UNIVERSITI PUTRA MALAYSIA

ENHANCE EFFICIENCY OF ANSWERING XML KEYWORD QUERY USING INCOMPACT STRUCTURE OF MCCTREE

UMMU SULAIM BINTI SAZALY

FSKTM 2013 3

ENHANCE EFFICIENCY OF ANSWERING XML KEYWORD QUERY USING INCOMPACT STRUCTURE OF MCCTREE

By

UMMU SULAIM BINTI SAZALY

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Master of Science

November 2012

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master of Science

ENHANCE EFFICIENCY OF ANSWERING XML KEYWORD QUERY USING INCOMPACT STRUCTURE OF MCCTREE

By

UMMU SULAIM BINTI SAZALY

November 2012

Chair : Associate Professor Mohd Hasan bin Selamat

Faculty : Computer Science and Information Technology

People nowadays live in cyber life where everything can be done by just typing through keyboard and system will complete the process. As the interaction is done through online, data sharing is the most important service to send and deliver information. Extended Markup Language (XML) has been chosen as the most important data sharing medium as it is very friendly for human and machine to interpret. Due to the importance of it, many studies have been done to increase the effectiveness of retrieving information from XML file. Many notions and techniques have been introduced especially to process query of information.

Compact Lowest Common Ancestor (CLCA) and Maximal Compact Lowest Common Ancestor (MCLCA) implemented in algorithms named CGTreeGenerator and MCCTreeGenerator has been proven in returning an accurate result in answering XML keyword query. CGTreeGenerator compacted the XML tree by eliminating irrelevant nodes based on CLCA notion, which produced Compact Global Tree (CGTree). MCCTreeGenerator used CGTree to select subtree called Maximal

Compact Connected Tree (MCCTree) as query result based on MCLCA notion. However, the MCCTree cannot be used directly in its ranking method because calculation in ranking method used the structure of subtree as before it has been compacted. If the result cannot be used directly by the ranking method, the algorithm has an ineffective process. Moreover, if the ineffective process requires re-examining the original tree, the efficiency of the process of the algorithm will be reduced. This study is a response to these weaknesses. This study proposes a new algorithm, namely XMCCTreeGenerator, to enhance the efficiency of the CGTree-MCCTreeGenerator.

This study identifies the effective processes needed in producing XML query result using MCLCA notion and without compacting it. Those processes constructed XMCCTreeGenerator algorithm which will produce the same subtree as MCCTree but difference in its structure. This new returned subtree called Extended MCCTree (XMCCTree) can be used directly by the ranking method because it is in an incompact structure. An experiment is run using XML datasets available in XML Data Repository from University of Washington's website. Two files are selected which consist of different data structure and divided into three ranges of size. Keywords are manually randomly selected from the files and executed between three to five numbers of keyword.

Two prototypes are developed which implement CGTree-MCCTreeGenerator and XMCCTreeGenerator. Since this study focuses on efficiency of the algorithm, elapsed time for each execution is collected from the experiment. In conclusion, the

proposed XMCCTreeGenerator is more efficient than the previous CGTree-MCCTreeGenerator in answering XML keyword query using MCLCA.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

TINGKATKAN KECEKAPAN MENJAWAB PENCARIAN KATA KUNCI XML MENGGUNAKAN STRUKTUR MCCTREE TIDAK PADAT

Oleh

UMMU SULAIM BINTI SAZALY

November 2012

Pengerusi : Profesor Madya Mohd Hasan bin Selamat

Fakulti : Komputer Sains dan Teknologi Maklumat

Masyarakat kini hidup dalam dunia siber di mana semua perkara boleh dilakukan dengan hanya menaip pada papan kekunci dan sistem akan menyelesaikan proses tersebut. Oleh kerana interaksi diselesaikan atas talian, perkongsian data adalah servis yang penting dalam menghantar dan menerima maklumat. Lanjutan Bahasa Penanda (XML) telah dipilih sebagai medium perkongsian data yang utama kerana ianya mudah untuk ditafsirkan oleh manusia dan mesin. Oleh kerana kepentingan itu, banyak kajian telah dilakukan untuk meningkatkan keberkesanan dalam mendapatkan maklumat dari fail XML. Pelbagai kaedah dan teknik telah diperkenalkan terutamanya untuk memproses pertanyaan maklumat.

Aturan Am Padat Terendah (CLCA) dan Aturan Am Padat Maksimum Terendah (MCLCA) digunakan dalam algoritma CGTreeGenerator dan MCCTreeGenerator telah terbukti dalam mengembalikan keputusan yang tepat dalam menjawab pertanyaan kata kunci XML. CGTreeGenerator memadatkan pepohon XML dengan menghapuskan nod yang tidak relevan berdasarkan anggapan CLCA, yang

menghasilkan Pepohon Global Padat (CGTree). MCCTreeGenerator menggunakan CGTree untuk memilih pohon dipanggil Pohon Bersambung Padat Maksima (MCCTree) sebagai keputusan pertanyaan berdasarkan anggapan MCLCA. Walau bagaimanapun, MCCTree tidak boleh digunakan secara terus dalam kaedah penarafan kerana pengiraan dalam kaedah penarafan menggunakan struktur pohon sebelum ianya dipadatkan. Jika keputusan tidak boleh digunakan terus oleh kaedah penarafan, proses tersebut adalah tidak efektif. Tambahan pula, jika proses yang tidak efektif perlukan pemeriksaan semula pepohon asal, kecekapan proses akan berkurangan. Kajian ini adalah tindakbalas kepada kelemahan tersebut. Kajian ini mencadangkan algoritma baru, dinamakan XMCCTreeGenerator, untuk meningkatkan kecekapan algoritma CGTree-MCCTreeGenerator.

Kajian ini mengenalpasti proses-proses efektif yang diperlukan dalam menghasilkan keputusan pertanyaan XML menggunakan MCLCA serta tanpa memadatkannya. Kesemua proses itu membina algoritma XMCCTreeGenerator yang menghasilkan pohon yang sama seperti MCCTree dalam struktur yang berlainan. Pohon baharu yang dikembalikan ini dinamakan lanjutan MCCTree (XMCCTree) boleh digunakan terus oleh kaedah penarafan kerana ianya berada dalam struktur yang tidak padat. Eksperimen dijalankan menggunakan dataset XML di Pangkalan Data XML di laman sesawang Universiti Washington. Dua fail dipilih yang mengandungi struktur data berlainan dan dibahagikan kepada tiga saiz. Kata kunci dipilih rawak secara manual dari fail-fail tersebut dan diproses antara tiga sehingga lima kata kunci.

Dua prototaip dibangunkan menggunakan CGTree-MCCTreeGenerator dan XMCCTree. Oleh kerana kajian ini memberi tumpuan kepada kecekapan algoritma,

tempoh masa setiap pemprosesan diambil daripada eksperimen. Kesimpulannya, XMCCTreeGenerator yang dicadangkan adalah lebih berkesan berbanding CGTree-MCCTreeGenerator yang sebelumnya dalam menjawab pertanyaan kata kunci XML menggunakan MCLCA.

TABLE OF CONTENTS