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

Institut Penyelidikan, Pembangunan dan Pengkomersilan (IRDC)
Institute of Research, Development and Commercialisation (IRDC)
(Sebelum ini dikenali sebagai Biro Penyelidikan dan Perundingan)
40450 Shah Alam, Malaysia
Website : http//: www.uitm.edu.my/brc

Surat Kami : 600-IRDC/ST 5/3/913
Tarikh : 30 Disember, 2004

Dekan
Fakuti Teknologi Maklumat dan Sains Kuantitatif
Universiti Teknologi MARA
40450 Shah Alam
SELANGOR

Tuan/ Puan

PERLANTIKAN BAGI MENJALANKAN PENYELIDIKAN

Merujuk kepada perkara di atas, bersama-sama ini dimajukan salinan surat kelulusan menjalankan penyelidikan serta ringkasan kos perbelanjaan bagi penyelidikan yang dijalankan oleh pensyarah dari Fakuti Teknologi Maklumat dan Sains Kuantitatif:

Tajuk Projek	: The Evaluation of Content-Oriented XML Document Retrieval: A Case Study of FTMSK Official Letter
Ketua Projek	: Cik Hayati Abdul Rahman
Kos Yang diluluskan	: RM 16,860.00
Jenis Geran	: Geran Dalaman

Sekian, terima kasih.

Yang benar

PROF MADYA DR ROSMIMAH MOHD ROSLIN
Ketua Penyelidikan (Sains Sosial dan Pengurusan)

s.k: 1. Prof Madya Dr Mohamad Alias Lazim
Ketua ProTRAD
Fakuti Teknologi Maklumat dan Sains Kuantitatif

~~2.~~ Cik Hayati Abdul Rahman
Ketua Projek
Fakuti Teknologi Maklumat dan Sains Kuantitatif

3. Encik Mohd Halil Marsuki
Penolong Akauntan
Unit Kewangan Zon 17

PENYELIDIKAN, PEMBANGUNAN DAN PENGKOMERSILAN LANDASAN KEWIBAWAAN DAN KECEMERLANGAN

elefon :

ong Naib Canselor (Penyelidikan)	: 03-55442094/5	Ketua Perundingan	: 03-55442100	Pegawai Eksekutif	: 03-55442057
Penyelidikan (Sains Sosial dan Pengurusan)	: 03-55442097	Ketua Pengkomersilan	: 03-55442750	Pejabat Am	: 03-55442093/2101
Penyelidikan (Sains dan Teknologi)	: 03-55442091	Ketua Harta Intelek	: 03-55442753	Fax	: 03-55442096
INFOREC	: 03-55442760	Penolong Pendaftar	: 03-55442092	Unit Kewangan Zon 17	: 03-55443440
Perundingan (Kewangan)	: 03-55442090	Pegawai Sains	: 03-55442098	Penolong Akauntan	: 03-55442099

QUALITY SYSTEM SIRIM UKAS 074
MS ISO 9001 REG NO. AR28

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

Page

CHAPTER 1

INTRODUCTION

1.1 Research Background

XML applications are now widely used to author documents accessible through the World Wide Web. An interesting feature of XML is the distinction drawn between presentation and content, so that the mark-up can represent information accurately. One of the advantages of using the XML is the content-oriented characteristic. As the project deals with document tagging, that characteristic facilitates the segmented contents of the document. This essentially relieves the query part of the document retrieval.

XML is the technology for creating markup languages to describe data of virtually any type in a structured manner. XML document structure contains of declarations, elements, comments, character references and others, which are indicated in the document by explicit markup.

One of the phases in information retrieval process is indexing. The purpose of indexing is to determine the subject matter of documents and express the subject matter in index terms to make subject retrieval possible. It is often implicitly assumed that the document will present its subject matter to the indexer and that the indexer