ระบบติดตามการคัดลอกเนื้อหาเว็บอัตโนมัติ โดยใช้วิธีการเลือกข้อความสำคัญ

Web Plagiarism Monitoring System Using Informative Text Selection Method

ศิริพร อ่วมมีเพียร (Siriporn Ummeepien)* และ สันติพงษ์ ไทยประยูร (Santipong Thaiprayoon)**

บทคัดย่อ

ในปัจจุบันมีเทคโนโลยีสารสนเทศและการสื่อสารที่ทันสมัย ทำให้ผู้ใช้สามารถแบ่งปันและเข้าถึงข้อมูลที่อยู่บนอินเทอร์เน็ตได้สะดวกและรวดเร็ว ในหลายครั้งข้อความจากเว็บเพจอาจถูกคัดลอกจากเพจหนึ่งไปยังอีกเพจโดยไม่มีการอ้างอิงแหล่งที่มา ไม่ว่าจะโดยเจตนาหรือไม่เจตนาก็ตาม ซึ่งส่งผลกระทบโดยตรงต่อเจ้าของผลงาน นักเขียนและบล็อกเกอร์ เป็นอย่างมาก เพื่อแก้ปัญหานี้ ผู้วิจัยจึงได้ออกแบบและพัฒนาระบบติดตามการคัดลอกเนื้อหาเว็บอัตโนมัติเพื่อช่วยนักเขียนคอยติดตาม ป้องกันและเฝ้าระวังเนื้อหาเว็บของตนเอง ระบบได้ประยุกต์ใช้อัลกอริธึมการเลือกข้อความสำคัญเพื่อช่วยลดเวลาในการตรวจสอบการคัดลอก ซึ่งผลการทดลองพบว่าวิธีการที่นำเสนอมีความความเร็วในการตรวจสอบการคัดลอกถึง 44 วินาที และค่า F-measure เท่ากับ 0.83 จากแบบประเมินความพึงพอใจด้านการใช้งานระบบโดยผู้ใช้ พบว่าค่าเฉลี่ยอยู่ที่ 4.24 ดังนั้นสามารถสรุปได้ว่าระบบที่พัฒนาขึ้นมีประสิทธิภาพอยู่ในระดับดี สามารถนำไปประยุกต์ใช้ในการตรวจสอบการคัดลอกเนื้อหาบนเว็บได้

คำสำคัญ: การโจรกรรมข้อมูลบนเว็บ การป้องกันเนื้อหาเว็บ เว็บเพจ ทีเอฟไอดีเอฟ

Abstract

With the current information and communication technology, access and sharing online content can be done at fingertips. Many times, content from one webpage can be copied to another page without owner's permission. This problem of web plagiarism has become a growing concern among content owner, writer and blogger. To solve the problem, we have designed and implemented as a tool to monitor and help protect owners' copyrighted content. To reduce the number of requests and increase execution time, our system applies the informative text selection algorithm to extract only informative text chunks. The experimental result was found the execute times is 44 seconds, a decrease of 46 seconds from the baseline method. The performance based on the F1 measure is equal to 70%. From evaluation of experiment found the mean of end users is 4.20. The results indicated that the efficiency of this system was a good level. It has been applied to monitor a web plagiarism.

Keywords: Web Plagiarism, Protect Content, Web Page, TF-IDF.

1. บทนำ

ปัจจุบันผู้ใช้สามารถสร้างและเผยแพร่เนื้อหา เขียนเล่าเรื่องราว ประสบการณ์ บทความ รูปภาพ และวิดีโอได้สะดวกโดยผ่านโซเชียลเน็ตเวิร์ค (Social Network) หรือเว็บไซต์ของตัวเอง ทำให้ข้อมูลที่เกิดขึ้นบนเครือข่ายอินเทอร์เน็ตเป็นไปอย่างรวดเร็วและมีจำนวนมหาศาล อย่างไรก็ตาม การเข้าถึงและคัดลอกข้อมูลเหล่านี้ก็ทำได้ง่ายเช่นกัน โดยผู้ใช้เพียงแค่คลิ๊กเปิดดูหน้าเว็บเพจแล้วทำการคัดลอกเนื้อหาบางส่วนหรือทั้งหมดมาเป็นงานของ

* สาขาวิชาคอมพิวเตอร์ธุรกิจ คณะบริหารธุรกิจ มหาวิทยาลัยราชพฤกษ์
* Business Computer, Faculty of Business Administration, Rajapruk University.
** ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ
** National Electronics and Computer Technology Center.

ตนเองโดยไม่มีการแก้ไขคำหรือปรับเปลี่ยนประโยคใหม่ อีกทั้งยังไม่มีการอ้างอิงแหล่งข้อมูลที่ได้มา ส่งผลให้เกิดการทุจริตคัดลอกเนื้อหาเว็บจากที่หนึ่งไปยังอีกที่หนึ่งทำได้ง่าย ไม่ว่าจะโดยเจตนาหรือไม่เจตนาก็ตาม การกระทำแบบนี้เรียกว่าการโจรกรรมข้อมูลบนเว็บ (Web Plagiarism) ปัญหานี้ถือว่าเป็นปัญหาสำคัญมากที่ส่งผลกระทบต่อเจ้าของผลงาน นักเขียนและบล็อกเกอร์โดยตรง เนื่องจากผู้ที่เป็นเจ้าของผลงานอาจสูญเสียผลประโยชน์ที่พึงได้รับ เช่น ยอดจำนวนคนดู ชื่อเสียง และคำติชม เป็นต้น ซึ่งผลประโยชน์เหล่านี้ผู้ที่เป็นเจ้าของผลงานอาจจะสามารถเปลี่ยนเป็นรายได้ที่เข้ามาให้กับตัวเจ้าของผลงานเองได้ โดยปกติผู้ที่เป็นเจ้าของผลงานสามารถตรวจสอบการคัดลอกเนื้อหาบนเว็บได้ด้วยมือ โดยการนำแต่ละประโยคไปค้นหาผ่านเครื่องมือสืบค้น ซึ่งถือเป็นงานต้องใช้เวลาและความละเอียดเป็นอย่างมากในการตรวจสอบ เพราะฉะนั้นเครื่องมือตรวจสอบการโจรกรรมการคัดลอกเนื้อหาบนเว็บแบบอัตโนมัติจึงเป็นส่วนที่สำคัญสำหรับช่วยเจ้าของผลงานในการหาแหล่งที่มาของเอกสารว่าถูกคัดลอกมาจากแหล่งใด

ซอฟต์แวร์ที่ใช้ช่วยค้นหา ติดตาม และตรวจสอบการคัดลอกเนื้อหาบนเว็บบนอินเทอร์เน็ตมีเป็นจำนวนมาก ซึ่งมีทั้งแบบเสียค่าใช้จ่ายและใช้งานฟรี เช่น Copyscape, CopyGator, Plagium และ PlagSpotter เป็นต้น เครื่องมือเหล่านี้ถูกพัฒนาขึ้นมาเพื่อตรวจสอบการคัดลอกเนื้อหาบนเว็บซึ่งช่วยลดเวลาและอำนวยความสะดวกในการตรวจสอบเว็บเพจที่คาดว่าจะมีความคล้ายกันให้กับผู้ที่เป็นเจ้าของผลงาน บล็อกเกอร์ นักเขียน นักข่าว และบรรณาธิการ แต่ซอฟแวร์เหล่านี้ยังไม่สนับสนุนการตรวจสอบภาษาไทย ไม่มีการบอกเปอร์เซ็นความคล้ายและทำแถบสีข้อความที่เหมือนกันให้แก่ผู้ใช้ ทำให้ยากต่อการพิจารณา อีกทั้งโดยส่วนมากระบบตรวจสอบการคัดลอกเนื้อหาเว็บจะใช้เทคนิคการตรวจสอบแบบละเอียดกล่าวคือ เลือกใช้ข้อความทั้งหมดที่อยู่ในเอกสารเปรียบเทียบกับเว็บเพจบนอินเทอร์เน็ต ซึ่งวิธีนี้ต้องใช้เวลาในการเข้าถึงและสกัดข้อความทั้งหมดของแต่ละเว็บเพจบนเว็บไซต์มาเปรียบเทียบ ทำให้เสียเวลาในการสกัดข้อความเป็นอย่างมาก ยิ่งถ้าแต่ละเว็บเซิร์ฟเวอร์มีความเร็วอินเทอร์เน็ตช้าก็จะทำให้การสกัดข้อความช้าตามไปด้วย

ดังนั้นผู้วิจัยจึงได้ออกแบบและพัฒนาระบบติดตามการคัดลอกเนื้อหาเว็บอัตโนมัติที่สามารถตรวจสอบเว็บเพจได้ทั้งภาษาไทยและภาษาอังกฤษ [1], [2] เพื่อช่วยนักเขียนคอยติดตาม ป้องกันและเฝ้าระวังเนื้อหาเว็บของตนเองว่าถูกใครขโมยหรือคัดลอกไปโดยไม่มีการอ้างอิงแหล่งที่มา พร้อมทั้งจะทำการแจ้งเตือนผ่านเว็บแอพพลิเคชั่นและอีเมลโดยอัตโนมัติ เมื่อระบบตรวจพบว่าเว็บเพจของเจ้าของผลงานถูกคัดลอก โดยระบบจะทำการแบ่งเนื้อหาเว็บเพจออกเป็นข้อความย่อย แล้วส่งข้อความย่อยเหล่านี้ไปค้นหาผ่านเครื่องมือสืบค้นเชิงพาณิชย์ (Commercial Search Engine) เช่น Google, Yahoo และ Bing เป็นต้น แล้วนำผลลัพธ์การสืบค้นไปคำนวณหาคะแนนความคล้ายอย่างละเอียดต่อไป เพื่อลดจำนวนคำสอบถาม (Query) ที่ร้องขอไปยังเครื่องมือสืบค้นและลดเวลาในการประมวลผล ระบบได้ประยุกต์ใช้อัลกอริธึมการเลือกข้อความสำคัญเพื่อสกัดเฉพาะข้อความย่อยที่สำคัญเท่านั้น

ในส่วนที่เหลือของบทความ ผู้วิจัยได้จัดหัวข้อไว้ดังนี้ ในส่วนถัดไปจะอธิบายเกี่ยวกับทฤษฎีที่เกี่ยวข้องและวิจารณ์งานวิจัยที่เกี่ยวกับการตรวจสอบการคัดลอกเนื้อหาบนเว็บ และอธิบายถึงเทคนิคต่างๆ ที่ใช้ในการตรวจสอบการคัดลอก ส่วนที่ 3 จะนำเสนภาพรวมของระบบและขั้นตอนการวิเคราะห์และคำนวณการคัดลอกเนื้อหาเว็บ ส่วนที่ 4 อธิบายข้อมูลที่ใช้ในการทดลอง การวัดผลการทดลอง วิธีการทดลอง ผลการทดลองและอภิปรายผล และในส่วนสุดท้ายสรุปผลการทดลอง งานวิจัยในอนาคตและข้อเสนอแนะ

2. ทฤษฎีที่เกี่ยวข้อง

ในการวิจัยเรื่องระบบติดตามการคัดลอกเนื้อหาเว็บอัตโนมัติ ผู้วิจัยได้ศึกษาหลักการของทฤษฎีและเทคโนโลยีต่าง ๆ ที่เกี่ยวข้องกับการพัฒนาระบบที่สามารถนำมาประยุกต์ใช้กับงานได้โดยแบ่งออกเป็นหัวข้อดังต่อไปนี้

2.1 การประมวลผลภาษาธรรมชาติ

ภาษาคอมพิวเตอร์คือศาสตร์ที่เกี่ยวข้องกับการทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์ หรือเรียกอีกอย่างว่าการประมวลผลภาษาธรรมชาติ (Natural Language Processing) เป็นการนำความรู้ทางภาษาศาสตร์การวิเคราะห์เกี่ยวกับรูปแบบหรือโครงสร้างของคำ การลดรูปคำ การวิเคราะห์หา

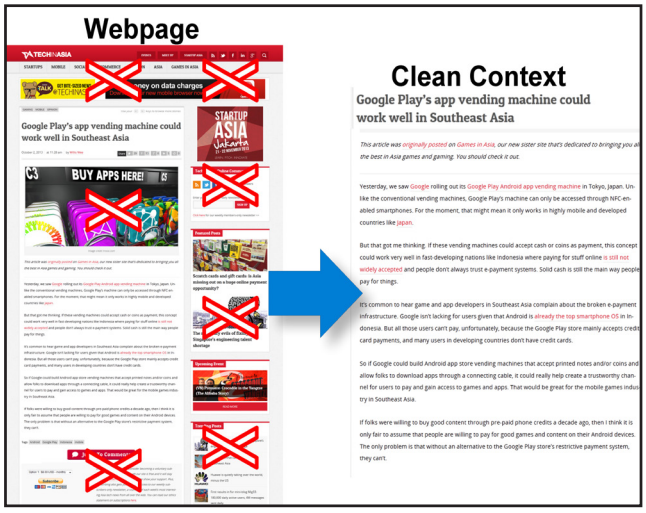
ชนิดของคำ (Parts of Speech Tagging) การสกัดหานิพจน์ระบุนาม (Name Entity Extracting) และการวิเคราะห์หาคำที่มีความหมายคล้ายกัน เป็นต้น มาจัดเก็บด้วยระบบคอมพิวเตอร์เป็นฐานความรู้ โดยระบบจะเรียกใช้ฐานความรู้ ตีความหมาย ถ่ายทอดความรู้ และตอบโต้ด้วยภาษาธรรมชาติ

2.2 การค้นคืนสารสนเทศ

การค้นคืนสารสนเทศ (Information Retrieval) เป็นกระบวนการเกี่ยวข้องกับการจัดรูปแบบการนำเสนอ การจัดเก็บ และการเข้าถึง ตัวเอกสาร หรือข้อมูลในเอกสาร ระบบการค้นคืนสารสนเทศเป็นอุปกรณ์หรือเครื่องมือที่เชื่อมระหว่างผู้ใช้ที่ต้องการข้อมูลและกลุ่มข้อมูล ทั้งนี้จุดมุ่งหมายของระบบคือ คัดเลือกเอาเฉพาะข้อมูลที่ผู้ใช้ต้องการใช้ และกรองเอาข้อมูลที่ผู้ใช้ไม่ต้องการออกไป การทำงานของการค้นคืนสารสนเทศ แบ่งออกได้เป็น 2 ขั้นตอน 1) ขั้นเตรียมฐานข้อมูล คือ การนำข้อมูลที่มีทั้งหมดมาแปลง (Representation) ให้อยู่ในรูปแบบที่ต้องการ ได้แก่ การตัดคำที่ไม่จำเป็น (Stopword Removal) การลดรูปคำ (Stemming) และการทำข้อความให้อยู่ในรูปปกติ (Normalization) แล้วจึงนำเอาเอกสารที่ได้มาทำการสร้างดัชนี (Indexing) 2) ขั้นตอนการค้นหาเอกสารที่ผู้ใช้ต้องการ คือ การนำคำที่ผู้ใช้ต้องการหา (Query) ไปแทนให้อยู่ในรูปแบบเดียวกับข้อ 1 จากนั้นไปค้นหาเอกสารที่ต้องการฐานข้อมูลที่ได้ทำไว้ (Searching) แล้วค้นคืนให้กับผู้ใช้ โดยมีการจัดทำลำดับของเอกสาร (Ranking) ตามลำดับความเหมือนของเอกสารกับข้อความที่ต้องการค้นหา (Similarity)

2.3 การสกัดข้อความบนเว็บ (Web Scraping)

การสกัดข้อความบนเว็บเพจ [3] หรือเรียกอีกอย่างว่า Web harvesting หรือ Web data extraction เป็นเทคนิคในการสกัดสารสนเทศจากเว็บไซต์ วัตถุประสงค์หลักคือการแปลงข้อมูลไร้โครงสร้าง (Unstructured Data) เป็นข้อมูลที่มีโครงสร้าง (Structured Data) เพื่อนำเนื้อหาเว็บที่สกัดได้ไปใช้ประโยชน์ต่อไป เช่น เปรียบเทียบราคาสินค้าออนไลน์ พยากรณ์อากาศ และตรวจสอบการเปลี่ยนแปลงหน้าเว็บ เป็นต้น โดยเทคนิคนี้จะเลือกสกัดเฉพาะข้อมูลที่มีความสำคัญและสามารถนำไปใช้ประโยชน์ได้เท่านั้นและตัดส่วนที่ไม่สำคัญบนหน้าเว็บเพจออก เช่น ป้ายโฆษณา แบนเนอร์ รูปภาพ และลิงค์ต่างๆ เป็นต้น สามารถแสดงได้ดังภาพที่ 1



ภาพที่ 1 *แสดงการสกัดข้อความบนเพจ*

2.4 การเลื่อนกรอบข้อความ (Sliding Window)

ผู้วิจัยได้นำเสนอเทคนิคการเลื่อนกรอบข้อความ ซึ่งวิธีการนี้จะช่วยเพิ่มประสิทธิภาพการตรวจสอบการคัดลอกข้อความให้มีความถูกต้องสูงขึ้น ในกรณีที่ผู้ใช้มีการเปลี่ยนแปลงคำในข้อความที่ต้องการตรวจสอบ โดยมีพารามิเตอร์ 2 ตัว ได้แก่ กำหนดจำนวนคำของประโยค (Window Size) และกำหนดจำนวนคำที่ต้องการเลื่อนย้อนกลับ (Sliding Size) ดังแสดงในภาพที่ 2 โดย T คือข้อความต้นฉบับ TT คือ ข้อความที่ตัดคำแล้ว และ C คือ ข้อความแต่ละหน่วย โดยกำหนด Window Size = 10 และ Sliding Size = 4

T = “การใช้สารอินทรีย์ในวงจรอิเล็กทรอนิกส์ ซึ่งจะมีต้นทุนในการผลิตถูกยิ่งกว่าการผลิต วงจรจากซิลิกอนที่ใช้ในปัจจุบันและสามารถนำไปใช้ได้”

TT = “การ|ใช้|สาร|อินทรีย์|ใน|วงจร|อิเล็กทรอนิกส์| |ซึ่ง|จะ|มี|ต้น|ทุน|ใน|การ|ผลิต|ถูก|ยิ่ง|กว่า|การ|ผลิต||วงจร|จาก|ซิลิกอน|ที่|ใช้|ใน|ปัจจุบัน|และ|สามารถ|นำ|ไป|ใช้|ได้|”

C1 = “การ|ใช้|สาร|อินทรีย์|ใน|วงจร|อิเล็กทรอนิกส์| |ซึ่ง|จะ|”

C2 = “อิเล็กทรอนิกส์| |ซึ่ง|จะ|มี|ต้น|ทุน|ใน|การ| ผลิต|”

C3 = “ทุน|ใน|การ|ผลิต|ถูก|ยิ่ง|กว่า|การ|ผลิต| |”

C4 = “กว่า|การ|ผลิต| |วงจร|จาก|ซิลิกอน|ที่|ใช้|ใน|”

C5 = “ที่|ใช้|ใน|ปัจจุบัน|และ|สามารถ|นำ|ไป|ใช้|ได้|”

ภาพที่ 2 *การแบ่งข้อความด้วยเทคนิคการเลื่อนกรอบข้อความ*

2.5 การตรวจสอบความคล้ายกันของเอกสาร

ประกอบด้วยการทำงานหลัก 2 ส่วน ได้แก่ 1) การหาแหล่งต้นฉบับที่คาดว่ามีความคล้าย (Candidate Source Retrieval) เป็นการค้นหาเอกสารที่คาดว่าจะมีความคล้ายในคลังเอกสารขนาดใหญ่ โดยมีงานวิจัย [4] ก่อนหน้านี้ได้นำเสนอวิธีการสร้างคำสอบถาม (Query Generation) เพื่อเตรียมคำสอบถามสำหรับการค้นคืนเอกสารที่เกี่ยวข้องในคลังข้อมูล สำหรับระบบนี้จะใช้เทคนิคการสร้างคำสอบถามแบบเป็นข้อความย่อย (Chunking Queries) 2) การเปรียบเทียบความคล้ายคลึง (Detail Similarity Comparison) มีงานวิจัยได้นำเสนอวิธีการเปรียบเทียบหาข้อความที่เหมือนกันระหว่าง 2 ข้อความ โดยใช้เทคนิค Text Difference หลังจากได้ข้อความความที่เหมือนกันแล้ว จะต้องนำข้อความมาคำนวณค่าความคล้ายคลึงด้วยวิธี Containment Coefficient [5] เนื่องจากวิธีการวัดนี้เหมาะสมสำหรับคู่เอกสารที่มีความยาวของเอกสารแตกต่างกันมาก โดยกำหนดให้ S(A) คือจำนวนคำในข้อความต้นฉบับและ S(B) คือจำนวนคำในข้อความที่นำมาตรวจสอบ สามารถแสดงได้ดังสมการที่ 11

$$Containment(A,B) = \frac{|S(A) \cap S(B)|}{|S(A)|} \quad (1)$$

2.6 การให้ค่าน้ำหนักคำ (Term Weighting)

การให้ค่าน้ำหนักหรือการกำหนดน้ำหนักคำ [6] โดยคำที่มีความสำคัญหรือใช้เป็นตัวแทนของเอกสารที่ดี ควรจะปรากฏอยู่เป็นจำนวนมากในเนื้อหาของเอกสารเฉพาะฉบับนั้น และปรากฏอยู่น้อยในชุดของเอกสารที่เหลือทั้งหมด แต่ถ้าคำนั้นปรากฏเป็นจำนวนมากในทุกๆ เอกสาร แสดงว่าคำดังกล่าวไม่สามารถเป็นตัวแทนของเอกสารใดๆ ได้ ซึ่งคำเหล่านั้นเรียกว่าคำหยุด (Stop Word) เช่น a, and, the เป็นต้น ดังนั้นการให้น้ำหนักคำๆ หนึ่งในเอกสารฉบับหนึ่งจะพิจารณาจากความถี่ของคำ ที่ปรากฏในเอกสารนั้นๆ และจำนวนของเอกสารทั้งหมดที่มีคำๆ นั้นปรากฏอยู่ วิธีการให้น้ำหนักของคำที่นิยมใช้กันมากคือ TF-IDF (Term Frequency-Inverted Document Frequency) โดย TF คือ ความถี่ของคำที่ปรากฏในเอกสาร สามารถแสดงได้ดังสมการที่ 2

$$tf_{t,d} = \frac{0.5 \times f_{t,d}}{\max\{f_{t,d} : t \in d\}} \quad (2)$$

โดยที่ $tf_{t,d}$ คือ จำนวนครั้งที่คำ t ปรากฏในเอกสารหารด้วยคำที่มีความถี่มากที่สุดในเอกสาร

ส่วน IDF คือ ส่วนกลับความถี่ของเอกสาร สามารถแสดงได้ดังสมการที่ 3

$$idf_{t,d} = \log \frac{N}{|\{d \in D : t \in d\}|} \quad (3)$$

โดยที่ N คือจำนวนเอกสารทั้งหมดที่อยู่ในคลังเอกสารหารด้วยจำนวนเอกสารที่มีคำ t ปรากฏอยู่และ TF-IDF สามารถแสดงได้ดังสมการที่ 4

$$w_{t,d,D} = tf_{t,d} \times idf_{t,d} \quad (4)$$

2.7 งานวิจัยที่เกี่ยวข้อง

ในปัจจุบันมีการศึกษาและพัฒนาระบบตรวจสอบการคัดลอกเนื้อหาเว็บอัตโนมัติโดยนักวิจัยหลายคน ดังนี้

งานวิจัยของ Petri Sirkkala [7] ได้นำเสนอระบบที่ชื่อว่า Nalkki ซึ่งเป็นเว็บแอปพลิเคชั่นที่ช่วยตรวจสอบงานของนักเรียนที่เสนอเข้ามา ซึ่งสามารถตรวจสอบเอกสารที่มีขนาดกลางไม่ใหญ่มาก ขั้นตอนการตรวจสอบเอกสารที่ถูกคัดลอกของระบบ Nalkki ประกอบด้วย 4 ขั้นตอน คือ ขั้นตอนที่หนึ่งอับโหลดเอกสารที่ประกอบด้วยไฟล์เดียวหรือรูปแบบไฟล์ Zip ที่ประกอบด้วยไฟล์เอกสารที่ต้องการตรวจสอบ ขั้นตอนที่สอง เข้าสู่กระบวนการเปรียบเทียบ (Comparison Process) เอกสาร ซึ่งประกอบด้วยในส่วนของแปลงเอกสารให้อยู่ในรูปที่เหมาะสมและสร้างคำค้นเพื่อส่งไปหาต้นฉบับ ขั้นตอนที่สาม แสดงผลลัพธ์ ขั้นตอนสุดท้ายตรวจสอบรายละเอียดผลลัพธ์ของแต่ละเอกสารที่น่าสงสัย

งานวิจัยของ Yi-Ting Liu [8] ได้พัฒนาระบบตรวจสอบการคัดลอกเอกสารออนไลน์ร่วมกับเครื่องมือสืบค้นเพื่อตรวจสอบเอกสารที่น่าสงสัย โดยการออกแบบของระบบนี้ได้ทำการแบ่งข้อความที่ต้องการตรวจสอบออกเป็นประโยคย่อยๆ และจัดลำดับความสำคัญโดยให้ค่าน้ำหนักของแต่ละประโยคโดย หลังจากนั้นประโยคที่มีค่าน้ำหนักสูงจะถูกสืบค้นกับเครื่องมือสืบค้นที่พัฒนาขึ้นเพื่อหาเอกสารที่คาดว่าจะมีความคล้าย

งานวิจัยของ Sebastian [9] ได้ออกแบบเครื่องมือตรวจสอบการคัดลอกชื่อว่า SNITCH โดยสามารถตรวจสอบได้ถูกต้องและรวดเร็วผ่าน Google Web API โดยอัลกอริธึมของเครื่องมือนี้ได้ประยุกต์ใช้เทคนิคการเลื่อนกรอบข้อความ (Sliding Windows) ร่วมกับการวัดความยาวตัวอักษรเฉลี่ยต่อ 1 คำ เพื่อระบุข้อความที่มีการคัดลอก

งานวิจัยของ Chow Kok Ken [10] ได้นำเสนอวิธีการตรวจสอบการคัดลอกเอกสารข้ามภาษา (Cross Language Plagiarism Detection) โดยทำการตรวจสอบการคัดลอกเอกสารข้ามภาษาระหว่างภาษาอังกฤษกับมลายู การทำงานของวิธีนี้คือส่งเอกสารที่เป็นภาษามลายูเข้าไปในระบบหลังจากนั้นระบบจะแปลเป็นภาษาอังกฤษโดยผ่าน Google Translate API เอกสารที่ถูกแปลจะถูกแบ่งออกเป็นส่วนๆ เพื่อตรวจสอบกับเอกสารออนไลน์บนอินเทอร์เน็ตโดยผ่าน Google AJAX Search API และจะเลือก 10 เอกสารที่สืบค้นเจอบ่อยที่สุดเป็นเอกสารที่น่าสงสัย หลังจากนั้นก็ส่งผลลัพธ์ให้กับผู้ใช้

ในงานวิจัยที่นำเสนอมีความแตกต่างจากงานก่อนหน้านี้ คือ ได้มีการนำเสนอเทคนิคการเลือกข้อความสำคัญ โดยอิงวิธี TF-IDF เพื่อให้น้ำหนักความสำคัญของคำในประโยคย่อย โดยวิธีนี้จะเลือกสกัดเฉพาะข้อความย่อยที่สำคัญเท่านั้น เพื่อลดจำนวนคำสอบถาม (Query) ที่ร้องขอไปยังเครื่องมือสืบค้นและลดเวลาในการประมวลผล

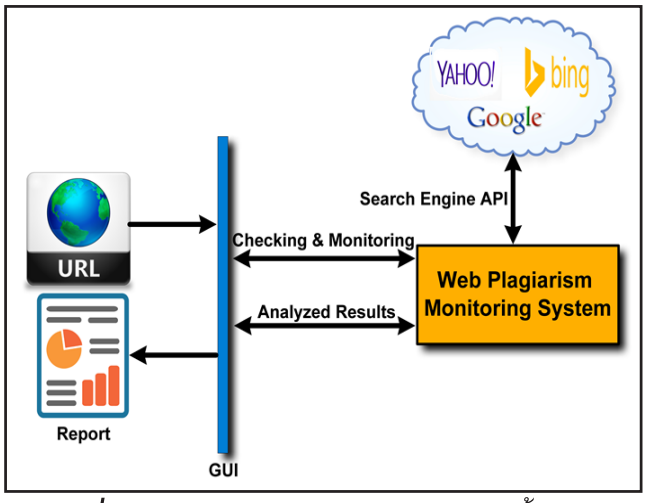
3. วิธีดำเนินการวิจัย

ในส่วนนี้ ผู้วิจัยจะอธิบายเกี่ยวกับภาพรวมของระบบออกแบบและพัฒนาขั้นตอนการทำงานของระบบติดตามการคัดลอกเนื้อหาเว็บอัตโนมัติ และสร้างแบบประเมินความพึงพอใจการใช้งานระบบ โดยมีรายละเอียดดังนี้

3.1 ภาพรวมการทำงานของระบบ

ภาพรวมการทำงานของระบบติดตามการคัดลอกเนื้อหาเว็บอัตโนมัติ สามารถแสดงได้ดังภาพที่ 3

ระบบถูกพัฒนาด้วยภาษาจาวาร่วมกับ JDK 6 โดยผู้ใช้สามารถใช้งานระบบได้ด้วยการพิมพ์ URL ที่ต้องการผ่านเว็บแอพพลิเคชั่นเพื่อทำการตรวจสอบและติดตามการคัดลอก



ภาพที่ 3 ภาพรวมระบบติดตามการคัดลอกเนื้อหาเว็บ

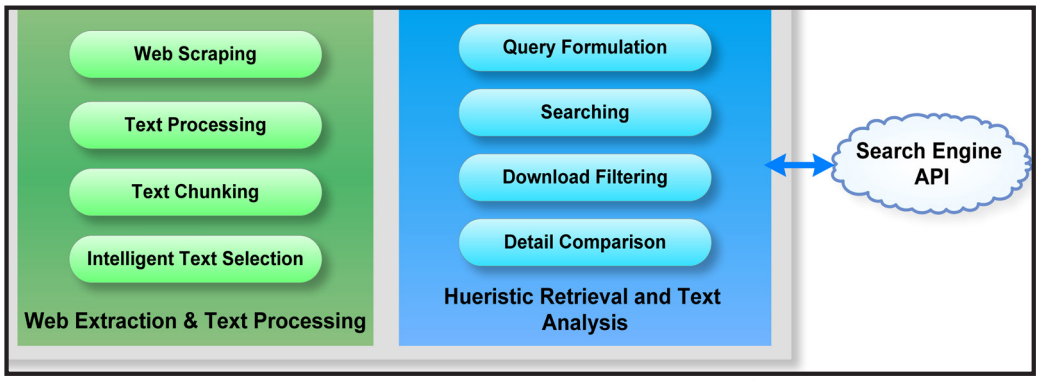
เมื่อระบบตรวจพบเว็บเพจที่เข้าข่ายว่ามีการคัดลอกบนอินเทอร์เน็ต ระบบจะทำการวิเคราะห์และแสดงผลเป็นค่าเปอร์เซ็นต์ความคล้ายกันของเว็บเพจ พร้อมทั้งระบุแหล่งข้อมูลที่พบและทำแถบสีข้อความในส่วนที่คล้ายกัน เพื่อคืนออกเป็นรูปแบบรายงานให้แก่ผู้ใช้

นอกจากนี้ผู้ใช้ยังสามารถกำหนดช่วงเวลาเพื่อให้ระบบคอยติดตามและตรวจสอบเนื้อหาเว็บเพจของตัวเองได้ เช่น รายวัน รายสัปดาห์ หรือรายเดือน เป็นต้น

3.2 เฟรมเวิร์กการทำงานของระบบ

การตรวจสอบการคัดลอกเนื้อหาเว็บอัตโนมัติ ประกอบด้วย 3 คอมโพเน้นท์ (Component) และหลายโมดูล (Module) ด้วยกัน ดังนี้ สามารถแสดงได้ดังภาพที่ 4

1) การสกัดเนื้อหาเว็บและการประมวลผลข้อความ (Web Extraction and Text Processing) โดยคอมโพเน้นนี้ทำหน้าที่สกัดข้อความจากหน้าเว็บเพจและประมวลผลข้อความพื้นฐานทั้งภาษาไทยและอังกฤษ ซึ่งประกอบด้วยโมดูลหลัก 4 โมดูล ดังนี้



ภาพที่ 4 เฟรมเวิร์กของระบบติดตามการคัดลอกเนื้อหาเว็บ

การสกัดข้อความบนเว็บ (Web Scraping) ในโมดูลนี้จะทำการเลือกสกัดเฉพาะเนื้อหาที่เกี่ยวข้องกับหน้าเว็บเพจเท่านั้น โดยระบบจะกำจัดส่วนของเนื้อหาที่ไม่จำเป็นออก เช่น รูปภาพ โลโก้ โฆษณาแบนเนอร์ (Banner) ข้อความหัวกระดาษ (Header) ข้อความท้ายกระดาษ (Footer) ป้ายนำทาง (Navigation) และความคิดเห็น (Comments) เป็นต้น

การประมวลผลข้อความ (Text Processing) ในโมดูลนี้ทำหน้าที่จัดการข้อความให้เหมาะสม ได้แก่ การทำข้อความให้อยู่ในรูปปกติ เช่น การกำจัดสัญลักษณ์พิเศษ ตัวเลข และการหาคำพ้องความหมาย (Synonym) เป็นต้น การแบ่งคำ (Tokenization) การตัดคำที่ไม่จำเป็น และการลดรูปคำ เป็นต้น

การแบ่งข้อความออกเป็นส่วนย่อย (Text Chunking) ในโมดูลนี้จะทำหน้าที่แบ่งข้อความออกเป็นข้อความย่อย ๆ (Chunk)

2) การเลือกข้อความที่สำคัญ (Informative Text Selection) ในคอมโพเน้นนี้ทำหน้าที่เลือกเฉพาะประโยคหรือข้อความย่อยที่สำคัญเพื่อลดจำนวนคำสอบถามในการส่งไปสืบค้นผ่านเครื่องมือสืบค้นและลดเวลาการประมวลผลในการตรวจสอบการคัดลอก โดยในส่วนการเลือกข้อความย่อยที่สำคัญประกอบด้วย 2 โมดูล ดังนี้

2) การเลือกข้อความที่สำคัญ (Informative Text Selection) ในคอมโพเน้นนี้ทำหน้าที่เลือกเฉพาะประโยคหรือข้อความย่อยที่สำคัญเพื่อลดจำนวนคำสอบถามในการส่งไปสืบค้นผ่านเครื่องมือสืบค้นและลดเวลาการประมวลผลในการตรวจสอบการคัดลอก โดยในส่วนการเลือกข้อความย่อยที่สำคัญประกอบด้วย 2 โมดูล ดังนี้

การคำนวณค่าน้ำหนักของคำโดยอิงวิธี TF-IDF (Term Weighting Calculation Based on TF-IDF) เพื่อให้น้ำหนักความสำคัญของคำในแต่ละประโยคย่อย
กำจัดประโยคหรือข้อความย่อยที่เหมือนกันในเอกสาร โดยวัดความคล้ายด้วยวิธี Containment Coefficient

โดยที่การคำนวณค่าน้ำหนักของแต่ละประโยคย่อยสามารถแสดงได้ดังสมการที่ 5

$$\inf ormationscore_{c,d} = \sum_{i=1}^{n} w_{i,c} \quad (5)$$

โดยที่ c คือข้อความย่อย ซึ่งแต่ละข้อความย่อยจะประกอบด้วยคำ

d คือเอกสารหรือเว็บเพจ

w คือค่าน้ำหนักของคำที่อยู่ใน c

3) การค้นคืนแหล่งที่คาดว่ามีความคล้ายและการวิเคราะห์ข้อความ (Candidate Source Retrieval and Text Analysis) ในคอมโพเน้นนี้ทำหน้าที่ค้นคืนแหล่งที่คาดว่ามีความคล้ายเพื่อนำแหล่งข้อมูลเหล่านั้นมาเปรียบเทียบความคล้ายโดยละเอียดอีกครั้ง โดยในส่วนนี้ประกอบด้วย 4 โมดูล ดังนี้

การจัดรูปแบบคำสอบถาม (Query Formulation) เป็นการสร้างตัวแทนคำสอบถามเพื่อให้การค้นคืนเอกสารได้ผลลัพธ์สอดคล้องตามความต้องการ สำหรับงานวิจัยครั้งนี้เราได้เลือกคำสำคัญ 12 คำเรียงต่อกันเพื่อเป็นคำสอบถามในการค้นคืนเอกสาร

การสืบค้น (Searching) เป็นการนำคำสอบถามไปค้นคืนเอกสารในฐานดัชนีที่จัดเตรียมไว้ โดยเลือก 10 ผลลัพธ์แรกของสืบค้นแต่ละครั้ง โดยในการทดลองนี้ผู้วิจัยใช้ Bing Search API

การคัดกรองผลลัพธ์ที่จะดาวน์โหลด (Download Filtering) เป็นการคัดเลือกผลลัพธ์ที่ได้จากการสืบค้นเพื่อนำผลลัพธ์เหลือนั้นมาสกัดข้อความ โดยขั้นตอนนี้ใช้วิธีคัดเลือก 20 ผลลัพธ์ที่ปรากฏบ่อยที่สุด

การเปรียบเทียบความคล้าย (Detail Comparison) เป็นการนำข้อความที่สกัดได้มาเปรียบเทียบความคล้ายคลึงกับเอกสารต้นฉบับเพื่อคำนวณคะแนนความคล้าย

3.3 ออกแบบระบบโดยใช้กระบวนการพัฒนาระบบแบบการพัฒนาด้วยการทำต้นแบบ

1) พัฒนาต้นแบบระบบติดตามการคัดลอก

2) ประเมินความพึงพอใจที่มีต่อระบบโดยใช้แบบสอบถามความพึงพอใจ

3) วิเคราะห์ความพึงพอใจต่อการใช้ระบบรวมถึง ปัญหาที่พบเจอในการใช้ระบบ โดยใช้สถิติเชิงพรรณนา ได้แก่ การแจกแจงค่าความถี่ (Frequency) ค่าร้อยละ (Percentage) ค่าเฉลี่ย (Mean) และค่าเบี่ยงเบนมาตรฐาน (Standard Division)

กำหนดเกณฑ์การประเมินคุณภาพของระบบ และความพึงพอใจของผู้ใช้ที่มีต่อระบบเป็นแบบมาตราส่วนประมาณค่า (Rating Scale) ซึ่งมีเกณฑ์ในการกำหนดค่าน้ำหนักของการประเมินเป็น 5 ระดับ ตามแบบของลิเคิร์ท

(Linkert's Scale) โดยมีหลักเกณฑ์ดังนี้ [11]

5 หมายถึง ความพึงพอใจมากที่สุด

4 หมายถึง ความพึงพอใจมาก

3 หมายถึง ความพึงพอใจปานกลาง

2 หมายถึง ความพึงพอใจน้อย

1 หมายถึง ความพึงพอใจน้อยที่สุด

เกณฑ์การแปลความหมายเพื่อจัดระดับคะแนนเฉลี่ยความพึงพอใจของผู้ใช้มีเกณฑ์การให้คะแนนในแต่ละระดับ ดังนี้

ค่าเฉลี่ย 4.50-5.00 หมายถึง ความพึงพอใจมากที่สุด

ค่าเฉลี่ย 3.50-4.49 หมายถึง มีความพึงพอใจมาก

ค่าเฉลี่ย 2.50-3.49 หมายถึง มีความพึงพอใจปานกลาง

ค่าเฉลี่ย 1.50-2.49 หมายถึง มีความพึงพอใจน้อย

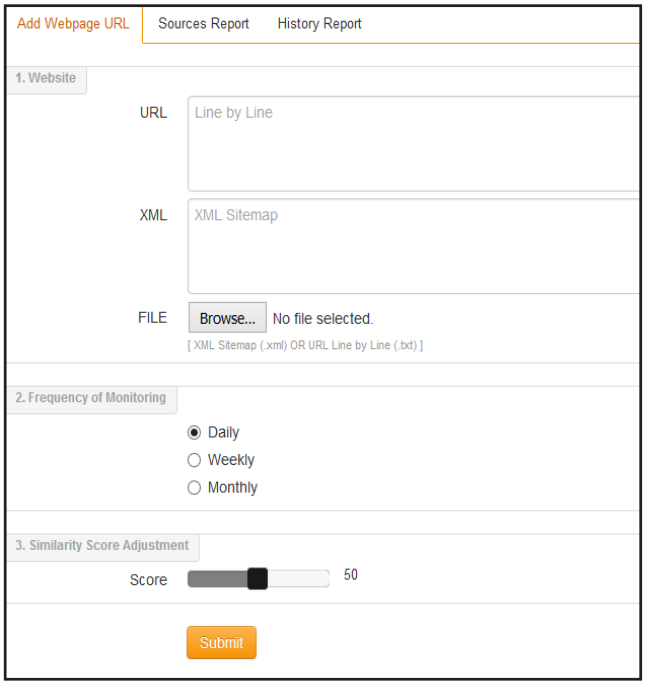
ค่าเฉลี่ย 1.00-1.49 หมายถึง มีความพึงพอใจน้อยที่สุด

4. ผลการดำเนินการวิจัยและอภิปรายผล

ผู้วิจัยได้ทำการพัฒนาระบบตามฟังก์ชันที่ได้ออกแบบไว้และประเมินประสิทธิภาพการทำงานของระบบ โดยแบ่งออกเป็น 2 ส่วน คือ 1) ผลการพัฒนาระบบ 2) ผลการประเมินประสิทธิภาพของระบบ

4.1 ผลการพัฒนาระบบ

หน้าจอกำหนดช่วงเวลาเพื่อให้ระบบคอยติดตามและตรวจสอบเนื้อหาเว็บเพจของตัวเองได้ เช่น รายวัน รายสัปดาห์ หรือรายเดือน เป็นต้นสามารถแสดงได้ดังภาพที่ 5



ภาพที่ 5 *หน้าจอการกำหนดเวลาการติดตาม*

หน้าจอแสดงผลลัพธ์การตรวจการคัดลอกเนื้อหาบนเว็บโดยระบบจะแสดงแถบสีบนข้อความที่ซ้ำพร้อมทั้งแหล่งที่พบและคะแนนความคล้าย สามารถแสดงได้ดังภาพที่ 6

ภาพที่ 6 *หน้าจอแสดงผลการคัดลอกเนื้อหาบนเว็บ*

หน้าจอแสดงประวัติการตรวจสอบการคัดลอก โดยระบบจะแสดงประวัติการตรวจสอบแยกเป็นรายวันและรายเดือนตามลำดับ สามารถแสดงได้ดังภาพที่ 77

ภาพที่ 7 *หน้าจอแสดงประวัติการตรวจสอบ*

4.2 ผลการประเมินประสิทธิภาพของระบบ

4.2.1 การประเมินความเร็วและความถูกต้องในการเลือกข้อความสำคัญ

ผู้วิจัยได้ทำการใช้ชุดทดสอบจำนวน 100 เว็บเพจ ซึ่งถูกสร้างโดยผู้เชี่ยวชาญทางด้านภาษาศาสตร์ โดยชุดทดสอบถูกสร้างมาจากหลายหัวเรื่อง เช่น ข่าว บันเทิง และสุขภาพ เป็นต้น

สำหรับการทดลองนี้ได้ทดสอบประสิทธิภาพด้วยเครื่องคอมพิวเตอร์ CPU Intel Dual Core Xeon @64 bit 2.67 GHz และระบบปฏิบัติการ Ubuntu 14.04

ผู้วิจัยได้ทดสอบวัดประสิทธิภาพความเร็วและความถูกต้องในการเลือกข้อความสำคัญโดยใช้ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และอัตราการรู้จำ (F-measure) ผลการทดสอบสามารถแสดงได้ดังตารางที่ 1

ตารางที่ 1 *ผลการทดสอบการเลือกข้อความสำคัญ*

Approach	P	R	F1	Response Time (sec)
Baseline	1.00	0.78	0.88	90
Random Text	0.68	0.64	0.66	58
Text Selection	0.90	0.76	0.83	44

จากตารางสามารถสรุปได้ว่าวิธีของผู้วิจัยที่นำเสนอให้ความเร็วเฉลี่ยในการตรวจสอบการคัดลอกดีที่สุดเนื่องจากมีการคัดเลือกเฉพาะข้อความที่สำคัญนำไปตรวจสอบเท่านั้น

แต่ค่าความแม่นยำและค่าความระลึกยังต่ำกว่าค่าพื้นฐาน เนื่องจากมีความผิดพลาดในการเลือกข้อความสำคัญ กล่าวคือในการคัดเลือกข้อความบางครั้งระบบไม่ได้เลือกข้อความที่มีคำนามปรากฏมาด้วย ส่วนวิธีดั้งเดิม (Baseline) ให้ค่าเฉลี่ยอัตราการรู้จำมากที่สุด เนื่องจากเลือกทุกข้อความนำไปตรวจสอบ แต่ใช้เวลาเฉลี่ยนานที่สุด

4.2.2 การประเมินความพึงพอใจของผู้ใช้งานที่มีต่อระบบ

การประเมินความพึงพอใจของผู้ใช้งานที่มีต่อระบบติดตามการคัดลอกเนื้อหาเว็บอัตโนมัติ ผู้วิจัยได้ให้ผู้ใช้งานจำนวน 30 คน ทำการประเมินระบบ โดยใช้วิธีการแบบแบล็คบ๊อกซ์ ผลการประเมินความพึงพอใจด้านการใช้งานระบบจากผู้ใช้ สามารถแสดงได้ดังตารางที่ 2

ตารางที่ 2 *ผลการประเมินความพึงพอใจด้านการใช้งานระบบจากผู้ใช้*

รายการประเมิน	ระดับประสิทธิภาพ		
	$\bar{x}$	S.D.	แปล
1. ระบบสามารถติดตามการคัดลอกเนื้อเว็บได้รวดเร็วและถูกต้องเหมาะสม	4.20	0.45	ดี
2. ความสามารถในการวิเคราะห์คำนวณหาค่าความคล้ายคลึง	4.20	0.45	ดี
3. การออกแบบและแสดงผลในส่วนของผลลัพธ์การตรวจสอบ สามารถใช้งานได้สะดวกเข้าใจง่าย และแสดงผลลัพธ์ถูกต้องเหมาะสม	4.20	0.45	ดี
4. การออกแบบและจัดวางองค์ประกอบของระบบ รวมทั้งสีที่ใช้สวยงามและเหมาะสม	4.20	0.45	ดี
5. ภาพรวมทั้งหมดของระบบ	4.40	0.55	ดี
สรุปรายการประเมิน	**4.24**	**0.47**	**ดี**

จากข้อมูลตารางที่ 2 ผลการแสดงความพึงพอใจต่อระบบของผู้ใช้โดยรวมมีค่าเฉลี่ยอยู่ที่ระดับดี โดยมีค่าเฉลี่ยเท่ากับ 4.24 และค่าเบี่ยงเบนมาตรฐานเท่ากับ0.47 ซึ่งอยู่ในช่วง 0 ถึง 1 ถือได้ว่าเป็น ข้อมูลที่ น่าเชื่อถือจึงสามารถสรุปได้ว่า กลุ่มตัวอย่างมีความคิดเห็นที่ค่อนข้างใกล้เคียงกันมีความพึงพอใจอยู่ในระดับดี

5. สรุปผลงานวิจัยและข้อเสนอแนะ

ในบทความนี้ทีมวิจัยได้ได้ออกแบบและพัฒนาระบบติดตามการคัดลอกเนื้อหาเว็บอัตโนมัติที่สามารถตรวจสอบเว็บเพจได้ทั้งภาษาไทยและภาษาอังกฤษเพื่อช่วยนักเขียนคอยติดตาม ป้องกันและเฝ้าระวังเนื้อหาเว็บของตนเองว่าถูกใครขโมยหรือคัดลอกไปโดยไม่มีการอ้างอิงแหล่งที่มา พร้อมทั้งจะทำการแจ้งเตือนผ่านเว็บแอปพลิเคชันและอีเมลโดยอัตโนมัติ ระบบได้ประยุกต์ใช้อัลกอริธึมการเลือกข้อความสำคัญเพื่อช่วยลดเวลาในการตรวจสอบการคัดลอก ซึ่งผลการทดลองพบว่าวิธีการที่นำเสนอมีความความเร็วในการตรวจสอบการคัดลอกถึง 44 วินาที และค่า F-measure เท่ากับ 0.83 จากแบบประเมินความพึงพอใจด้านการใช้งานระบบโดยผู้ใช้ พบว่าค่าเฉลี่ยอยู่ที่ 4.24 สำหรับการพัฒนางานในอนาคตทีมวิจัยจะประยุกต์ใช้วิธีการระบุหน้าที่ของคำในประโยค (POS Tagging) เพื่อให้ระบบสามารถเลือกข้อความสำคัญได้ถูกต้องมากยิ่งขึ้น

6. เอกสารอ้างอิง

[1] S. Thaiprayoon and C. Haruechaiyasak. “Web Plagiarism Detection Based on Search Result Snippets.” *International Technical Conference on Circuit/Systems Computers and Communications*, pp. 1097-1100. 2010.

[2] S. Thaiprayoon. “Automatic Plagiarism Detection Based on Information Retrieval and Natural Language Processing.” Applied Computer Technology and Information Systems, 2013.

[3] “Web scraping.” Available Online at https://en.wikipedia.org/wiki/Web_scraping

[4] Š. Suchomel and M. Brandejs. “Heterogeneous queries for synoptic and phrasal search.” *PAN*, 2014.

[5] A. Z. Broder. “On the resemblance and containment of documents.” *In Proceedings of the Compression and Complexity of Sequences*, pp. 21-29, 1997.

[6] G. Salton and C. Buckley. “Term-weighing approaches in automatic text retrieval.” *In Information Processing & Management*, Vol. 24, No. 5, pp. 513-523, 1988.

[7] P. Sirkkala and P. Nalkki. “Nalkki-project - tool for plagiarism detection using the web.” *In Proceedings of the Seventh Baltic Sea Conference on Computing Education Research*, pp. 229-230, 2007.

[8] Y. T. Liu, H. R. Zhang, T. W. Chen, and W.-G Teng.

"Extending Web Search for Online Plagiarism Detection." *Proceedings of the IEEE Int. Conf. on Information Reuse and Integration (IRI 2007)*, pp. 164-169, 2007.

[9] S. Niezgoda and T. P Way. "SNITCH: A Software Tool for Detecting Cut and Paste Plagiarism." *Proceedings of the 37th SIGCSE Technical Symposium on Computer Science Education*, pp. 51-55, 2006.

[10] C. K. Kent and N. Salim. "Web Based Cross Language Plagiarism Detection." *Second International Conference on Computational Intelligence*, pp. 199-204, 2010.

[11] T. Silpcharu. *Research and statistical analysis with spss, Bangkok s-r-printing-mass product*, 2008.