ชื่อ : นายนิเวศ จิระวิชิตชัย
ชื่อวิทยานิพนธ์ : การจัดหมวดหมู่เอกสารภาษาไทยโดยใช้เทคนิคการถ่วงน้ำหนักอัตราส่วนของคำแบบกำกวม
สาขาวิชา : เทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก : ผู้ช่วยศาสตราจารย์ ดร.ปริญญา สงวนสัตย์
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม : ผู้ช่วยศาสตราจารย์ ดร.พยุง มีสัจ
ปีการศึกษา : 2553

บทคัดย่อ

งานวิจัยด้านการจัดหมวดหมู่เอกสารภาษาไทยที่ผ่านมานั้น มุ่งพัฒนาทางด้านการสกัดคุณลักษณะ การลดมิติของคุณลักษณะ ตลอดจนการนำหลักไวยากรณ์ทางภาษาเข้ามาประยุกต์ใช้ในการจัดหมวดหมู่เอกสารภาษาไทย โดยงานวิจัยด้านจัดหมวดหมู่เอกสารส่วนใหญ่คำนวณค่าน้ำหนักให้กับคุณลักษณะที่สกัดได้บนพื้นฐานของความถี่ของคำ ซึ่งวิธีการคำนวณค่าน้ำหนักดังกล่าวไม่สามารถสะท้อนอำนาจจำแนกที่แท้จริงของคำและปัญหาคำกำกวมที่พบมากในการจัดหมวดหมู่เอกสารภาษาไทยได้ ทำให้ประสิทธิภาพในการจัดหมวดหมู่เอกสารภาษาไทยลดลง

งานวิจัยนี้มีวัตถุประสงค์ที่จะพัฒนาวิธีการคำนวณค่าน้ำหนักแบบใหม่ที่แก้ปัญหาความกำกวมของคำและถ่วงน้ำหนักคำใหม่ เพื่อสะท้อนถึงอำนาจจำแนกที่แท้จริงของคำที่สกัดได้ โดยนำเสนอวิธีการคำนวณค่าน้ำหนักให้กับดัชนีชื่อ Ambiguity Ratio Term Weighting (ARTW) เพื่อพัฒนาประสิทธิภาพในการจำแนกเอกสารบนแบบจำลองการจัดหมวดหมู่เอกสารภาษาไทย ผลจากการทดลอง เมื่อวัดประสิทธิภาพด้วยค่าเฉลี่ย F-Measure วิธีการคำนวณค่าน้ำหนักให้กับดัชนี ARTW ที่นำเสนอ ร่วมกับลดคุณลักษณะด้วยค่าสถิติไคแสควร์ (Chi-square) พบว่าอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพดีสุดที่ระดับ 99.9% รองลงมาเป็นอัลกอริทึม Naïve-Bayes ให้ประสิทธิภาพดีสุดที่ 97.7% และสุดท้ายอัลกอริทึม Decision Tree ให้ประสิทธิภาพดีสุดที่ 92.6% และเมื่อใช้วิธีการคำนวณค่าน้ำหนัก ARTW ร่วมกับค่าการเพิ่มของข้อมูล (Information Gain) พบว่าอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพดีสุดที่ระดับ 99.9% รองลงมาอัลกอริทึม Naïve-Bayes ให้ประสิทธิภาพดีสุดที่ 97.6% และอัลกอริทึม Decision Tree ให้ประสิทธิภาพดีสุดที่ 92.8% และสุดท้ายเมื่อใช้วิธีการคำนวณค่าน้ำหนัก ARTW ร่วมกับลดคุณลักษณะด้วยค่าความถี่เอกสาร (DF) พบว่าอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพดีสุดที่ระดับ 99.9% รองลงมาอัลกอริทึม Naïve-Bayes ให้ประสิทธิภาพดีสุดที่ระดับ 97.3% และอัลกอริทึม Decision Tree ให้ประสิทธิภาพดีสุดที่ระดับ 92.7%

ผลจากการทดลองสรุปได้ว่าวิธีการคำนวณค่าน้ำหนักให้กับดัชนี ARTW ที่พัฒนาขึ้นให้ประสิทธิภาพในจำแนกหมวดหมู่เอกสารภาษาไทยสูงกว่าวิธีการคำนวณน้ำหนักแบบอื่นๆทุกประเภทและทุกอัลกอริทึมที่นำมาเรียนรู้เพื่อสร้างตัวจำแนกเอกสารอย่างชัดเจน

เมื่อพิจารณาพารามิเตอร์ที่ส่งผลให้ประสิทธิภาพด้านการจัดหมวดหมู่เอกสารภาษาไทยดีที่สุดด้วยวิธีการคำนวณค่าน้ำหนักด้วย ARTW ร่วมกับอัลกอริทึม Support Vector Machine พบว่าที่จำนวน 5,000 คุณลักษณะ ให้ประสิทธิภาพดีสุดที่ระดับ 99.9% ซึ่งสามารถลดคุณลักษณะลงได้มากถึง 82.78% โดยไม่กระทบต่อประสิทธิภาพในการจัดหมวดหมู่เอกสารแต่อย่างใด

Name : Mr. Nivet Chirawichitchai
Thesis Title : Thai Document Categorization Using Ambiguity Ratio Term Weighting Technique
Major Field : Information Technology
King Mongkut's University of Technology North Bangkok
Thesis Advisor : Assistant Professor Dr.Parinya Sa-nguansat
Co-Advisor : Assistant Professor Dr.Phayung Meesad
Academic Year : 2010

Abstract

Referring to traditional Thai documents, categorization research usually focuses on feature extraction and feature selection of the document, language grammar is used for classifying Thai documents. Most research to calculate the term weighting of feature extraction is based on the frequency of words. But his method of calculating term weighting is not effective and leads too many problems commonly found with ambiguity in classifying Thai documents.

This research aims to develop a new method to calculate term weighting that solves the ambiguity issues and reflects the actual discrimination of isolated words. The new presented method of calculating the term weighting is called Ambiguity Ratio Term Weighting (ARTW), this will improve efficiency of Thai document categorization framework.

The experimental results showed that reducing the feature by Chi-square method and process Ambiguity Ratio Term Weighting (ARTW) with support vector machines will yielded a very high classification with F-Measure equaling 99.9%. Followed by Naïve-Bayes which yielded performance equaling 97.7%, and finally Decision Tree which yielded performance equaling 92.6%, respectively. When calculating the weight with ARTW and Information Gain methods, the Support Vector Machine yielded a very high classification with the F-Measure equaling 99.9%. Followed by Naïve-Bayes which yielded performance equaling 97.6% and Decision Tree performance equaling 92.8%. Finally, ARTW weighting with document frequency found that Support Vector Machine yielded performance equaling 99.9%. Followed by Naïve-Bayes which yielded performance equaling 97.3%, and finally Decision Tree which yielded performance equaling 92.7%, respectively. The experiments concluded that the ARTW weighting method clearly improved the efficiency of Thai document categorization over other weighting methods and algorithms. Based on the parameters affecting the efficiency of Thai document categorization with Support Vector Machine found the number of 5000 features the best performing at 99.9%.Thus, feature can be reduced to 82.78% without affecting the efficiency of Thai document categorization.