# A TABLE-BASED CLUSTERING STRATEGY FOR XML DATA STORAGE AND QUERYING

**NATTAPONG THAWORNKOOL**

**A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2005**

**ISBN 974-04-6337-1**
**COPYRIGHT OF MAHIDOL UNIVERSITY**

กลยุทธ์การจัดกลุ่มข้อมูลแบบเทเบิลเบสสำหรับการจัดเก็บและการสอบถามข้อมูลเอกซ์เอ็มแอล
(A TABLE-BASED CLUSTERING STRATEGY FOR XML DATA STORAGE AND QUERYING)

ณัฐพงศ์ ถาวรกูล 4337388 SCCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการควบคุมวิทยานิพนธ์ : ชาญยศ ปลื้มปีติวิริยะเวช,Ph.D, ดำรัส วงศ์สว่าง,Ph.D.

## บทคัดย่อ

เอกซ์เอ็มแอลกลายเป็นสื่อกลางในการแลกเปลี่ยนและรับส่งข้อมูลระหว่างแหล่งข้อมูลต่างๆ เมื่อการประมวลผลข้อมูลเอกซ์เอ็มแอลเพิ่มมากขึ้น ความจำเป็นในการจัดเก็บและการสอบถามข้อมูลเอกซ์เอ็มแอลอย่างมีประสิทธิภาพกลายเป็นประเด็นที่ได้รับความสนใจและมีการศึกษากันอย่างกว้างขวาง ซึ่งแบ่งออกได้เป็น 2 แนวทางด้วยกัน คือ นอนเนทีฟ เอกซ์เอ็มแอลและเนทีฟ เอกซ์เอ็ม ในวิทยานิพนธ์นี้ได้มุ่งเน้นไปที่เนทีฟ เอกซ์เอ็มแอลซึ่งสามารถแบ่งกลยุทธ์ตามการจัดกลุ่มข้อมูลได้เป็น 2 แบบคือ เอลิเม็นเบสซึ่งจัดกลุ่มข้อมูล ตามชนิดของเอลิเม็นและ ซับทรีเบส ซึ่งจะทำการจัดกลุ่มข้อมูลตามความสัมพันธ์ของเอลิเม็นซึ่งเกิดจากการแตกทรีออกเป็นทรีย่อยๆ

วิทยานิพนธ์นี้อธิบายกลยุทธ์การจัดกลุ่มข้อมูลแบบเทเบิลเบส เพื่อรองรับการจัดเก็บแบบเนทีฟและการสอบถามข้อมูลเอกซ์เอ็มแอล กลยุทธ์ดังกล่าวซึ่งมีลักษณะผสมผสานระหว่าง การจัดกลุ่มแบบเอลิเม็นเบสและซับทรีเบส กลยุทธ์เทเบิลเบสทำการจัดกลุ่มข้อมูลตามชนิดของเอลิเม็นและแบ่งแยกกลุ่มของข้อมูลตามโครงสร้างต้นไม้ ซึ่งแบ่งแยกตามความสัมพันธ์เชิงจำนวนระหว่างเอลิเม็นและจะทำการจัดเก็บข้อมูลเอกซ์เอ็มแอลโดยใช้โครงสร้างแบบตาราง จากผลการทดลอง พบว่ากลยุทธ์การจัดเก็บแบบเทเบิลเบสสามารถลดขนาดเอกสารเอกซ์เอ็มแอลได้ประมาณครึ่งหนึ่ง นอกจากนั้นการสอบถามข้อมูลเอกซ์เอ็มแอลทำได้อย่างมีประสิทธิภาพโดยเฉพาะการสอบถามข้อมูลที่เกี่ยวกับเอลิเม็นหลายๆชนิด

79 หน้า ISBN 974-04-6337-1

A TABLE-BASED CLUSTERING STRATEGY FOR XML DATA STORAGE AND QUERYING

NATTAPONG THAWORNKOOL 4337388 SCCS/M

M.Sc. (COMPUTER SCIENCE) MAJOR IN COMPUTER SCIENCE

THESIS ADVISORS : CHARNYOTE PLUEMPITIWIRIYAWEJ, Ph.D,
DAMRAS WONGSAWANG, Ph.D.

ABSTRACT

XML, a data format, has become the de facto standard for interchanging data among sources. As the processing of XML data is increasing, the requirement of efficient XML data storage and querying is apparent. Native and non-native XML storage are two aspects of storing XML data that facilitate efficient XML query processing. The element-based and the subtree-based clustering strategies are used to cluster XML data in the native storage with respect to the type of XML data elements and the sub-tree structure, respectively.

In this thesis, a new *table-based clustering strategy* to support efficient native storage and querying of XML data is described. The strategy combines element-based clustering and the subtree-based clustering. It categorizes the XML data with respect to the type of data elements and classifies the tree-like structure of XML data regarding the cardinalities of relationships between the data elements. The clustering technique stores XML data into a set of tables. With table-based clustering, the size of XML files can be reduced by half. Additionally, table-based clustering has been evaluated with a good performance for complex queries.

KEY WORDS : EXTENSIBLE MARKUP LANGUAGE / DATA CLUSTERING /
XML DATA STORAGE AND QUERYING /
TABLE-BASED CLUSTERING STRATEGY

79 P. ISBN 974-04-6337-1