การพัฒนาระบบต้นแบบการสกัดสารสนเทศสำหรับเอกสารราชการไทยแบบหลายโดเมนด้วยนิพจน์ปรกติบนฐานคุณลักษณะของเอกสาร

The Prototype System Development of Information Extraction for the Multi-Domain Thai Official Documents using Regular Expression based on the Document Features

อุษานาฎ เอื้ออภิสิทธิ์วงศ์, [1] ธีรพงษ์ สังข์ศรี [2]
Usanad Ua-apisitwong, [1] Teerapong Sungsri [2]

บทคัดย่อ

ในปัจจุบันระบบการจัดการเอกสารอิเล็กทรอนิกส์ (E-Document System) ได้เข้ามาอำนวยความสะดวกในการจัดเก็บและค้นหาเอกสารในรูปแบบไฟล์อิเล็กทรอนิกส์ในองค์กรต่างๆ อย่างไรก็ตามปัญหาที่เกิดขึ้นคือการจัดเก็บและค้นหาไฟล์ในระบบดังกล่าวนี้จะได้เพียงไฟล์ข้อมูลต้นฉบับทั้งไฟล์ โดยถ้าต้องการวิเคราะห์หรือคัดเลือกสารสนเทศจากเอกสารอิเล็กทรอนิกส์จะต้องอาศัยมนุษย์ในการทำงานซึ่งสิ้นเปลืองเวลาและแรงงานอย่างมาก งานวิจัยชิ้นนี้นำเสนอระบบต้นแบบการสกัดสารสนเทศสำหรับเอกสารราชการไทยแบบหลายโดเมนบนฐานคุณลักษณะของเอกสารร่วมกับการประยุกต์ใช้นิพจน์ปรกติเพื่อช่วยให้สามารถคัดเลือกสารสนเทศที่ต้องการจากเอกสารราชการไทยในรูปแบบไฟล์อิเล็กทรอนิกส์ได้อย่างอัตโนมัติ การดำเนินงานในการวิจัยครั้งนี้ประกอบด้วย 5 ขั้นตอนคือ คือ การศึกษาและเก็บรวบรวมข้อมูล การจัดเตรียมข้อมูล การจำแนกประเภทของเอกสารราชการ การสกัดสารสนเทศเอกสารราชการ และการพัฒนาระบบต้นแบบการสกัดสารสนเทศสำหรับเอกสารราชการ ผลการทดลองด้านภาพรวมของประสิทธิภาพความถูกต้องในการสกัดสารสนเทศเท่ากับร้อยละ 91.67 และผลประเมินความพึงพอใจจากผู้ใช้งานอยู่ในระดับพอใช้ (ค่าเฉลี่ยของคะแนนเท่ากับ 3.37)

คำสำคัญ: การสกัดสารสนเทศ เอกสารอิเล็กทรอนิกส์ นิพจน์ปรกติ

Abstract

Nowadays, an electronic document Management System (E-Document System) is a convenient software for collecting and searching on an electronic file format in an organization or university. However, a problem of this system is an operation of system can only operate with the raw data and human stills work for extracting or analyzing any information in documents. This research proposes the prototype of Information Extraction System of the Multi-Domain Official Documents based on the Document Features using a Regular Expression for automatic extracting any information of electronic documents. This research has 5 steps: 1) to study and collect all data. 2) to prepare data. 3) Document classification. 4) Information Extraction for the interested information and 5) to implement the Information Extraction System. The result of experiment in the extracting

[1] ผู้ช่วยศาสตราจารย์, [2] อาจารย์, โปรแกรมวิชาวิทยาการสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครราชสีมา อำเภอเมือง จังหวัดนครราชสีมา 30000

[1] Asst. Prof., [2] Lecturer, Informatics Program, Faculty of Science and Technology, Nakhonratchasima Rajabhat University, Mueang District, Nakhonratchasima 30000

accuracy is 91.67% and the overall of satisfaction with users is the fair level (the average of overall score is 3.37).

Keywords: Information Extraction, Electronic Document, Regular Expression

บทนำ

ในปัจจุบันปริมาณของข้อมูลมีการเพิ่มขึ้นอย่างรวดเร็วและมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่องพบว่าในปี ค.ศ. 2015 ปริมาณข้อมูลที่ถูกจัดเก็บไว้ในองค์กรมีขนาดสูงถึง 1600 เอกซะไบต์ (Exabytes) โดยมีข้อมูลประเภทไม่มีโครงสร้าง (Unstructured Data) อยู่ถึงประมาณร้อยละ 90 ของข้อมูลที่ใช้ในองค์กรทั้งหมด ข้อมูลจึงเปรียบเสมือนแหล่งทรัพยากรที่สำคัญสำหรับการสร้างองค์ความรู้และสารสนเทศเพื่อนำไปพัฒนาศักยภาพการทำงานของแต่ละองค์กร ซึ่งส่วนใหญ่ได้มีการจัดเก็บและประมวลผลข้อมูลในรูปแบบของเอกสารอิเล็กทรอนิกส์ (Electronic Document: E-Document) [1]

เอกสารราชการมีลักษณะเป็นข้อมูลแบบไม่มีโครงสร้างที่ถูกใช้โดยทั่วไปในองค์กรและหน่วยงานทางราชการ เช่น คำสั่ง ประกาศ บันทึกข้อความ รายงานการประชุม เป็นต้น โดยรูปแบบการสร้างเอกสารราชการแต่ละประเภทจะมีลักษณะการสร้างเป็นเอกสารอิเล็กทรอนิกส์ (Electronic Files) และแต่ละองค์กรจะมีการใช้ระบบสารสนเทศสำหรับการจัดเก็บและสืบค้นไฟล์เอกสารอิเล็กทรอนิกส์ต่างๆ เพื่อช่วยอำนวยความสะดวกในการทำงานของบุคลากรในองค์กร อย่างไรก็ตามจะพบว่าข้อจำกัดของระบบดังกล่าวคือระบบมีความสามารถในการจัดเก็บและสืบค้นเอกสารอิเล็กทรอนิกส์ที่เป็นรูปแบบไฟล์เท่านั้น แต่ไม่สามารถสืบคืนสารสนเทศที่มีอยู่ในเอกสารอิเล็กทรอนิกส์ได้ รวมทั้งยังไม่สามารถสกัดสารสนเทศที่ต้องการจากไฟล์เอกสารอิเล็กทรอนิกส์ได้ โดยงานส่วนดังกล่าวนี้ยังจำเป็นต้องใช้มนุษย์ในการทำงานอยู่ จากปัญหาดังกล่าวจึงได้มีงานวิจัยที่ประยุกต์ใช้เทคนิคการสกัดสารสนเทศเข้ามาช่วย เพื่อการคัดเลือกข้อมูลหรือสารสนเทศในเอกสารอิเล็กทรอนิกส์ เช่น การใช้กฎสำหรับการสกัดสารสนเทศร่วมกับโครงสร้างเอกสาร XML เพื่อสกัดสารสนเทศจากเอกสารราชการไทยซึ่งมีความซับซ้อนในการทำงานด้วยรูปแบบเอกสารต้องเป็นโครงสร้างเอกสารแบบ XML ทำให้ไม่สะดวกต่อผู้ใช้งาน [2] การใช้เทคนิคการสกัดสารสนเทศร่วมกับออนโทโลยีเพื่อสกัดสารสนเทศจากเอกสารธุรกรรมแบบอิเล็กทรอนิกส์โดยมีการออกแบบ ออนโทโลยีของคำที่ใช้บ่งบอกลักษณะหัวข้อที่ต้องการสกัดสารสนเทศทำให้ต้องมีการออกแบบโครงสร้างออนโทโลยีที่มีขนาดใหญ่สำหรับการสกัดสารสนเทศจากเอกสารหลายประเภท [3]

จากปัญหาดังกล่าว งานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อพัฒนาระบบต้นแบบสำหรับการสกัดสารสนเทศจากเอกสารราชการไทยโดยใช้รูปแบบนิพจน์ปรกติร่วมกับคุณลักษณะของเอกสารอิเล็กทรอนิกส์เพื่อสกัดสารสนเทศที่ต้องการสำหรับเอกสารราชการแต่ละประเภทที่มีรูปแบบเฉพาะลงในฐานข้อมูลอย่างอัตโนมัติ โดยประยุกต์ใช้คุณสมบัติเฉพาะของเอกสารราชการในการจำแนกประเภทเอกสารราชการและเทคนิคการสกัดสารสนเทศสำหรับการสกัดสารสนเทศที่ต้องการจากเอกสารราชการแต่ละประเภทซึ่งสามารถใช้งานกับเอกสารราชการในรูปแบบไฟล์เอกสารอิเล็กทรอนิกส์ทั้งรูปแบบ .doc .docx และ .pdf

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1. การสกัดสารสนเทศ [1]

การสกัดสารสนเทศ (Information Extraction) เป็นกระบวนการที่สำคัญอย่างหนึ่งในการทำงานด้านเหมืองข้อความ โดยเป็นขั้นตอนในการคัดเลือกหรือสืบค้นคำที่ต้องการออกจากเอกสารแบบไม่มีโครงสร้างได้แบบอัตโนมัติ วัตถุประสงค์สำคัญสำหรับการสกัดสารสนเทศ คือการแปลงเอกสารแบบไม่มีโครงสร้างให้อยู่ในรูปแบบเอกสารแบบมีโครงสร้างเพื่อให้สะดวกต่อการนำไปวิเคราะห์หรือจัดเก็บลงฐานข้อมูล ปัจจุบันมีการประยุกต์ใช้เทคนิคการสกัดสารสนเทศในงานด้านต่างๆ เช่น การระบุคำชื่อเฉพาะในเอกสาร (Relation Extraction) การระบุความสัมพันธ์ระหว่างสิ่งที่สนใจใน

เอกสาร (Name Entity Recognition) เป็นต้น สามารถแสดงการสกัดสารสนเทศ ได้ดัง Figure1

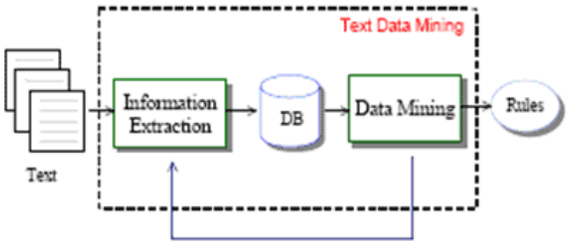

Figure 1 การสกัดสารสนเทศในการทำเหมืองข้อความ [9]

จาก Figure 1 แสดงขั้นตอนการสกัดสารสนเทศ ซึ่งเป็นกระบวนการทำงานย่อยของการทำเหมืองข้อความ เมื่อมีข้อมูลเอกสารข้อความที่มีลักษณะไม่มีโครงสร้างจะทำการคัดเลือกข้อมูลที่ต้องการเพื่อจัดเก็บในฐานข้อมูลที่มีรูปแบบข้อมูลแบบมีโครงสร้าง จากนั้นจะนำข้อมูลในฐานข้อมูลไปวิเคราะห์ต่อไป การสกัดสารสนเทศสามารถแบ่งวิธีการที่นิยมทำงานออกได้เป็น 3 วิธี ดังนี้

1.1 ฐานกฎ (Rule Learning based) วิธีการดังกล่าวแบ่งออกได้เป็น 3 เทคนิคคือ Dictionary based method เป็นเทคนิคที่อ้างอิงการสกัดสารสนเทศบนฐานความรู้ของดิชชันนารี เทคนิค Rule based method เป็นเทคนิคที่นำโครงสร้างของเอกสารที่มี มาวิเคราะห์เพื่อสร้างเป็นกฎการสกัดสารสนเทศในหัวข้อที่ต้องการ และเทคนิค Wrapper induction method เป็นเทคนิคที่มีการประยุกต์ใช้การเรียนรู้ของเครื่องเข้ามาช่วยให้การสกัดสารสนเทศ โดยมุ่งเน้นไปที่เอกสารแบบไม่มีโครงสร้างและเอกสารแบบกึ่งมีโครงสร้างเป็นหลัก

1.2 ฐานการเรียนรู้ของเครื่อง (Machine Learning based) วิธีการดังกล่าวเป็นวิธีการที่นำเทคนิคด้านการจำแนกชนิดข้อมูล (Classification) ที่มีการเรียนรู้แบบมีผู้สอน (Supervised Learning) เข้ามาช่วยในการสกัดสารสนเทศออกจากเอกสาร เช่น เทคนิค SVM (Support Vector Machine) เทคนิคต้นไม้ตัดสินใจ (Decision Tree)

1.3 ฐานโครงสร้างประโยค (Sequential based) วิธีการดังกล่าวเป็นการสกัดสารสนเทศด้วยการกำหนดหน้าที่ของคำ เช่น คำนามและคำกริยา โดยมีการประยุกต์เทคนิคการประมวลผลภาษา ธรรมชาติเข้ามาช่วยในการกำหนดหน้าที่ของคำ (Part of Speech Tagging: POS Tagging)

และขั้นตอนในการสกัดสารสนเทศสามารถแบ่งออกได้เป็นสองขั้นตอนคือ การเรียนรู้ (Training) และการสกัดสารสนเทศด้วยโมเดลการเรียนรู้ (Extraction) สามารถแสดงตัวอย่างกระบวนการสกัดสารสนเทศได้ดัง Figure 2

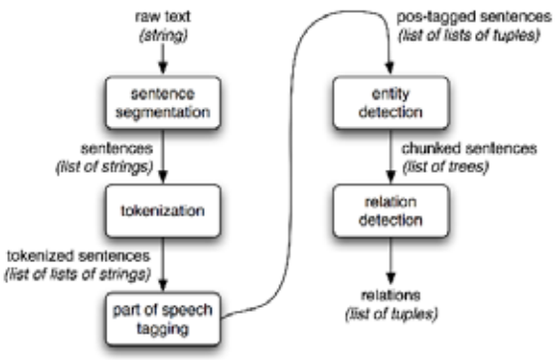

Figure 2 การสกัดสารสนเทศสำหรับความสัมพันธ์ [8]

จาก Figure 2 แสดงกระบวนการการสกัดสารสนเทศประเภทความสัมพันธ์ของชื่อคำนามโดยกระบวนการที่สำคัญสำหรับการสกัดสารสนเทศจากเอกสารคือเทคนิคด้านการประมวลผลภาษาธรรมชาติเริ่มตั้งแต่การตัดประโยคและตัดคำ (Sentence Segmentation and Tokenization) และการจำแนกหน้าที่ของคำ (Part of Speech Tagging)

2. งานวิจัยที่เกี่ยวข้อง

การสกัดสารสนเทศจากเอกสารราชการแบบหลายโดเมนบนฐานคุณลักษณะของเอกสาร ผู้วิจัยได้มีการทบทวนและศึกษางานวิจัยที่เกี่ยวข้อง ดังนี้

งานวิจัยในกลุ่มของการสกัดสารสนเทศในรูปแบบของชื่อเฉพาะ (Name Entity) เป็นงานวิจัยที่นำเสนอเทคนิคที่ใช้สำหรับการสกัดชื่อเฉพาะจากเอกสารเพื่อระบุคำนามที่ปรากฏขึ้นในเอกสารว่าสื่อความหมายในสิ่งใดเช่น ชื่อบุคคล ชื่อสถานที่ ชื่อแม่น้ำ เป็นต้น ซึ่งได้มีงานวิจัยที่กล่าวถึงเทคนิคที่ใช้สำหรับการสกัดหรือรู้จำชื่อเฉพาะของเอกสารโดยวิธีการที่นิยมใช้คือการสร้างกฎสำหรับการสกัดชื่อเฉพาะโดยอาศัย

ผู้เชี่ยวชาญหรือการศึกษารูปแบบของเอกสารอย่างเชี่ยวชาญ [4-6] สำหรับงานวิจัยที่มีการประยุกต์ใช้การสกัดสารสนเทศที่เป็นเอกสารภาษาไทยจะมีการใช้กฎที่สร้างขึ้นจากโครงสร้าง XML เพื่อนำมาสกัดสารสนเทศที่ต้องการ [10] และงานวิจัยที่นำบริบทโดยรอบของเอกสารมาช่วยในการสกัดสารสนเทศที่ต้องการจากเอกสารประเภทเว็บไซต์ที่มีลักษณะเป็นกึ่งโครงสร้าง (Semi-Structured Data) จากการศึกษาทำให้ผู้วิจัยพบว่าเอกสารราชการเป็นเอกสารประเภทไม่มีโครงสร้างที่มีจุดเด่นเฉพาะคือความคงที่ของรูปแบบเอกสารที่ชัดเจนจึงเหมาะสำหรับการสร้างกฎในการสกัดสารสนเทศที่ต้องการ แต่จำเป็นต้องมีการเพิ่มความยืดหยุ่นในการสกัดสารสนเทศแบบหลายโดเมนอีกด้วย

วิธีการดำเนินงานวิจัย

งานวิจัยนี้ได้แบ่งการดำเนินงานออกเป็น 5 ขั้นตอนคือ การศึกษาและเก็บรวบรวมข้อมูล การจัดเตรียมข้อมูล การจำแนกประเภทของเอกสารราชการ การสกัดสารสนเทศเอกสารราชการ และการพัฒนาระบบการสกัดสารสนเทศสำหรับเอกสารราชการไทย สำหรับดำเนินการวิจัยและมีวิธีดำเนินการวิจัยออกเป็นขั้นตอน ดังต่อไปนี้

1. การศึกษาและเก็บรวบรวมข้อมูล

งานวิจัยชิ้นนี้ได้มีการเก็บรวบรวมข้อมูลโดยใช้ข้อมูลเอกสารราชการของคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครราชสีมา จำนวนทั้งสิ้น 300 เอกสาร โดยมีรูปแบบเป็นเอกสารอิเล็กทรอนิกส์ประกอบด้วยทั้ง .doc .docx และ .pdf และสามารถแยกเอกสารราชการตามแต่ละประเภทเอกสาร กรอบแนวคิดในการดำเนินงานวิจัยประกอบด้วยขั้นตอนหลักคือ 1) ขั้นตอนการเก็บรวบรวมข้อมูลเอกสารราชการที่เป็นเอกสารราชการภาษาไทยเท่านั้น 2) ขั้นตอนการจัดเตรียมข้อมูลก่อนการประมวลผลเพื่อให้ข้อมูลมีความเหมาะสมก่อนนำไปสร้างกฎสำหรับการสกัดสารสนเทศและ 3) ขั้นตอนการประมวลผลประกอบด้วยการจำแนกประเภทเอกสารราชการและการสกัดสารสนเทศจากเอกสารราชการโดยใช้คุณลักษณะของเอกสารร่วมกับนิพจน์ปรกติ สามารถแสดงกรอบแนวคิดการทำงานได้ ดัง Figure 3

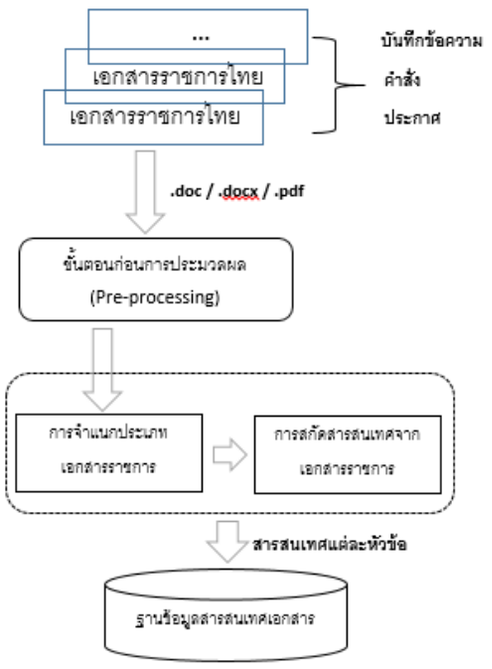

Figure 3 กรอบแนวคิดการทำวิจัย

2. การจัดเตรียมข้อมูล

เอกสารราชการทั้งหมดที่เก็บรวบรวมมาในรูปแบบไฟล์อิเล็กทรอนิกส์จะถูกนำมาผ่านขั้นตอนก่อนการประมวลผล (pre-processing) เพื่อให้ข้อมูลมีความเหมาะสมสำหรับนำไปทำงานในขั้นตอนการจำแนกประเภทเอกสารราชการ และขั้นตอนการสกัดสารสนเทศสำหรับเอกสารราชการ การจัดเตรียมข้อมูลเอกสารดังกล่าวจะทำการแปลงเอกสารอิเล็กทรอนิกส์ในรูปแบบไฟล์ต่างๆ (.doc .docx .pdf) ให้เป็นรูปแบบลำดับย่อหน้า (Paragraph) ของเอกสารโดยใช้ Apache POI library และมีการตัดย่อหน้าที่ว่างเปล่าออกไป จากนั้นประยุกต์ใช้โครงสร้างข้อมูลแบบแถวลำดับสองมิติมาช่วยในการจัดเก็บข้อมูลเอกสารที่สมบูรณ์ ดัง Figure 4

ย่อหน้า1	ย่อหน้า2	ย่อหน้า3	...	ย่อหน้า k

Figure 4 แถวลำดับสองมิติสำหรับโครงสร้างเอกสารราชการ

จาก Figure 4 แสดงการประยุกต์ใช้โครงสร้างข้อมูลแบบแถวลำดับมาช่วยในการจัดเก็บข้อมูลเอกสารราชการ โดยข้อมูลที่จัดเก็บในแต่ละดัชนีในแถวลำดับคือข้อมูลแต่ละย่อหน้าของเอกสารราชการหนึ่งเอกสาร และมีการตัดย่อหน้าที่ว่างเปล่าออกไป

3. การจำแนกประเภทของเอกสารราชการ

การจำแนกประเภทของเอกสารราชการในงานวิจัยนี้เป็นการประยุกต์ใช้คุณลักษณะของเอกสารคือลำดับของย่อหน้า โดยผู้วิจัยได้ตั้งสมมุติฐานสำหรับการจำแนกประเภทของเอกสารราชการคือ คำสำคัญที่แสดงประเภทของเอกสารราชการจะอยู่ในย่อหน้าแรกของไฟล์เอกสารราชการแบบอิเล็กทรอนิกส์เสมอ ดังนั้นในขั้นตอนการจำแนกประเภทของเอกสารราชการจะตรวจสอบคำสำคัญในดัชนีแรกของแถวลำดับที่จัดเก็บข้อมูลของเอกสารราชการด้วยนิพจน์ปรกติ (Regular Expression) ป้องกันข้อผิดพลาดการเทียบคำทั้งหมด

4. การสกัดสารสนเทศของเอกสารราชการ

งานวิจัยชิ้นนี้ได้แบ่งหัวข้อสารสนเทศที่ต้องการสกัดจากเอกสารราชการยกตัวอย่างเอกสารราชการประเภทบันทึกข้อความประกอบด้วยหัวข้อคือ 1) ชื่อประเภทเอกสาร 2) ส่วนราชการ 3) ที่และวันที่ 4) เรื่อง 5) เรียน 6) เนื้อหา และ 7) ชื่อผู้บันทึก โดยมีลักษณะเด่นของเอกสารราชการคือทุกเอกสารราชการในกลุ่มประเภทเดียวกันจะมีโครงสร้างที่แน่นอน ดังนั้นผู้วิจัยจึงประยุกต์ใช้เทคนิคการสกัดสารสนเทศด้วยกฎที่ถูกสร้างขึ้นจากโครงสร้างของเอกสารราชการแต่ละประเภทบนฐานคุณลักษณะของเอกสารประกอบด้วยลำดับย่อหน้า ขนาดของย่อหน้า ลักษณะคำในย่อหน้า และจำนวนย่อหน้าทั้งหมด ซึ่งกฎสำหรับการสกัดสารสนเทศจากเอกสารราชการแต่ละประเภททางผู้วิจัยได้กำหนดขึ้นมาตามหัวข้อสารสนเทศที่ต้องการสกัดในเอกสารราชการแต่ละประเภท เช่น บันทึกข้อความประกอบด้วยวันที่ เรียน เรื่อง เนื้อหา ลงชื่อ เป็นต้น สามารถแสดงกระบวนการสกัดสารสนเทศจากเอกสารราชการได้ ดัง Figure 5 และ Figure 6

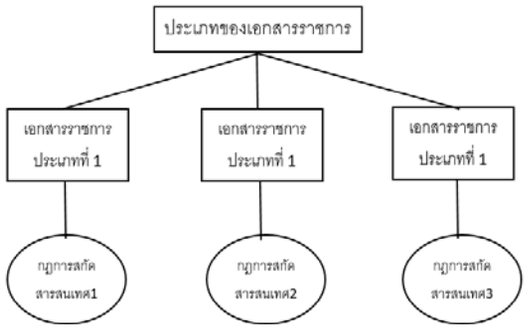

Figure 5 กระบวนการสกัดสารสนเทศจากเอกสารราชการ

จาก Figure 5 แสดงกระบวนการสกัดสารสนเทศสำหรับเอกสารราชการโดยใช้กฎที่ถูกสร้างขึ้นจากคุณลักษณะของเอกสารราชการในแต่ละประเภทของเอกสารราชการ การทำงานจะเริ่มจากการรู้ประเภทของเอกสารราชการที่จำแนกได้ในขั้นตอนการจำแนกประเภทเอกสารราชการ จากนั้นจะทำการเลือกกฎสำหรับการสกัดสารสนเทศที่ต้องการในแต่ละหัวสำหรับกลุ่มประเภทของเอกสารราชการ

ย่อหน้า	Required Information
1	ชื่อประเภทเอกสารราชการ (บันทึกข้อความ)
2	ส่วนราชการ
3	ที่ และ วันที่
4	เรื่อง
5	
6	เรียน
7	รายละเอียดเนื้อความ
…	
…	
k	ชื่อผู้บันทึก

Figure 6 ตัวอย่างโครงสร้างแถวลำดับสำหรับการสกัดสารสนเทศจากเอกสารราชการประเภทบันทึกข้อความ

จาก Figure 6 แสดงตัวอย่างโครงสร้างเอกสารที่ผ่านขั้นตอนก่อนการประมวลผลแล้วเพื่อนำมาสร้างกฎการสกัดสารสนเทศสำหรับเอกสารราชการประเภทบันทึกข้อความ โดยกฎการสกัดสารสนเทศสำหรับบันทึกข้อความ หัวข้อสารสนเทศที่ 1 ถึง 5 ประกอบด้วยชื่อประเภทเอกสาร ส่วนราชการ ที่และวันที่ เรื่อง เรียน จะใช้หมายเลขลำดับย่อหน้าของ

เอกสารที่คงที่ (ตำแหน่งดัชนีในแถวลำดับ) เช่น การสกัดสารสนเทศในหัวข้อส่วนราชการจะสกัดได้จากการดึงข้อมูลในตำแหน่งย่อหน้าที่ 2 ของเอกสาร(ตำแหน่งดัชนีที่ 1 ในแถวลำดับ)

สำหรับหัวข้อสารสนเทศที่ 6 และ 7 ประกอบด้วยเนื้อหาเอกสารและลงชื่อผู้บันทึกจะไม่สามารถใช้ลำดับหมายเลขย่อหน้าของเอกสารได้เนื่องด้วยเนื้อหาสามารถมีจำนวนย่อหน้าได้ไม่คงที่ซึ่งผู้วิจัยได้กำหนดกฎการสกัดสารสนเทศในหัวข้อดังกล่าวคือ ย่อหน้าทั้งหมดตั้งแต่ย่อหน้าลำดับที่ 7 (ตำแหน่งดัชนีที่ 6 ในแถวลำดับ) จากนั้นจึงใช้กฎของนิพจน์ปรกติที่ถูกสร้างขึ้นเพื่อแยกสกัดหัวข้อสารสนเทศของชื่อผู้บันทึกออกมาอีกครั้ง โดยนิพจน์ปรกติที่ใช้ในการสกัดสารสนเทศที่เป็นชื่อผู้บันทึกมีรูปแบบดังนี้

5. การพัฒนาระบบต้นแบบการสกัดสารสนเทศ

สำหรับในขั้นตอนการพัฒนาระบบต้นแบบจะถูกออกแบบให้ทำงานบน Window application ถูกพัฒนาขึ้นด้วยภาษาจาวา โดยเครื่องมือในการพัฒนาเป็นโปรแกรม Netbeans IDE 8.0 ระบบต้นแบบที่ถูกออกแบบมามีส่วนการทำงานหลักประกอบด้วย ส่วนแรกเป็นส่วนการนำเอกสารราชการเข้าสู่โปรแกรมซึ่งสามารถเลือกการนำเข้าได้ทั้งแบบไฟล์เดี่ยวและแบบหลายไฟล์ ส่วนที่สองเป็นส่วนการสกัดข้อมูลหรือสารสนเทศออกจากเอกสารราชการ (ใช้แนวคิดการสกัดสารสนเทศดังที่กล่าวมาในขั้นตอนก่อนหน้านี้) โดยส่วนการทำงานนี้จะทำการแยกชนิดของเอกสารราชการไทยก่อนจึงเริ่มสกัดสารสนเทศออกมาตามลำดับ ดัง Figure 7 และ 8

Figure 7 หน้าจอหลักการสกัดสารสนเทศจากเอกสารราชการ

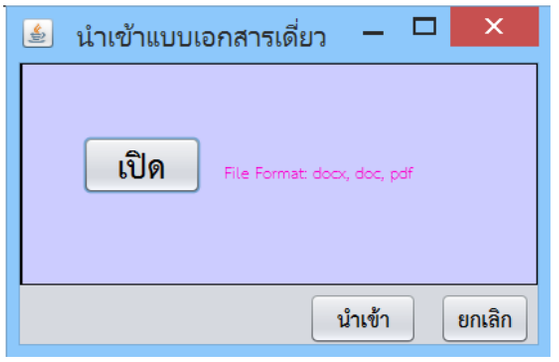

Figure 8 หน้าจอการนำข้อมูลเข้าแบบเอกสารเดี่ยว

จาก Figure 7 และ 8 แสดงตัวอย่างหน้าจอการทำงานของระบบต้นแบบการสกัดสารสนเทศจากเอกสารราชการ โดยลักษณะการทำงานจะเริ่มจากการเลือกไฟล์เอกสารราชการที่ต้องการจะสกัดสารสนเทศออกมา

ผลการทดลอง

งานวิจัยชิ้นนี้สามารถสรุปผลการทดลองได้ โดยแบ่งเป็น 2 ส่วน ดังนี้

1. ผลประเมินประสิทธิภาพของการสกัดสารสนเทศ

การประเมินประสิทธิภาพด้วยความถูกต้องของการสกัดสารสนเทศจากเอกสารราชการไทยแบบหลายโดเมน ดังสมการที่ 1 โดยการทดสอบจะใช้ไฟล์เอกสารราชการไทยจำนวน 300 เอกสารประกอบด้วยเอกสารราชการประเภทบันทึกข้อความ ประเภทคำสั่งและประเภทประกาศ

$$\text{ความถูกต้อง} = \frac{\text{จำนวนเอกสารที่สกัดถูกต้อง}}{\text{จำนวนเอกสารทั้งหมด}} \times 100 \quad (1)$$

โดยการวัดความถูกต้องในการสกัดสารสนเทศของเอกสารราชการแต่ละประเภทจะมีการบันทึกผลการทดลองแยกเป็นแต่ละหัวข้อและภาพรวมความถูกต้องในการสกัดสารสนเทศซึ่งมีการนำผลการสกัดสารสนเทศด้วยโปรแกรมไปเทียบกับผลการสกัดสารสนเทศของเจ้าหน้าที่ปฏิบัติงานเพื่อดูความถูกต้อง ดัง Table 1

Table 1 ตัวอย่างผลการวัดความถูกต้องการสกัดสารสนเทศจากเอกสารราชการประเภทบันทึกข้อความ

หัวข้อ	ความถูกต้อง (%)
ชื่อประเภทเอกสาร	100
ส่วนราชการ	100
ที่และวันที่	100
เรื่อง	96.67
เรียน	96.67
รายละเอียดเนื้อหา	91.67
ชื่อผู้บันทึก	93.33
ภาพรวมของเอกสาร	91.67

จาก Table 1 แสดงผลการทดลองการสกัดสารสนเทศจากเอกสารราชการด้วยตัวอย่างประเภทบันทึกข้อความซึ่งการสกัดสารสนเทศหัวข้อชื่อประเภทเอกสาร ส่วนราชการและ ที่และวันที่จะมีค่าความถูกต้องเท่ากับร้อยละ 100 โดยภาพรวมความถูกต้องในการสกัดสารสนเทศ (ต้องมีการสกัดสารสนเทศได้ถูกต้องทุกหัวข้อเมื่อนำไปเทียบกับผลการสกัดสารสนเทศของเจ้าหน้าที่) ได้ค่าความถูกต้องเท่ากับร้อยละ 91.67 จากผลการทดลองพบว่าเอกสารราชการแต่ละประเภทส่วนใหญ่มีรูปแบบแน่นอนในแต่ละหัวข้อทำให้สามารถใช้คุณลักษณะเฉพาะของเอกสารเพื่อสกัดสารสนเทศที่ต้องการออกมาได้ด้วยค่าความถูกต้องที่สูง อย่างไรก็ตามยังคงมีข้อผิดพลาดที่เกิดขึ้นในการสกัดสารสนเทศจากเอกสารที่สามารถเกิดขึ้นได้จากขั้นตอนการสร้างไฟล์เอกสารราชการ

2. ส่วนการประเมินความพึงพอใจของระบบ

การทดสอบระบบเพื่อหาความพึงพอใจในการใช้งานระบบ การประเมินความพอใจของระบบใช้การประเมินความพึงพอใจด้วยมาตราอันดับ 5 อันดับ และแปลความหมายของค่าเฉลี่ยเป็นแบบ 5 ช่วงคะแนน[7] ดัง Table 2 โดยมีการทดสอบกับผู้ใช้จำนวน 15 คน ประกอบด้วยเจ้าหน้าที่ที่เกี่ยวข้องกับการออกเอกสารราชการจำนวน 10 คนและอาจารย์แต่ละคณะจำนวน 5 คน

Table 2 เกณฑ์การแปลผลการประเมินความพึงพอใจ

เกณฑ์ค่าเฉลี่ย	การแปลความหมายของระดับคะแนน
4.51–5.00	เกณฑ์ดีมาก
3.51–4.50	เกณฑ์ดี
2.51–3.50	เกณฑ์พอใช้
1.51–2.50	เกณฑ์ปรับปรุง
1.00–1.50	เกณฑ์ไม่เหมาะสม

ซึ่งสถิติที่ใช้ในการวิเคราะห์ข้อมูลความพึงพอใจต่อระบบมี 2 สมการ คือ ค่าเฉลี่ยเลขคณิต (Mean) ดังสมการที่ 2 และส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) ดังสมการที่ 3

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (2)$$

$$S.D. = \sqrt{\frac{n\sum fx_i^2 - (\bar{x})^2}{n(n-1)}} \quad (3)$$

ผลการประเมินความพึงพอใจของผู้ใช้งาน มีความพึงพอใจต่อระบบการสกัดสารสนเทศสำหรับเอกสารราชการแบบหลายโดเมนในภาพรวมของระบบอยู่ในเกณฑ์พอใช้เท่ากับ 3.37 และมีส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.69 ดัง Table 3 โดยถ้าพิจารณาเฉพาะด้านอินเตอร์เฟสการทำงานผู้ใช้มีความพึงพอใจในระดับดี สำหรับด้านความสะดวกในการเลือกใช้เมนู ส่วนด้านการใช้ระบบผู้ใช้มีความพึงพอใจในความสะดวกสำหรับการสกัดข้อมูลจากเอกสารราชการอยู่ในเกณฑ์ดีเนื่องด้วยผู้ใช้เห็นว่าการนำรูปเอกสารแบบไฟล์ word หรือ pdf มีความสะดวกในการทำงาน อย่างไรก็ตามด้านความเหมาะสมในการนำไปใช้ประโยชน์มีระดับความพึงพอใจอยู่ในเกณฑ์พอใช้เนื่องด้วยผู้ใช้ให้ข้อคิดเห็นว่าข้อมูลที่ได้จากการสกัดสารสนเทศในเอกสารราชการยังคงไม่ได้ถูกนำไปใช้งานต่อสำหรับองค์กรโดยภาพรวม

Table 3 ผลประเมินความพึงพอใจของผู้ใช้งาน

รายละเอียด / ค่าเฉลี่ยความพอใจ	ระบบ	S.D.
ส่วนอินเตอร์เฟสการทำงาน		
ความสะดวกสำหรับการเลือกใช้เมนูการทำงาน	3.60	0.74
ความสวยงามของฟอร์มการทำงานต่างๆ ในระบบ	2.86	0.92
ส่วนการใช้งานระบบ		
ความสะดวกในการสกัดข้อมูลจากเอกสาร	3.70	0.59
ความเหมาะสมในการนำไปใช้ประโยชน์ในหน่วยงาน	3.33	0.49
ค่าเฉลี่ยโดยรวมของระบบ	3.37	0.69

สรุปผลและวิจารณ์ผลการทดลอง

งานวิจัยนี้เป็นการนำเสนอแนวคิดสำหรับการสกัดสารสนเทศจากเอกสารราชการไทยที่ประกอบด้วยหลายโดเมนและพัฒนาเป็นระบบต้นแบบสำหรับการสกัดสารสนเทศจากเอกสารราชการประเภทบันทึกข้อความ คำสั่งและประกาศ โดยขั้นตอนการสกัดสารสนเทศของเอกสารราชการเริ่มจากการจำแนกประเภทของเอกสารราชการแต่ละประเภทด้วยสมมุติฐานที่ตั้งไว้คือ หัวข้อในย่อหน้าแรกของเอกสารเป็นสิ่งที่บอกถึงประเภทของเอกสารราชการ จากนั้นในขั้นตอนการสกัดสารสนเทศจะใช้กฎสำหรับการสกัดที่ถูกต้องขึ้นในแต่ละประเภทของเอกสารด้วยคุณลักษณะของเอกสารและประยุกต์ร่วมกับนิพจน์ปรกติเพื่อลดปัญหาข้อผิดพลาดในการสกัดสารสนเทศที่เกิดจากการพิมพ์ข้อความผิด ผลการทดลองด้านความถูกต้องในการสกัดสารสนเทศจากเอกสารราชการแบบหลายโดเมนมีค่าความถูกต้องเท่ากับร้อยละ 91.67 ซึ่งอยู่ในระดับที่สูง อย่างไรก็ตามข้อผิดพลาดที่ทำให้ระบบต้นแบบไม่สามารถสกัดสารสนเทศที่ถูกต้องออกมาได้เกิดจากลักษณะการสร้างไฟล์เอกสารราชการแบบอิเล็กทรอนิกส์ที่อาจมีการขึ้นย่อหน้าใหม่ในหัวข้อเดิม ปัญหาอีกประการหนึ่งคือกฎการสกัดสารสนเทศที่สร้างขึ้นด้วยนิพจน์ปรกติที่สามารถเกิดข้อผิดพลาดขึ้นได้ง่าย และผลด้านประเมินความพึงพอใจโดยภาพรวมของระบบเท่ากับ 3.37 อยู่ในเกณฑ์พอใช้ ข้อควรปรับปรุงในการพัฒนาต่อคือ การสร้างกฎสำหรับการสกัดสารสนเทศจากเอกสารจำเป็นต้องใช้ผู้เชี่ยวชาญทางด้านคอมพิวเตอร์ในการสร้างทำให้เกิดข้อจำกัดในการนำไปประยุกต์ใช้กับเอกสารประเภทอื่นเช่น เอกสารประกันคุณภาพการศึกษาที่มีรูปแบบแน่นอนเช่นกัน ดังนั้นในอนาคตจะมีการพัฒนาโมดูลในการสร้างกฎการสกัดจากการเอกสารตัวอย่างแบบอัตโนมัติต่อไป

กิตติกรรมประกาศ

งานวิจัยชิ้นนี้สำเร็จลุล่วงไปได้ด้วยดี เนื่องด้วยความอนุเคราะห์หลายๆ ด้านของทางมหาวิทยาลัยราชภัฏนครราชสีมา ขอขอบพระคุณท่านผู้ช่วยศาสตราจารย์รุจิรา อุปวานิช คณบดีคณะวิทยาศาสตร์และเทคโนโลยีที่ให้ความอนุเคราะห์ผู้วิจัยเก็บข้อมูลตัวอย่างเอกสารราชการจากทางคณะวิทยาศาสตร์และเทคโนโลยี ขอขอบพระคุณท่านคณบดีอาจารย์เจ้าหน้าที่ทุกท่านจากคณะต่างๆ ที่อนุเคราะห์ช่วยทดลองใช้ระบบเพื่อมีผลความพึงพอใจของผู้ใช้

เอกสารอ้างอิง

1. สุพัฒนวรี ทิพย์เจริญ. การใช้เทคนิคการระบุชื่อเฉพาะในการสกัดสารสนเทศเพื่อใช้สร้างฐานความรู้เฉพาะด้าน. วารสารวิชาการมหาวิทยาลัยฟาร์อีสเทอร์น 2552; 2(2):51-56.
2. อุษานาฏ เอื้ออภิสิทธิ์วงศ์. ระบบต้นแบบการสร้างฐานข้อมูลสารสนเทศบนฐานกฎแบบอัตโนมัติสำหรับเอกสารราชการไทย. ใน: รายงานต่อเนื่องการประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 11 (NCCIT2015). คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยพระจอมเกล้าพระนครเหนือ. กรุงเทพฯ; 2558. หน้า 583-588.
3. ธนดล วัฒนะสุทธิวงศ์. การติดตามความรู้บนเว็บแบบหลายมุมมองโดยใช้เทคโนโลยีสกัดข้อสนเทศและออนโทโลจี. ใน: The Second National Conference on Electronic Business. มหาวิทยาลัยเกษตรศาสตร์. กรุงเทพฯ; 2546.

4. สุฤดี ฉัตรไตรมงคล. การรู้จำและจำแนกประเภทของชื่อเฉพาะภาษาไทย. วิทยานิพนธ์ ปริญญาอักษรศาสตรมหาบัณฑิต มหาวิทยาลัยเกษตรศาสตร์. 2548.
5. อมรเทพ พวงไธสง. การสกัดนิพจน์ระบุนามสำหรับเหมืองข้อมูลข่าวภาษาไทยตามกิจกรรมของบุคคล. วิทยานิพนธ์ ปริญญาวิทยาศาสตรมหาบัณฑิต มหาวิทยาลัยธรรมศาสตร์. 2553.
6. ณัฐดาพร เลิศชีวะ. การรู้จำชื่อเฉพาะ: การศึกษาชื่อผลิตภัณฑ์ในข่าวเศรษฐกิต. วิทยานิพนธ์ ปริญญาวิทยาศาสตรมหาบัณฑิต มหาวิทยาลัยธรรมศาสตร์. 2553.

ทัศนวรรณ แก้วใส และ สุพจน์ นิตย์สุวัฒน์, "ระบบแนะนำภาพยนตร์ด้วยเทคนิคตัวกรองเชิงร่วมมือร่วมกับวิธีเคมีน (Movies Recommender System using Collaborative Filtering and K-Means)," *The 5th National Conference on Computing ang Information Technology (NCCIT 2009)*, พฤษภาคม 2552, หน้า 502-507.

7. Bird S, Klein E and Loper E. Natural Language Processing with Python [Internet]. Natural Language Toolkit [http://nltk.org/]; 2014 [cited 2016 Aug 5]. Available from: http://www.nltk.org/book/ch07.html.
8. Mooney RJ and Nahm UY. Text Mining with Information Extraction. In: Proceeding of the 4th International MIDP Colloquium. South Africa; 2005. P.141-160.
9. Ua-apisitwong U and Sungsri T. The Automatic Rule-based Information Extraction for Unknown Official Document. International Journal of Intelligent Information Processing (IJIIP) 2013; 4(4):46-51.