THE EVALUATION OF CONTEXT-ORIENTED XML DOCUMENT RETRIEVAL: A CASE STUDY OF OFFICIAL LETTER

INSTITUT PENYELIDIKAN, PEMBANGUNAN DAN PENGKOMERSILAN
UNIVERSITI TEKNOLOGI MARA
40450 SHAH ALAM, SELANGOR
MALAYSIA

BY:

HAYATI ABD RAHMAN

FEBRUARY 2006

Tarikh : 27 Januari 2006

No. Fail Projek: **600-IRDC/ST 5/3/913**

Penolong Naib Canselor (Penyelidikan)
Institut Penyelidikan, Pembangunan dan Pengkomersilan
Universiti Teknologi MARA
40450 Shah Alam

Ybhg. Prof.,

LAPORAN AKHIR PENYELIDIKAN "THE EVALUATION OF CONTENT-ORIENTED XML DOCUMENT RETRIEVAL: A CASE STUDY OF OFFICIAL LETTER"

Merujuk kepada perkara di atas bersama-sama ini disertakan 3 (tiga) naskah Laporan Akhir Penyelidikan bertajuk "THE EVALUATION OF CONTENT-ORIENTED XML DOCUMENT RETRIEVAL: A CASE STUDY OF OFFICIAL LETTER".

Sekian, terima Kasih.

Yang benar,

HAYATI ABD RAHMAN
Ketua
Projek Penyelidikan

PENGHARGAAN

Setinggi-tinggi penghargaan dan ribuan terima kasih diucapkan kepada semua pihak yang terlibat secara langsung dan tidak langsung bagi membolehkan penyelidikan ini disiapkan dengan sempurna.

Diantaranya:

Prof Madya Dr Adnan Ahmad
(Dekan Fakulti Teknologi Maklumat dan Sains Kuantitatif)

Prof Dr Zainab Abu Bakar
(Pengerusi Pusat Pengajian Sains Komputer)

dan

Cik Siti Fadhilah Md Yasin
(Pembantu Penyelidik)

ABSTRACT

The research applies the process of document segmentation in which document is separated into many parts. The term segmentation is usually used in which the document retrieval is significant. It is important since the content of documents appear as one big part. Later in the retrieval development, the segmentation would be used for the indexing part. The letter document has their own format, which consists of many parts. The prototype has been developed to allow the segmentation and the existence of content-based to the letter document. The documents are divided into smaller, recognized labels that are intensive and flexible for managing, editing, and extracting. The target of this thesis is to apply the standard of official letter for the system, as well as to develop the algorithm which will segment the letter documents, and convert to XML documents. The software used for this prototype is Visual Basic 6.0. More over, the information retrieval makes the retrieval of document or collection of data in the storage media more efficient, effective, relevant, faster and more reliable than before. Such indexing techniques may influence the effectiveness of retrieval itself. The extension component within the indexing structure may also influence the performance of the retrieval process. This research is to develop a prototype for indexing algorithm considering tag weighting for the XML document and also to test the indexer with the existing document. In order to perform efficient retrieval on documents, appropriate index structure or algorithm must be used which include the structural information. The inverted file method has been used for the indexing techniques to develop the indexing algorithm of the FTMSK official letter. The relevancy of the document for the retrieval by using the algorithm has been successful achieved and it can prove that the prototype can increase the relevancy of document retrieval.

TABLE OF CONTENTS