บทคัดย่อ

วิทยานิพนธ์นี้เสนอการจัดกลุ่มเอกสารภาษาไทยออกเป็นกลุ่มๆ ด้วยระบบคอมพิวเตอร์ ซึ่งวิทยานิพนธ์ฉบับนี้จะแบ่งออกเป็น 2 ส่วน ในส่วนแรกจะเป็นการตัดคำภาษาไทย ส่วนที่ 2 จะเป็นวิธีการจัดกลุ่มเอกสารภาษาไทย โดยระบบจะต้องทำการตัดคำภาษาไทยออกจากประโยคก่อน และแก้ความคลุมเครือของการตัดคำภาษาไทยซึ่งจะใช้วิธี Bigram จากนั้นระบบจะนำเอาคำที่ได้ไปเทียบหากลุ่มเอกสาร เอกสารที่คล้ายกันจะอยู่ในกลุ่มเดียวกัน ในการพิจารณาจัดกลุ่มเอกสารภาษาไทย จะพิจารณาระยะห่างระหว่างกลุ่มคำที่เหมือนกัน, จำนวนความถี่คำ,การเหมือนกันของคำระหว่างเอกสาร และจำนวนเอกสารหรือความถี่เอกสารที่มีคำเหมือนกันปรากฏอยู่ ทั้งหมดนี้ถูกนำมาวัดผลลัพธ์การจัดกลุ่มด้วยเครื่องมือวัดคือ ค่า Precision, Recall และ ค่า F-measure เปรียบเทียบกับการใช้มนุษย์จัดกลุ่มเอกสาร ในวิทยานิพนธ์นี้จะเน้นศึกษาดูการเหมือนกันของคำระหว่างเอกสาร รวมกับการหาระยะห่างระหว่างกลุ่มคำที่เหมือนกัน และศึกษาจำนวนเอกสารที่มีคำเหมือนกันปรากฏอยู่เป็นหลักเปรียบเทียบกับวิธีอื่น ผลของการจัดกลุ่มที่ได้วิธีพิจารณาการเหมือนกันของคำระหว่างเอกสารร่วมกับการหาระยะห่างระหว่างกลุ่มคำที่เหมือนกัน เอกสารในกลุ่มจะมีเนื้อหาใกล้เคียงกันมากหรือน้อยตามค่าเทรดโฮลที่กำหนด ถ้ากำหนดค่าเทรดโฮลน้อยเอกสารจะมีเนื้อหาใกล้เคียงกันมาก แต่จะได้กลุ่มเอกสารจำนวนมากด้วย และระบบมักจะไม่หยุดทำงานเนื่องจากเอกสารมีการเปลี่ยนกลุ่มไปมา ในขณะที่การจัดกลุ่มแบบพิจารณาตามจำนวนเอกสารที่มีคำเหมือนกันปรากฏอยู่ ระบบจะหยุดทำงานได้และการจัดกลุ่มก็ดีกว่าทุกวิธีที่เปรียบเทียบ

ABSTRACT

This thesis presents a Thai document clustering by the computer. It is divided into two parts. First part shows a Thai word parsing and second part shows a clustering. The system has to parse Thai words from sentences and modify an ambiguity on Thai word parsing by Bigram before it brings the acquired keywords to find out groups and to gather in a cluster. That is the same document is in the same group. This research considers Thai document clustering from keywords distance, keywords frequency, similarity of keywords and quantity of documents or documents frequency which has the same keywords. All of these considerations are measured by Precision, Recall and F-measure in comparison with a clustering from human. This thesis emphasizes the study of similarity of keywords in company with keywords distance, and documents frequency which has the same keywords compared with the other methods. The results are that the similarity of keywords in company with keywords distance, the documents are comparable in contents more or less after assigned by threshold. If the threshold value is less, the contents will be very comparable and have many groups too. But sometimes the system does not stop because the documents are changed all the time. In the meanwhile, the consideration of documents frequency which has the same keywords can stop the system and the document clustering is the best.