การศึกษาเปรียบเทียบเทคนิคการเลือกคำสำคัญโดยอัตโนมัติสำหรับงานสารบรรณอิเล็กทรอนิกส์

จุฑาวรรณ สิทธิโชคสถาพร

คณะแพทยศาสตร์ มหาวิทยาลัยสงขลานครินทร์

sjuthawa@medicine.psu.ac.th

บทคัดย่อ

แนวโน้มของการบริหารจัดการเอกสารในสำนักงาน มีการเปลี่ยนแปลงไปสู่ยุคของเอกสารอิเล็กทรอนิกส์มากขึ้นเรื่อยๆ องค์กรภาครัฐหลายแห่งเห็นความสำคัญของการนำเอารูปแบบการใช้งานเอกสารอิเล็กทรอนิกส์มาปรับใช้ในการบริหารจัดการเอกสารตัวอย่างเช่น งานสารบรรณ ซึ่งเป็นหน่วยงานส่วนกลางที่ทำหน้าที่หลักในการรับและส่งเอกสารของทั้งองค์กร เพื่อช่วยให้เกิดความสะดวกรวดเร็ว ลดปริมาณการใช้กระดาษ และลดพื้นที่ในการจัดเก็บเอกสาร แต่ระเบียบการจัดเก็บเอกสารของระบบราชการก่อนการทำลายนั้นมีระยะเวลานาน ทำให้เอกสารที่จำเป็นต้องจัดเก็บมีปริมาณมากขึ้น ซึ่งยังคงเป็นปัญหาในการสืบค้นข้อมูล/เอกสารต่างๆ เนื่องจากไม่สามารถค้นหาเอกสารตามเนื้อหา/คำสำคัญที่ผู้ใช้ต้องการได้ งานวิจัยชิ้นนี้ได้ศึกษาเปรียบเทียบเทคนิคการเลือกเนื้อหา/ คำสำคัญมาเป็นตัวแทนของเอกสาร ที่เหมาะสมสำหรับงานสารบรรณอิเล็กทรอนิกส์ให้โดยอัตโนมัติ เพื่อช่วยให้การสืบค้นข้อมูลสะดวก รวดเร็ว ถูกต้องตรงความต้องการของผู้ใช้

Abstract

Nowadays, electronic document is increasingly used to manage document in many workplaces. Not only in public organization, many government organizations are converting to electronic document management systems. Although it can reduce paper usage and storage space because all documents can be stored in the system, the searching of the bulk of electronic documents still hard. In this paper, we apply a technique of SEO (Search Engine Optimization) to develop the filtering of Thai electronic document system based on user-subjecting criteria. So it can make a convenience, accuracy and easy to use.

คำสำคัญ

คำสำคัญ, การเลือกคำสำคัญอัตโนมัติ, เทคนิคการเลือกคำสำคัญ, งานสารบรรณอิเล็กทรอนิกส์

1. บทนำ

เนื่องจากการบริหารจัดการเอกสารที่เข้าสู่ยุคเทคโนโลยีสารสนเทศ ทำให้รูปแบบการจัดเก็บเอกสารเปลี่ยนแปลงไปจากกระดาษสู่เอกสารอิเล็กทรอนิกส์ ซึ่งช่วยให้การบริหารข้อมูล/เอกสารต่างๆ มีประสิทธิภาพยิ่งขึ้น สามารถเก็บรวบรวมไว้ในที่เดียวกันได้จำนวนมาก ไม่กระจัดกระจาย ส่งผลให้ปริมาณเอกสารที่ถูกจัดเก็บมีแนวโน้มเพิ่มมากขึ้น อีกทั้งความสามารถในการจัดเก็บยาวนานขึ้น ปัญหาที่ตามมาคือการสืบค้นข้อมูล/ เอกสารทั้งในอดีตและปัจจุบันที่มีปริมาณมาก ให้รวดเร็วและตรงกับความต้องการของผู้ใช้งานที่หลากหลายนั้น ยังทำได้ยาก ผู้วิจัยจึงได้ศึกษาเปรียบเทียบวิธีการเลือกเนื้อหา/คำสำคัญในเอกสารด้วยรูปแบบต่างๆ เพื่อนำมาเป็นตัวแทนของเอกสารนั้น เพื่อใช้แก้ปัญหาการค้นคืนเอกสารที่ไม่ตรงตามความต้องการของผู้ใช้

จากงานของ Anette Hulth, Jussi Karlgren Anna Jonsson Henrik Bostrom และ Lars Asker (2553) ได้ศึกษาการกำหนดคำสำคัญให้กับเอกสารโดยเทียบเคียงกับการคัดเลือกคำสำคัญที่อาศัยมนุษย์เป็นผู้กำหนด ซึ่งทีมงานวิจัยเชื่อว่าการทำดัชนีคำสำคัญ (Keyword Indexing) เป็นส่วนสำคัญที่จะช่วยเสริมให้ระบบการค้นคืนเอกสารทำงานได้ดีขึ้น โดยทดลองกับการจัดเก็บเอกสารภายในห้องสมุด ซึ่งมีข้อบกพร่องสำหรับการค้นหา คือ ผู้ใช้จำเป็นจะต้องรู้ว่าหนังสือที่ตนต้องการหานั้นน่าจะถูกจัดเก็บไว้ในหมวดหมู่ไหน และบ่อยครั้งที่เอกสารเดียวกันถูกเก็บไว้หลายหมวดหมู่ จึงได้แก้ไขได้โดยการจัดทำ index ให้กับเอกสาร ซึ่งใช้การกำหนด Keyword ที่มีความสำคัญในการบ่งชี้หัวข้อและเนื้อสำคัญใน

เอกสาร และพบว่าการทำโดยมนุษย์ (Manual) นั้น อาจมีความผิดพลาดได้ เนื่องจากขาดความแม่นยำและการที่ใช้ผู้ประเมินไม่เหมือนกัน จึงนำเอา Domain Knowledge มาใช้ในการตัดคำสำคัญอัตโนมัติ เพื่อนำมาจัดทำดัชนี (Index) ให้กับเอกสารที่ถูกจัดเก็บในห้องสมุดของรัฐสภาสวีเดน มีการทดลองโดยรวบรวมเอกสารจากห้องสมุดของรัฐสภาสวีเดนที่ได้มีการจัดทำดัชนี แบบ Manual โดยผู้เชี่ยวชาญ และใช้ทั้ง Machine Learning Algorithm และ Morphological Pre-processing Tools ในการพัฒนา Domain ที่เหมาะสมที่สุด พบว่าสามารถใช้สร้างรายการคำสำคัญได้สอดคล้อง/ ถูกต้องตรงกับตัวอย่างที่ได้จากวิธีการแบบ Manual

ลิขสิทธิ์ ทำนอง ละอองดาว มาดี และวิวัฒน์ ศรีภูมิ (2550) ได้พัฒนาระบบสรุปข้อมูลสำหรับเอกสารภาษาไทยเพื่ออำนวยความสะดวกแก่ผู้อ่านเอกสารที่มีจำนวนหลายหน้าแต่มีเวลาที่จำกัด ซึ่งอาจทำให้ผู้อ่านรับทราบข้อมูลที่ไม่ครบถ้วน เนื่องจากอ่านไม่จบหรืออ่านด้วยความเร่งรีบ ทำให้ได้ใจความสำคัญของเอกสารไม่ครบถ้วน ซึ่งทีมผู้วิจัยได้เลือกคำสำคัญที่ปรากฏในเอกสารเพื่อมาเป็นตัวแทนในการจัดกลุ่มโดยใช้อัลกอริทึม K-means แบ่งข้อมูลออกเป็น K กลุ่ม จากนั้นคำนวณค่ากึ่งกลางของแต่ละกลุ่ม ข้อมูลจะถูกจัดเข้ากลุ่มที่อยู่ใกล้ที่สุด จากนั้นจะคำนวณค่าจุดกึ่งกลางใหม่ กระบวนการจะดำเนินไปจนกระทั่งข้อมูลทั้งหมดถูกจัดเข้ากลุ่มและไม่มีข้อมูลใดต้องเปลี่ยนกลุ่มอีก อัลกอริทึมจะหยุดทำงานและรายงานผลการจัดกลุ่ม จากผลการทดลองพบว่าการนำทฤษฎีทางด้านปัญญาประดิษฐ์เข้ามาช่วยในการสร้างใจความสำคัญ ทำให้ระบบมีการเรียนรู้จากตัวอย่างที่ถูกต้องและนำไปใช้ทำนายกับเอกสารจริง เพื่อช่วยเพิ่มประสิทธิภาพในการสรุปใจความสำคัญของเอกสาร

Gonenc Ercan และ Ilyas Cicekli (2548) นำเทคนิคการหาความสัมพันธ์ระหว่างคำในเอกสาร ซึ่งเป็นคุณสมบัติของ Lexical Chain มาในใช้พิจารณาเลือกคำสำคัญ ซึ่งส่วนมากคุณสมบัติดังกล่าวจะถูกใช้สำหรับการสรุปใจความสำคัญ (Text Summarization) มากกว่านำมาใช้แก้ปัญหาเรื่องการสกัดคำหรือวลีสำคัญ โดยทีมวิจัยได้ศึกษาเปรียบเทียบวิธีการสกัดคำสำคัญทั้งแบบที่ใช้และไม่ใช้ Lexical Chain ผลที่ได้คือ กระบวนการที่ใช้ Lexical Chain นั้นมีความแม่นยำขึ้น และสามารถสรุปเอาเนื้อหาที่สำคัญในเอกสารเพื่อนำมาเป็นตัวแทนของเอกสารได้อย่างชัดเจน

ถิรนันท์ ดำรงค์สอน และพีรวัฒน์ วัฒนพงศ์ (2545) ได้เสนอการสกัดคำหรือวลีสำคัญแบบอัตโนมัติโดยใช้โครงข่ายประสาทเทียม โดยสร้างแบบจำลองโครงข่ายประสาทเทียมจากการเรียนรู้จากเอกสารตัวอย่างที่มีวลีสำคัญของเขียนกำกับอยู่ และนำแบบจำลองที่ได้ไปใช้ในการพิจารณาเลือกคำสำคัญจากเอกสารที่เข้ามาใหม่ จากนั้นจัดเรียงลำดับความสำคัญของกลุ่มคำหรือวลีที่ได้มา โดยพิจารณาจากค่าความถี่และตำแหน่ง ซึ่งคำหรือวลีที่อยู่ในลำดับต้นๆ จะถูกเลือกมาเป็นคำสำคัญของเอกสาร จากผลการทดลองพบว่าการนำแบบจำลองโครงข่ายประสาทเทียมมาใช้สามารถเลือกคำสำคัญได้ถูกต้องแม่นยำมากยิ่งขึ้น

2. ทฤษฎีและหลักการ

2.1 ระบบค้นคืนข้อมูลสารสนเทศ (Information Retrieval: IR System)

มีเป้าหมายเพื่อค้นคืนเอกสารตามคำค้น (Query) ของผู้ใช้ (User) จากคลังเอกสารจำนวนมาก โดยระบบจะมีหน้าที่หลัก คือ ประมวลผลเอกสาร (Document Operations) สร้างตัวแทนเอกสารหรือดัชนี (Index or Document Representation) ประมวลผลคำค้น (Query Operations) สร้างตัวแทนคำค้น (Query Representation) และค้นคืนเอกสาร (Searching)

ส่วนสำคัญของงานวิจัยชิ้นนี้ คือ การสร้างตัวแทนเอกสารหรือดัชนี เพื่อใช้เปรียบเทียบความเหมือนของตัวแทนคำค้นกับตัวแทนเอกสารในการค้นคืนเอกสาร โดยประสิทธิภาพของระบบจะวัดที่ความถูกต้องของเอกสารที่ถูกนำกลับมา เป็นค่า Precision และ Recall

Precision คือ อัตราส่วนระหว่างจำนวนเอกสารที่ถูกค้นคืนกลับมาแล้วถูกต้องกับจำนวนเอกสารทั้งหมดที่ถูกค้นคืนมาได้

Recall คือ อัตราส่วนระหว่างจำนวนเอกสารที่ถูกค้นคืนกลับมาแล้วถูกต้องกับจำนวนเอกสารที่ถูกต้องทั้งหมดของระบบ

2.2 การสกัดคำหรือวลีสำคัญ (Keyword extraction)

เป็นการพิจารณาค่าความสำคัญของคำหรือวลีที่มีความเหมาะสมและสามารถนำมาเป็นตัวแทนของเอกสารแต่ละฉบับได้ โดยหลักการเบื้องต้นในการให้ความสำคัญของคำ คือ การให้น้ำหนักของคำ (Term Weighting) และกระบวนการวิเคราะห์เนื้อหาของเอกสาร ซึ่งวิธีการที่มีใช้อยู่ในปัจจุบันสามารถแบ่งออกเป็น 3 วิธีหลักๆ ได้ดังนี้

1) วิธีการทางสถิติอย่างง่าย (Simple Statistic Approach) เป็นการนำเอาสูตรทางสถิติมาหาค่าสำหรับประเมินหาคำสำคัญที่นำมาใช้เป็นตัวแทนของเอกสาร ตัวอย่างเช่น การหาค่าความน่าจะเป็นของการเกิดคำใดๆ (N-gram Model) การหาค่าความถี่ของคำในเอกสาร (Word Frequency) การหาค่าความถี่เอกสารระยะความถี่ผกผัน (TF-IDF: Term Frequency-Inverse Document Frequency) การหาลำดับความเหมือนที่ยาวที่สุดในอักขระ (LCS: Longest Common Substring) เป็นต้น ข้อดีของวิธีการทางสถิติ คือ ใช้งานง่าย เวลาในกระบวนการดำเนินการน้อย และให้ผลลัพธ์ที่ดี

2) วิธีการพิจารณาลักษณะทางภาษาของคำ (Linguistics Approaches) คือ วิธีการที่ให้ความสำคัญกับลักษณะทางภาษาของคำ หรือประโยค เช่น หน้าที่ของคำในประโยค โครงสร้างทางไวยกรณ์ รวมทั้งความหมายของคำ ตัวอย่างเช่น Lexical Theasuarus Ontology เป็นต้น

3) วิธีการด้านปัญญาประดิษฐ์ (Machine Learning Approaches) คือ การสร้างการเรียนรู้จากข้อมูลตัวอย่างให้กับระบบคอมพิวเตอร์จนสามารถตัดสินใจเลือกคำสำคัญที่เหมาะสมได้เอง ตัวอย่างเช่น สูตรของ Naïve Bayes การแทนค่าด้วย Vector (Vector Space Model) โครงข่ายประสาทเทียม (Neural Network)

3. ผลการดำเนินงาน

1. รวบรวมเอกสารประเภทต่างๆ ที่นำเข้าในระบบ เช่น เอกสารจากการ scan, pdf files และ doc files จากนั้นแปลงเอกสารให้อยู่ในรูปที่สามารถประมวลผลข้อความได้
2. ตัดคำในเอกสารและหาค่าคล้ายคลึงโดยเปรียบเทียบกับพจนานุกรมหรือคลังข้อมูล
3. ศึกษาเทคนิคการสกัดคำหรือวลีคำสำคัญ ตามแนวทางหลัก 3 ประเภท คือ
   - 3.1) Simple statistics approaches คือ word frequency, TF/IDF และ co-occurrence
   - 3.2) Linguistics approaches คือ lexical analysis, syntactic analysis และ ontology/domain
   - 3.3) Machine learning approaches คือ Naive Bayes และ support vector machine “SVM”
4. นำคำค้นมาเปรียบเทียบกับตารางดัชนี (inverted index) เพื่อเทียบเคียงคำค้นกับดัชนีที่ถูกเก็บไว้ในเอกสารต่างๆ โดยใช้เทคนิค VSM
5. ทดสอบค่าความแม่นยำและความถูกต้องของการค้นคืนเอกสารในแต่ละเทคนิค ด้วยค่า Precision และ Recall

4. เอกสารอ้างอิง

[1] ลิขสิทธิ์ทำนอง, ละอองดาว มาดี, วิวัฒน์ ศรีภูมิ การสรุปข้อมูลเอกสารโดยใช้การแบ่งกลุ่มข้อมูล สาขาวิชาเทคโนโลยีสารสนเทศและการสื่อสาร คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม

[2] Anette Hulth, Jussi Karlgren Anna Jonsson Henrik Bostrom และ Lars Asker, 2553, Automatic Keyword Extraction Using Domain Knowledge

[3] Todsanai Chumwatana, Kok Wai Wong, Hong Xie, 2552, An automatic indexing technique for Thai texts using frequent max substring, Eighth

International Symposium on Natural Language Processing

[4] An automatic document-based indexing system for meeting retrieval, Multimed Tools Appl, 2550 37:135–167 DOI 10.1007/s11042-007-0137-4

[5] Blaz Fortuna, Dunja Mladenic, Marko Grobelnik, 2548, Semi-automatic construction of topic ontology, department of knowledge technologies Jozed Stefan Institue, Slovenia.

[6] Gonenc Ercan, Ilyas Cicekli, 2550, Using Lexical Chains for Keyword Extraction, Department of computer engineering, Bilkent University, Turkey.

[7] Xiaoyuan Wu, Alvaro Bolivar, 2551, Keyword Extraction for Contextual advertisement

[8] Yi Wang, Hu Jin, 2552, Chinese keywords clustering based on SOM, Fourth International Conference on National Computation

[9] His-Cheng, Chiun-Chieh Hsu, 2548, Using topic keyword clusters for automatic document clustering, Third International Conference on Information Technology and Applications (ICITA)