UNIVERSITI PUTRA MALAYSIA

DIRECT APPROACH FOR MINING ASSOCIATION RULES FROM STRUCTURED XML DATA

ASHRAF RIAD ABAZEED

FSKTM 2012 21

DIRECT APPROACH FOR MINING ASSOCIATION RULES FROM STRUCTURED XML DATA

ASHRAF RIAD ABAZEED

DOCTOR OF PHILOSOPHY

UNIVERSITY PUTRA MALAYSIA

2012

DIRECT APPROACH FOR MINING ASSOCIATION RULES FROM STRUCTURED XML DATA

ASHRAF RIAD ABAZEED

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Doctor of Philosophy

JANUARY 2011

Abstract of thesis presented to Senate of University Putra Malaysia in fulfillment of the requirement for the degree of Doctor of Philosophy

DIRECT APPROACH FOR MINING ASSOCIATION RULES FROM STRUCTURED XML DATA

By

ASHRAF RIAD ABAZEED

January 2011

Chairman : Associate Professor Ali Mamat , PhD

Faculty : Computer Science and Information Technology

XML has become the standard for data representation on the internet. This expansion in reputation has prompt the need for a technique to access XML documents for particular information and to manipulate repositories of documents represented in XML to find specific documents. Having the ability to extract information from XML data would answer the problem of mining the web contents which is a very useful and required power nowadays. Efforts are made to develop a new tool or method for extracting information from XML data directly without any preprocessing or post processing of the XML documents.

Association rules express the probability of the existing of a set of items when another set of items exists. It searches for similarities among large database. “Web mining” refer to how we can apply the traditional mining techniques that works on relational data and

bind it to new data input represented in XML data which might be semi structure or unstructured.

There are several techniques to mine association rules from XML data. The basic approach is to map the XML documents to relational data model and to store them in a relational database. This allows us to apply the standard tools that are in use to perform rule mining from relational databases. Even though it makes use of the existing technology, this approach is often time consuming and involves manual intervention because of the mapping process.

The focus of this study is to propose an enhancement on memory consumption by reducing the number of candidates generated for the existing FLEX algorithm which will reduce the amount of memory needed to execute the algorithm. Another aim of this study is to do an enhancement on the current structure of FLEX algorithm in terms of elimination of the candidate generation step. The thesis also provides a two different implementation of the modified FLEX algorithm using a java based parsers and XQuery implementation.

The thesis outlines the two different implementation techniques of the existing FLEX algorithm using java based parsers and using a query language for XML. The implementation details shows the difference in accessing and manipulating XML

documents using java based parsers and query languages for XML and the steps needed to access an XML document until we produce a list of association rules .

The proposed method, XiFLEX has been implemented using two different techniques (java based & XQuery) and compared with the original FLEX algorithm in its basic implementation and the Apriori algorithm for frequent patterns generation.

The experiments were conducted on self generated data sets (7 different sets) and well known datasets (Mushroom & Cars Data set). The results have shows that the proposed method, XiFLEX, has a better performance in terms of the time it takes to generate frequent patterns and the number of candidates generated (memory consumption).

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

PENDEKATAN LANGSUNG UNTUK PERATURAN PERSATUAN PERLOMBONGAN DARI DATA XML BERSTRUKTUR

Oleh

ASHRAF ABAZEED

Januari 2011

Chairman : Profesor Madya Ali Mamat , PhD

Faculty : Sains Komputer dan Teknology Maklumat

XML telah menjadi standard untuk perwakilan data di internet.Ini pengembangan dalam reputasi mempunyai segera keperluan untuk teknik untuk mengakses dokumen XML untuk maklumat tertentu dan untuk memanipulasi repositori dokumen yang diwakili dalam XML untuk mencari dokumen-dokumen tertentu. Mempunyai kebolehan untuk mengekstrak maklumat daripada data XML akan menjawab masalah perlombongan kandungan web yang merupakan satu kuasa yang sangat berguna dan diperlukan di zaman sekarang. Usaha-usaha dibuat untuk membina alat baru atau kaedah untuk mendapatkan maklumat daripada data XML secara langsung tanpa apa-apa pra pemprosesan atau pemprosesan pos dokumen XML.

Peraturan persatuan menerangkan kebarangkalian yang sedia ada bagi satu set item apabila satu lagi set barangan wujud. Ia mencari persamaan antara pangkalan data yang

besar. "Perlombongan Web" merujuk kepada bagaimana kita boleh menggunakan teknik perlombongan tradisional yang bekerja pada data hubungan dan mengikat kepada input data baru yang diwakili dalam data XML yang mungkin struktur separa atau tidak berstruktur

Terdapat beberapa teknik untuk lombong persatuan peraturan dari data XML. Pendekatan asas adalah untuk memetakan dokumen XML kepada model data hubungan dan untuk menyimpan mereka dalam pangkalan data hubungan. Ini membolehkan kita untuk memohon alat standard yang digunakan untuk melaksanakan perlombongan pemerintahan daripada pangkalan data hubungan. Walaupun ia menggunakan teknologi yang sedia ada, pendekatan ini selalunya memakan masa dan melibatkan campur tangan manual kerana proses pemetaan.

Fokus kajian ini adalah untuk mencadangkan satu peningkatan pada penggunaan memori dengan mengurangkan bilangan calon yang dijana untuk algoritma FLEX sedia ada yang akan mengurangkan jumlah ingatan yang diperlukan untuk melaksanakan algoritma. Dan untuk mencadangkan satu peningkatan pada struktur semasa FLEX algoritma dari segi penghapusan langkah generasi calon. Tesis ini juga menyediakan dua pelaksanaan yang berbeza algoritma FLEX diubahsuai menggunakan parsers java berasaskan dan XQuery pelaksanaan.

Tesis menggariskan dua pelaksanaan teknik yang berbeza algoritma FLEX sedia ada menggunakan java parsers berasaskan dan menggunakan bahasa pertanyaan untuk

XML. Butir-butir pelaksanaan menunjukkan perbezaan dalam mengakses dan memanipulasi dokumen XML menggunakan java parsers berasaskan dan bahasa pertanyaan untuk XML dan langkah-langkah yang diperlukan untuk mengakses dokumen XML sehingga kita menghasilkan senarai peraturan persatuan.

Kaedah dicadangkan XiFLEX telah dilaksanakan dengan menggunakan dua teknik yang berbeza (java berasaskan & XQuery) dan berbanding dengan algoritma asal FLEX dalam pelaksanaan asas dan algoritma apriori untuk corak generasi yang kerap.

Kajian ini telah dijalankan ke atas set data diri yang dijana (7 set) dan terkenal dataset (Data Cendawan & Kereta set). Keputusan menunjukkan bahawa kaedah yang dicadangkan, XiFLEX, mempunyai prestasi yang lebih baik dari segi masa yang diambil untuk menjana corak kerap dan bilangan calon yang dijana (penggunaan memori).

ACKNOWLEDGEMENTS

In the name of Allah, the most Gracious, the most merciful, I thank Allah for granting me the perseverance and the strength that I need to complete my thesis

I would like to express the deepest appreciation to my committee chair, Associate Prof. Dr. Ali Mamat, who has the attitude and the substance of a genius: he continually and convincingly conveyed a spirit of adventure in regard to research.

I would like to thank my committee members,. Associate Prof. Dr. Md. Nasir Sulaiman and, Associate Prof. Dr. Hamidah Ibrahim who's always help me in direction, assistance, and guidance thought my study. And their recommendations and suggestions have been invaluable during my study

I would like to thank many people I have met during my stay in Malaysia for their help , encourage and support.

Finally, words alone cannot express the thanks I owe to my father Riad Abazeed and my mother Faridah Shekany for their encouragement and assistance during all these years.

Ashraf Riad Abazeed

January 2012

DECLARATION

I declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently submitted for any other degree at University Putra Malaysia or other institutions.

_____

ASHRAF RIAD ABAZEED

Date: 20 January 2012

TABLE OF CONTENTS