บทคัดย่อ

ปัจจุบันเอกสารในรูปแบบอิเล็กทรอนิกส์มีปริมาณและเนื้อหาที่หลากหลายมากขึ้น การสืบค้นและการจัดการเอกสารจะง่าย และเป็นไปตามความต้องการ ต้องอาศัยการจัดแบ่งเอกสารเป็นกลุ่มหรือหมวดหมู่ ให้สอดคล้องและตรงกับดัชนี เพื่อให้จัดเก็บและค้นคืนเอกสารได้อย่างรวดเร็วและมีประสิทธิภาพ งานวิจัยนี้ประสงค์เพื่อพัฒนาเครื่องมือในการจำแนกหมวดหมู่เอกสารข้อความภาษาไทยด้วยอัลกอริทึม Feature Projection Text Categorization (FPTC) ซึ่งเป็นอัลกอริทึมที่ปรับมาจาก k-Nearest Neighbor ลักษณะเด่นของ FPTC คือ การแทนคุณลักษณะในแบบภาพฉายของแต่ละคุณลักษณะ การจำแนกหมวดหมู่จะใช้วิธีการเปรียบเทียบความคล้ายของคำที่ปรากฏในเอกสารที่ใช้ทดสอบกับคำที่ปรากฏในเอกสารที่ใช้ในกระบวนการเรียนรู้ เพื่อหาเอกสารที่คล้ายกับเอกสารทดสอบมากที่สุด และกำหนดหมวดหมู่ของเอกสารนั้นให้กับเอกสารทดสอบ โดยจะใช้เอกสารข่าวภาษาไทยจากหนังสือพิมพ์ออนไลน์เป็นกรณีศึกษา

จากผลการทดสอบพบว่า การจำแนกหมวดหมู่ด้วยอัลกอริทึม FPTC สามารถจำแนกหมวดหมู่เอกสารภาษาไทยได้อย่างมีประสิทธิผลดี สำหรับข้อมูลที่มีการกระจายตัวของหมวดหมู่เท่ากัน

Abstract

Amounts of Electronic documents have more increased and their contexts have more various. It needs competent management and retrieval systems for fast and most satisfying retrieval that required efficient indexing include document categorization. This research experiment Thai text document categorization according to prearranged categories by using feature projection text categorization (FPTC) in training and classification. Classification perform on contents of documents by comparing similarities of terms presented in test documents with similarities of terms presented in training documents