

ปัจจุบันเอกสารเป็นแหล่งเก็บความรู้ที่สำคัญของมนุษย์ ซึ่งความรู้ต่าง ๆ ได้กระจายอยู่ในเอกสารและมีการเติบโตอย่างรวดเร็ว แต่ผู้ใช้จะต้องเสียเวลาในการอ่านเอกสารจำนวนมากเพื่อค้นหาความรู้และนำความรู้ นั้นไปใช้ประโยชน์ ในงานวิจัยนี้จึงมีวัตถุประสงค์ในการพัฒนาเทคนิคสำหรับสกัดความรู้และเข้าถึงความรู้จาก เอกสารภาษาไทย การสกัดความรู้หมายถึงการสกัดเฉพาะใจความสำคัญในขอบเขตที่ผู้ใช้สนใจ โดยในงานวิจัย นี้จะสนใจเฉพาะความรู้ที่เกี่ยวกับคุณสมบัติของอ็อบเจกต์ เนื่องจากคุณสมบัติของอ็อบเจกต์สามารถใช้ระบุถึง อ็อบเจกต์ที่สนใจได้ ตัวอย่างเช่น ในโดเมนการเกษตร การสืบค้นชื่อของแมลงศัตรูพืชสามารถสืบค้นได้จาก คุณสมบัติของแมลง เป็นต้น ปัญหาการสกัดความรู้จากเอกสารภาษาไทยนั้นจะมีปัญหาที่เกี่ยวกับการ ประมวลผลภาษา ได้แก่ การใช้สิ่งอ้างอิงร่วมแบบสูญรูปและการละคำ ซึ่งทำให้องค์ประกอบของความรู้ที่ ปรากฏในรูปประโยคมีความไม่สมบูรณ์ และในการเข้าถึงความรู้นั้นจะมีปัญหาเกี่ยวกับความไม่ชัดเจนของค่า คุณสมบัติในการสืบค้นอ็อบเจกต์ เนื่องจากค่าคุณสมบัติอาจจะอยู่ในรูปของค่าที่เป็นนามบัญญัติและส่วนขยาย เช่น “เขียวเข้ม”, “แดงอ่อน” เป็นต้น วิทยานิพนธ์นี้จึงเสนอเทคนิคสกัดความรู้เกี่ยวกับคุณสมบัติของอ็อบเจกต์ รวมทั้งการเข้าถึงความรู้ด้วยคอมพิวเตอร์ โดยในส่วนสกัดความรู้ การประมวลผลภาษาระดับบทความสำหรับการ ประมวลผลสิ่งอ้างอิงร่วมแบบสูญรูปและการประมวลผลการละคำ เป็นสิ่งจำเป็นเพื่อทำรูปประโยคให้ สมบูรณ์และใช้เทคนิคการเปรียบเทียบแม่แบบเพื่อทำการสกัดความรู้ และได้มีการใช้ทฤษฎีฟิชชีและการวัด ความคล้ายคุณสมบัติแบบฟิชชี ในส่วนของการเข้าถึงความรู้

การวัดผลของงานวิจัยนี้ประกอบด้วยการวัดผลใน 4 ส่วน ได้แก่ อัลกอริทึมประมวลผลการใช้สิ่ง อ้างอิงร่วมแบบสูญรูป อัลกอริทึมประมวลผลการละคำ การสกัดความรู้ และ การสืบค้นอ็อบเจกต์ โดยเน้น โดเมนการเกษตรซึ่งมีขนาดคลังเอกสารประมาณ 2,000 ประโยค โดยผลค่าความถูกต้องของการประมวลผลสิ่ง อ้างอิงร่วมแบบสูญรูปอยู่ที่ 82.14% ค่าความระลึก 71.42% และค่า F-measure 76.40% ผลค่าความถูกต้องของ การประมวลผลการละคำอยู่ที่ 96.96% ค่าความระลึก 96.96% และค่า F-measure 96.96% ผลค่าความถูกต้อง ของเทคนิคสกัดความรู้อยู่ที่ 88.88% ค่าความระลึก 47.50% และค่า F-measure 61.53% และผลค่าความถูกต้อง ของเทคนิคการสืบค้นอ็อบเจกต์อยู่ที่ 81.81% ค่าความระลึก 90.90% และค่า F-measure 86.11%

Nowadays, documents are the important storage of human knowledge. The knowledge is spread through the document and also grows rapidly. However, the users have to take a lot of time to read the document to find the knowledge and utilize it. This thesis has the objective to develop the technique to extract the knowledge from Thai texts and also to access the knowledge. The knowledge extraction is the tool to assist the user extract the salient of the document in the interested domain. This research attended to the knowledge about the properties of the object. The properties of the object are able to identify the interested object such as to search the pest by given the pest properties. To extract the knowledge from Thai texts, there are problems involved with the linguistic phenomena such as the zero anaphora and textual ellipsis that make the components of knowledge incomplete in the sentence. To access the knowledge, there are the problems about the obscurity of the property value such as "dark green", "light red". Accordingly, this research proposed the technique for extracting the knowledge about the object properties from Thai texts and the technique for accessing the knowledge by computer. For extracting the knowledge, the discourse processing for solving the zero anaphora and textual ellipsis are needed in order to make the sentence complete and template matching is used for extracting the knowledge from the sentence. For accessing knowledge about the property of the object, the fuzzy theory and fuzzy property similarity measurement are utilized.

Four major components of the system, which are the zero anaphora resolution, the textual ellipsis resolution, the knowledge extraction module, and the query module, were evaluated with 2,000 sentences from corpus in agricultural domain. In the zero anaphora resolution, the precision, the recall and F-measurement are 82.14%, 71.42% and 76.40% respectively. In the textual ellipsis resolution, those are 96.96%, 96.96% and 96.96%. In the knowledge extraction module, those are 88.88%, 47.50% and 61.53%. And finally, in the query module, those are 81.81%, 90.90% and 86.11%.