ประสิทธิภาพการสกัดคำและการค้นคืนข้อมูลมะเร็งจากเว็บไซต์ภาษาไทย

สุภาพร วีระพันธ์ยานนท์[1*] และ พยุง มีสัจ[2]

บทคัดย่อ

งานวิจัยนี้นำเสนอเทคนิคการสกัดคำเกี่ยวกับมะเร็งและการค้นคืนข้อมูลมะเร็งจากเว็บไซต์ภาษาไทย ในการสกัดคำ ผู้วิจัยได้นำเสนอเทคนิค TH-OnSeg ซึ่งประยุกต์ใช้อัลกอริทึมเล็กซ์โตร่วมกับพจนานุกรมมะเร็งและออนโทโลยีมะเร็งเพื่อสกัดคำเกี่ยวกับมะเร็ง ซึ่งใช้เป็นดัชนีเอกสารของเว็บไซต์มะเร็ง ในการวิจัยได้ทดลองเปรียบเทียบกับการสกัดคำโดยอัลกอริทึมเล็กซ์โตร่วมกับพจนานุกรมสื่ออิเล็กทรอนิกส์ไทย ผลการวิจัยพบว่าเทคนิค TH-OnSeg มีประสิทธิภาพในการสกัดคำได้ดีกว่าทั้งประเภทของคำที่ไม่รู้จัก คำที่รู้จักและคำกำกวม นอกจากนี้ผู้วิจัยได้นำเสนอเทคนิคเว็บเชิงความหมายร่วมกับเอ็นแกรมส์ในการค้นคืนข้อมูลมะเร็ง ในการวิจัยได้ทดลองเปรียบเทียบเทคนิคที่นำเสนอกับวิธีค้นคืนข้อมูลโดยทั่วไปในฐานข้อมูล และการใช้เฉพาะเทคนิคเว็บเชิงความหมาย ผลการวิจัยพบว่าการใช้เทคนิคเว็บเชิงความหมายร่วมกับเอ็นแกรมส์ ให้ผลลัพธ์จำนวนข้อมูลเว็บไซต์มะเร็งได้มากที่สุด และมีค่าความครบถ้วนสูงสุดไม่ต่ำกว่า 0.9 ในทุกการทดลองทั้งกรณีของคำสำคัญที่สะกดถูกและคำสำคัญที่สะกดผิด

คำสำคัญ: มะเร็ง, การสกัดคำ, การค้นคืนข้อมูล, ออนโทโลยี, TH-OnSeg

รับพิจารณา: 27 กันยายน 2561
แก้ไข: 5 กุมภาพันธ์ 2563
ตอบรับ: 11 มีนาคม 2563

[1] นักศึกษาปริญญาเอก ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
[2] รองศาสตราจารย์ ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
[*] ผู้นิพนธ์ประสานงาน โทร. +668 2687 8444 อีเมล: wee.supaporn@gmail.com

Technical Education Journal : King Mongkut's University of Technology North Bangkok
Vol. 11 No. 3 September – December, 2020

Effectiveness of Word Extraction and Information Retrieval on Cancer from Thai Website

Supaporn Weeraphanyanont[1*] and Phayung Meesad[2]

Abstract

This article proposes word extraction and cancer information retrieval from the Thai website. For word extraction, TH-OnSeg is proposed as a words segmentation based on LexTo algorithm with cancer dictionary and cancer oncology. TH-Onseg is used to extract cancer related words to be used as document indexing for cancer websites. The experiments were conducted by comparing the word extraction with LexTo words segment algorithm based on Thai electronic dictionary. The results show that the TH-OnSeg technique has higher efficiency; it can extract more words than LexTo for unknown words, known words, and ambiguous words. In addition, we propose a semantic web-based technique combined with n-grams for cancer information retrieval. The experiments were conducted by comparing the proposed technique with information retrieval methods in database. The results show that the use of semantic web techniques combined with N-gram for cancer information retrieval yields the highest number of cancer websites. The highest recall is not less than 0.9 in all experimental cases of both misspellings and misspellings.

Keywords: cancer, word segmentation, information retrieval, ontology, TH-OnSeg

Received: September 27, 2018
Revised: February 5, 2020
Accepted: March 11, 2020

[1] Ph.D. Student, Department of Information Technology, Faculty of Information Technology, King Mongkut's University of Technology North Bangkok
[2] Associate Professor, Department of Information Technology, Faculty of Information Technology, King Mongkut's University of Technology North Bangkok
[3] Corresponding Author Tel. +668 2687 8444 e-Mail: wee.supaporn@gmail.com

1. บทนำ

ภาวะปัจจุบันมะเร็งเป็นหนึ่งในโรคที่มีผู้ป่วยมากที่สุด โดยเห็นได้จากผลการสำรวจของสำนักงานสถิติแห่งชาติ [1] มะเร็งมีผลกระทบต่อบุคคลเป็นจำนวนมากทั้งผู้ป่วยมะเร็งและญาติ ดังนั้นจึงมีผู้สนใจหาข้อมูลเกี่ยวกับมะเร็งเพื่อการดูแลสุขภาพ หรือรักษาพยาบาลเป็นอย่างมาก ซึ่งหนทางหนึ่งที่เป็นที่นิยมในการค้นหาข้อมูลเกี่ยวกับมะเร็งนั้น คือ อินเทอร์เน็ต แต่ข้อมูลที่เผยแพร่นั้นมาจากหลายแหล่งซึ่งมีปัญหาในด้านความน่าเชื่อถือของข้อมูล

ดังนั้นจึงต้องพิจารณาถึงความน่าเชื่อถือของเว็บไซต์ในต่างประเทศจะมีมูลนิธิฮอน (Health On the Net Foundation: HON) [2] ให้การรับรองเนื้อหาเว็บไซต์เกี่ยวกับการแพทย์และสุขภาพ แต่ในประเทศไทยยังไม่มีหน่วยงานให้การรับรอง สำหรับการพิจารณาความน่าเชื่อถือของเว็บไซต์โดยเฉพาะเกี่ยวกับมะเร็ง จึงพิจารณาจากแหล่งที่มาของข้อมูล ซึ่งหากมาจากแพทย์หรือหน่วยงานด้านการรักษาพยาบาล ก็จะมีความน่าเชื่อถือกว่าบริษัท ร้านค้าหรือตัวแทนขายประกัน [3]

นอกจากนั้น เนื่องจากข้อมูลบนเว็บไซต์มีจำนวนมาก การที่จะพิจารณาว่าเนื้อหาของเว็บไซต์มุ่งเน้นนำเสนอในเรื่องใดเป็นการยากที่จะระบุได้ ดังนั้นจึงได้มีการนำเทคโนโลยีสารสนเทศมาช่วยในการพิจารณาคำสำคัญที่เป็นตัวแทนของเว็บไซต์ ซึ่งวิธีการที่ได้รับความสนใจ คือ การสร้างดัชนีเอกสาร แต่จำเป็นที่จะต้องใช้เทคนิคในการสกัดคำที่มีประสิทธิภาพเพื่อให้ได้คำที่สื่อความหมายชัดเจนเข้าสู่กระบวนการสร้างดัชนีเอกสารเพื่อหาตัวแทนคำสำคัญของเว็บไซต์ โดยที่หากเป็นเว็บไซต์ที่มีลักษณะเฉพาะด้าน อย่างเช่นมะเร็ง จะมีความซับซ้อนมากยิ่งขึ้น อีกทั้งข้อมูลที่เป็นภาษาไทยจะมีความยุ่งยาก ขึ้นเนื่องจากไม่มีการแบ่งวรรคคำเหมือนภาษาอังกฤษ

ส่วนการค้นหาข้อมูลมะเร็งในเว็บไซต์ที่เป็นภาษาไทยก็ยังพบปัญหา เช่น การเลือกใช้คำค้นเกี่ยวกับประเภทของมะเร็ง การสะกดคำค้นที่ไม่ถูกต้อง ซึ่งส่งผลให้ได้ผลลัพธ์ที่ไม่ครบถ้วนหรือไม่ตรงตามที่ต้องการ เป็นต้น

จากปัญหาและข้อจำกัดดังกล่าว งานวิจัยนี้จึงได้นำเสนอเทคนิค TH-OnSeg ซึ่งเป็นแนวทางเพิ่มประสิทธิภาพในการสกัดคำเกี่ยวกับมะเร็ง โดยได้ใช้อัลกอริทึมเล็กซ์โต (LexTO) และพจนานุกรมมะเร็ง (CancerDic+) ร่วมกับออนโทโลยีมะเร็ง (Cancer Ontology) และการเพิ่มประสิทธิภาพในการค้นหาข้อมูลเกี่ยวกับมะเร็งในเว็บไซต์ด้วยเทคนิคเว็บเชิงความหมายร่วมกับเอ็นแกรมส์ (N-Gram)

2. วัตถุประสงค์งานวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอการเพิ่มประสิทธิภาพการสกัดคำเกี่ยวกับมะเร็ง โดยใช้เทคนิค TH-OnSeg และการเพิ่มประสิทธิภาพค้นคืนข้อมูลมะเร็งจากเว็บไซต์ภาษาไทย โดยใช้เทคนิคเว็บเชิงความหมายร่วมกับเอ็นแกรมส์

3. วรรณกรรมที่เกี่ยวข้อง

3.1 ความน่าเชื่อถือของเว็บไซต์

ความน่าเชื่อถือของเว็บไซต์เป็นปัจจัยที่สำคัญต่อการนำข้อมูลบนเว็บไซต์ไปใช้ประโยชน์ สำหรับความน่าเชื่อถือสามารถพิจารณาเป็น 2 แนวทางหลัก [4] คือ 1) ความไว้วางใจ (Trustworthiness) และ 2) ความชำนาญหรือคุณภาพ (Expertise or Quality) สำหรับข้อมูลทางการแพทย์และสุขภาพนั้น ในต่างประเทศมีองค์การรับรองความน่าเชื่อถือของเนื้อหาภายในเว็บไซต์ คือ มูลนิธิฮอน (Health On the Net Foundation: HON) ซึ่งเว็บไซต์ที่ได้รับการรับรองความน่าเชื่อถือจะได้สัญลักษณ์ HONcode กำกับในเว็บไซต์ แต่ในประเทศไทยไม่มีหน่วยงานในลักษณะดังกล่าว ดังนั้นงานวิจัยนี้จึงพิจารณาความน่าเชื่อถือของเว็บไซต์ จากแนวคิดของงานวิจัย [3] ผู้วิจัยได้ทำการจำแนกเว็บไซต์ที่มีเนื้อหาเกี่ยวกับมะเร็งโดยใช้การเรียนรู้ของเครื่อง ซึ่งจะพิจารณาจากแหล่งที่มาของข้อมูล โดยแหล่งข้อมูลที่น่าเชื่อถือมาจาก แพทย์ ผู้เชี่ยวชาญ หรือ หน่วยงานทางสาธารณสุข ส่วนแหล่งที่ไม่เชื่อถือมาจากการโฆษณาสินค้าของบริษัทหรือการขายประกันสุขภาพ

3.2 การสร้างดัชนีเอกสาร (Document Indexing)

เมื่อได้เว็บไซต์มะเร็งที่น่าเชื่อถือแล้ว จะสร้างดัชนีเอกสารของเว็บไซต์มะเร็ง เพื่อทำให้การค้นหาข้อมูลมะเร็งตรงประเด็นและรวดเร็วขึ้น ซึ่งการสร้างดัชนีเอกสารนั้นเป็นการแปลงเอกสารจากภาษาธรรมชาติให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถประมวลผลได้ โดยจะสร้างตัวแทนข้อมูลเนื้อหาในรูปแบบเวกเตอร์ของน้ำหนักคำ (Term Weighting) เพื่อนำมาสร้างดัชนี [5] สำหรับการคำนวณค่าน้ำหนักให้แก่ดัชนีใช้วิธีการ TF-

IDF Weighting (Term Frequency-Inverse Document Frequency) [6] ซึ่งการสร้างดัชนีเอกสารที่ดีนั้นจำเป็นต้องอาศัยวิธีการสกัดคำที่ความเหมาะสม เพื่อให้ได้คำที่ชัดเจนและสื่อความหมายเหมาะสมกับการนำไป ใช้งาน

3.3 การสกัดคำ (Word Extraction)

การสกัดคำเป็นการวิเคราะห์คำที่เป็นตัวอักษรออกจากข้อมูลข่าวสารต่าง ๆ โดยผลลัพธ์ที่ได้สามารถนำไปใช้ประโยชน์ [7] เช่น การทำเหมืองข้อความ เป็นต้น สำหรับการสกัดคำสามารถแบ่งเป็น 2 ขั้นตอน [8] คือ 1) การตัดคำ (Word Segmentation) จะแยกคำในเอกสารแต่ละคำออกจากกัน โดยยังต้องมีความหมายถูกต้องสมบูรณ์ สำหรับวิธีที่นิยมใช้ในการตัดคำภาษาไทย ได้แก่ การใช้กฎ (Rule-based Approach) การใช้พจนานุกรม (Dictionary-based Approach) และการใช้คลังข้อความ (Corpus-based Approach) [9] และ 2) การกำจัดคำหยุด (Stop Word) เป็นการตัดคำที่ไม่สื่อความหมายสำคัญออกจากเอกสาร ซึ่งจะใช้ฐานข้อมูลคำศัพท์ที่ไม่มีความหมายเป็นส่วนประกอบ [10]จากนั้นจะนำคำที่ได้ไปหาค่าน้ำหนักคำที่เป็นตัวแทนในการสร้างดัชนีเอกสาร

สำหรับงานวิจัยเกี่ยวกับการสกัดคำ ได้แก่ งานวิจัย [6], [11], [12], [13] และ [14] ซึ่งพบว่า พจนานุกรมเป็นส่วนประกอบสำคัญต่อการสกัดคำ ดังนั้นงานวิจัยนี้จึงได้ใช้พจนานุกรมโดยเฉพาะคำศัพท์ด้านมะเร็งเพื่อใช้เป็นองค์ประกอบในการสกัดคำ

3.4 ออนโทโลยี (Ontology)

การให้ความหมายหรือประกาศคุณลักษณะที่ชัดเจนของคำศัพท์ เพื่ออธิบายความหมายจากหลากหลายแนวคิดให้เป็นไปในแนวทางเดียวกัน [15] โดยที่โครงสร้างและความสัมพันธ์ระหว่างคำศัพท์เหล่านั้นประกอบด้วย คลาส (Class) คุณสมบัติของคลาส (Properties) ที่มีการถ่ายทอดพฤติกรรมบางอย่างต่อกันไปทั้งแบบเป็นลำดับชั้น (Taxonomic relation) และแบบไม่เป็นลำดับชั้น (Non-taxonomic relation) เช่น CauseBy, ResultOf เป็นต้น [16]

ตัวอย่างงานวิจัยที่ใช้ออนโทโลยี เช่น งานวิจัย [17], [18] ซึ่งจะเห็นได้ว่าออนโทโลยีมีประโยชน์ต่อการสกัดคำโดยเฉพาะการแก้ปัญหาคำกำกวม ดังนั้นงานวิจัยนี้จึงได้นำออนโทโลยีมาใช้เพื่อการเพิ่มประสิทธิภาพความถูกต้องในการสกัดคำเกี่ยวกับมะเร็งที่เป็นภาษาไทยด้วย

3.5 เว็บเชิงความหมาย (Semantic Web)

เว็บเชิงความหมายเป็นมุมมองการพัฒนาเว็บไซต์ ยุค 3.0 ตามมาตรฐานที่กำหนดโดย W3C [19] หมายถึงเว็บไซต์ที่กำหนดโครงสร้างในการอธิบายความหมายของข้อมูล เป็นมาตรฐานเดียวกัน ทำให้คอมพิวเตอร์เข้าใจโครงสร้างและเข้าถึงเว็บไซต์ที่มีความสัมพันธ์กันได้ ซึ่งผู้ใช้งานจะสามารถสร้างคำศัพท์ฐานความรู้ออนโทโลยีเพื่อใช้สืบค้นข้อมูลเว็บไซต์ตามความหมาย โดยใช้เทคนิคต่าง ๆ เช่น RDF และ OWL เป็นต้น [20]

ตัวอย่างงานวิจัย เช่น การวิจัย [21] และ [22] โดยในงานวิจัยที่ได้นำเสนอนี้จะใช้เว็บเชิงความหมายร่วมกับออนโทโลยีมะเร็งที่ได้พัฒนาไว้เพื่อเพิ่มประสิทธิภาพในการค้นคืนข้อมูลจากเว็บไซต์

3.6 เอ็นแกรมส์ (N-Gram)

เอ็นแกรมส์เป็นแบบจำลองความน่าจะเป็นของลำดับตัวอักษรที่จะรวมเป็นคำ (Character Sequence) หรือคำที่จะรวมเป็นประโยค (Word Sequence) โดยความน่าจะเป็นของตัวอักษรและคำนั้นจะได้จากคลังคำศัพท์ที่กำหนดไว้ [23] และ [24] โดยที่ตัวอักษรหรือคำที่ใช้จะมีหน่วยเป็น แกรมส์ และใช้การเปรียบเทียบจับคู่ตัวอักษรซึ่งขนาดของแกรมส์มีตั้งแต่ 1 ถึง N โดยถ้าขนาดของแกรมส์มากขึ้นก็จะเพิ่มความซับซ้อนในการประมวลผล หลังจากนั้นจะเปรียบเทียบความคล้ายคลึงกันของข้อมูล [25]

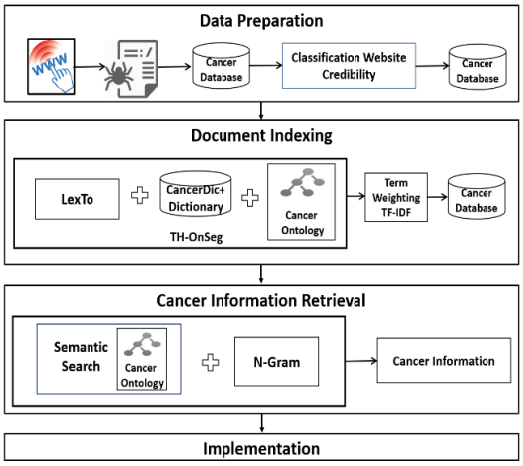
ตัวอย่างงานวิจัยเช่น งานวิจัย [26], [27] และ [28] เมื่อพิจารณาตัวอย่างงานวิจัยที่นำเสนอจะเห็นได้ว่าเทคนิคเอ็นแกรมส์สามารถงานใช้ได้กับหลายภาษา และช่วยเพิ่มประสิทธิภาพการค้นคืนข้อมูลได้ดียิ่งขึ้น ดังนั้นในงานวิจัยนี้จึงได้มีแนวคิดนำเอ็นแกรมส์มาใช้ร่วมกับเทคโนโลยีเว็บเชิงความหมายและออนโทโลยีมะเร็ง เพื่อการค้นคืนข้อมูลด้านมะเร็งที่มีประสิทธิภาพมากขึ้น

4. การวิจัย

แนวทางการวิจัยที่นำเสนอ แสดงได้ดังรูปที่ 1 ซึ่งแบ่งการวิจัยออกเป็น 4 ส่วนดังนี้

4.1 การเตรียมข้อมูล (Data Preparation)

ใช้โปรแกรมคลอเลอร์รวบรวมข้อมูลจากเว็บไซต์ที่สืบค้นด้วยคำสำคัญที่เกี่ยวกับมะเร็ง เช่น อาการของมะเร็ง อาหารสำหรับผู้ป่วยมะเร็ง สมุนไพรรักษามะเร็ง เป็นต้น จากนั้นจะจำแนกประเภทความน่าเชื่อถือ โดยการพิจารณาจากแหล่งที่มาของเว็บไซต์ เพื่อบันทึกจัดเก็บลงฐานข้อมูล



รูปที่ 1 แนวทางการวิจัย

4.2 การสร้างดัชนีเอกสาร (Document Indexing)

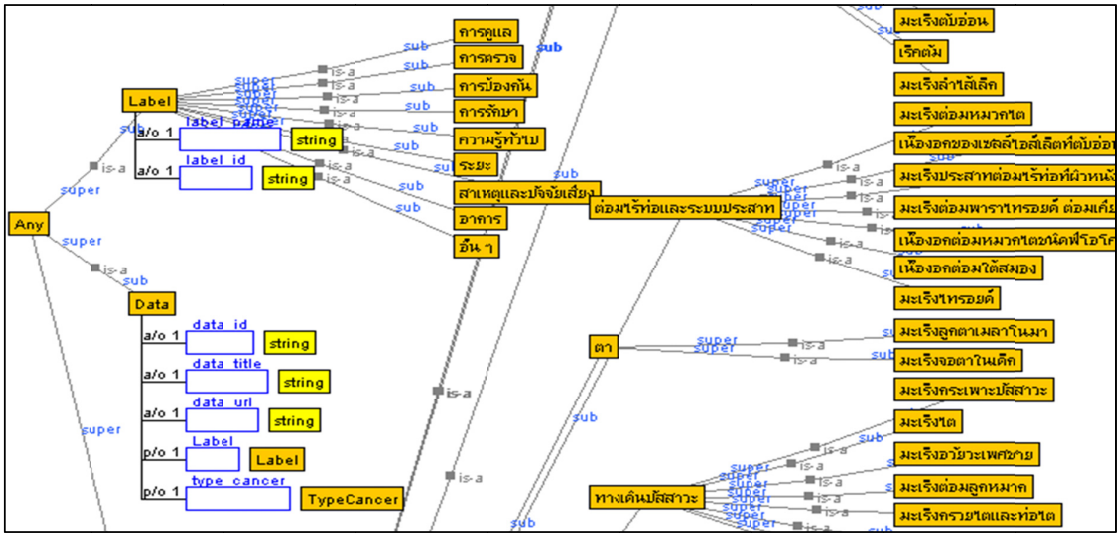
การสร้างดัชนีเอกสารของเว็บไซต์มะเร็ง จำเป็นต้องใช้วิธีการสกัดคำที่มีความเหมาะสม ซึ่งงานวิจัยนี้ได้นำเสนอเทคนิค TH-OnSeg โดยนำ 3 วิธีการมารวมกัน ได้แก่

4.2.1 อัลกอริทึมเล็กซ์โต (Thai Lexeme Tokenizer: LexTo) [29] เป็นโปรแกรมตัดคำจากข้อความภาษาไทย โดยเทียบกับคำที่ยาวที่สุดที่พบในพจนานุกรม

4.2.2 พจนานุกรมมะเร็ง (Cancer Dictionary Plus: CancerDic+) [3] เป็นพจนานุกรมคำศัพท์เฉพาะด้านมะเร็งที่ผู้วิจัยได้พัฒนาขึ้น

4.2.3 ออนโทโลยีมะเร็ง (Cancer Ontology) ได้ออกแบบออนโทโลยีมะเร็งตามคำแนะนำของผู้เชี่ยวชาญ โดยมีรูปแบบจากบนลงล่าง (Top Down) และกำหนดความสัมพันธ์แบบลำดับชั้น (Is-a hierarchy) โดยที่คลาสหลักได้แก่ อวัยวะต่าง ๆ 15 อวัยวะ คลาสย่อย ได้แก่ มะเร็งที่เกิดขึ้นในอวัยวะนั้น ๆ ซึ่งอ้างอิงข้อมูลจากสถาบันมะเร็งแห่งชาติประเทศสหรัฐอเมริกา [30]

เมื่อกำหนดข้อมูลมะเร็งตามอวัยวะที่เกิดแล้ว ได้ออกแบบฐานข้อมูลและออนโทโลยีมะเร็ง จากนั้นจึงใช้โปรแกรมสนับสนุนการพัฒนาออนโทโลยี Hozo [31] สร้างออนโทโลยีมะเร็ง แสดงได้ดังรูปที่ 2 ซึ่งเมื่อสกัดคำได้แล้ว ก็จะคำนวณค่าน้ำหนักของคำในเอกสารด้วยวิธีการ TF-IDF Weighting เพื่อให้ได้คำสำคัญที่เป็นตัวแทนของเอกสาร ที่นำไปใช้สำหรับการค้นคืนข้อมูลเว็บไซต์มะเร็ง



รูปที่ 2 ออนโทโลยีมะเร็งที่สร้างด้วย HoZo Tool

4.3 การค้นคืนข้อมูลมะเร็ง (Cancer Data Retrieval)

ได้มุ่งเน้นค้นคืนข้อมูลจากเว็บไซต์ภาษาไทย ซึ่งได้นำเสนอ 2 วิธีการผสมกัน คือ เว็บเชิงความหมายที่ใช้ออนโทโลยีมะเร็งร่วมกับเอ็นแกรมส์ เพื่อที่จะสามารถค้นคืนข้อมูลจากเว็บไซต์มะเร็งได้อย่างถูกต้อง รวดเร็วถึงแม้ผู้ใช้จะป้อนคำสำคัญผิด

4.4 การประยุกต์ใช้ (Implement)

ขั้นตอนสุดท้าย การสร้างต้นแบบเว็บท่าด้านมะเร็ง (Cancer Web Portal) โดยใช้เทคนิคการสืบค้นเชิงความหมายร่วมกับออนโทโลยีมะเร็งและเอ็นแกรมส์

5. การออกแบบการทดลอง

จากแนวทางการวิจัยที่นำเสนอนั้น ได้ออกแบบการทดลองโดยแบ่งเป็นข้อมูลที่ใช้ในการทดลอง และ การวัดประสิทธิภาพของวิธีการที่นำเสนอ ซึ่งมีรายละเอียดดังนี้

5.1 ข้อมูลที่ใช้ในการทดลอง

เมื่อกำหนดคำค้นหาเกี่ยวกับมะเร็งแล้ว จึงได้ใช้โปรแกรมคลอเลอร์เว็บไซต์จากอินเทอร์เน็ต ซึ่งมีจำนวนทั้งสิ้น 640 เว็บไซต์ และพิจารณาความน่าเชื่อถือของเว็บไซต์ตามหลักการแหล่งที่มาของข้อมูล โดยที่ข้อมูลจากเว็บไซต์ที่น่าเชื่อถือจะถูกนำมาใช้ในการวิจัย

5.2 การวัดประสิทธิภาพวิธีการที่นำเสนอ

วัดประสิทธิภาพของการสกัดคำ และการค้นคืนข้อมูลเว็บไซต์ โดยพิจารณาจากค่าความครบถ้วน (Recall) ซึ่งแสดงดังสมการที่ 1 [32]

$$recall = \frac{|\{relevant\ pages\} \cap \{retrieved\ pages\}|}{|\{relevant\ pages\}|} \quad (1)$$

เมื่อ

กำหนดให้ $relevant\ pages$ หมายถึง หน้าเอกสารที่ได้รับจากการค้นคืน ประกอบด้วยคำที่มีความหมายเกี่ยวข้องและสอดคล้องกับคำค้น

กำหนดให้ $retrieved\ pages$ หมายถึง หน้าเอกสารทั้งหมดที่ได้รับจากการสืบค้น การเปรียบเทียบกับปัจจัยต่าง ๆ สามารถทำได้ดังนี้

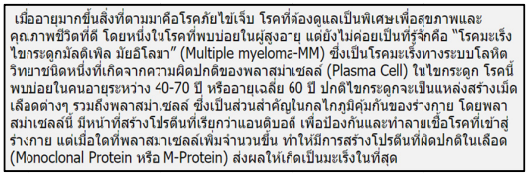
5.2.1 การวัดประสิทธิภาพการสกัดคำจากเว็บไซต์ ได้เปรียบเทียบเทคนิค TH-OnSeg กับการใช้อัลกอริทึมเล็กซ์โตร่วมกับพจนานุกรมสื่ออิเล็กทรอนิกส์ไทย (LEXiTRON) [33]

5.2.2 การวัดประสิทธิภาพการค้นคืนข้อมูลในเว็บไซต์ ได้เปรียบเทียบ 3 ลักษณะ คือ 1) การค้นคืนแบบทั่วไป 2) การค้นคืนแบบเว็บเชิงความหมาย และ 3) การค้นคืนแบบเว็บเชิงความหมายร่วมกับเทคนิคเอ็นแกรมส์ โดยกำหนดให้ N = 2 เพื่อลดความซับซ้อนในการประมวลผล

6. สรุปผลการวิจัย

6.1 ประสิทธิภาพการสกัดคำจากเว็บไซต์

ผู้วิจัยได้นำข้อมูลเว็บไซต์เกี่ยวกับมะเร็งที่น่าเชื่อถือมาทดลองวัดประสิทธิภาพการสกัดคำ ดังเช่นในรูปที่ 3


เมื่ออายุมากขึ้นสิ่งที่ตามมาคือโรคภัยไข้เจ็บ โรคที่ต้องดูแลเป็นพิเศษเพื่อสุขภาพและคุณภาพชีวิตที่ดี โดยหนึ่งในโรคที่พบบ่อยในผู้สูงอายุ แต่ยังไม่ค่อยเป็นที่รู้จักคือ "โรคมะเร็งไขกระดูกมัลติเพิล มัยอีโลมา" (Multiple myeloma-MM) ซึ่งเป็นโรคมะเร็งทางระบบโลหิตวิทยาชนิดหนึ่งที่เกิดจากความผิดปกติของพลาสมาเซลล์ (Plasma Cell) ในไขกระดูก โรคนี้พบบ่อยในคนอายุระหว่าง 40-70 ปี หรืออายุเฉลี่ย 60 ปี ปกติไขกระดูกจะเป็นแหล่งสร้างเม็ดเลือดต่างๆ รวมถึงพลาสม่าเซลล์ ซึ่งเป็นส่วนสำคัญในกลไกภูมิคุ้มกันของร่างกาย โดยพลาสม่าเซลล์นี้ มีหน้าที่สร้างโปรตีนที่เรียกว่าแอนติบอดี เพื่อป้องกันและทำลายเชื้อโรคที่เข้าสู่ร่างกาย แต่เมื่อใดที่พลาสมาเซลล์เพิ่มจำนวนขึ้น ทำให้มีการสร้างโปรตีนที่ผิดปกติในเลือด (Monoclonal Protein หรือ M-Protein) ส่งผลให้เกิดเป็นมะเร็งในที่สุด

รูปที่ 3 ตัวอย่างเว็บไซต์ที่มีข้อความเกี่ยวกับมะเร็ง [34]

การเปรียบเทียบประสิทธิภาพเทคนิคที่นำเสนอแสดงผลลัพธ์ได้ดังตารางที่ 1

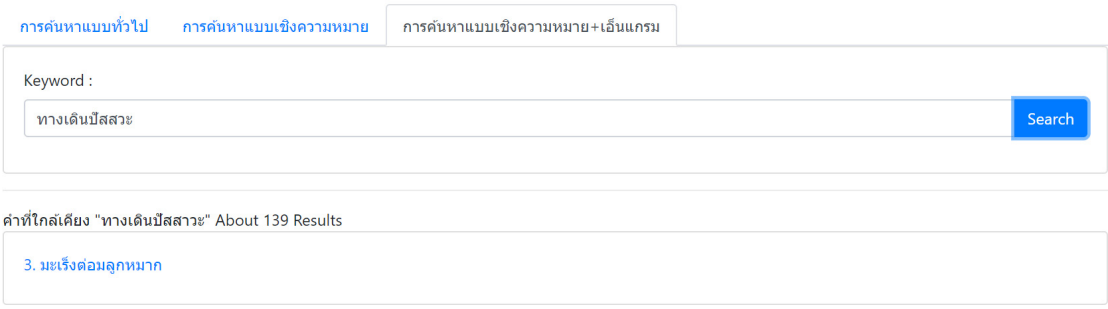
ตารางที่ 1 ผลลัพธ์การสกัดคำ

วิธีการสกัดคำ	LexTo & LEXiTRON	Th-OnSeg
คำที่ไม่รู้จัก	19	5
คำที่รู้จัก	622	647
คำกำกวม	11	ไม่ปรากฏ

จากตารางที่ 1 เห็นได้ว่าวิธี Th-OnSeg ให้ผลลัพธ์ดีกว่า ซึ่งเป็นผลมาจากการใช้พจนานุกรมมะเร็ง และออนโทโลยีมะเร็ง ที่สามารถสกัดคำที่สอดคล้องกับเนื้อหาของเว็บไซต์ได้มากกว่า

6.2 ประสิทธิภาพการค้นคืนข้อมูลในเว็บไซต์

หลังจากที่สกัดคำด้วยวิธี Th-OnSeg แล้วจึงสร้างดัชนีเอกสารของเว็บไซต์มะเร็ง จากนั้นผู้วิจัยได้สร้างต้นแบบเว็บท่าด้านมะเร็งเพื่อทดสอบประสิทธิภาพการค้นคืนข้อมูลในเว็บไซต์ ซึ่งพิจารณาตัวอย่างการค้นคืนข้อมูลเว็บไซต์ เมื่อใช้คำว่า "ทางเดินปัสสวะ" ซึ่งเป็นคำที่สะกดผิด ผลลัพธ์ที่ได้มีเฉพาะการค้นหาแบบเชิงความหมายร่วมกับเอ็นแกรมส์เท่านั้นที่สามารถค้นหาข้อมูลเว็บไซต์จากคำใกล้เคียงได้ แสดงดังรูปที่ 4



รูปที่ 4 ผลลัพธ์การค้นหาแบบเชิงความหมายร่วมกับเอ็นแกรมด้วยคำว่า “ทางเดินปัสสวะ”

สำหรับการเปรียบเทียบประสิทธิภาพค่าความครบถ้วนในการค้นคืนข้อมูลเว็บไซต์มะเร็งนั้นได้กำหนดคำสำคัญที่ใช้ในการค้นคืนข้อมูล 2 กลุ่ม คือ คำที่สะกดถูกและคำที่สะกดผิด ดังตารางที่ 2 โดยที่ค่าความครบถ้วนในการค้นคืนข้อมูล แสดงในตารางที่ 3 และตารางที่ 4

ตารางที่ 2 คำสำคัญที่ใช้ในการค้นคืนข้อมูล

คำที่สะกดถูก	คำที่สะกดผิด
นรีเวช	นารีเวช
โรคคุชชิง	โรคกุดชิง
เจิมเซลล์ตัวอ่อน	เจิมเซลล์ตัวอ่อล
ยูวิง	ยูวิว
โลหิตวิทยา	โลหิดวิดยา

ตารางที่ 3 การเปรียบเทียบค่าความครบถ้วนของการค้นคืนข้อมูลสำหรับคำสำคัญที่สะกดถูก

คำสำคัญที่สะกดถูก		นรีเวช	โรคคุชชิง	เจิมเซลล์ตัวอ่อน	ยูวิง	โลหิตวิทยา
Normal	ค้นพบ	24	0	0	1	30
	Recall	0.95	0	0	1	1
Semantic	ค้นพบ	141	0	0	1	35
	Recall	1	0	0	1	1
Semantic and N- Gram	ค้นพบ	141	2	12	3	35
	Recall	1	1	0.91	1	1

ตารางที่ 4 การเปรียบเทียบค่าความครบถ้วนของการค้นคืนข้อมูลสำหรับคำสำคัญที่สะกดผิด

คำสำคัญที่สะกดผิด		นารีเวช	โรคกุดชิง	เจิมเซลล์ตัวอ่อล	ยูวิว	โลหิดวิดยา
Normal	ค้นพบ	1	0	0	0	0
	Recall	1	0	0	0	0
Semantic	ค้นพบ	1	0	0	0	0
	Recall	1	0	0	0	0
Semantic and N- Gram	ค้นพบ	141	2	12	3	35
	Recall	1	1	0.91	1	1

จะเห็นได้ว่าการค้นหาแบบเชิงความหมายร่วมกับเอ็นแกรมส์ให้ผลลัพธ์ที่ดีที่สุดทั้ง 2 กรณี ในทุกตัวอย่างคำสำคัญ เนื่องจากการค้นหาแบบเชิงความหมายเป็นการค้นคืนข้อมูลโดยผ่านออนโทโลยีมะเร็งที่ออกแบบไว้ ส่วนเอ็นแกรมส์จะช่วยแบ่งคำสำคัญซึ่งแก้ปัญหาการสะกดคำผิดได้

7. สรุป อภิปรายผล และข้อเสนอแนะ

งานวิจัยนี้ได้นำเสนอเทคนิค TH-OnSeg และการค้นคืนข้อมูลมะเร็งจากเว็บไซต์ โดยเทคนิค TH-OnSeg เป็นการสกัดคำเกี่ยวกับมะเร็งโดยใช้อัลกอริทึมเล็กซ์โตร่วมกับพจนานุกรมมะเร็งและออนโทโลยีมะเร็ง ส่วนการค้นคืนข้อมูลมะเร็งได้ใช้เทคนิคเว็บเชิงความหมายร่วมกับเอ็นแกรมส์ ซึ่งผลการวิจัยพบว่า เทคนิค TH-

OnSeg สามารสกัดคำได้ดีในทุกประเภทของคำ และการใช้เทคนิคเว็บเชิงความหมายร่วมกับเอ็นแกรมส์สามารถค้นคืนข้อมูลเว็บไซต์มะเร็งโดยมีค่าความครบถ้วนสูงสุดไม่ต่ำกว่า 0.9 สำหรับงานวิจัยในอนาคตผู้วิจัยมีเป้าหมายที่จะจัดกลุ่มข้อมูลมะเร็งด้วยเทคนิคเหมืองข้อความ และพัฒนาเว็บท่าที่สามารถแสดงการจัดกลุ่มข้อมูลมะเร็งในลักษณะกราฟวิชวลไลเซชั่น

8. เอกสารอ้างอิง

[1] National Statistics Office, "Death rates per 100,000 population by leading cause of death and sex, whole kingdom : 2007 - 2014," 2014. [Online]. Available: http://service.nso.go.th/nso/web/statseries/statseries09.html. [Accessed 10 January 2017]. (in Thai)

[2] Health On the Net Foundation, "HONcode certification," 1995. [Online]. Available: https://www.hon.ch. [Accessed 10 January 2017].

[3] S. Kertkid, A. Aun-Anan and P. Meesad, "Classification of Reliable Content on Cancer Thai Website using CancerDic+," *Journal of information science and technology (JIST),* vol. 5, no. 2, pp. 34-43, 2015. (in Thai)

[4] W. Teppabutre, "Credibility-Enhancing Communication Framework for Rajabhat Universities' Website," *SDU Research Journal,* vol. 9, no. 2, pp. 187-198, 2013. (in Thai)

[5] S. Ummeepien and S. Thaiprayoon, "Web Plagiarism Monitoring System Using Informative Text Selection Method," *Information Technology Journal,* vol. 12, no. 2, pp. 1-9, 2016. (in Thai)

[6] N. Chirawichitchai, "The application of modeling to automatic classification of Thai document," *JIT,* vol. 2013, no. 1, pp. 141-149, 2013. (in Thai)

[7] J. R. Quinlan, "Induction of Decision Trees in Machine Learning," 1986. [Online]. Available: http://hunch.net/~coms-4771/quinlan.pdf. [Accessed 10 January 2017].

[8] N. Chirawichitchai, P. Sanguansa and P. Meesad, "Effective Automatic Thai Document Categorization," *NIDA Development Journal,* vol. 51, no. 3, pp. 187-205, 2011. (in Thai)

[9] S. Tepdang, Improving Thai word segmentation with named entity recognition, Pathum Thani: Thammasat University, 2018. (in Thai)

[10] S. Bualerng and W. Songpan, "Question Classification for Answer Searching Using Semantic Web and Data Mining," in *The 10th National Conference on Computing and Information Technology*, Phuket, 2014. (in Thai)

[11] W. Aroonmanakun, "Collocation and Thai Word Segmentation," in *SNLP-oriental COCOSDA 2002*, Thailand, 2002. (in Thai)

[12] P. Urathamakun and K. Runapongsa, "Improved Rule-Based and New Dictionary for Thai Word Segmentation," *JCSSE,* vol. 2006, pp. 4-40, 2006. (in Thai)

[13] C. Mahatthanachai, PTTSF word parsing techniques, Chiang Mai: Chiang Mai Rajabhat University, 2012. (in Thai)

[14] C. Haruechaiyasak, C. Sangkeetrakarn, P. Palingoon, S. Kongyoung and C. Damrongrat, "A Collaborative Framework for Collecting Thai Unknown words from the web," in *Proceeding COLING-ACL 2006*, Sydney, 2006.

[15] S. Kertkid and P. Meesad, "Improvement of Search Performance for Cancer Contents Using Semantic Search," in *The 8th National Conference on Information Technology*, Krabi, 2015. (in Thai)

[16] C. Chongchorhor, "Using Ontology Tools for Information Services," 2012. [Online]. Available: http://bls.buu.ac.th/~f55361/05Jul11/%a1%d2%c3%e3%aa%bb%c3%d0%e2%c2%aa%b9%a8%d2%a1.pdf. [Accessed 10 February 2017]. (in Thai)

[17] L. Liu, C. Wang, L. Bai and H. Chen, "Study of Ontology Technology in Field Word Segmentation System of Digital Library," in *the 14th International Conference on Computer Supported Cooperative Work in Design 2010*, Shanghai, 2010.

[18] D. W. Wang, "A new field word segmentation model based on ontology in digital library," *IJACT,* vol. 4, no. 17, pp. 418-425, 2012.

[19] S. Niwattanakul, " Access to Agricultural Knowledge by Semantic Web Technologies," Suranaree University of Technology, Nakhon ratchasima, 2013. (in Thai)

[20] W. Chotirat, P. Boonrawd and S. Na Wichian, "Developing an Ontology Knowledge Based for Automatic Online News Analysis," *Information Technology Journal,* vol. 7, no. 14, pp. 13-18, 2011. (in Thai)

[21] P. Nilaphruek and R. Khanankhoaw, "The Enhancement of Efficiency in e-Recruitment System using Semantic Matching Technique," *Science and Technology RMUTT Journal,* vol. 5, no. 1, p. 83–99, 2015. (in Thai)

[22] P. Butte and W. Puarungroj, "Development of Semantic Web for Searching Cultural Information In Loei Province," *Information Technology Journal,* vol. 12, no. 2, pp. 33-41, 2016. (in Thai)

[23] S. Pumikong, The design and development of an algorithm for safety-related news extraction, Nakhon ratchasima: Suranaree University of Technology, 2012. (in Thai)

[24] A. Ekwonganan, Identification of Thai and transliterated words by N-Gram Models, Bangkok: Chulalongkorn University, 2005. (in Thai)

[25] S. PhiaKoksong and N. Chamnongsri, "A Knowledge Navigation System for Accessing Contents in Printed Materials," *Journal of Information Science,* vol. 28, no. 3, pp. 9-20, 2010. (in Thai)

[26] G. Liu and Z. Chen, "Chinese Error Correction of Searching Engine under N-Gram Statistic Model," in *The 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM)*, Chengdu, 2010.

[27] S. Ismail and M. S. Rahman, "Bangla word clustering based on N-gram language model," in *2014 International Conference on Electrical Engineering and Information & Communication Technology*, Dhaka, 2014.

[28] K. Nur Hossain, K. Md. Farukuzzaman, I. Md. Mojahidul, R. Md. Habibur and S. Bappa, "Verification of Bangla Sentence Structure using N-Gram," *Global Journal Inc,* vol. 14, no. 1, pp. 1-5, 2014.

[29] The National Electronics and Computer Technology Center, "Thai Lexeme Tokenizer," 2016. [Online]. Available: http://www.sansarn.com/lexto. [Accessed 10 January 2017]. (in Thai)

[30] National Cancer Institute, 1971. [Online]. Available: https://www.cancer.gov/. [Accessed 10 January 2017]. (in Thai)

[31] The National Electronics and Computer Technology Center, "Hozo Ontology Editor," 2010. [Online]. Available: http://text.hlt.nectec.or.th/ontology/. [Accessed 10 January 2017]. (in Thai)

[32] S. Suguna, V. Sundaravadivelu and B. Gomathi, "A Novel Semantic Approach in E-learning Information Retrieval System," in *The 2nd IEEE ICETECH*, Coimbatore, 2016.

[33] The National Electronics and Computer Technology Center, "Thai-English Electronic Dictionary," 2016. [Online]. Available: https://www.nectec.or.th/innovation/innovation-software/lexitron.html. [Accessed 10 January 2017]. (in Thai)

[34] S. Issaragrisil, "Coping "Multiple myeloma-MM", 2012. [Online]. Available: http://oknation.nationtv.tv/blog/loongjame/2012/07/02/entry-4. [Accessed 10 February 2017]. (in Thai)