บทคัดย่อ

การศึกษาส่วนใหญ่ของเดต้าไมนิ่ง (Data Mining) มุ่งเน้นเกี่ยวกับข้อมูลที่มีโครงสร้าง (Structured Data) แต่อย่างไรก็ตามข้อมูลเอกสารประเภทต่างๆ เช่น ข่าว บทความ เอกสารงานวิจัย เป็นต้น ซึ่งจะจัดเก็บ ข้อมูลที่เก็บอยู่ในฐานข้อมูลเอกสาร (Document Databases) การจัดกลุ่มเอกสารเหล่านี้จึงเป็นเรื่องที่มีความสำคัญสำหรับการวิเคราะห์การจัดแบ่งกลุ่ม วิทยานิพนธ์นี้จึงนำเสนออัลกอริทึมการจัดกลุ่มเอกสารโดยใช้ เท็กอะแดปทีพเรโซแน้นเทียรีนิวรอลเน็ตเวิร์ค (Text Adaptive Resonance Theory Neural Network) วิธีการของเท็กอะแดปทีพเรโซแน้นเทียรีนิวรอลเน็ตเวิร์ค (Text Adaptive Resonance Theory Neural Network) ถูกออกแบบเพื่อแบ่งกลุ่มข้อมูลที่เป็นข้อความโดยตรงและไม่มีการแปลงข้อความเป็นตัวเลข ในการทดลองการทำงานของโครงข่ายประสาทเทียมที่ได้รับการออกแบบใหม่นี้ ได้ใช้ชุดข้อมูล 2 ชุด ได้แก่ ข้อมูลสังเคราะห์ (Synthesized Dataset) ซึ่งสร้างขึ้นในห้องปฏิบัติการ Data Mining & Data Exploration Laboratory และ ข้อมูลข่าวรอยเตอร์ (Reuters-21578 Distribution 1.0) ซึ่งถูกรวบรวมจากข้อมูลข่าวจริงของรอยเตอร์ (Reuters) ปี ค.ศ. 1987 ค่า Entropy และ F-Measure ถูกใช้เพื่อวัดผลลัพธ์ความถูกต้องของวิธีการใหม่ จากผลการทดลองโครงข่ายประสาทเทียมที่ได้รับการปรับแต่งนี้สามารถรับข้อมูลที่เป็นข้อความได้โดยตรง และทำการจัดกลุ่มข้อมูลที่มีคุณสมบัติมีค่าเป็นข้อความได้เป็นอย่างดี

ABSTRACT

The most studies of data mining have focused on structured data such as relational, transactional, and data warehouse data. However, the most available data that consist of large amounts of text documents such as news, articles, and research papers is stored in document database. The ability to group these documents is an important requirement for clustering analysis. This research proposes A Text Adaptive Resonance Theory Neural Network for document clustering. A Text Adaptive Resonance Theory Neural Network is designed to cluster on that data set that has a non-numerical feature value. Consequently, the proposed learning algorithm works directly on textual information without text transformation into a numerical value. The experiments are conducted on 2 datasets. The first dataset is a synthesized dataset, which is generated by the Data Mining & Data Exploration Laboratory and the second dataset is a Reuter-21578 Distribution 1.0 which is collected from the documents appeared on the Reuters newswire in 1987. The Entropy and F-Measure is used to measure the effectiveness of the proposed technique. According to the experimental results, the proposed neural network has shown good performance in clustering textual data.