รายงานการวิจัยเรื่อง	ส่วนประกอบในการค้นหากฎความเกี่ยวข้องสำหรับข้อมูลในรูปแบบ XML
หัวหน้าโครงการวิจัย	นางสาวอารยา โพธิสรณ์
ผู้ร่วมโครงการวิจัย	นายวิชิต สมบัติ
	นายธนกร ลิ้มสุวรรณ

คณะวิศวกรรมศาสตร์	มหาวิทยาลัยอุบลราชธานี
ปีงบประมาณ	2546
งบประมาณที่ได้รับ	58,200.- บาท
คำสำคัญ	กฎความเกี่ยวข้อง, เหมืองข้อมูล, จาวา

บทคัดย่อ

ข้อมูลเป็นสิ่งสำคัญในการตัดสินใจ อย่างไรก็ดีถ้าข้อมูลมีมากแต่ไม่มีเครื่องมือช่วยวิเคราะห์ ข้อมูลนั้นก็ใช้ประโยชน์ได้น้อยหรือไม่มีประโยชน์เลย กฎความสัมพันธ์เป็นเครื่องมือที่ใช้ในการค้นหาความสัมพันธ์ของข้อมูลเพื่อช่วยให้เราสามารถตัดสินใจได้อย่างมีเหตุผล

กฎความสัมพันธ์เป็นกระบวนการที่ใช้กันอย่างแพร่หลายในวงการธุรกิจ การเงิน การค้า การจัดการข้อมูลของแม่ข่าย การวิจัยทางสังคมและชีววิทยา และยังนำไปใช้ทางด้านต่างๆอีกมาก เช่น ช่วยวิเคราะห์หาความสัมพันธ์ของสินค้าที่ลูกค้าซื้อจากร้านขายของชำ เพื่อช่วยในการสั่งสินค้าเพิ่ม หรือกลยุทธ์ในการลดราคา

XML เป็นรูปแบบเอกสารมาตรฐานที่มีหน่วยงานสนับสนุนมากมาย ระบบฐานข้อมูลในปัจจุบันก็มีการพัฒนาให้สามารถรองรับรูปแบบเอกสาร XML ดังนั้นระบบที่สามารถดึงข้อมูลในรูปแบบดังกล่าวจึงจะสามารถนำมาใช้ประโยชน์ได้อย่างสูง

ระบบที่พัฒนาขึ้นในการวิจัยครั้งนี้ใช้เทคโนโลยีของจาวาที่เป็นที่ยอมรับกันอย่างแพร่หลาย โดยระบบจะใช้กระบวนวิธี Apriori ที่พัฒนาขึ้นโดยกลุ่มผู้พัฒนาของบริษัท IBMในการค้นหากฎความสัมพันธ์จากข้อมูลนำเข้าในรูปแบบ XML แล้วเขียนเป็นข้อมูลในรูปแบบของ PMML ซึ่งเป็นรูปแบบที่เหมาะสำหรับการสืบค้นกฎความสัมพันธ์ของข้อมูล เนื่องจาก PMML เป็นรูปแบบที่พัฒนาขึ้นโดยกลุ่ม Data Mining Group มีมีบริษัทใหญ่อย่าง Microsoft, IBM, SPSS, Oracle Corporations เป็นสมาชิกหลัก

ระบบที่พัฒนามีความสมบูรณ์ กล่าวคือสามารถเขียนผลลัพธ์ที่ได้จากการค้นหาความสัมพันธ์ของแฟ้มข้อมูลในรูปแบบ XML ให้เป็นข้อมูลในรูปแบบ PMML ทั้งนี้ระบบที่พัฒนายังสามารถพัฒนาต่อให้สามารถใช้งานผ่านเครือข่ายได้ ผู้ใช้สามารถกำหนดจำนวนข้อมูลสนับสนุนและระดับความน่าเชื่อถือให้กับระบบได้ ในกรณีที่ผู้ใช้ไม่ระบุระบบจะใช้ระดับความน่าเชื่อถือเป็นร้อยละ 30 และจำนวนข้อมูลสนับสนุนเป็นร้อยละ 20

Association rule extraction component for XML data

Head of Project Ms. Araya Pothisorn

Co-researchers Mr. Wichit Sombat

Mr. Thanakorn Limsuwan

Faculty of Engineering, Ubon Rajathanee University

In Finance Year 2002 for 58,200.- Bath

Keyword Association rule, PMML, Java-XML, Data Mining, Apriori

Abstract

Data is important assisting in making decision as well as a proper operation. However, having too much data without tools for refining the knowledge from the data is less useful or even useless. As one of such tools, association rules are among the most popular representations for local patterns in data mining. It can summarize the association of the information in order to clarify the characteristics of the data which we then can use to assist in making decision.

Association rules are used widespread in many aspects such as business, finance, marketing, server administration, social and biological research. Also, the applications are still extending to other fields. For example, the grocery, the result may be something similar to that most of the customers who buy beer often buy snacks. From this knowledge, the grocery can decide to stock more snacks when beer price is reduced.

Nowadays, because of its flexibility, XML is the standard for exchange the data and it is well supported by many large organizations. Also, many commercial database products are heading to support XML, so the tool for extracting the knowledge from data in form of XML will be substantially useful.

This project uses the widely accepted Java technology to develop a system for extracting association rules using A Priori algorithm developed by the IBM research team from XML-format input to product output file of PMML, a more suitable format for data mining process. PMML is specifications of the Data Mining Group:DMG which has IBM Corp. Microsoft, SPSS Inc., Oracle Corporations and fifteen software company as full members.

The project successfully implements the system. Output files in PMML format are produced corresponding to XML input format. The system runs in command-line mode. but could easily be ported to standalone GUI application or even an on-line version. User can specify minimum support and confidential level through configuration files. The system uses the default values for minimum support and confidential level of 20% and 30% respectively.