DOCUMENT CLUSTERING: COMPARISON OF WARD'S CLUSTERING AND KOHONEN NETWORK PERFORMANCE

PREPARED BY:

NOOR SURIANA ABU BAKAR
NURUL NISA MOHD NASIR

MAY 2012

Tarikh : 10 Mei 2012
No. Fail Projek : 600-RMI/SSP/ 5/3/Dsp (297/2011)

Penolong Naib Canselor (Penyelidikan)
Institut Pengurusan Penyelidikan (RMI)
UiTM, Shah Alam

Puan,

LAPORAN AKHIR PENYELIDIKAN 'DOCUMENT CLUSTERING: COMPARISON OF WARD'S CLUSTERING AND KOHONEN NETWORK PERFORMANCE'

Merujuk kepada perkara di atas, bersama-sama ini disertakan 3 (tiga) naskah Laporan Akhir Penyelidikan bertajuk 'Document Clustering: Comparison of Ward's Clustering and Kohonen Network Performance' oleh kumpulan Penyelidik dari Fakulti Sains Komputer dan Matematik untuk makluman pihak puan.

Sekian, terima kasih.

Yang benar,

NOOR SURIANA ABU BAKAR
Ketua
Projek Penyelidikan

UNIVERSITI TEKNOLOGI MARA

PEMENANG Anugerah Kualiti Perdana Menteri 2008

Surat Kami . 600-RMI/SSP/DANA 5/3/Dsp (297/2011)
Tarikh . 10 Jun 2011

Pn Nor Suriana Abu Bakar
Fakulti Sains Komputer dan Matematik
Universiti Teknologi MARA Cawangan Melaka
KM. 26, Jalan Lendu
78000 Alor Gajah, Melaka

Y. Brs. Profesor./Tuan/Puan

KELULUSAN PERMOHONAN DANA KECEMERLANGAN 06/2011

Tajuk Projek	:	Document Clustering : Comparison of Ward's Clustering and Kohonen Network Performance
Kod Projek	:	600-RMI/SSP/DANA 5/3/Dsp (297 /2011)
Kategori Projek	:	Kategori F (2011)
Tempoh	:	15 Jun 2011 – 14 Jun 2012 (12 bulan)
Jumlah Peruntukan	:	RM 5,000.00
Ketua Projek	:	Pn Nor Suriana Abu Bakar

Dengan hormatnya perkara di atas adalah dirujuk.

2. Sukacita dimaklumkan pihak Universiti telah meluluskan cadangan penyelidikan Y. Brs Profesor/tuan/puan untuk membiayai projek penyelidikan di bawah Dana Kecemerlangan UiTM.

3. Bagi pihak Universiti kami mengucapkan tahniah kepada Y. Brs. Profesor/tuan/puan kerana kejayaan ini dan seterusnya diharapkan berjaya menyiapkan projek ini dengan cemerlang.

4. Peruntukan kewangan akan disalurkan melalui tiga (3) peringkat berdasarkan kepada laporan kemajuan serta kewangan yang mencapai perbelanjaan lebih kurang 50% dari peruntukan yang diterima.

Peringkat Pertama	20%
Peringkat Kedua	40%
Peringkat Ketiga	40%

5. Untuk tujuan mengemaskini, pihak Y. Brs. Profesor/tuan/puan adalah diminta untuk melengkapkan semula kertas cadangan penyelidikan sekiranya perlu, mengisi borang setuju terima projek penyelidikan dan menyusun perancangan semula bajet yang baru seperti yang diluluskan. Sila lihat lampiran bagi tatacara tambahan untuk pengurusan projek.

Sekian, harap maklum.

"SELAMAT MENJALANKAN PENYELIDIKAN DENGAN JAYANYA"

Yang benar

DR OSKAR HASDINOR HASSAN
Ketua Penyelidikan (Sains Sosial dan Pengurusan)

/RS...

Penolong Naib Canselor (Penyelidikan) : 603-5544 2094/2095
Bahagian Penyelidikan : 603-5544 2097/2091/2101/5521 1462
Bahagian Perundingan : 603-5544 2100/2787/2092/2093
Bahagian Inovasi : 603-5544 2750/2747/2748
Bahagian Penerbitan : 603-5544 1425/2785
Bahagian Sokongan ICT : 603-5544 3097/2104/5521 1461
Bahagian Sains : 603-5544 2098/5521 1463
Pejabat Am : 603-5544 2559/2057/5521 1636
Penolong Pentadbiran : 603-5544 2090
Fax : 603-5544 2096/2767
Unit Kewangan Zon 17 : 603-5544 3404
603-5521 1386

5. REPORT

5.1 Proposed Executive Summary

Document clustering has been investigated for use in a number of different areas of information retrieval. The aimed of research in the field is to improve efficiency and effectiveness of retrieval. Since the clusters perform best quality, Hierarchical clustering is most commonly used in document clustering.

Recently, there exist researches that apply neural network in information retrieval. However research in neural network based document clustering still less frequent. Therefore this study will apply hierarchical based document clustering and neural network based document clustering in terms of suggestion supervisor and examiner for thesis.

The methodology used in this research are thesis collection and digitization, stop removal, Porter Stemming, document vector representation, compare the SOM's and Ward's Clustering and get the solution.

The results from these two techniques will then compare with manual system to find out whether hierarchical based or neural network based performed better. The collection of theses will be used and employed the pre-processing including stop word removal and stemming further measure the document similarity before apply the clustering techniques. The result will give some insight whether neural network is better for suggestion supervisor and examiner.

5.2 Enhanced Executive Summary

Document clustering has been investigated for use in a number of different areas of information retrieval. The aimed of research in the field is to improve efficiency and effectiveness of retrieval. Since the clusters perform best quality, Hierarchical clustering is most commonly used in document clustering. Recently, there exist researches that apply NN in IR. However research in NN based document clustering still less frequent. Therefore this study will apply hierarchical based document clustering and NN based document clustering in terms of suggestion supervisor and examiner for thesis. The results from these two techniques will then compare with manual system to find out whether hierarchical based or NN based performed better. The collection of theses will be used and employed the pre-processing including stopword removal and stemming further measure the document similarity before apply the clustering techniques. The result will give some insight whether NN is better for suggestion supervisor and examiner.