ระบบการค้นคืนสารสนเทศโดยใช้เทคนิค N-Gram
Information Retrieval System Using N-Gram Technique

สิทธิโชค ปัญญาฤกษ์ชัย[2] และ ศิศาณี นุชิตประสิทธิ์ชัย [1]

[1,2]ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ 1518 ถนนพิบูลสงคราม บางซื่อ กรุงเทพฯ 10800

Mobile : 086-803-1935[1],084-075-0150[2]

E-Mail : ochin2808@hotmail.com[1], sittichoke_m303@hotmail.com[2]

บทคัดย่อ

บทความวิจัยนี้มีวัตถุประสงค์ เพื่อพัฒนาระบบการค้นคืนสารสนเทศโดยใช้เทคนิค N-Gram ซึ่งเป็นระบบการค้นคืนสารสนเทศ เพื่อให้ได้ข้อมูลสารสนเทศที่ตรงตามความต้องการของผู้ใช้งาน เป็นระบบการจัดการสารสนเทศและการค้นคืนสารสนเทศ ทำการประเมินประสิทธิภาพการใช้ค่าความแม่นยำ Precision และค่าความถูกต้อง Recall ในการค้นคืน โดยผลที่ได้จากการประเมินประสิทธิภาพของระบบ พบว่ามีค่าเฉลี่ยความถูกต้อง 78 เปอร์เซ็นต์ จากทฤษฎีค่าความถูกต้องมีมากกว่า 60 เปอร์เซ็นต์ แสดงให้เห็นว่าระบบมีการค้นคืนโดยรวมอยู่ในระดับดีและระบบสามารถนำไปใช้งานได้จริง

คำสำคัญ: เอ็นแกรม, การค้นคืนสารสนเทศ

Abstract

The purpose of this research is to develop and improve Information Retrieval system proposed the design of system which developed by using PHP, N-Gram technique and MySQL as database management. The performance of the system has been assessed. We used precision and recall to evaluate performance of the system. From testing by users in terms of information retrieval, the performance of the search has average accuracy about 78 percent. This shows that the system information retrieval is in the level of "good" performance.

Keyword: N-Gram, Information Retrieval

1. บทนำ

ในปัจจุบันมีข้อมูลมากมายที่น่าสนใจทั้งบนอินเทอร์เน็ตอินทราเน็ตและตามองค์กรหลาย ๆ แห่ง ทำให้ต้องมีการใช้เทคนิคการค้นคืนข้อมูลต่าง ๆ เพื่อความสะดวกรวดเร็วในการค้นหาข้อมูลสำคัญ ๆ ที่ผู้ใช้ต้องการ ข้อมูลสำหรับการใช้งานหรือการศึกษา

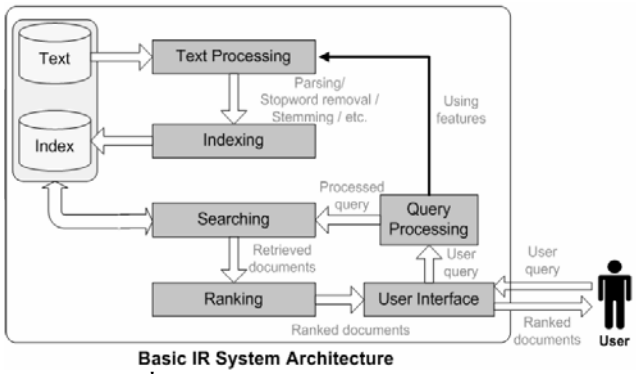
การค้นคืนสารสนเทศ (Information Retrieval) ในการประมวลผลข้อความ (Text Processing) สิ่งที่เป็นพื้นฐานที่จำเป็นอย่างยิ่งคือ "หน่วยคำ" ดังนั้นการหาขอบเขตของแต่ละคำจึงเป็นสิ่งแรกที่ต้องคำนึงถึง เพราะหากเลือกการหาขอบเขตคำไม่เหมาะสมอาจนำมาสู่ระบบการประมวลผลข้อความที่ไม่ถูกต้อง สำหรับภาษาไทยการหาขอบเขตคำค่อนข้างเป็นปัญหาเนื่องจากลักษณะการเขียนภาษาไทยนั้นไม่มีการใช้ตัวอักษรหรือสัญลักษณ์ที่นำมาใช้คั่นระหว่างคำหรือว่ามีการวรรคระหว่างคำเหมือนภาษาอังกฤษ งานต่างๆ ในด้านการประมวลผลภาษาไทยนั้น จึงจำเป็นอย่างยิ่งที่ต้องทราบขอบเขตของคำ นั่นคือต้องมีกระบวนการตัดคำ (Word Segmentation) ที่เหมาะสมก่อนเป็นอันดับแรก

จากข้างต้น ผู้วิจัยจึงจัดทำระบบการค้นคืนสารสนเทศโดยใช้เทคนิค N-Gram ซึ่งเป็นเทคนิคการตัดคำที่ต้องการค้นหาในการนำไปใช้ในการทำงานให้กับผู้ใช้งานทำการค้นคืนข้อมูลโดยนำทฤษฎีเกี่ยวกับการค้นคืนสารสนเทศ มาเป็นเทคนิคในการค้นคืนสารสนเทศ เพื่อให้ผู้ใช้งานต่าง ๆ สามารถเข้าใช้ได้อย่างสะดวก เข้าถึงข้อมูลได้อย่างรวดเร็ว

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎี

การค้นคืนสารสนเทศ (Information Retrieval) เป็นกระบวนการที่เกี่ยวข้องกับการจัดรูปแบบการนำเสนอ การจัดเก็บ และการเข้าถึง ตัวเอกสาร หรือ ข้อมูลในเอกสาร ระบบการค้นคืนสารสนเทศ เป็น อุปกรณ์หรือเครื่องมือที่เชื่อมระหว่างผู้ใช้ที่ต้องการข้อมูลและกลุ่มข้อมูล ทั้งนี้ จุดมุ่งหมายของระบบคือ คัดเลือกเอาเฉพาะข้อมูลที่ผู้ใช้ต้องการใช้ และกรองเอาข้อมูลที่ผู้ใช้ไม่ต้องการออกไป [1] การทำงานของการค้นคืนสารสนเทศ แบ่งออกได้เป็น 2 ขั้นตอน 1) ขั้นเตรียมฐานข้อมูลคือการนำข้อมูลที่มีทั้งหมดมาแปลง (Representation) ให้อยู่ในรูปแบบที่ต้องการ ได้แก่ การตัดคำที่ไม่จำเป็น (Stop Word) การทำให้อยู่ในรูปของราก (Stem) และการหาคำเหมือน (Normalization) แล้วจึงนำเอาเอกสารที่ได้มาทำดัชนี (Indexing) ดังภาพที่ 1



ภาพที่ 1: การค้นคืนจากฐานดัชนี (Retrieval)

2) ขั้นค้นหาเอกสารที่ผู้ใช้ต้องการ คือ การนำคำที่ผู้ใช้ต้องการหา (Query) ไปแทนให้อยู่ในรูปแบบเดียวกับข้อ 1 จากนั้นไปค้นหาเอกสารที่ต้องการฐานข้อมูลที่ได้ทำไว้ (Searching) แล้วค้นคืนให้กับผู้ใช้ โดยมีการจัดทำลำดับของเอกสาร (Ranking) ตามลำดับความเหมือนของเอกสารกับข้อความที่ต้องการค้นหา (Similarity)

N-Gram คือ แบบจำลองที่ใช้คำนวณค่าความน่าจะเป็นของชุดอักขระ (Character Sequence) ที่เกิดขึ้นร่วมกันเป็นคำ หรือค่าความน่าจะเป็นของคำที่เขียนเรียงกัน (Word Sequence) ที่เกิดขึ้นร่วมกันเป็นประโยค โดยค่าความน่าจะเป็นของชุดอักขระหรือคำ ประมาณได้จากคลังข้อมูลที่สร้างไว้ซึ่ง N-Gram ได้ใช้หลักการของสถิติในหลาย ๆ ด้านมาประยุกต์ใช้ [3]

2.2 การศึกษางานวิจัยที่เกี่ยวข้อง

ชูชาติ หฤไชยะศักดิ์ [1] ได้นำเสนองานวิจัยและพัฒนาโครงสร้างพื้นฐานสารสนเทศอัจฉริยะ ฝ่ายวิจัย และพัฒนาเทคโนโลยีสารสนเทศ (RDI) ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ ได้นำเสนอเทคนิคการสร้างดัชนีของลูซีน ซึ่งเป็นแบบ Inverted File Index วิธีการสร้างดัชนีแบบเริ่มจากการนำเอกสารมาวิเคราะห์และแบ่งเป็นรายการของคำ ซึ่งมีการกรองคำที่ไม่มีความหมายสำคัญออก หรือแปลงคำให้เป็นรากศัพท์ จากนั้นจึงนำมาเก็บเป็นไฟล์ดัชนีซึ่งเป็นคำต่าง ๆ พร้อมทั้งมีจำนวนเอกสารที่คำนั้นๆ ปรากฏอยู่ แต่ละคำก็จะมีหมายเลขของเอกสารพร้อมทั้งจำนวนคำที่ปรากฏอยู่ในเอกสารนั้น ๆ ผู้ใช้สามารถส่งคิวรี่ให้กับ หน่วยประมวลคิวรี่ หน่วยค้นคืนจะเปรียบเทียบคำที่ค้นคืนกับค่าดัชนี แล้วจะจัดลำดับผลลัพธ์การค้นคืนให้กับผู้ใช้

ธนารักษ์ ธีระมั่นคง และคณะ [2] ได้นำเสนองานวิจัยโครงการวิจัยพัฒนาต้นแบบระบบฐานความรู้ด้านการแพทย์ในประเทศไทย ได้นำเทคนิคการใช้งาน N-Gram มาใช้และนำเทคนิค Ontology มาใช้ในการจำแนกกลุ่มของโรคและสมุนไพรจำแนกออกเป็นชนิด

กิตติชน แม้นสมุทร [3] ได้นำเสนอการสร้าง Search Engine โครงงานวิชา Information Retrival เป็นการสอนสร้าง Search Engine โดยใช้เทคนิค N-Gram เข้ามาช่วยในการค้นคืนข้อมูลที่ผู้ใช้ต้องการ สรุปการสกัดคำออกจากเว็บเพจที่ได้ทำการ Crawler นั้นยังทำได้ไม่ดีพอ ด้านการตัดคำภาษาไทยโดยใช้ N-Gram พบว่าใช้ 3-Gram และ 4-Gram จะทำให้ได้ประสิทธิภาพดี การตัดคำ (Word Segmentation) การเขียนในภาษาไทยนั้นจะมีความแตกต่างกับภาษาอังกฤษอย่างเด่นชัด เนื่องจากภาษาอังกฤษจะมีช่องว่างในการระบุคำแต่ละคำ ซึ่งการตัดคำของภาษาไทยส่วนใหญ่นั้นจะอาศัยโปรแกรมตัดคำ โดยใช้พจนานุกรมในการตัดคำแต่ก็ไม่ได้มีประสิทธิภาพที่ดี 100% เนื่องจากมีความเป็นไปได้ที่คำที่ปรากฏในเอกสาร อาจจะไม่ปรากฏในพจนานุกรม และไม่สามารถตัดคำที่เป็นประโยคได้

Brown [4] ได้นำเสนอการสร้างโมเดลที่ใช้ N-Gram เป็นตัววิเคราะห์การสร้างโมเดลภาษา ไวยากรณ์ ที่เกี่ยวข้องกับการพูดและภาษาพูด ซึ่งนำมาผสมผสานกับการทำงานของ XML และกฎของไวยากรณ์ต่าง ๆ มาสร้างเป็นโมเดล

อัครพล เอกวงศ์อนันต์ [5] ได้นำเสนอ ผลงานเรื่อง การระบุคำไทยและคำศัพท์ด้วยแบบจำลองเอ็นแกรม เป็นวิทยานิพนธ์จากจุฬาลงกรณ์มหาวิทยาลัย ในปี 2548 ซึ่งตัวผลงานกล่าวถึงวิธีระบุทำที่เป็นภาษาไทย เนื่องจากภาษาไทยนั้นเขียนติดกันไม่มีการเว้นวรรคเหมือนภาษาอังกฤษจึงทำให้ยากต่อการตัดคำเป็นอย่างมาก เอ็นแกรมเป็นตัวช่วยในการตัดคำตามจำนวนตัวอักษรซึ่งเอ็นแกรมเพียงอย่างเดียวก็ไม่สามารถระบุได้จึงต้องมีกฎของคำศัพท์ภาษาไทยเข้ามาช่วยในการระบุด้วย

นิพนธ์ เจริญกิจการ [6] ได้นำเสนอระบบการค้นคืนเอกสารภาษาไทยด้วยเทคนิคขั้นสูง ระยะที่ 2 ได้นำเสนองานวิจัยในการค้นคืนเอกสารภาษาไทย การวิจัยนี้จะเน้นไปที่การค้นคืนในแบบแนวความคิด เทคนิคที่ได้รับการยอมรับอย่างกว้าง ขวางเช่น vector space และ latent semantic indexing

Sasiporn Usanavasin [7] ได้นำเสนองานวิจัย Non-Dictionary-Based Thai Word Segmentation Using Decision Trees การตัดคำภาษาไทยโดยใช้ Decision Trees ตัดสินใจแบบไม่มีพจนานุกรม เป็นงานวิจัยที่ใช้การตัดคำโดยใช้เทคนิค Decision Trees ในการเลือกคำภาษาไทย

ตารางที่ 1 การเปรียบเทียบเทคนิคในงานวิจัยจากผลการค้นคว้า

ชื่อเรื่อง	เทคนิคที่ใช้	ข้อดี	ข้อด้อย
(ศันสนีย์ เอื้อพันธ์วิริยะกุล และคณะ) ต้นแบบการสร้างระบบค้นหาโรงแรมโดยใช้เทคนิคการหาเหตุผลโดยฟัซซีซีมีน	เทคนิคการหาเหตุผลโดยประมาณและฟัซซีซีมีน	1. ง่ายในการค้นคืน 2. มีความรวดเร็วในการค้นหา	ในการทำงานครั้งแรกอาจมีความล่าช้าในการตรวจสอบคีย์เวิร์ดที่ User ต้องการกับผลลัพท์ที่ได้
(ชูชาติ หฤไชยะศักดิ์) งานวิจัยและพัฒนาโครงสร้างพื้นฐานสารสนเทศอัจฉริยะ	เทคนิคการสร้างดัชนีของลูซีน ซึ่งเป็นแบบ Inverted File Index	1. ง่ายในการค้นคืน 2. มีความรวดเร็วในการค้นหา 3. มีดัชนีไว้สำหรับการค้นหาครั้งต่อไป	การค้นหาจะทำการเก็บดัชนี ซึ่งจะทำให้ข้อมูลมากขึ้นเรื่อย ๆ
(ธนารักษ์	1. เทคนิค N-Gram 2. เทคนิค Ontology	สามารถ	ในการค้นหา
ธีระมั่นคง และคณะ) โครงการวิจัยพัฒนาต้นแบบระบบฐานความรู้ด้านการแพทย์ในประเทศไทย		แบ่งกลุ่มเนื้อหาได้อย่างชัดเจน	หากมีคำที่ไม่จัดอยู่ในกลุ่มการแพทย์จะทำให้หาไม่พบ
(กิตติชน แม้นสมุทร) การสร้าง Search Engine	1.เทคนิค N-Gram	ด้านการตัดคำพบว่าใช้ 4-Gram จะทำให้ได้ประสิทธิภาพดี	การสกัดคำออกจากเว็บเพจที่ได้ทำการ Crawler นั้นยังไม่ดีพอ
(Brown) การสร้างโมเดลที่ใช้ N-Gram เป็นตัววิเคราะห์การสร้างโมเดล	1.เทคนิค N-Gram 2.XML	ทำให้ได้ภาษาที่ใกล้เคียงภาษาธรรมชาติ	โมเดลบางโมเดลอาจจะใช้งานได้ไม่ดีพบ
(อัครพล เอกวงศ์อนันต์) การระบุคำไทยและคำศัพท์ด้วยแบบจำลองเอ็นแกรม	1.แบบจำลองเอ็นแกรม	สามารถระบุคำไทยได้อย่างชัดเจน ที่มีอยู่ในศัพท์	ภาษาไทยนั้นเขียนติดกันไม่มีการเว้นวรรคเหมือนภาษาอังกฤษจึงทำให้ยากต่อการตัดคำ

จากการศึกษาทฤษฎีและงานวิจัยข้างต้น พบว่าแนวทางในการแก้ปัญหาเรื่องของความต้องการของผู้ใช้เพื่อให้ได้ข้อมูลที่ตรงกับความต้องการสามารถแก้ไขได้ด้วยเทคนิค N-Gram ผสมกับเทคนิคการทำฐานดัชนี

3. วิธีดำเนินการวิจัย

วิธีการดำเนินงานเพื่อพัฒนาระบบการค้นคืนสารสนเทศโดยใช้เทคนิค N-Gram ทางผู้พัฒนาได้กำหนดวิธีการดำเนินงานออกเป็น 4 ขั้นตอนดังนี้

3.1 ศึกษาและวิเคราะห์ปัญหาระบบงาน

จากการศึกษาปัญหาและความต้องการของระบบ เป็นการศึกษาถึงความสำคัญในการใช้งาน Search engine ในปัจจุบันที่มีความสำคัญในการดำเนินชีวิตประจำวัน เป็นข้อมูล

ที่ถูกเก็บไว้ในหลาย ๆ ที่ซึ่งทำให้การค้นหานั้นย่อมต้องมีความยากลำบากมากกว่าเดิม เพื่อนำไปสู่การวิเคราะห์และการออกแบบระบบงานในขั้นตอนต่อไป จากเทคนิคการค้นหาทั่วไปที่พบเห็นทำให้ต้องการทราบถึงความสามารถของเทคนิค N-Gram ว่ามีความสามารถเทียบเท่ากับเทคนิคในปัจจุบันที่มีความนิยมกันอยู่และเทคนิค N-Gram มีการทำงานในรูปแบบที่สามารถใช้งานได้อย่างมีประสิทธิภาพจึงจะมีสามารถทำการค้นคืนสารสนเทศได้อย่างดี

เทคนิค N-Gram ที่ได้ทำการศึกษามาแล้วนำมาใช้ในการค้นคืนเอกสารตามคำที่ผู้ใช้ต้องการโดยมีหลักการทำงานของ N-Gram การประมาณค่าความน่าจะเป็นของชุดอักขระ โดยการใช้เอ็น แกรมดังที่กล่าวมา คือ การใช้สมมติฐานของมาร์คอฟ (Markov assumption) ว่า การปรากฏของตัวอักษรตัวหนึ่งขึ้นกับตัวอักษรก่อนหน้าเพียง n-1 ตัว ซึ่งวิธีนี้มักนิยมใช้ในงานระบุภาษาของข้อความกันมาก เนื่องจากสามารถใช้เพื่อระบุภาษาได้อย่างมีประสิทธิภาพและเรียบง่ายกว่า โดยสามารถประมาณได้ดังนี้ ไบแกรม ไตรแกรมและควอดิแกรม โดยนำค่าที่ได้จากการค้นหาในแต่ล่ะ Gram มาคำนวณหาค่าที่มีความถี่มากที่สุดในแต่ล่ะช่วงเพื่อนำมาเปรียบเทียบกับค่าใน Gram อื่น ๆ เพื่อหาค่าที่มีความถี่มากที่สุดของ Gram ทั้งหมด แล้วจึงนำค่าของ Gram ที่ได้จากการคำนวณมาหาค่าของ Gram ที่มีความสัมพันธ์กับเอกสารที่มีอยู่ทั้งหมดเพื่อคำนวณหาเอกสารที่มีความถี่ในการค้นหาข้อมูลมากที่สุด จากการคำนวณค่าของ N-Gram ทั้งหมดเทคนิค N-Gram จะแบ่งคำออกเป็นจำนวน 1-4 Gram ดังภาพที่ 2

Keyword = computer

1-Gram :
c o m p u t e r

2-Gram :
c co om mp pu ut te er r

3-Gram :
c co com omp mpu put ute ter er r

4-Gram :
c co com comp ompu mput pute uter ter er r

ภาพที่ 2: การแบ่งคำด้วยเทคนิค N-Gram

ค่าความถี่ที่ได้ค่ามากที่สุดในการคำนวณคือ 4 Gram ในขั้นตอนการสร้างดัชนีจากเอกสาร วิธีการในการจัดทำดัชนีของคำหลักที่พบภายในเอกสาร โดยการกำหนดความสำคัญของคำเมื่อเทียบกับเอกสารทั้งระบบ โดยใช้ค่า TF/IDF (Term Frequency/Inverse Document Frequency) มีสูตรดังภาพที่ 3

$$w_{ij} = tf_{ij} \times \log_2 \frac{N}{n_j}$$

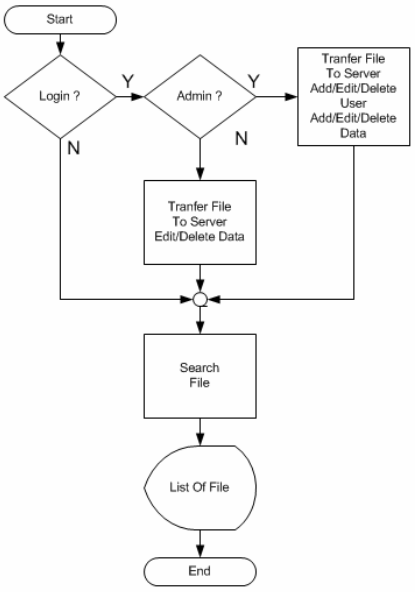
โดยที่

w_{ij}	คือค่าน้ำหนักของคำที่ i ในเอกสาร D_j
tf_{ij}	คือความถี่ของคำที่ i ในเอกสาร D_j
N	คือจำนวนเอกสารทั้งหมด
n_j	คือจำนวนเอกสารที่มีคำที่ i ปรากฏอยู่อย่างน้อย 1 ครั้ง

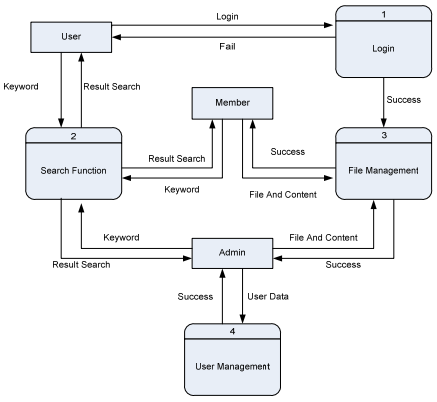
ภาพที่ 3: สูตรการคำนวณในการสร้างดัชนี

3.2 การออกแบบระบบงาน

จากการวิเคราะห์ระบบโดยรวมในเบื้องต้น จากนั้นได้ทำการออกแบบระบบเกี่ยวกับการดำเนินงานภายในระบบดังภาพที่ 4 และการไหลของข้อมูลภายในระบบทั้งหมด (Data Flow diagram) เพื่อแสดงรายละเอียดการทำงานของระบบดังภาพที่ 5



ภาพที่ 4: ขั้นตอนการทำงานของระบบ



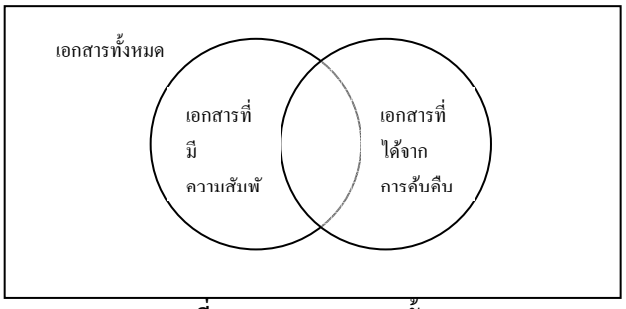
ภาพที่ 5: การไหลของข้อมูลในระบบ

3.3 การพัฒนาระบบ

จากการศึกษาและวิเคราะห์ระบบงาน ได้ทำการพัฒนาระบบตามที่ได้ทำการออกแบบไว้ทำการสร้างฐานข้อมูลโดยใช้ MySQL ในการเก็บข้อมูลต่าง ๆ ภายในระบบงานที่ได้พัฒนาขึ้นโดยการวางโครงสร้างฐานข้อมูลของระบบและทำการพัฒนาโปรแกรมด้วยภาษา HTML และ PHP เป็น Web-Application

3.4 การทดสอบระบบ

สำหรับการทดสอบประสิทธิภาพของระบบนั้นจะใช้การวัดความเที่ยง (Precision) และการเรียกกลับ (Recall) จุดประสงค์เพื่อใช้ตรวจสอบว่าระบบที่พัฒนาขึ้นสามารถค้นคืนเอกสารได้อย่างมีประสิทธิภาพ ดังภาพที่ 6



ภาพที่ 6: แสดงเซตเอกสารทั้งหมด

ความเที่ยงคืออัตราส่วนระหว่างเซตของเอกสารที่ได้จากการค้นคืนดังสมการ

$$precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved} \qquad (3\text{-}1)$$

จากสมการที่ 3-1 คำนวณหาค่าความแม่นยำได้จากเอกสารที่ค้นคืนทั้งหมดหารกับจำนวนเอกสารทั้งหมดที่มีความเกี่ยวข้อง

ความถูกต้องคืออัตราส่วนระหว่างเซตของเอกสารจากการค้นคืนเอกสารดังสมการที่ (3-2)

$$recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents} \qquad (3\text{-}2)$$

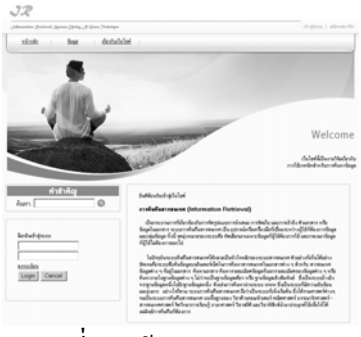
จากสมการที่ 3-2 คำนวณหาค่าความถูกต้องได้จากเอกสารที่ค้นคืนและเกี่ยวข้องทั้งหมดหารกับจำนวนเอกสารทั้งหมดที่ค้นคืนได้

จากรายละเอียดเกณฑ์ในประเมินประสิทธิภาพของระบบคือค่าของการเรียกกลับ (Recall) ต้องมีค่ามากกว่า 60 % จึงจะถือว่าระบบที่พัฒนาขึ้นมีประสิทธิภาพในระดับดี

4. ผลการดำเนินงาน

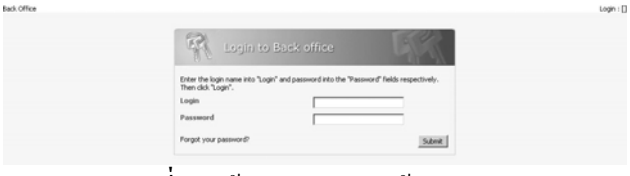
4.1 ผลการพัฒนาระบบ

สามารถพัฒนาระบบการค้นคืนสารสนเทศโดยใช้เทคนิค N-Gram ได้ผลลัพธ์ดังนี้



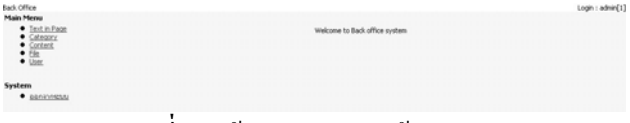
ภาพที่ 7: หน้าจอแรกของระบบ

จากภาพที่ 7 แสดงหน้าจอแรกที่ผู้ใช้งานระบบเข้ามา โดยแบ่งระดับผู้ใช้งานเป็น 3 ประเภท ได้แก่ ผู้ดูแลระบบ ผู้ใช้งานที่เป็นสมาชิกและผู้ใช้งานทั่วไป หน้าจอหลักสำหรับผู้ดูแลระบบจะแยกออกกับผู้ใช้งาน ดังภาพที่ 8



ภาพที่ 8: หน้าจอแรกสำหรับผู้ดูแลระบบ

เมื่อผู้ดูแลระบบทำการล็อกออนเข้าสู่ระบบจะเข้าสู่เมนูการปรับแต่งรายละเอียด ดังภาพที่ 9



ภาพที่ 9: หน้าจอเมนูสำหรับผู้ดูแลระบบ

และหน้าจอสำหรับผู้ใช้งานเมื่อทำการล็อกออนเข้าสู่ระบบจะปรากฏเมนูสำหรับผู้ใช้งาน ดังภาพที่ 10

ภาพที่ 10: หน้าจอเมนูสำหรับผู้ใช้งาน

และหน้าจอสำหรับการค้นคืนสารสนเทศจะทำการแสดงรายละเอียดการคำนวณด้วยเทคนิค **N**-Gram ผู้ใช้งานได้ทราบถึงวิธีการทำงานของระบบ ดังภาพที่ 11

ภาพที่ 11: หน้าจอแสดงรายละเอียดการคำนวณ

4.2 ผลการวัดประสิทธิภาพของระบบ

การทดสอบเพื่อวัดประสิทธิภาพของระบบการค้นคืนสารสนเทศโดยใช้เทคนิค N-Gram ได้ทำการทดสอบการค้นหาคำโดยวัดประสิทธิภาพด้วยค่าความถูกต้อง (Recall) และค่าความแม่นยำ (Precision) ซึ่งทดสอบจากผู้ใช้ทั่วไปจำนวน 30 คน ผลของการทดสอบการค้นหาคำสำคัญสามารถแสดงได้ดังนี้ ผลการประเมินประสิทธิภาพของระบบ โดยผู้ใช้ทั่วไปจำนวน 30 คน นั้นพบว่ามีคำที่ผู้ใช้ทั่วไปใช้ในการค้นหามีซ้ำกันมากซึ่งบางคำอาจจะใช้ภาษาต่างกันแต่ให้ความหมายเหมือนกัน ได้ผลดังตารางที่ 1

ตารางที่ 2 ผลการประเมินประสิทธิภาพของระบบ

คำที่เกิดขึ้น	เอกสาร			Precision	Recall
	ค้นคืน	เกี่ยวข้อง	ค้นคืนและเกี่ยวข้อง		
คอมพิวเตอร์	46	40	35	0.76	0.87
เทคโนโลยี	33	31	26	0.78	0.83
โปรแกรม	25	27	19	0.76	0.70
เทคนิค	17	21	16	0.94	0.76
สารสนเทศ	31	27	21	0.67	0.77
สรุป				0.78	0.78

จากตารางที่ 2 สรุปผลการวัดประสิทธิภาพจากการทดสอบโดยผู้ใช้ทั่วไปจากคำในการค้นหาสารสนเทศพบว่าประสิทธิภาพของการค้นหามีค่าเฉลี่ยความถูกต้อง 78 เปอร์เซ็นต์ ค่าเฉลี่ยความแม่นยำ 78 เปอร์เซ็น ดังนั้นระบบที่พัฒนาขึ้นมีประสิทธิภาพในการค้นคืนอยู่ในระดับดี

5. สรุปผล

การพัฒนาระบบการค้นคืนสารสนเทศโดยใช้เทคนิค N-Gram ถือเป็นการผสมผสานเทคนิคการค้นคืนสารเทศ (Information Retrieval) กับการใช้เทคนิคการตัดคำของ N-Gram เป็นการนำเทคนิคมาเพิ่มช่วยเพิ่มความสามารถในการค้นคืนสารสนเทศให้ตรงตามความต้องการของผู้ใช้งานมากยิ่งขึ้น โดยผลสรุปของการประเมินประสิทธิภาพระบบการค้นคืนอยู่ในระดับดี

6. เอกสารอ้างอิง

[1] ชูชาติ หฤไชยะศักดิ์.โปรแกรมสำหรับพัฒนาระบบค้นคืนสารสนเทศภาษาไทย, National and Computer Technology Center (NECTEC), Available online http://sansarn.com, 2537.

[2] ธนารักษ์ ธีระมั่นคง. “โครงการวิจัยพัฒนาต้นแบบระบบฐานความรู้ด้านการแพทย์ในประเทศไทย”, สถาบันเทคโนโลยีนานาชาติสิรินธร มหาวิทยาลัยธรรมศาสตร์. Available online http://kindml.siit.tu.ac.th/~mdproject/MDweb/main.html, 2539.

[3] กิตติชน แม้นสมุทร. การสร้าง Search Engine, 2550.

[4] Michael K. Brown. Stochastic Language Models (N-Gram) Specification, W3C, 2001.

[5] อัครพล เอกวงศ์อนันต์. การระบุคำไทยและคำทับศัพท์ด้วยแบบจำลองเอ็นแกรม, วิทยานิพนธ์ อส.ม., กรุงเทพฯ : จุฬาลงกรณ์มหาวิทยาลัย, 2548.

[6] นิพนธ์ เจริญกิจการ. ระบบการค้นคืนเอกสารภาษาไทยด้วยเทคนิคชั้นสูง ระยะที่ 2, Advanced Thai Text Retrieval System, 2544.

[7] Sasiporn Usanavasin. Non-Dictionary-Based Thai Word Segmentation Using Decision Trees, Sinrindhorn International Institute of Technology, 2003.