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

_21-0060(001-172).indd 11 29/1/2564 BE 13:07

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

_21-0060(001-172).indd 12 29/1/2564 BE 13:07