan kewujudan istilah dalam dokumen tersebut. Memandangkan bilangan istilah dalam sesuatu dokumen biasanya begitu sedikit berbanding dengan bilangan istilah dalam seluruh koleksi dokumen, matriks A adalah sangat jarang.

Nilai a_{ij} digantikan dengan nilai pemberat tertentu untuk meningkatkan pencapaian. LSI menggunakan skema pemberat tempatan dan global untuk meningkatkan atau mengurangkan kepentingan relatif untuk istilah di dalam dokumen dan dalam seluruh koleksi dokumen. Hasil darab fungsi pemberat tempatan dan global dilaksana untuk setiap elemen bukan sifar matriks A, iaitu:

$$a_{ij} = L(i, j) * G(i), \tag{1}$$

Dengan $L(i, j)$ ialah fungsi pemberat tempatan untuk istilah i dalam dokumen j, dan $G(i)$ ialah fungsi pemberat global untuk istilah i. Dumais (1991) mendapati skema pemberat log-entropi menghasilkan keputusan 40% lebih baik daripada frekuensi istilah atas koleksi dokumen ujian piawai. Fungsi pemberat tempatan dan global log-entropi adalah seperti di bawah:

$$\text{Log} \quad : \; L(i,j) = \log_2 (tf_{ij} + 1).$$

$$\text{Entropi} : \; G(i) = 1 - \sum_j \frac{p_{ij} \log_2 (p_{ij})}{\log_2 (ndocs)}, p_{ij} = \frac{tf_{ij}}{gf_i} \tag{2}$$

dengan :

tf_{ij} = frekuensi istilah i dalam dokumen j
gf_i = frekuensi global istilah i
$ndocs$ = bilangan dokumen di dalam koleksi

Penghuraian Nilai Singular

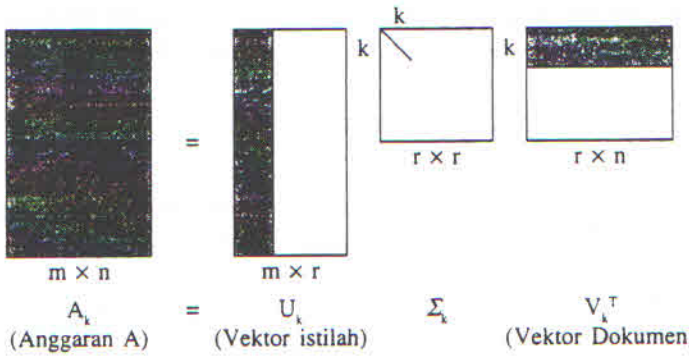
Daripada matriks $m \times n$, A satu matriks anggaran untuk A, A_k yang berpangkat k dikira menggunakan penghuraian nilai singular (SVD) dan nilai $k << \min(m,n)$. SVD untuk matriks A ditakrifkan sebagai hasil darab tiga matriks, iaitu:

$$A = U \Sigma V^T. \tag{3}$$

V^T ialah vektor transposisi daripada vektor V. U dan V^T ialah vektor singular kiri dan kanan dan Σ ialah matriks pepenjuru dengan elemen pepenjuru Σ ialah nilai singular matriks A mengikut tertib menurun. Lajur U dan V ialah ortogon, sehingga $U^T U = V^T V = I_r$ dan r ialah pangkat matriks A. Sebanyak k lajur pertama bagi matriks U dan V serta sebanyak k nilai singular A terbesar diguna untuk membina anggaran untuk A pada pangkat k melalui:

$$A_k = U_k S_k V_k^T. \tag{4}$$

Gambaran SVD ditunjukkan dalam Rajah 1. Kawasan gelap U, V dan garisan dalam Σ mewakili A_k.



RAJAH 1. Gambaran SVD

KAEDAH RUANG VEKTOR

Kaedah ruang vektor digunakan dalam penyelidikan dapatan semula maklumat lebih 30 tahun lalu [Salton dan McGill 1983]. Selepas matriks istilah-dokumen dibina, pengiraan persamaan boleh dilakukan antara dua objek teks. Satu objek teks q diwakilkan dengan satu vektor $n \times 1$, seperti satu lajur daripada matriks A dan dengan jenis pemberat istilah yang sama digunakan. Seterusnya persamaan antara objek teks q_1 dan q_2 dapat dikira dengan nilai kosinus, iaitu:

$$sim(q_1, q_2) = \frac{q_1^T q_2}{\sqrt{q_1^T q_1 \cdot q_2^T q^2}} \tag{5}$$

dengan:

q_1 dan q_2 : objek-objek teks yang dinilai kesamaannya

q_1^T dan q_2^T : vektor-vektor transposisi untuk vektor-vektor q_1 dan q_2 masing-masing.

PENGINDEKSAN SEMANTIK TERPENDAM

Pengindeksan semantik terpendam direka bentuk untuk mengatasi masalah dalam kaedah vektor [Deerwester et al. 1990; Berry, Dumais dan O'Brien 1995]. Idea utama LSI ialah pemetaan setiap dokumen dan pertanyaan ke dalam ruang dimensi yang lebih rendah yang berkaitan dengan konsep. LSI

juga bermula dengan pembentukan matriks istilah-dokumen. Kemudian matriks ini dianalisis menggunakan penghuraian nilai singular untuk mengekstrak struktur berkenaan hubung kait antara dokumen-dokumen dan istilah-istilah. Proses ini boleh mengenalpasti, misalnya "car" dan "automobile" bila digunakan dalam konteks yang sama dalam seluruh koleksi, dan maklumat ini boleh diguna untuk meningkatkan dapatan semula.

Pertanyaan dibentuk menjadi dokumen-pseudo yang memberikan lokasi pertanyaan dalam ruang istilah-dokumen yang dikecilkan. Dengan q, satu vektor elemen bukan sifar mengandungi pemberat (menggunakan skema pemberat tempatan dan global yang sama dengan koleksi dokumen) frekuensi istilah dalam pertanyaan, dokumen-pseudo, $\hat{q}$ diwakilkan sebagai:

$$\hat{q} = q^T U_k \sum_k^{-1} \tag{6}$$

dengan:

q^T : vektor transposisi untuk vektor q,

U_k : vektor istilah daripada persamaan (4),

$\sum_k^{-1}$: matriks songsang untuk matriks S_k daripada persamaan (4).

Persamaan di antara dua dokumen di dalam koleksi dokumen adalah menggunakan vektor dokumen, V_k iaitu:

$$A_k^T A_k = V_k \sum_k^2 V_k^T \tag{7}$$

dengan:

A_k^T : matriks transposisi untuk matriks A_k,

V_k^T : vektor transposisi untuk vektor V_k daripada persamaan (4),

$\sum_k^2$: kuasa dua untuk matriks S_k daripada persamaan (4).

Eksperimen Menggunakan Kaedah Vektor dan LSI

Eksperimen dijalankan terhadap satu koleksi dokumen bahasa Inggeris dan dokumen bahasa Melayu. Capaian kepada koleksi dokumen ini dibuat melalui pertanyaan daripada bahasa Melayu dan Inggeris dengan menggunakan kaedah vektor dan LSI.

Koleksi dokumen bahasa Melayu terdiri dari teks terjemahan Al-Quran bahasa Melayu yang mempunyai 6,236 dokumen. Koleksi dokumen ini mengandungi 196,071 patah perkataan dengan 7,526 perkataan yang unik. Manakala koleksi dokumen bahasa Inggeris terdiri daripada teks terjemahan Al-Quran bahasa Inggeris yang terdiri mempunyai 6,236 dokumen yang setara dengan dokumen-dokumen bahasa Melayu. Koleksi dokumen ini mengandungi 167,477 patah perkataan dengan 6,383 perkataan yang unik.

Set pertanyaan bahasa Melayu mengandungi 36 pertanyaan dalam bentuk bahasa tabii. Manakala pertanyaan bahasa Inggeris adalah 36 pertanyaan dalam bentuk bahasa tabii yang merupakan terjemahan daripada pertanyaan bahasa Melayu.

Empat eksperimen dijalankan terhadap setiap koleksi dokumen iaitu:

a) eksperimen menggunakan kaedah vektor tanpa penggunaan pengakar bahasa
b) eksperimen menggunakan kaedah LSI tanpa penggunaan pengakar bahasa
c) eksperimen menggunakan kaedah vektor dengan penggunaan pengakar bahasa
d) eksperimen menggunakan kaedah LSI dengan penggunaan pengakar bahasa

Eksperimen dijalankan bermula dengan pembentukan matriks istilah-dokumen dan matriks istilah-pertanyaan. Elemen bagi matriks ini ialah frekuensi istilah dalam dokumen atau pertanyaan yang berkenaan. Seterusnya diumpukkan nilai pemberat log-entropi kepada setiap elemen bukan sifar matriks tersebut.

Capaian menggunakan kaedah vektor melibatkan langkah seperti berikut:

a) kira persamaan setiap pertanyaan dengan setiap dokumen menggunakan nilai kosinus
b) susun senarai dokumen berkenaan mengikut nilai kosinus dalam tertib menurun untuk setiap pertanyaan

Manakala capaian menggunakan kaedah LSI pula melibatkan langkah seperti berikut:

a) kira SVD untuk matriks istilah-dokumen pada nilai k sebanyak 200
b) bina dokumen-pseudo untuk matriks istilah-pertanyaan
c) gabung dokumen-pseudo dengan vektor dokumen
d) kira persamaan dokumen-pseudo dengan setiap dokumen
e) susun senarai dokumen berkenaan mengikut nilai persamaan dalam tertib menurun untuk setiap pertanyaan

Seterusnya eksperimen diulang dengan menggunakan pengakar bahasa yang berkenaan atas indeks koleksi dokumen dan set pertanyaan. Pengakar Ahmad (1995) digunakan untuk koleksi bahasa Melayu dan pengakar Porter (1980) digunakan untuk koleksi bahasa Inggeris.

KEPUTUSAN DAN PERBINCANGAN

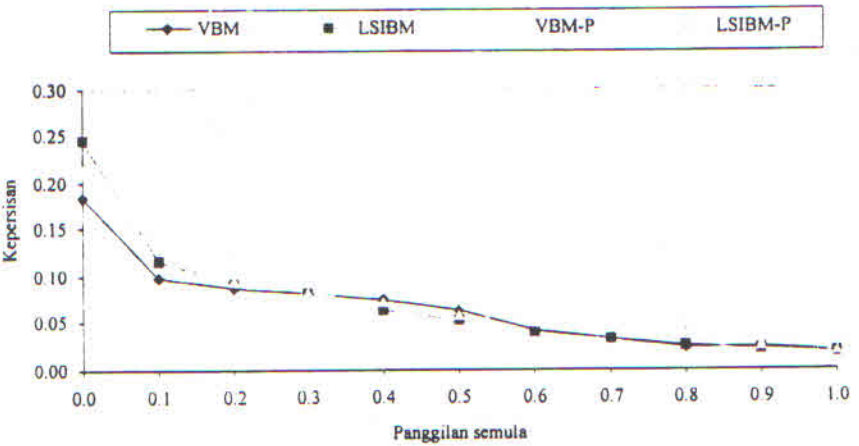
Keputusan yang diperolehi daripada empat eksperimen untuk dua koleksi dokumen ditunjukkan dalam bahagian berikut. Keputusan untuk dapatan semula maklumat bahasa Melayu ditunjukkan pada Jadual 1. Jadual ini menunjukkan kepersisan untuk capaian dokumen bahasa Melayu menggunakan kaedah ruang vektor (VBM); capaian dokumen bahasa Melayu menggunakan kaedah pengindeksan semantik terpendam (LSIBM); capaian dokumen bahasa Melayu menggunakan kaedah ruang vektor dan pengakar (VBM-P); dan capaian dokumen bahasa Melayu menggunakan kaedah pengindeksan semantik terpendam dan pengakar (LSIBM-P).

JADUAL 1. Keputusan dapatan semula maklumat bahasa Melayu

Panggilan semula	Kepersisan			
	VBM	LSIBM	VBM-P	LSIBM-P
0.0	0.183584	0.245733	0.213627	0.224134
0.1	0.098444	0.116278	0.139822	0.125747
0.2	0.088175	0.091869	0.094857	0.107330
0.3	0.080917	0.082287	0.081724	0.097803
0.4	0.075425	0.064352	0.072881	0.081417
0.5	0.062735	0.052383	0.058078	0.069830
0.6	0.041278	0.040215	0.051244	0.055403
0.7	0.032769	0.033473	0.043785	0.049949
0.8	0.023738	0.025660	0.039536	0.042928
0.9	0.022735	0.021481	0.025456	0.028933
1.0	0.019459	0.017616	0.020815	0.021366
Purata kepersisan	0.066296	0.071941	0.076530	0.082258

Graf kepersisan lawan panggilan semula untuk capaian dokumen bahasa Melayu diplotkan pada Rajah 2. Secara purata didapati tanpa menggunakan pengakar, capaian menggunakan kaedah LSI ialah 8.5% lebih tinggi kepersisannya berbanding kaedah vektor. Manakala capaian menggunakan kaedah LSI ialah 7.5% lebih tinggi kepersisannya berbanding kaedah vektor dengan menggunakan pengakar. Didapati penggunaan pengakar telah dapat meningkatkan kepersisan sebanyak 15.4% untuk kaedah vektor dan kepersisan meningkat sebanyak 14.3% bagi kaedah LSI.

Keputusan untuk dapatan semula maklumat bahasa Inggeris ditunjukkan pada Jadual 2. Jadual ini menunjukkan kepersisan untuk capaian dokumen bahasa Inggeris menggunakan kaedah ruang vektor (VBI); capaian dokumen bahasa Inggeris menggunakan kaedah pengindeksan semantik terpendam (LSIBI); capaian dokumen bahasa Inggeris menggunakan kaedah ruang vektor



RAJAH 2. Kepersisan dan panggilan semula capaian bahasa Melayu

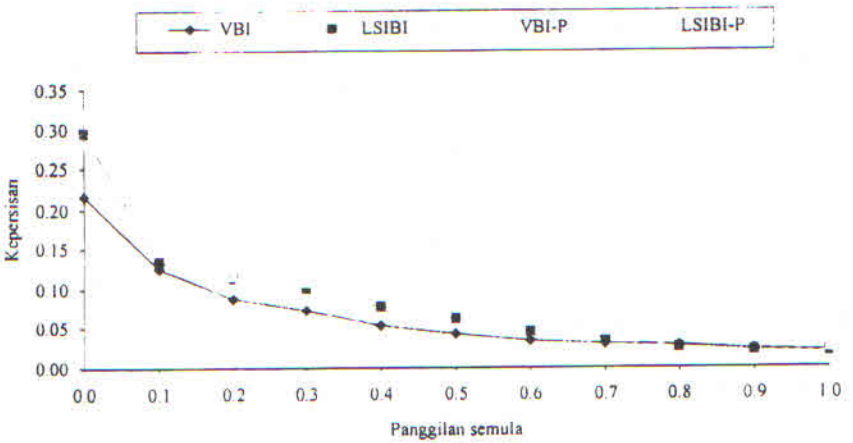
dan pengakar (VBI-P); dan capaian dokumen bahasa Inggeris menggunakan kaedah pengindeksan semantik terpendam dan pengakar (LSIBI-P).

JADUAL 2. Keputusan dapatan semula maklumat bahasa Inggeris

Panggilan semula	Kepersisan			
	VBI	LSIBI	VBI-P	LSIBI-P
0.0	0.215590	0.294146	0.317394	0.299287
0.1	0.124611	0.134528	0.152468	0.185682
0.2	0.087193	0.112053	0.116215	0.144249
0.3	0.073112	0.098917	0.107622	0.131102
0.4	0.053866	0.077101	0.090707	0.114065
0.5	0.042872	0.060540	0.079616	0.093251
0.6	0.033519	0.044170	0.062438	0.077318
0.7	0.030013	0.035175	0.046988	0.061935
0.8	0.026897	0.025384	0.042324	0.042234
0.9	0.023613	0.021058	0.036778	0.033830
1.0	0.021260	0.018352	0.025310	0.020645
Purata kepersisan	0.066595	0.083766	0.097987	0.109418

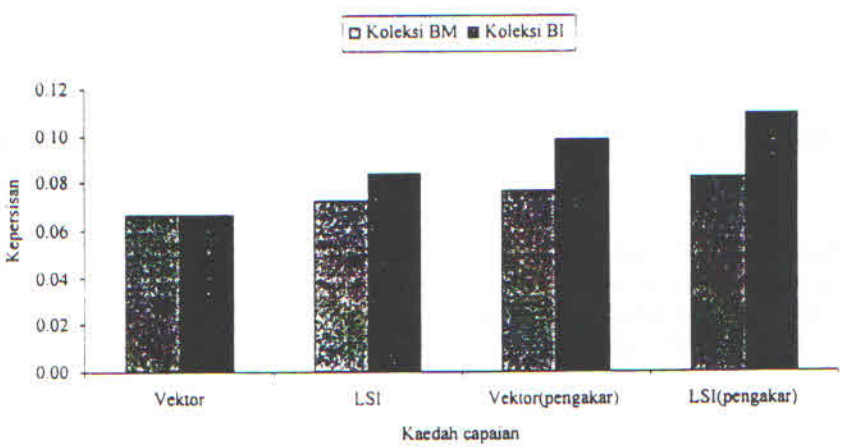
Seterusnya graf kepersisan lawan panggilan semula untuk capaian dokumen bahasa Inggeris diplotkan pada Rajah 3. Daripada nilai purata didapati tanpa menggunakan pengakar, capaian menggunakan kaedah LSI ialah 25.8% lebih tinggi kepersisannya berbanding kaedah vektor. Manakala capaian menggunakan kaedah LSI ialah 11.7% lebih tinggi kepersisannya

berbanding kaedah vektor dengan menggunakan pengakar. Di samping itu, didapati penggunaan pengakar dapat meningkatkan kepersisan sebanyak 47.1% untuk kaedah vektor dan kepersisan meningkat sebanyak 30.6% bagi kaedah LSI.



RAJAH 3. Kepersisan dan panggilan semula capaian bahasa Inggeris

Perbandingan hasil eksperimen untuk koleksi dokumen bahasa Melayu dan Inggeris ditunjukkan pada Rajah 4.



RAJAH 4. Kepersisan Mengikut Kaedah Capaian

Daripada perbandingan ini didapati capaian menggunakan kaedah LSI menghasilkan kepersisan lebih tinggi berbanding kaedah vektor sama ada

tanpa menggunakan pengakar atau dengan menggunakan pengakar. Keputusan ini adalah selaras untuk koleksi dokumen kedua-dua bahasa.

KESIMPULAN

Hasil kajian ini menunjukkan bahawa penggunaan kaedah LSI dalam sistem dapatan semula maklumat dapat meningkatkan lagi kepersisan berbanding kaedah vektor. Dengan ini pendekatan baru dalam membangunkan sistem dapatan semula maklumat bahasa Melayu harus memanfaatkannya. Penggunaan kaedah LSI juga perlu menggabungkan penggunaan pengakar bahasa kerana penggunaannya terbukti dapat meningkatkan purata kepersisan yang ketara.

RUJUKAN

Ahmad, F. 1995. *A Malay language document retrieval system: an experimental approach and analysis*. Tesis Ijazah Doktor Falsafah. Universiti Kebangsaan Malaysia.

Baeza-Yates, R. and Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. New York: Addison-Wesley.

Berry, M.W.; Dumais, S.T.; and O'Brien, G.W. 1995. Using linear algebra for intelligent information retrieval. *SIAM Review* 37(4):573-595.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R.A. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391-407.

Dumais, S. 1991. Using the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23(2):229-236.

Grefenstette, G. 1998. The Problem of Cross-Language Information Retrieval. *Dalam Cross-Language Information Retrival*, ed. G. Grefenstette, ms 1-9. Boston: Kluwer Academic Publishers.

Porter, M.F. 1980. An Algorithm for Suffix Stripping. *Program* 14(3):130-137.

Salton, G. dan McGill, M.J. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

MAKLUMAT PENGARANG

Muhammad Taufik Abdullah,
Fatimah Ahmad, Ramlan Mahmod
Fakulti Sains Komputer dan Teknologi Maklumat
Universiti Putra Malaysia
43400 UPM Serdang, Selangor
Emel: {taufik,fatimah,ramlan}@fsktm.upm.edu.my

Tengku Mohammed Tengku Sembok
Fakulti Teknologi dan Sains Maklumat
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor
Emel: tmts@ftsm.ukm.my