การจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริงแนวใหม่

A NEW SUFFIX TREE CLUSTERING ALGORITHM

จงกล จันทร์เรือง
JONGKOL JANRUANG

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ
บัณฑิตวิทยาลัย
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2549
ISBN 974-9809-02-2

การจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งแนวใหม่

A NEW SUFFIX TREE CLUSTERING ALGORITHM

จงกล จันทร์เรือง

JONGKOL JANRUANG

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2549

ISBN 974-8308-02-2

A NEW SUFFIX TREE CLUSTERING ALGORITHM

JONGKOL JANRUANG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2006

ISBN 974-8308-02-2

หัวข้อวิทยานิพนธ์	การจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งแนวใหม่
นักศึกษา	นางสาวจงกล จันทร์เรือง
รหัสนักศึกษา	46066728
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
พ.ศ.	2549
อาจารย์ที่ปรึกษาวิทยานิพนธ์	รศ.ดร.วรพจน์ กรีสุระเดช

บทคัดย่อ

การจัดกลุ่มผลการสืบค้นด้วยเทคนิค Suffix Tree Clustering (STC) เป็นเทคนิคการจัดกลุ่มผลการสืบค้นที่ทำงานเร็วและมีความต่อเนื่อง (Fast Incremental Clustering) ซึ่งเหมาะสำหรับการจัดกลุ่มผลการสืบค้น ที่ต้องการเปลี่ยนแปลงของผลการจัดกลุ่มตามผลการสืบค้นที่มีการเปลี่ยนแปลงตลอดเวลา (Dynamic or on-the-fly Clustering) แต่อย่างไรก็ตามเทคนิค Suffix Tree Clustering (STC) ก็ยังไม่เพียงพอสำหรับผลการจัดกลุ่มและป้ายชื่อกลุ่ม (Cluster Label) ที่ถูกตัดขาดจากการทำงานร่วมกับเทคนิค n-gram รายงานฉบับนี้ เราได้นำเสนอแนวทางการจัดกลุ่มผลการสืบค้นด้วย Suffix Tree Clustering (STC) แนวใหม่ ด้วยการรวมกลุ่มพื้นฐานจากการเชื่อมป้ายชื่อกลุ่มเอกสารจากการพิจารณาการมีอยู่ร่วมกันของคำที่ปรากฏในป้ายชื่อกลุ่มของคู่กลุ่มเอกสาร และพิจารณาความคล้ายคลึง (similarity) ของกลุ่มเอกสารเพื่อเชื่อมกลุ่มจากเงื่อนไข " สมาชิกของกลุ่มย่อยมากกว่าหนึ่งเอกสารเป็นสมาชิกของกลุ่มใหญ่ในลักษณะที่ตำแหน่งของคำเชื่อมต่อกัน " ซึ่งเป็นเทคนิคที่จะช่วยเพิ่มจำนวนเอกสารที่ตรงกับความต้องการ (Relevance Web Document) และลดจำนวนเอกสารที่ทับซ้อนกัน (Overlap Web Document) ของกลุ่มเอกสารได้มากกว่าการทำงานในรูปแบบเดิมของการจัดกลุ่มผลการสืบค้นด้วยเทคนิค Suffix Tree Clustering (STC)

Thesis Title	A New Suffix Tree Clustering Algorithm
Student	Miss. Jongkol Janruang
Student ID.	46066728
Degree	Master of Science
Program	Information Technology
Year	2006
Thesis Advisor	Assoc. Prof. Dr. Worapoj Kreesuradej

ABSTRACT

Suffix Tree Clustering (STC) is incremental and linear time (in the document collection size) algorithm for dynamic or on-the-fly web search results clustering. However, STC is inadequate since they generate interrupted cluster label due to using n-gram technique. In this paper, we propose a new Suffix Tree Clustering algorithm that combining base cluster by cluster label combination to become new cluster and cluster label, consider the combined cluster label from word which joint appear within the cluster label of pair cluster is subset of member document within cluster and consider to adjoin of word location in document using word together. It's provides more relevance and less overlap of web document clusters than conventional Suffix Tree Clustering algorithms.

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จลุล่วงได้เป็นอย่างดี ด้วยความกรุณาให้คำปรึกษา พร้อมทั้งคำชี้แนะแนวทางการแก้ปัญหาของงานวิจัยตั้งแต่เริ่มต้นจนเสร็จสมบูรณ์ ตลอดจนให้ความรู้และประสบการณ์ที่ดีแก่ข้าพเจ้า จาก รศ.ดร. วรพจน์ กรีสุระเดช ซึ่งเป็นอาจารย์ผู้ควบคุมวิทยานิพนธ์ ผู้วิจัยรู้สึกซาบซึ้งในความอนุเคราะห์ของท่านและขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณบิดา มารดา คุณตา คุณยาย ญาติพี่น้อง และเพื่อนๆ ของข้าพเจ้า ซึ่งให้กำลังใจและความช่วยเหลือโดยตลอด

ขอขอบคุณ พี่ๆ น้องๆ และเพื่อนๆ คณะเทคโนโลยีสารสนเทศที่ให้กำลังใจ คำปรึกษา และให้ความช่วยเหลือในการทำวิทยานิพนธ์นี้โดยตลอด

สุดท้ายขอขอบคุณคณาจารย์ และเจ้าหน้าที่คณะเทคโนโลยีสารสนเทศที่ให้ความรู้ และความช่วยเหลือในทุกๆ ด้าน

สำหรับคุณค่าและประโยชน์อันใดที่เกิดจากวิทยานิพนธ์นี้ ข้าพเจ้าขอมอบให้กับบิดา มารดา คุณตา คุณยาย และอาจารย์ที่ปรึกษา ซึ่งเป็นที่รักและเคารพยิ่ง ตลอดจนผู้มีพระคุณทุกท่าน

จงกล จันทร์เรือง

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันระบบเครือข่ายใยแมงมุม (World Wide Web) เต็มไปด้วยข้อมูลที่มีการเปลี่ยนแปลง (Dynamic Hypertext Information) จำนวนมหาศาล ซึ่งข้อมูลเหล่านั้นมีความสัมพันธ์กันกระจายอยู่เป็นล้านๆเอกสาร (Web Pages) และมีอิทธิพลต่อชีวิตเราเพิ่มมากขึ้น ส่งผลให้ต้องมีการพัฒนาและปรับปรุงระบบสืบค้น (Search Engine) เพื่อบริการและอำนวยความสะดวกให้กับผู้สืบค้น แต่ระบบสืบค้นหลายระบบ เช่น Google.com , Yahoo.com และ Msn.com ให้ผลการสืบค้นเป็นลำดับรายการยาว ๆ ซึ่งประกอบด้วย หัวข้อ (Titles) , คำอธิบายสั้น ๆ (Snippets) และชื่อเรียกที่อยู่ของเว็บ (URL : Uniform Research Location) อย่างเป็นลำดับ จากจำนวนเอกสารที่เพิ่มขึ้นอย่างมหาศาลนี้ (Stanislaw Osinski(2003) : ระบุว่า ในปี 2001 google มีจำนวนเอกสารมากกว่า 1.35 พันล้านหน้า และในปี 2003 มีจำนวนเอกสารเพิ่มมากขึ้นมากกว่า 3.08 พันล้านหน้า) ส่งผลให้ลำดับรายการของผลการสืบค้นมีจำนวนเพิ่มมากขึ้นด้วย ดังตัวอย่างการค้นหาคำว่า " data mining " ระบบ google.com แสดงผลจำนวน 54,800,000 รายการ , yahoo.com แสดงผลจำนวน 34,900,000 รายการ และ msn.com แสดงผลจำนวน 2,794,005 รายการ ถ้าผู้สืบค้นต้องการเอกสารคำว่า " data mining " ที่เกี่ยวข้องกับคำว่า " CRM " ผู้สืบค้นต้องไปค้นหาเอกสารในลำดับรายการดังกล่าว ซึ่งอาจอยู่ในลำดับรายการที่ 34 , 36, 108, หรือ 117 จากลำดับรายการผลการสืบค้นเหล่านี้ ทำให้ผู้สืบค้นต้องเสียเวลาในการค้นหาข้อมูลที่ต้องการหลังจากที่ระบบสืบค้นได้แสดงผลการสืบค้นและโดยปกติผู้สืบค้นมักจะค้นหาข้อมูลตามลำดับรายการต้น ๆ เท่านั้น ทำให้ผู้สืบค้นอาจไม่พบข้อมูลที่ต้องการ นอกจากปัญหาลำดับรายการผลการสืบค้นมีจำนวนมหาศาลแล้ว ระบบสืบค้นในลักษณะของ Directories Search Engine เช่น yahoo.com หรือ dmoz.com ซึ่งให้บริการสืบค้นตามหัวข้อที่สนใจ โดยมีการจัดเตรียมป้ายชื่อกลุ่มไว้ให้บริการแก่ผู้สืบค้น นั่นคือผู้สืบค้นต้องทราบว่าเอกสารที่ต้องการนั้นอยู่ในหัวข้อใด และป้ายชื่อกลุ่มเอกสารที่จัดเตรียมไว้ให้นั้นไม่ใช่คำหรือวลีที่เป็นตัวแทนจากเอกสารโดยตรง

จากปัญหาการเข้าถึงผลการสืบค้น และปัญหาป้ายชื่อกลุ่มเอกสารของ Directories Search Engine ไม่ใช่ตัวแทนของเอกสารโดยตรงและมีลักษณะเป็น Static Clustering จึงมีการพัฒนาระบบการจัดกลุ่มผลการสืบค้นแบบอัตโนมัติ (Dynamic Clustering) โดยใช้ snippets เป็นตัวแทนเอกสาร เช่นระบบจัดกลุ่มผลการสืบค้นชื่อ Suffix Tree Clustering (STC) [1][2][3][4][5] ซึ่งมีจุดเด่นในลักษณะที่เป็น Fast Incremental and Dynamic Clustering แต่ยังมีข้อด้อยซึ่งประกอบด้วย

1. ขนาดความสูงของ Suffix Tree มีระดับไม่ชัดเจนและควบคุมไม่ได้ เนื่องจากจำนวนคำของแต่ละ snippets ไม่เท่ากันและไม่คงที่ ทำให้การสร้าง suffix tree มีความยุ่งยาก และเมื่อมีการนำกระบวนการแตกประโยคเพื่อสร้างรายการลำดับของคำ n คำ (n-gram) มาใช้ร่วมกับการสร้าง suffix tree [5] เพื่อควบคุมความสูงของ suffix tree ให้ง่ายต่อการพัฒนาและการจัดการ ทำให้มีผลกระทบต่อความถูกต้องของป้ายชื่อกลุ่มเพราะมันถูกตัดขาดด้วยขนาดของ n-gram

2. การค้นหา Final Cluster ด้วยการรวมกลุ่มพื้นฐาน (Base Cluster) โดยการคำนวณค่าความคล้ายคลึง (Similarity) ของกลุ่มเอกสารเป็นเปอร์เซ็นต์จากจำนวนสมาชิกที่ปรากฏเหมือนกันของคู่กลุ่มเอกสาร ดังนั้นความคล้ายคลึงของกลุ่มจะขึ้นอยู่กับขนาดของกลุ่มที่ใกล้เคียงกัน ส่งผลให้ได้รับกลุ่มจำนวนมากและมีขนาดใหญ่ และอัตราการทับซ้อนของเอกสารสูงเกินไป จนเป็นเหตุให้มีอัตราของเอกสารที่ไม่ตรงกับความต้องการ (Irrelevant document) ภายในกลุ่มสูงตามไปด้วย

จากปัญหาการจัดกลุ่มผลการสืบค้นด้วย STC นี้ จึงมีการพัฒนาการจัดกลุ่มผลการสืบค้นด้วย Suffix Tree Clustering (STC) แนวใหม่ โดยมุ่งที่จะพัฒนาเทคนิคการจัดกลุ่มผลการสืบค้นด้วยแนวทางของ STC ให้สามารถรวมกลุ่มพื้นฐานจากการเชื่อมป้ายชื่อกลุ่มที่ได้จากการสร้าง suffix tree ร่วมกับเทคนิค n-gram เพื่อแก้ปัญหาป้ายชื่อกลุ่มไม่สมบูรณ์ อันเนื่องมาจากถูกตัดขาดด้วยขนาดของ n-gram และแก้ปัญหาการรวมกลุ่มพื้นฐานที่พิจารณาเฉพาะจำนวนเอกสารภายในกลุ่มตามแนวทางของ STC ที่ทำให้ได้รับกลุ่มจำนวนมากและกลุ่มมีขนาดใหญ่ ส่งผลให้กลุ่มมีอัตราการทับซ้อนของเอกสารสูงเกินไป

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

งานวิจัยนี้ มีวัตถุประสงค์เพื่อพัฒนาการทำงานของระบบการจัดกลุ่มผลการสืบค้นแบบ Suffix Tree Clustering (STC) โดยการนำเทคนิคการแตกประโยคเพื่อสร้างรายการลำดับของคำ n คำ (n-gram) มาช่วยในการสร้าง suffix tree ให้ง่ายและสะดวกต่อการจัดการ พร้อมกับการรวมกลุ่มพื้นฐาน (Base Cluster) ด้วยการพิจารณาการมีอยู่ร่วมกันของคำที่ปรากฏในป้ายชื่อกลุ่ม (Join Cluster Label) ของคู่กลุ่มเอกสาร และพิจารณาความคล้ายคลึงของกลุ่มเอกสารเพื่อการเชื่อมป้ายชื่อกลุ่มตามเงื่อนไขจำนวนเอกสารที่ใช้กำหนดกลุ่มพื้นฐาน ร่วมกับการพิจารณาตำแหน่งของคำที่เชื่อมต่อกันในแต่ละเอกสารที่เป็นสมาชิกร่วมกันของทั้งสองกลุ่ม เพื่อค้นหาป้ายชื่อกลุ่มที่ถูกตัดขาดด้วยขนาดของ n-gram และลดจำนวนพร้อมกับขนาดของกลุ่มเอกสาร เพื่อให้อัตราการทับซ้อนของเอกสารลดลง ทำให้จำนวนเอกสารที่ไม่ตรงกับความต้องการ (Irrelevant document) ภายในกลุ่มเอกสารลดลง ส่งผลให้คุณภาพกลุ่มสูงขึ้น

1.3 ทฤษฎีหรือแนวคิดที่ใช้ในการวิจัย

งานวิจัยนี้ได้นำทฤษฎีและเทคนิคต่างๆมาประยุกต์ใช้ประกอบด้วย

1. เทคนิคการจัดกลุ่มผลการสืบค้นชื่อ Suffix Tree Clustering (STC) ซึ่งมีคุณสมบัติที่ดีในด้านความเร็ว คือเป็น linear time ตามขนาดของข้อมูล , ด้านความเป็น overlap และ incremental clustering คือสามารถจัดกลุ่มผลการสืบค้นเพิ่มขึ้นได้อย่างต่อเนื่อง และเอกสารที่เข้ามานั้นสามารถปรากฏได้ในหลายกลุ่มเอกสาร และผลการจัดกลุ่มจะเปลี่ยนแปลงตามผลการการสืบค้นที่ได้รับ คือมีลักษณะเป็น Dynamic หรือ on-the-fly Clustering

2. เทคนิคการแตกประโยคเพื่อสร้างรายการลำดับของคำ n คำ (n-gram) ที่ช่วยให้การสร้าง suffix tree ใช้หน่วยความจำน้อยลง และสามารถควบคุมความสูงของ tree ได้ ทำให้ง่ายต่อการพัฒนาและการจัดการกับโครงสร้างข้อมูลชื่อ Suffix Tree

1.4 ขอบเขตของการวิจัย

งานวิจัยนี้นำเสนอโมเดลที่ประยุกต์ใช้โครงสร้างข้อมูลชื่อ Suffix Tree ร่วมกับเทคนิคของ n-gram เพื่อให้สามารถพัฒนาและจัดการกับ suffix tree ได้ง่ายขึ้นและลดปริมาณการใช้หน่วยความจำ พร้อมกับการค้นหาป้ายชื่อกลุ่มที่ถูกตัดขาดด้วยขนาดของ n-gram โดยการรวมกลุ่มพื้นฐานจากการเชื่อมป้ายชื่อกลุ่ม ตามเงื่อนไขการมีอยู่ร่วมกันของคำและตำแหน่งคำในเอกสารเดียวกันที่เชื่อมต่อกันได้ของคู่กลุ่มเอกสาร

1.5 ขั้นตอนของการศึกษา

ในขั้นตอนของการศึกษานี้ ได้แสดงลำดับการทำงานตั้งแต่เริ่มต้นจนถึงสิ้นสุดการทำงานดังรายละเอียดต่อไปนี้

1.5.1 ศึกษาทฤษฎีและงานวิจัยจากเอกสารบทความต่าง ๆ ที่เกี่ยวข้องกับการทำงานวิจัย

1.5.2 กำหนดหัวข้อ วัตถุประสงค์ และขอบเขตการทำงานวิจัย

1.5.3 วิเคราะห์อัลกอริทึมและออกแบบโมเดลใหม่

1.5.4 พัฒนาโปรแกรม โดยใช้ซอฟท์แวร์ Java Netbeans

1.5.5 เตรียมข้อมูลที่ใช้งานจริง เพื่อนำมาทดสอบการทำงานโมเดล

1.5.6 ทดลองกับข้อมูลที่ใช้งานจริง พร้อมทั้งวัดประสิทธิภาพการทำงานของโมเดล

1.5.7 รวบรวมผลการทดลองจากการทำงานของโมเดล

1.5.8 วิเคราะห์และสรุปผลการทดลอง

1.5.9 เรียบเรียงเอกสารประกอบวิทยานิพนธ์

การทำเว็บไมนิ่ง (Web Mining) หรือ การค้นพบประโยชน์โดยอัตโนมัติ (Automatic discovery) ของข้อมูลที่น่าสนใจ ซึ่งอาจซ่อนอยู่ในเอกสารที่ปรากฏบนระบบเครือข่ายใยแมงมุม ได้กลายมาเป็นสิ่งสำคัญในการสกัดหาความรู้จากข้อมูล (Data Mining) จึงมีการคิดและพัฒนากระบวนการการจัดการหลายชนิด เช่น การค้นหา (Crawling) , การจัดกลุ่ม(Clustering) และการจำแนกประเภท (Classification) เพื่อให้ได้ความรู้ที่เป็นประโยชน์จากข้อมูลเอกสารที่เผยแพร่บนเครือข่ายใยแมงมุม (Dynamic Hypertext Information) [6] งานวิจัยนี้มุ่งพัฒนาด้านการจัดกลุ่มผลการสืบค้น โดยประยุกต์ใช้หลักการทำงานของ Suffix Tree Clustering ซึ่งมีลักษณะการจัดกลุ่มแบบเปลี่ยนแปลงตามผลการสืบค้น (Dynamic Clustering หรือ on-the-fly Clustering) ร่วมกับเทคนิคการแตกประโยคเพื่อสร้างรายการลำดับของคำ n คำ (n-gram)

ในบทที่ 2 จะกล่าวถึง ทฤษฎีต่างๆ ที่เกี่ยวข้องกับการทำงานวิจัยที่นำเสนอนี้

บทที่ 2
ทฤษฎีพื้นฐาน และงานวิจัยที่เกี่ยวข้อง

การจัดกลุ่มผลการสืบค้นบนเครือข่ายใยแมงมุม (Web Search Results Clustering) มีวัตถุประสงค์ เพื่ออำนวยความสะดวกกับผู้สืบค้นให้สามารถเข้าถึงข้อมูลที่ตรงกับความต้องการได้สะดวกรวดเร็วยิ่งขึ้น ประกอบด้วยรายละเอียดดังต่อไปนี้

2.1 โครงสร้างสถาปัตยกรรมระบบจัดกลุ่มผลการสืบค้น

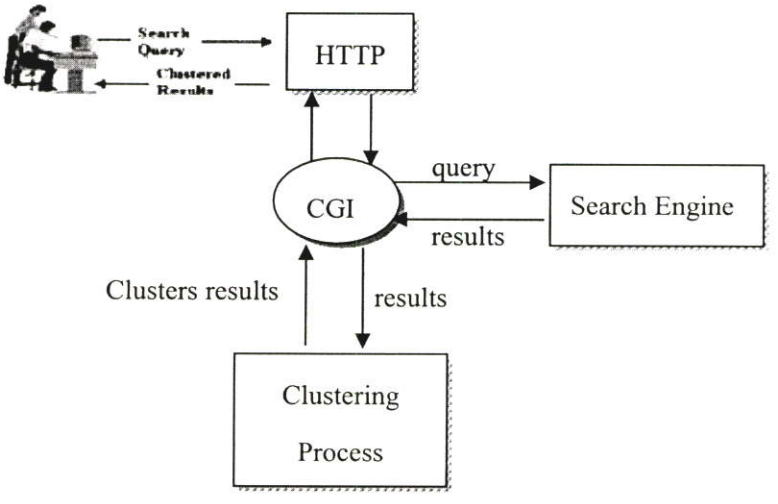
โครงสร้างสถาปัตยกรรมของระบบจัดกลุ่มผลการสืบค้น ประกอบด้วย

1. HTTP Server หรือ Web Server จะทำหน้าที่ในการรับคำสืบค้น (Search Query) จากผู้สืบค้น และส่งผลการจัดกลุ่มผลการสืบค้นให้กับผู้สืบค้น

2. Common Gateway Interface (CGI) เป็นโปรแกรมที่ทำหน้าที่ควบคุมดูแลการเข้าถึงข้อมูลและเป็นตัวกลางบนระบบเครือข่ายใยแมงมุม

3. Search Engine เป็นระบบสืบค้นที่ระบบจัดกลุ่มผลการสืบค้นต้องการนำผลการสืบค้นมาทำการจัดกลุ่มผลการสืบค้น เช่น Google , Yahoo เป็นต้น

4. Clustering Process เป็นกระบวนการทำงานของการจัดกลุ่มผลการสืบค้น ด้วยเทคนิคต่างๆ เช่น Suffix Tree Clustering (STC) , Lingo และ Semantic, Hierarchical, Online Clustering of Web Search (SHOC) เป็นต้น



รูปที่ 2.1 แสดงโครงสร้างทิศทางการไหลของข้อมูลในระบบการจัดกลุ่มผลการสืบค้น

2.2 การจัดกลุ่มผลการสืบค้น

การจัดกลุ่มผลการสืบค้น (Web Search Results Clustering) บนระบบเครือข่ายใยแมงมุม คือการนำผลการสืบค้นที่มีความสัมพันธ์หรือเหมือนกัน (Similarity) มารวมอยู่ในกลุ่มเดียวกัน เพื่ออำนวยความสะดวกในการเข้าถึงข้อมูลที่ตรงกับความต้องการให้กับผู้สืบค้น และช่วยให้ผู้สืบค้นค้นพบประโยชน์ที่ซ่อนอยู่ในผลการสืบค้น แบ่งออกเป็น 4 ลักษณะ [7] คือ

2.2.1 การจัดกลุ่มด้วยคำเดี่ยวในลักษณะลำดับรายการ

การจัดกลุ่มด้วยคำเดี่ยวในลักษณะลำดับรายการ (Single Word and Flat Clustering) คือ ลักษณะการนำคำเดี่ยว (Single Word) มาเป็นตัวกำหนดคุณลักษณะหรือตัวแทนเอกสาร เพื่อนำไปจัดกลุ่มในลักษณะของลำดับรายการ (Flat Clustering) เช่น WEBCAT [8] ใช้เทคนิคของ data mining มาช่วยในการจัดกลุ่มเอกสารที่มีคุณสมบัติเหมือนกัน และใช้ keyword ที่มีความสัมพันธ์โดยตรงกับ query เป็นตัวกำหนดกลุ่ม มี 3 ขั้นตอนการทำงาน [9] คือ

1. Pre-processing เป็นการเตรียมข้อมูลเพื่อนำไปใช้ในการจัดกลุ่ม ประกอบด้วย
   1.1 การกำจัดตัวเลข เครื่องหมายวรรคตอน และสัญลักษณ์ต่าง
   1.2 การกำจัดคำที่ไม่มีความหมาย เช่น คำนำหน้านามในภาษาอังกฤษ (เช่น An , The , A) คำบุพบท (เช่น " for ", " of ") คำสรรพนาม (เช่น " you " , " this ") เป็นต้น
   1.3 การแปลงคำให้อยู่ในรูปของรากศัพท์เดิม เทคนิคที่นิยมใช้คือ Stemming Algorithm เป็นการตัด prefix , suffix , การทำให้อยู่ในรูปเอกพจน์
2. กระบวนการจัดกลุ่มใช้เทคนิค Transactional k-Means
3. การแสดงผลให้กับผู้สืบค้น แบ่งออกเป็น 2 ส่วนคือ
   3.1 Summary Table เป็นตารางแสดงผลการจัดกลุ่มทั้งหมด
   3.2 Clusters คือกลุ่มเอกสารที่สามารถค้นพบเอกสารในลำดับรายการของคำ

การใช้คำเพียงคำเดียวในการจัดกลุ่ม ทำให้กลุ่มมีขนาดใหญ่ และได้รับกลุ่มจำนวนมาก เพราะคำเดี่ยวมีอำนาจในการจำแนกกลุ่มน้อย และยากที่ผู้สืบค้นจะเข้าใจความหมายของกลุ่ม

2.2.2 การจัดกลุ่มด้วยวลีในลักษณะลำดับรายการ

การจัดกลุ่มด้วยวลีในลักษณะลำดับรายการ (Sentence and Flat Clustering) คือ ลักษณะการนำ Sentence หรือ Phrase มาเป็นตัวกำหนดคุณลักษณะหรือตัวแทนเอกสาร เพื่อนำไปจัดกลุ่มตามลักษณะของ Flat Clustering ปัจจุบันมีหลายระบบที่มีการพัฒนาในลักษณะ Sentence and Flat Clustering ดังตัวอย่างงานวิจัยต่อไปนี้

2.2.2.1 Carrot^2 and Language Properties in Web Search Results Clustering [10] เป็นการพัฒนาโครงร่างระบบจัดกลุ่มผลการสืบค้น โดยการนำเทคนิคการจัดกลุ่มผลการสืบค้นมากกว่าหนึ่งเทคนิคมาประกอบอยู่ในโครงร่างเดียวกันประกอบด้วย Suffix Tree Clustering (STC) , Agglomerative Hierarchical Clustering(AHC) และ Lingo เพื่อให้ผู้สืบค้นสามารถเลือกระบบจัดกลุ่มที่ต้องการได้ ระบบ Carrot^2 จะทำงานตามแนวทางของของระบบ Carrot ที่ใช้ Suffix Tree Clustering (STC) และทำงานบนพื้นฐานของภาษาอังกฤษกับภาษาโปแลนด์ ดังแสดงในรูปที่ 2.2

รูปที่ 2.2 แสดงผลการทำงานของระบบ Carrot^2 [11]

2.2.2.2 An Algorithm for Clustering of Web Search Results พัฒนาระบบชื่อ Lingo ซึ่งใช้พื้นฐานแนวคิดแบบ description – oriented – algorithm คือการพัฒนาคำอธิบายกลุ่ม (Cluster Label) ให้มีความรัดกุม สามารถอ่านและเข้าใจได้ง่าย ลักษณะการทำงานจะต่างจากกระบวนการทำงานอื่น คือ ระบบจะค้นหาคำอธิบายกลุ่มก่อน แล้วจึงจัดกลุ่มตามคำอธิบายกลุ่มที่เลือกมาเป็นตัวแทนของกลุ่ม ประกอบด้วย 5 ขั้นตอน [12] คือ

1. Pre-processing ประกอบด้วย

1.1 Text Filtering เป็นการกรองสิ่งไม่ต้องการออกไป เช่น HTML tags , สัญลักษณ์ต่างๆ ที่ไม่ใช่คำที่มีความหมาย เช่น “#” , “.” , “/” เป็นต้น

1.2 Language Identification การระบุภาษาเพื่อเลือกใช้ Stemming ที่แตกต่างกันเมื่อภาษาของเอกสารที่เข้ามาต่างกัน เพราะระบบ Lingo พัฒนาเพื่อรองรับ 2 ภาษา คือภาษาอังกฤษและภาษาโปแลนด์

1.3 Stemming จะแยกทำงานตามภาษาที่ถูกระบุมาจากขั้นตอน Language Identification ถ้าเป็นภาษาอังกฤษจะใช้ Porter Stemmer Algorithm แต่ถ้าเป็นภาษาโปแลนด์จะ

ใช้ Lametyzator Algorithm ซึ่งเป็น Stemming สำหรับภาษาโปแลนด์

1.4 Stop Words Marking ระบบจะไม่ลบ Stop Words ทิ้ง เพราะ Stop Words อาจช่วยให้ผู้สืบค้นเข้าใจความหมายของ phrase มากขึ้น เช่น ถ้าเปรียบเทียบ “Chamber Commerce” กับ “ Chamber of Commerce ” ประโยคหลังจะให้ความหมายมากกว่า

2. Feature extraction การค้นหา phrase และ word โดยใช้ Suffix Array

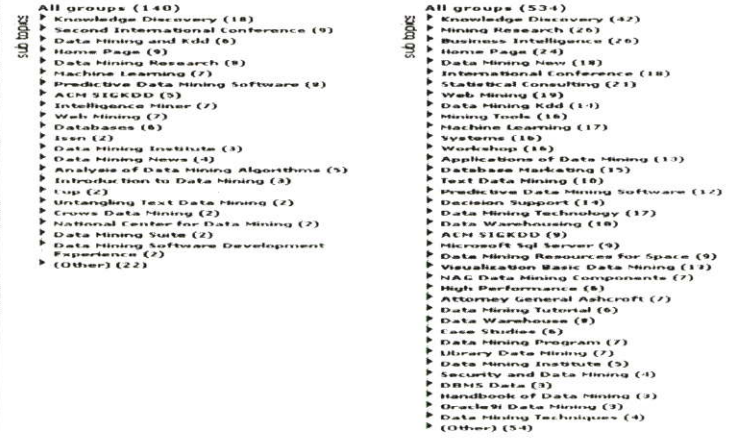
3. Cluster Label Induction คือการค้นหาป้ายชื่อกลุ่มที่แท้จริง โดยใช้เทคนิค Singular Value Decomposition(SVD) และ Latent Semantic Indexing(LSI) ใช้ค่า tfidf ของแต่ละ word และ phrase มาสร้าง matrix เพื่อคำนวณหาป้ายชื่อกลุ่มที่มีความหมาย

4. Cluster Content Discovery คือการจัดกลุ่มให้ Snippets เข้าไปตามป้ายชื่อกลุ่มที่สัมพันธ์กันด้วยเทคนิค Vector Space Model ถ้าค่าความเหมือนสูงกว่าค่าพื้นฐานจะนำ snippets เข้ากลุ่มในป้ายชื่อกลุ่มนั้น และสร้างกลุ่มอื่นๆ (other cluster) สำหรับ snippets ที่ไม่มีความสัมพันธ์กับป้ายชื่อใดเลย

5. Final Cluster Formation คำนวณคะแนนกลุ่มเพื่อจัดลำดับความสำคัญของกลุ่มในการแสดงผล ดังสมการ cluster_score = label_score*number_count

จากการศึกษารายงานวิจัยของ Lingo พบว่า

1. ระบบ Lingo มีการทับซ้อน (Overlap) ของเอกสารในแต่ละกลุ่ม ดังแสดงตัวอย่างในรูปที่ 2.3



	Left	Right
Query	*data mining*	*data mining*
Source search engine	Google	Google
Input snippets count	100	300

รูปที่ 2.3 แสดงผลการทับซ้อนของเอกสารในระบบ Lingo

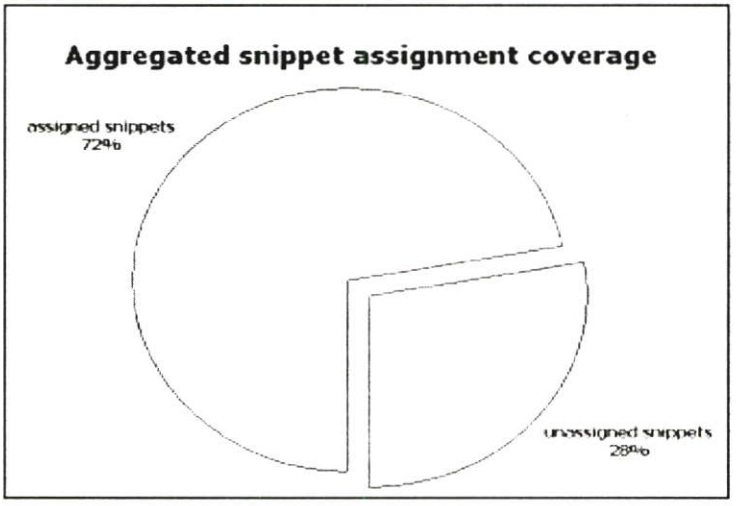
จากรูปที่ 2.3 คิดค่า Overlap ตามสูตร $v = a / s - 1$ จะได้

$$v = (140 / 100) - 1 = 0.40$$

เมื่อ a คือ จำนวนรวมของ snippet ที่ถูกจัดกลุ่มรวมทั้งกลุ่มอื่นๆด้วย

s คือ จำนวนรวมของ snippet ที่นำมาจัดกลุ่ม

2. ระบบ Lingo มีระดับของ Coverage ดังแสดงในรูปที่ 2.4



รูปที่ 2.4 แสดงผลความคลอบคลุมเอกสารในระบบ Lingo

จากรูปที่ 2.4 แสดงผลการจัดกลุ่มแบบ Lingo มีค่า Coverage เฉลี่ย 28% ดังสมการ

$$c = (s - o) / s$$

เมื่อ o คือ จำนวนรวมของ snippet ที่ถูกจัดกลุ่มในกลุ่มอื่นๆ

s คือ จำนวนรวมของ snippet ที่นำมาจัดกลุ่ม

นอกจากปัญหาเรื่อง Overlap และ Coverage แล้ว Lingo ยังมีข้อด้อยซึ่งประกอบด้วย

1. ลักษณะการจัดกลุ่มและแสดงผลเป็น Flat Clustering ซึ่งต้องปรับปรุงและออกแบบให้มีลักษณะเป็น Hierarchical Groupings เพราะมีบางกลุ่มที่มีขนาดใหญ่เกินไป

2. การทำงานของระบบยังไม่รองรับการจัดกลุ่มในลักษณะ Incremental Processing

3. การกำหนดจำนวนกลุ่มที่แน่นอน (Fixed Number of Cluster) เพราะระบบใช้การจัดกลุ่มที่ต้องกำหนดป้ายชื่อกลุ่มแล้วจึงทำการจัดกลุ่ม ทำให้การจัดกลุ่มมีปัญหาเรื่องจำนวนของกลุ่มที่ต้องกำหนดไว้ล่วงหน้า

4. อัตราการ Coverage มีหลายเอกสารที่ต้องถูกจัดให้อยู่ในกลุ่มอื่นๆ (Other Topic) เพราะมีบางเอกสารที่ไม่มีความสัมพันธ์กับป้ายชื่อกลุ่มที่ได้กำหนดไว้ล่วงหน้าแล้ว

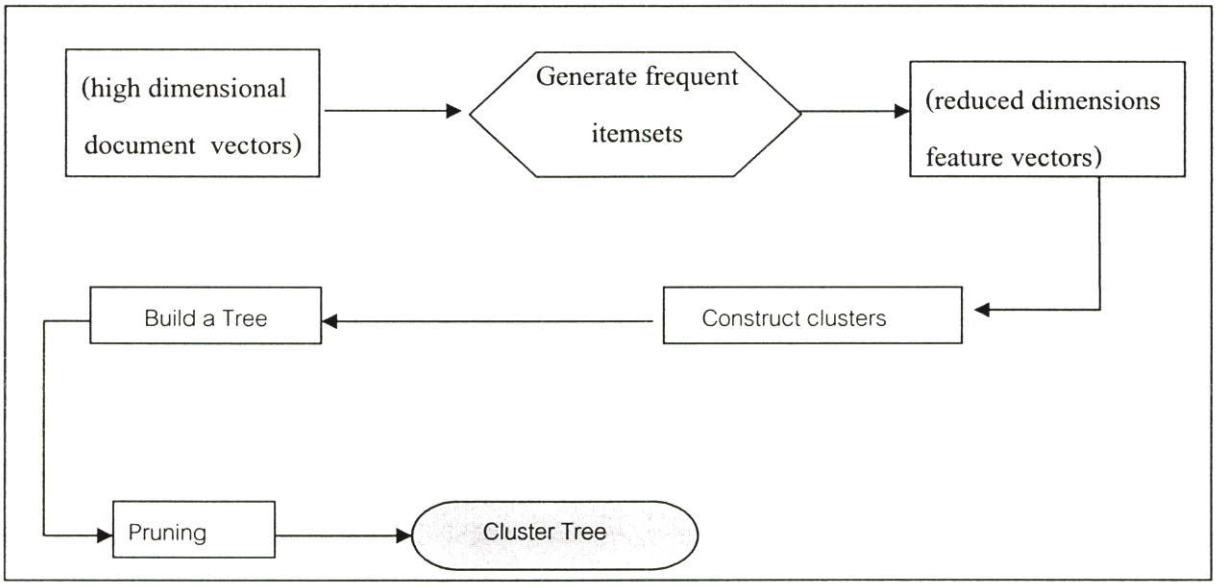
5. Overspecialised Cluster Label การจัดกลุ่มด้วยหลักการทำงานของ Vector Space Model (VSM) ในการจัดกลุ่มทำให้ภายในกลุ่มบางครั้งมีเอกสารที่ไม่ตรงกับความต้องการ

6. SVD จะใช้เวลามากในการทำงานถ้า snippets มีจำนวนมาก

2.2.3 การจัดกลุ่มด้วยคำเดี่ยวในลักษณะลำดับชั้น

การจัดกลุ่มด้วยคำเดี่ยวในลักษณะลำดับชั้น (Single Word and Hierarchical Clustering) คือลักษณะการใช้คำเดี่ยว (Single Word) เป็นตัวกำหนดคุณลักษณะหรือตัวแทนเอกสาร เพื่อนำไปจัดกลุ่มลักษณะของ Hierarchical Clustering ดังตัวอย่างระบบที่ยังมีการให้บริการแก่ผู้สืบค้นทั่วไปและผู้สืบค้นที่ไม่มีทักษะในการสืบค้นเท่านั้น คือระบบ CREDO [13] ซึ่งมีลักษณะการทำงานตามแนวทาง data mining ซึ่งผลการจัดกลุ่มจะแสดงป้ายชื่อกลุ่มเป็นคำเดี่ยวเท่านั้น แม้คำสืบค้นจะมีมากกว่าคำเดี่ยวก็ตาม และ ระบบ FIHC [14]

ตัวอย่างการทำงานของระบบ Frequent Item Hierarchical Clustering (FIHC) [13] โดยมีลักษณะการทำงานที่ใช้ความถี่ของคำในการจัดกลุ่ม ดังแสดงกระบวนการทำงานในรูปที่ 2.5



รูปที่ 2.5 แสดงกระบวนการทำงานของ FIHC

ก่อนการทำงานในขั้นตอนต่างๆ ระบบจะทำงานในขั้นตอน Pre-processing ก่อนเสมอเพื่อลดสิ่งที่ไม่มีความหมายในการจัดกลุ่ม ซึ่งประกอบด้วย การกำจัด stop words และการทำ stemming

words เพื่อกรองคำในเอกสารให้มีจำนวนลดลง และเป็นการลดเวลาในการประมวลผลแล้วจึงทำงานตามขั้นตอนการทำงานดังต่อไปนี้

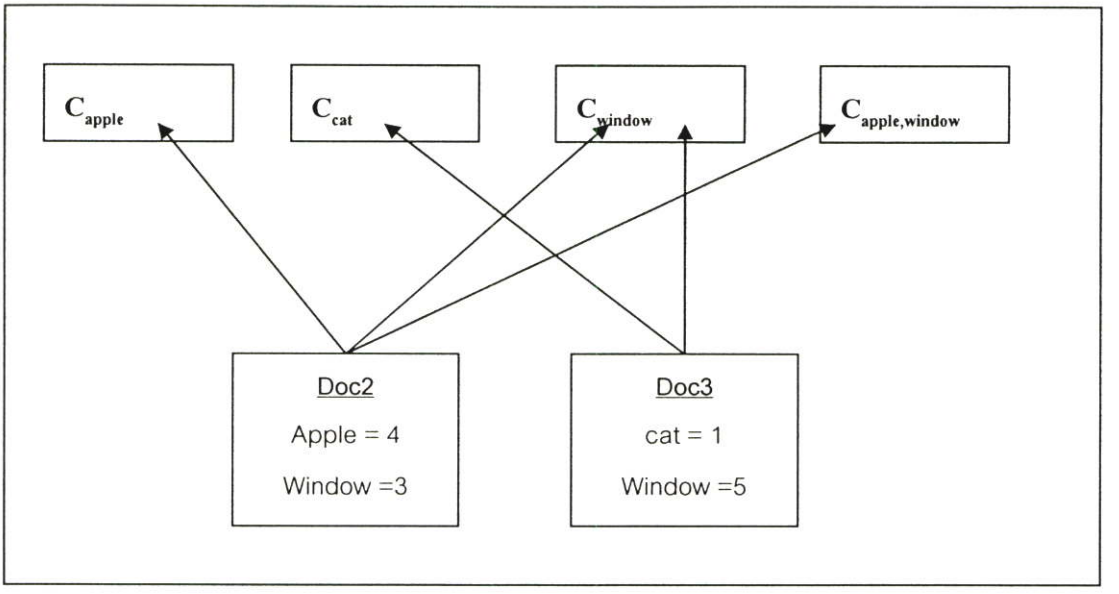
1 Generate Frequent Itemsets คือขั้นตอนการค้นหา set ของ item ที่มีค่าจำนวนความถี่การปรากฏภายในเอกสารมากกว่าหรือเท่ากับ minimum support ที่กำหนดไว้ เพื่อนำมาสร้างเป็น vector model และลดขนาดของมิติใน vector model โดยการขจัด item ที่มีค่าจำนวนความถี่ของคำในการปรากฏภายในเอกสารน้อยกว่า minimum support ที่กำหนด และมีการพิจารณาการปรากฏร่วมกันของแต่ละ item เป็น Global Frequent Itemset ดังแสดงตัวอย่าง ในตารางที่ 2.1

ตารางที่ 2.1 แสดงตัวอย่าง document vector

item / document	apple	you	cat	window
1	5	2	1	1
2	4	0	0	3
3	0	3	1	5
4	8	0	2	0
5	5	0	0	3
Global Support	80%	40%	60%	60%
Feature Vector	Apple , cat , window			
Global Frequent Itemset	{apple} , {window} , {apple window}			

2. Construct clusters ขั้นตอนการกำหนดกลุ่ม ประกอบด้วย

2.1 Constructing Initial Cluster เป็นลักษณะการสร้างกลุ่มเบื้องต้น โดยการนำทุกเอกสารมาจัดกลุ่มให้กับทุกกลุ่มที่มี item เดียวกับเอกสาร และ กำหนดให้ Feature Vector และ Global Frequent Itemset เป็นคำอธิบายกลุ่ม (cluster label) ในขั้นตอนนี้จะยังไม่แยกกลุ่มเพราะเอกสารหนึ่งเอกสารอาจประกอบอยู่ในหลายๆ Global frequent itemset และ จะลดการทับซ้อนกันในขั้นตอนต่อไป ดังแสดงตัวอย่างการจัดกลุ่มเอกสารลง Global Frequent Itemsetในรูปที่ 2.6

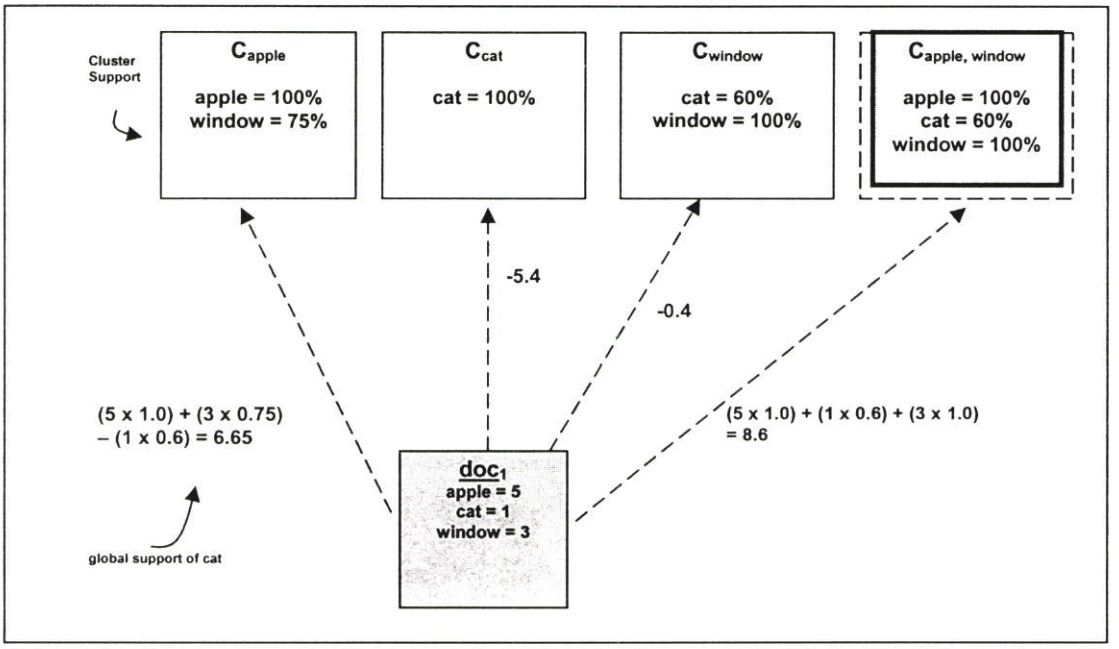


รูปที่ 2.6 แสดงตัวอย่างการจัดกลุ่มเอกสารลงใน Global Frequent Itemset

2.2 Marking Clusters Disjoint ขั้นตอนนี้จะทำการแยกกลุ่มเอกสารให้อยู่เฉพาะในกลุ่มที่ดีที่สุด โดยพิจารณาจากคะแนนความเป็นสมาชิกของเอกสาร(D_j) กับกลุ่ม(C_i) ถ้าเอกสารมีค่าคะแนนความเป็นสมาชิกในกลุ่มใดสูงที่สุดจะให้เอกสารเป็นสมาชิกในกลุ่มนั้น และลบเอกสารออกจากทุกกลุ่มที่เหลือ ถ้าคะแนนเท่ากันจะเลือกกลุ่มที่มีจำนวน item ใน cluster label มากกว่า หลังจากขั้นตอนนี้แล้วเอกสารจะปรากฏในกลุ่มใดกลุ่มหนึ่งเพียงกลุ่มเดียวเท่านั้น และสูตรคำนวณ

$$Score(c_i \leftarrow doc_j) = [\sum n(x) * cluster\_support(x)] - [\sum n(x') * global\_support(x')]$$

การจัดกลุ่มในลักษณะนี้จะให้หลักการว่า " ทุกๆเอกสารที่เป็นสมาชิกภายในกลุ่ม จะต้องประกอบด้วย items ทุกๆ items ภายในคำอธิบายกลุ่ม (cluster label) " การคำนวณคะแนนเลือกกลุ่ม แสดงตัวอย่างในรูปที่ 2.7



รูปที่ 2.7 แสดงตัวอย่างการคำนวณคะแนนการเลือกกลุ่ม

3 Build a Tree การสร้าง cluster tree อยู่บนพื้นฐานของการคำนวณค่าความเหมือน (similarity)ในกลุ่มของ cluster ตามลักษณะของ Hierarchical Clustering

4 pruning เป็นการปรับแต่ง folder label ในกรณีที่ Tree ประกอบด้วยหลายๆกลุ่ม เอกสารที่มีความสัมพันธ์กัน

2.2.4 การจัดกลุ่มด้วยวลีในลักษณะลำดับชั้น

การจัดกลุ่มด้วยวลีในลักษณะลำดับชั้น (Sentence and Hierarchical Clustering) คือ การนำประโยค (Sentence) หรือวลี (Phrase) มาเป็นตัวกำหนดคุณลักษณะหรือตัวแทนของเอกสาร เพื่อจัดกลุ่มตามแบบของลำดับชั้น (Hierarchical Clustering) ดังตัวอย่างระบบต่อไปนี้

2.2.4.1 ระบบจัดกลุ่มผลการสืบค้นชื่อ SNAKET [15] หรือ A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering [7] มีลักษณะการจัดกลุ่มแบบ on-the-fly หรือ Dynamic Clustering โดยแบ่งการทำงานออกเป็น 3 ส่วน

1. Sentence selection and Ranking คือการค้นหาความถี่ของ itemset ของแต่ละคู่คำที่ปรากฏร่วมกันใน snippets พร้อมกับหาค่า *itidf* ของคำจากการนำไปเปรียบเทียบหา ใน DMOZ Data ที่มีการจัดเตรียมกลุ่มไว้มากกว่า 3,500,000 site และมากกว่า 460,000 กลุ่ม และจัดการรวมคู่คำที่ปรากฏใน snippet เดียวกันและอยู่ใน fixed proximity window ใกล้กัน แต่ไม่จำเป็นต้องอยู่

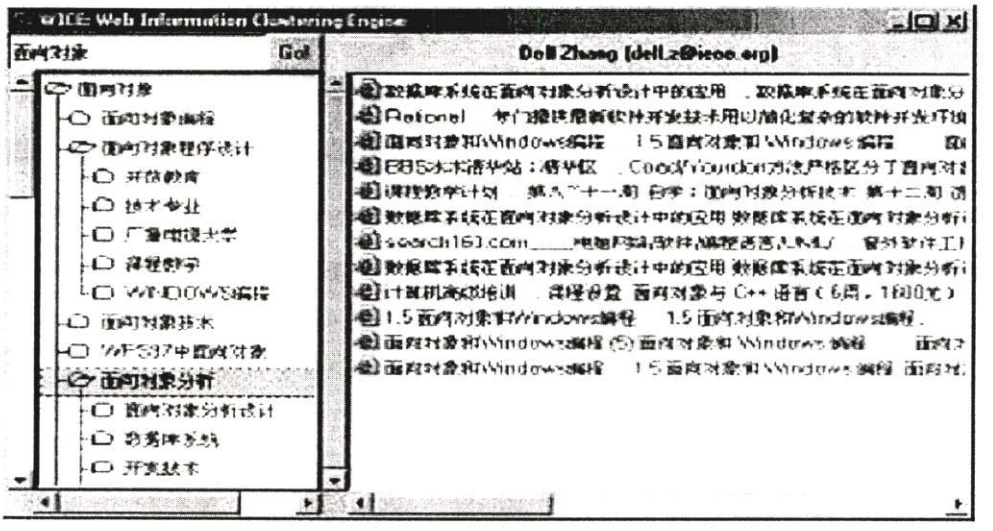
ติดกันเหมือนในโครงสร้าง suffix tree และ suffix array การรวมคู่คำจะรวมกันจนกว่า itemset นั้นจะมีความยาวเท่ากับ 8

2. Hierarchical Clustering and Labeling ใช้การจัดกลุ่มในลักษณะตามแนวทาง bottom-up hierarchical clustering โดยยินยอมให้มีการ overlap ของ folder label โดยแบ่งการทำงานเป็น การจัดรูปแบบ Parent ของกลุ่มที่ใช้ gapped sentence ร่วมกัน การจัดลำดับความสำคัญ (ranking phrase) จากค่าความสัมพันธ์ของ parent และ children label และขั้นตอนสุดท้ายคือการปรับแต่งกิ่ง (prining phrase) ด้วยการปรับตำแหน่งของ folder hierarchy

3. Persinalized Ranking เป็นส่วนงานที่ติดต่อกับผู้สืบค้น โดยผู้สืบค้นสามารถเลือกการทำงานของระบบได้ด้วยตนเอง เช่นการยุบ หรือขยาย folder และสามารถสืบค้นต่อไปจากป้ายชื่อกลุ่ม เป็นต้น

SNAKET จะจัดกลุ่มผลการสืบค้นจาก 16 ระบบสืบค้น (Meta search engine) ลงในลำดับชั้นของคำอธิบายกลุ่ม ซึ่งมีประสิทธิภาพและประสิทธิผลในการทำงานที่คล้ายกับ vivisimo.com ที่ไม่เปิดเผยลักษณะการทำงานแต่ระบบนี้จะเปิดเผยการทำงานของระบบ และ label folder มีลักษณะเป็น sentence ที่ไม่กำหนดความยาวที่แน่นอนทำให้ผู้สืบค้นอ่านและเข้าใจแก่นสารของกลุ่มได้ถึงแม้คำสืบค้นจะไม่สมบูรณ์

2.2.4.2 ระบบ Semantic, Hierarchical, Online Clustering of Web Search (SHOC) [16] พัฒนาระบบจัดกลุ่มผลการสืบค้นเพื่อรองรับการสืบค้นที่เป็นภาษาจีน โดยใช้โครงสร้างข้อมูลชื่อ suffix array ค้นหา phrase เพื่อสร้างคำอธิบายกลุ่ม (Cluster Label) แล้วนำมาคำอธิบายกลุ่มมาเป็นต้นแบบในการจัดกลุ่มด้วยเทคนิค Singular Value Decomposition (SVD) ร่วมกับ Latent Semantic Indexing (LSI) [17] ดังรูปที่ 2.8



รูปที่ 2.8 แสดงตัวอย่างผลการจัดกลุ่มของระบบ SHOC

การสร้างลำดับชั้นของ SHOC ใช้การเปรียบเทียบทีละคู่ของกลุ่ม ถ้าสามารถรวมเข้าเป็นกลุ่มเดียวกันได้ก็รวมเป็นกลุ่มเดียวกัน แต่ถ้ารวมไม่ได้ก็พิจารณาว่ากลุ่มใดควรเป็นลูกของอีกกลุ่ม ถ้าต้องรวมเป็นกลุ่มเดียวต้องคิดเปรียบเทียบว่าควรใช้คำอธิบายกลุ่มของกลุ่มใดโดยดูจากการเป็น substring ของทั้ง 2 กลุ่มที่ถูกนำมาเปรียบเทียบ ดังแสดงขั้นตอนในรูปที่ 2.9

```
if (|X ∩ Y|/|X ∪ Y| > t1){
   X and Y are merged into one cluster;
}else { if (|X| > |Y|){
            if (|X ∩ Y|/|Y| > t2){
               lat Y become X's child;
             }//end if
       } else { if (|X ∩ Y|/|X| > t2){
                   lat X become Y's child; }
                }//end if
         }//end else
}//end else
```

รูปที่ 2.9 แสดงกระบวนการรวม 2 กลุ่มเข้าด้วยกันของระบบ SHOC

การปรับแต่งคำอธิบายกลุ่มหรือป้ายชื่อกลุ่ม ระบบใช้การเปรียบเทียบการใช้คำร่วมกันของสองกลุ่มคือ X และ Y จากการพิจารณาคำที่ปรากฏเป็นป้ายชื่อกลุ่มของทั้งสองกลุ่มว่ามีความสัมพันธ์เกี่ยวข้องกันในลักษณะการใช้คำร่วมกัน จะทำให้ป้ายชื่อกลุ่มลดลงและจำนวนกลุ่มก็ลดลงด้วยเช่นกัน เพราะการทำงานของ SHOC ใช้แนวทางการค้นหาป้ายชื่อกลุ่มหรือคำอธิบายกลุ่มก่อน แล้วจึงนำ snippets หรือเอกสารมาเปรียบเทียบความคล้ายคลึงกันของกลุ่มและเอกสารตามแนวทางของ Vector Spec Model (VSM) ดังแสดงขั้นตอนการเปรียบเทียบเงื่อนไขการรวมกลุ่มของสองกลุ่มในรูปที่ 2.10

```
if(label_x is a substring of label_y){
   label_xy=label_y;
}else if(label_y is a substring of label_x){
        label_xy=label_x;
    }else {
        label_xy="label_x+label_y";
    }
```

รูปที่ 2.10 แสดงกระบวนการรวมคำอธิบายกลุ่มสองกลุ่มของระบบ SHOC

ระบบ SHOC พัฒนามาเพื่อแก้ปัญหาของ STC ด้านต่างๆดังนี้

1. ข้อจำกัดของ Suffix Tree ด้านความไม่เหมาะสมกับภาษาจีนหรือภาษาของชาวตะวันออก
2. การสร้างลำดับชั้น (Hierarchy) ของ STC โดยตรงเป็นการสร้างลำดับชั้นที่ไม่สมเหตุสมผล
3. การทำงานของ STC จะมองข้ามเรื่องของคำที่มีความหมายเหมือนกัน (synonymy) และคำหนึ่งคำอาจมีหลายความหมาย (polysemy) ซึ่งเป็นธรรมชาติของภาษา

จุดเด่นของ Hierarchical Clustering คือมีลักษณะการแสดงผลที่ให้ความสะดวกต่อการเข้าถึงผลการสืบค้นแก่ผู้สืบค้นอย่างแท้จริง เพราะมีลักษณะของผลการจัดกลุ่มเป็นลำดับชั้นตามความสัมพันธ์ของกลุ่มเอกสาร คือมีการแสดงกลุ่มย่อยในกลุ่มใหญ่ แต่มีข้อด้อยในเรื่องของการคำนวณที่ซับซ้อนและมหาศาล ในการค้นหาความสัมพันธ์ของกลุ่มแต่ละกลุ่ม และการปรับแต่งความสัมพันธ์ของแต่ละ folder label

2.3 การจัดกลุ่มผลการสืบค้นโดยใช้เทคนิคซัฟฟิกทรีคลัสเตอร์ริ่ง

Oren Zamir and Oren Etzioni. (1998) [1] กำหนดหัวใจสำคัญในการจัดกลุ่มผลการสืบค้นก่อนการแสดงผลให้กับผู้สืบค้น ต้องประกอบด้วย

1. ความสอดคล้องกันของกลุ่ม (Coherent Clusters) แต่ละกลุ่มต้องมีความสอดคล้องกัน คือการจัดกลุ่มจะต้องเป็นการนำเอกสารที่เหมือนกันมาอยู่ร่วมกันภายในกลุ่ม ซึ่งบางครั้งเอกสารหนึ่งอาจมีหลายหัวข้อ หลากหลายสาระสำคัญ และมันอาจจะไม่ปรากฏเพียงกลุ่มเดียวเท่านั้น แต่มันอาจจะไปปรากฏได้ในกลุ่มอื่นๆ นั่นคือแต่ละกลุ่มอาจมีสมาชิกทับซ้อนกันได้ หรือ เอกสารหนึ่งๆอาจปรากฏได้ในหลายๆกลุ่ม (overlapping cluster) เมื่อเอกสารมีความหลากหลายในเนื้อหาสาระและมีความสำคัญในหลายๆหัวข้อ

2. ประสิทธิภาพในการเข้าถึง (Efficiently Browsable) คือการอำนวยความสะดวกในการเข้าถึงผลการสืบค้นที่ตรงกับความต้องการ เพราะผู้สืบค้นต้องการขอบเขตและรายละเอียดของกลุ่ม ดังนั้นระบบจะต้องจัดเตรียมคำอธิบายกลุ่มที่ชัดเจน กะทัดรัดและได้ใจความ เพื่อบอกรายละเอียดและขอบเขตของกลุ่มให้กับผู้สืบค้น

3. ความเร็ว (Speed) ระบบการจัดกลุ่มผลการสืบค้นจะต้องเร็วในด้านการจัดกลุ่มผลการสืบค้นและแสดงผลการจัดกลุ่มให้กับผู้สืบค้นทันที ทำให้ระบบจะต้องเพิ่มความต้องการในเรื่องกระบวนการทำงานที่รวดเร็ว(Algorithm Speed) และ ยอมใช้ snippet มาจัดกลุ่มแทนการใช้เอกสารเต็ม (Snippet-Tolerance) เพื่อให้กระบวนการทำงานเร็วขึ้น ไม่ต้องเสียเวลาไปดึงข้อมูลที่เป็นเอกสารเต็มมาใช้งาน

ต่อมาในปี 1999 Oren Eli Zamir [2] ได้ระบุความต้องการพื้นฐานในการจัดกลุ่มผลการสืบค้น เพื่ออำนวยความสะดวกให้แก่ผู้สืบค้น ประกอบด้วย

1. Overlaping Cluster เอกสารหนึ่งอาจมีหลายหัวข้อหรือมีความสัมพันธ์กับเอกสารอื่นๆ นั่นหมายความว่าเอกสารหนึ่งสามารถปรากฏได้มากกว่าหนึ่งกลุ่มเอกสาร

2. Phrase การใช้วลี (phrase) ในการจัดกลุ่มจะทำให้คุณภาพของกลุ่มดีขึ้น เพราะใช้ข้อมูลข้อความในเอกสารมากกว่าคำหนึ่งคำ และวลี (phrase)เป็นคำอธิบายกลุ่มที่ถูกต้องและรัดกุมมากกว่าคำเดี่ยวที่มีอำนาจจำแนกน้อยกว่าวลี (phrase)

3. Simple Cluster Definition การจัดกลุ่มผลการสืบค้นต้องมีคำอธิบายกลุ่ม เพื่อช่วยให้ผู้สืบค้นเข้าใจกลุ่มง่ายขึ้น

จากหัวใจสำคัญและความต้องการพื้นฐานในการจัดกลุ่มผลการสืบค้น ทำให้มีการพัฒนาเทคนิคในการจัดกลุ่มผลการสืบค้นบนเครือข่ายใยแมงมุมหลายเทคนิค และเทคนิคที่ได้รับความนิยมในการนำมาจัดกลุ่มผลการสืบค้นคือเทคนิค Suffix Tree Clustering (STC) เทคนิคนี้ทำให้การจัดกลุ่มผลการสืบค้นมีความก้าวหน้า เพราะมีจุดเด่นคือ

1. มีลักษณะการทำงานเป็น Automatic Clustering and Labeling
2. เป็นเครื่องมือที่ให้ความสำคัญกับวลีมากกว่าคำ สามารถค้นหาวลีได้
3. ลักษณะการจัดกลุ่มเป็นแบบ incremental clustering คือสามารถจัดกลุ่มจากเอกสารที่เข้ามาใหม่ได้เรื่อยๆ โดยไม่ต้องเริ่มต้นกระบวนการจัดกลุ่มใหม่เมื่อมีเอกสารใหม่เข้ามา
4. ใช้เวลาในการทำงานเป็นเส้นตรงขึ้นอยู่กับจำนวนเอกสาร ($O(n)$)
5. สามารถทำงานกับคำอธิบายสั้นๆ (snippets) แทนเอกสารแบบเต็มรูปแบบ
6. มีลักษณะการทำงานที่ให้ผลการจัดกลุ่มเป็น overlap cluster คือเอกสารหนึ่งเอกสารสามารถปรากฏได้ในหลายๆกลุ่มเอกสาร เพื่อเพิ่มพื้นที่ในการเข้าถึงเอกสารให้กับผู้สืบค้น

ด้วยกระบวนการทำงานของโครงสร้างข้อมูลแบบ Suffix Tree ส่งผลให้ Suffix Tree Clustering ได้รับความนิยม เนื่องจากมีลักษณะเป็น Fast Incremental Clustering และมีลักษณะของผลการจัดกลุ่มที่มีการเปลี่ยนแปลงตามผลการสืบค้น (Dynamic Clustering หรือ on-the-fly Clustering) ดังตัวอย่างงานวิจัยต่อไปนี้

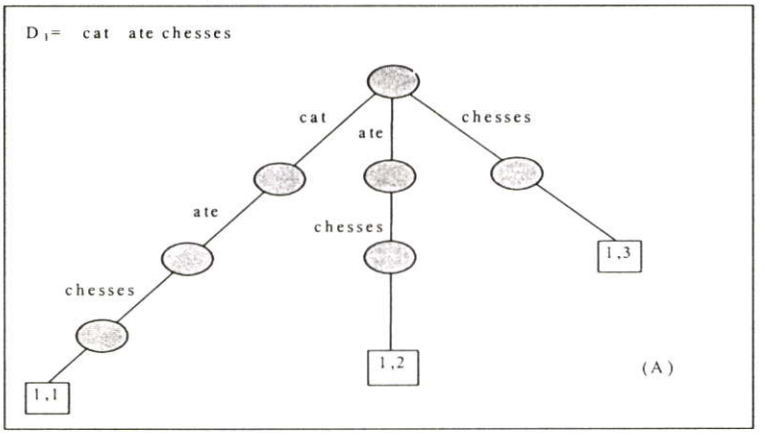
2.3.1 รายงานวิจัยเรื่อง Web Document Clustering : A Feasibility Demonstration

รายงานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อศึกษาประสิทธิภาพการทำงานของ STC ซึ่งจะทำการเปรียบเทียบกับ 6 กระบวนการทำงาน (Algorithm) คือ original list , Single-Pass , K-means , Buckshot , Fractionation , GAHC และเปรียบเทียบการจัดกลุ่มผลการสืบค้นโดยใช้ snippets กับ เอกสาร ซึ่งมีขั้นตอนการทำงานตามแบบของ STC [1] คือ

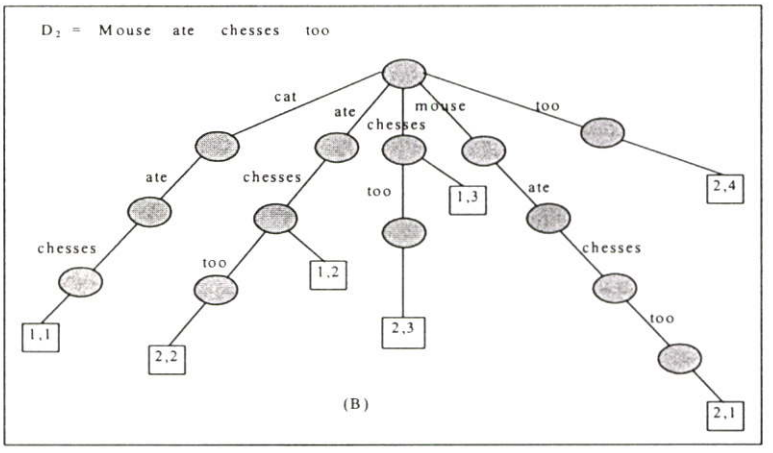
1. Document " cleaning " เป็นการเปลี่ยนแปลงตัวอักษรในแต่ละเอกสารโดยการใช้กระบวนการทำงานของ Stemming Algorithm คือลบตัวหน้า (prefixes) และตัวหลัง (suffixes) ของคำ และลดรูปของคำให้อยู่ในรูปแบบของคำเอกพจน์(singular) การลบคำที่ไม่มีความหมายไม่มีอำนาจจำแนก เช่น ตัวเลข (number) , HTML tags เป็นต้น การเปลี่ยนแปลงคำหรือตัวอักษรเหล่านี้เราสามารถที่จะแปลงมันกลับให้มาอยู่ในรูปเดิมเมื่อต้องการแสดงผลให้ผู้ใช้สามารถอ่านได้

2. Identifying Base Clusters เป็นการกำหนดกลุ่มพื้นฐาน (base cluster) โดยการใช้โครงสร้างข้อมูลที่เรียกว่า Suffix Tree มาช่วยในการสร้างและค้นหา common phrase ที่กลุ่มเอกสารใช้ร่วมกัน ประกอบด้วยขั้นตอนการทำงานดังต่อไปนี้

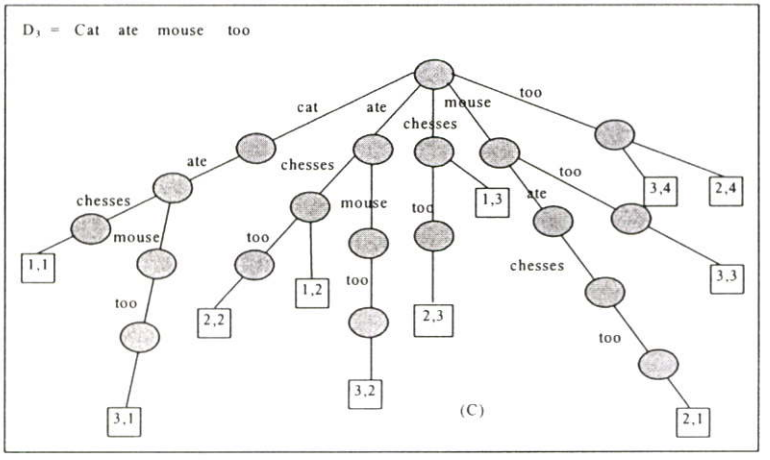
2.1 การสร้าง Suffix Tree เป็นการนำ snippets มาแตกประโยคตามลำดับของคำแล้วนำคำพร้อมลำดับคำ และหมายเลขเอกสาร มาสร้าง tree ดังตัวอย่างจาก 3 เอกสาร คือ D1 : Cat ate cheeses , D2 : Mouse ate chesses too และ D3 : Cat ate mouse too แสดงตัวอย่างการสร้าง tree ในรูปที่ 2.11 , 2.12 และ 2.13



รูปที่ 2.11 แสดงตัวอย่างการสร้าง Suffix Tree ของเอกสาร D1

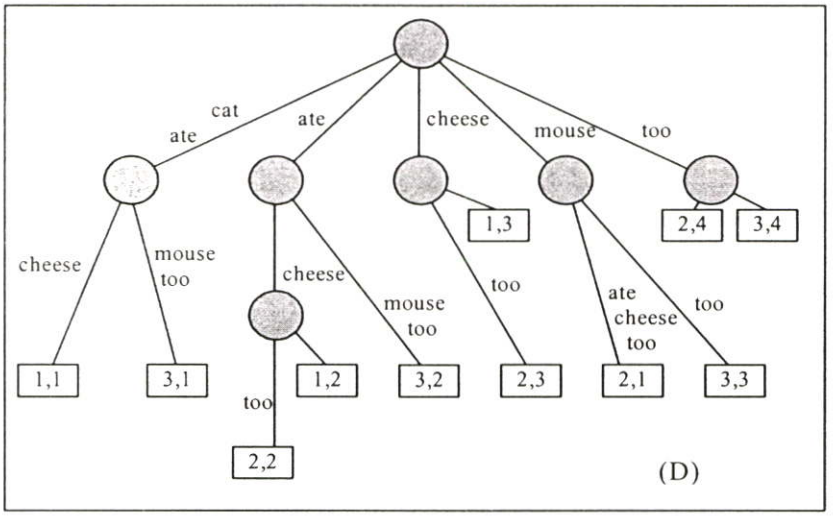


รูปที่ 2.12 แสดงตัวอย่างการสร้าง Suffix Tree ของเอกสาร D1 และ D2



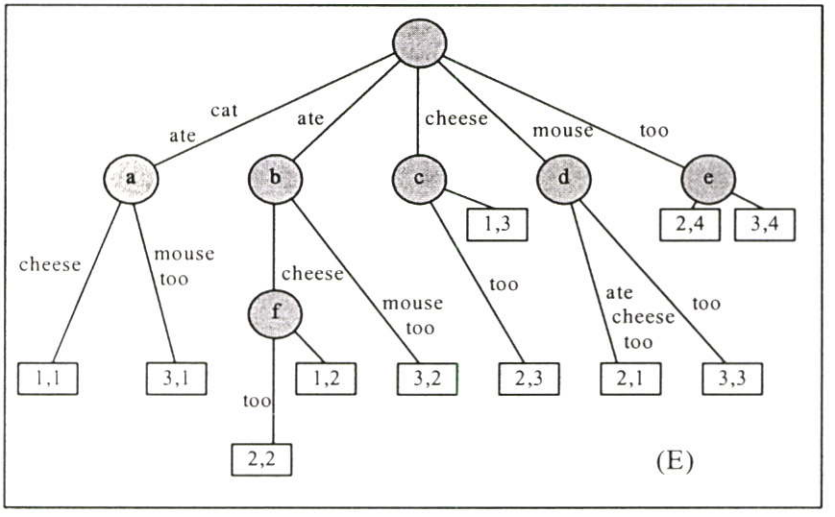
รูปที่ 2.13 แสดงตัวอย่างการสร้าง Suffix Tree ของเอกสาร D1, D2 และ D3

2.2 การยุบรวม node ที่ไม่มีเอกสาร และมี link = 1 เพื่อค้นหา common phrase ที่กลุ่มเอกสารใช้ร่วมกัน ดังแสดงตัวอย่างการยุบ node (จากรูปที่ 2.13 เมื่อยุบรวม node แล้วจะได้ผลดังแสดงในรูปที่ 2.14)



รูปที่ 2.14 แสดงขั้นตอนการยุบรวม node ของ suffix tree

2.3 การค้นหากลุ่มพื้นฐาน (base cluster) โดยพิจารณาจากการใช้ common phrase ร่วมกันของกลุ่มเอกสารนั่นคือแต่ละ node ที่จะได้รับเลือกเป็นกลุ่มพื้นฐานจะต้องมีเอกสารหรือ link เป็นส่วนประกอบอยู่ในมันมากกว่า 1 จึงจะกำหนดให้มันเป็นกลุ่มพื้นฐาน (base cluster) ดังแสดงตัวอย่างในรูปที่ 2.15



รูปที่ 2.15 แสดงกลุ่มพื้นฐานของเทคนิค STC

2.4 การคำนวณคะแนนกลุ่มพื้นฐาน เพื่อใช้ในการเลือกกลุ่มในการแสดงผลเป็น final cluster และใช้ในการคิดคะแนนรวมกลุ่มในการจัดลำดับกลุ่ม (ranking) ดังสมการ

$$S(B) = |B| * f(|P|)$$

$$f(|P|) = \begin{cases} 0, if\ |P| = 1 \\ |P|, if\ 2 \le |P| \le 6 \\ \alpha, if\ |P| > 6 \end{cases}$$

เมื่อ B คือ กลุ่มพื้นฐาน (Base Cluster)

$|B|$ คือ จำนวนของเอกสารใน B

P คือ Phrase ที่นำมาใช้เป็นป้ายชื่อของ Base Cluster

$|P|$ คือ จำนวนของคำที่อยู่บน Phrase

α คือ ค่าคงที่

การกำหนดกลุ่ม base cluster จะมีการเปรียบเทียบคำกับตาราง stop-list และปรับปรุงตาราง stop-list อยู่เสมอโดยการเพิ่มคำที่มีกำหนดในอินเตอร์เน็ท (เช่น "previous" , "java" , " frames" และ "mail") หรือถ้ามีคำๆนั้นปรากฏอยู่ใน stop-list หรือปรากฏน้อยมาก (เช่น 3 ครั้งหรือน้อยกว่านั้น) หรือปรากฏเป็นจำนวนมาก (เช่น ปรากฏมากกว่า 40 % ของทั้งหมด) และถ้าป้ายชื่อกลุ่มมีคำใน stop-list ประกอบอยู่ทั้งหมด base cluster นั้นจะมีคะแนนกลุ่มเป็นศูนย์ และจากการทำงานกับเอกสารทั้ง 3 เอกสารคือ D1 : Cat ate cheeses , D2 : Mouse ate chesses too และ D3 : Cat ate mouse too จะได้ผลการทำงานดังแสดงในตารางที่ 2.2

ตารางที่ 2.2 แสดงผลการกำหนดกลุ่มพื้นฐานของ STC

Node	*Phrase*	*Documents*	*S(b)*
A	cat ate	1,3	(2*2) = 4
B	ate	1,2,3	(3*0) = 0
C	cheese	1,2	(2*0) = 0
D	mouse	2,3	(2*0) = 0
E	too	2,3	(2*0) = 0
F	ate cheese	1,2	(2*2) = 4

จากตารางที่ 2.2 จะเห็นว่า suffix tree จะให้ผลคือกลุ่มพื้นฐานจำนวนมากและซ้ำซ้อน

3. Combining Base Clusters จากการได้ผลกลุ่มจำนวนมากและสมาชิกมีความซ้ำซ้อนกัน จึงมีขั้นตอนการรวมกลุ่ม based cluster ที่มีความคล้ายคลึงกัน โดยพิจารณาจากจำนวนสมาชิกที่เหมือนกัน ดังสมการ $|B_m \cap B_n| / |B_m| > 0.5$ and $|B_m \cap B_n| / |Bn| > 0.5$ ถ้าคำนวณแล้วผลเป็นจริงทั้งคู่ค่า similarity จะมีค่าเป็น 1 และ ถ้า base cluster คู่ใดมีค่า similarity เป็น 1 base cluster คู่นั้นจะถูกรวมกลุ่มเข้าด้วยกัน โดยเลือกกลุ่มที่มีคะแนน Final score หรือคะแนน S(c) มากกว่าเป็นป้ายชื่อกลุ่ม ดังสมาการคิดคะแนน S(c) คือ

$$s(c) = \sum_{b \in c} s_b$$

เมื่อ S(c) คือ คะแนน Final score

b คือ base cluster ที่ถูกรวมไว้ใน c

S_b คือ คะแนน S(b) ของแต่ละ base cluster ที่ถูกรวมไว้ใน c

เมื่อรวมกลุ่ม base cluster ตามเงื่อนไขการรวมกลุ่ม จะได้ผลการรวมกลุ่ม ดังตารางที่ 2.3

ตารางที่ 2.3 แสดงผลการรวมกลุ่มพื้นฐานและ คะแนน Final score

Phrase	*Documents*	**S(c)**
cat ate	1,2,3	4+0+0+0+0+4=8

รายงานวิจัยฉบับนี้ให้ความสำคัญกับ 10 ลำดับแรกของ final cluster และให้เหตุผลว่าเอกสารแต่ละเอกสารไม่จำเป็นที่จะปรากฏได้เพียงกลุ่มเดียวเท่านั้นเพราะเอกสารหนึ่งๆ อาจมีหลายหัวข้อ ถ้าบังคับให้เอกสารปรากฏได้เพียงกลุ่มเดียวอาจเป็นการลดประโยชน์การจัดกลุ่มได้ แต่ต้องมีความเหมือนกันภายในกลุ่มอย่างเหนียวแน่น ซึ่งอาจจะใช้เครื่องมือในการช่วยการตัดสินใจจากระบบ Information Retrieval ที่มีให้เลือกใช้มากมาย

ผลการทดลองของงานวิจัยฉบับนี้บ่งบอกว่า

1. STC มีค่าเฉลี่ยความถูกต้อง(average precision) สูงที่สุดเมื่อเปรียบเทียบกับการทำงานของ original list , Single-Pass , K-means , Buckshot , Fractionation , GAHC

2. STC ใช้เวลาในการทำงานน้อยที่สุดเมื่อเปรียบเทียบกับ original list , Single-Pass , K-means , Buckshot , Fractionation , GAHC

3. STC มีค่าเฉลี่ยความถูกต้อง (average precision) สูงกว่าถ้ายอมให้ในแต่ละกลุ่มมีการทับซ้อน(overlap)กันได้

4. STC มีค่าเฉลี่ยความถูกต้อง (average precision) ในการจัดกลุ่มโดยใช้วลีงกว่าการใช้คำเพียงคำเดียวในการจัดกลุ่ม

การเปรียบเทียบจำนวนเอกสารที่ตรงความต้องการและไม่ตรงกับความต้องการของ STC มีค่าเฉลี่ยสูงกว่าเมื่อเปรียบเทียบกับ K-means และ Buckshot ค่าเฉลี่ยทั้งสองค่านี้อาจเป็นผลมาจากการยอมให้มีการทับซ้อนกันในแต่ละกลุ่ม จะส่งทั้งผลดีและผลเสียให้กับกลุ่มและผู้สืบค้น ถ้าการทับซ้อนกันนั้นอยู่ในระดับที่พอเหมาะและเป็นข้อมูลที่ตรงกับความต้องการของผู้สืบค้น มันจะเป็นผลดีคือเพิ่มความหนาแน่นและขอบเขตการแสดงผลที่ตรงกับความต้องการ ทำให้ผู้สืบค้นสามารถเข้าถึงข้อมูลที่ต้องการได้ในหลายพื้นที่ แต่ถ้าระดับการทับซ้อนสูงและเป็นข้อมูลที่ไม่ตรงกับความต้องการ จะเป็นการลดคุณภาพของกลุ่มอย่างร้ายแรง ดังแสดงผลในตารางที่2.4

ตารางที่ 2.4 แสดงค่าเฉลี่ยของเอกสารที่ตรงกับความต้องการและไม่ตรงกับความต้องการ

	K-Means	Buckshot	STC
Avg. num of cluster: Relevant document	1.40	1.40	2.60
Avg. num of cluster: Irrelevant document	1.55	1.35	1.90
Ratio of the above	0.90	1.04	1.37

การจัดกลุ่มผลการการสืบค้นโดยใช้ snippets จะมีค่าเฉลี่ยความถูกต้อง (average precision) น้อยกว่าการใช้เอกสารแบบเต็มรูปแบบในการจัดกลุ่มผลการสืบค้น แต่การใช้ snippets จะเร็วกว่าการต้องไปเสียเวลา Download เอกสารเต็มมาใช้งาน

จุดเด่นของรายงานวิจัยฉบับนี้คือ

1. STC มีลักษณะการทำงานที่เป็น Automatic Clustering คือ ระบบสามารถจัดกลุ่มเอกสารพร้อมคำอธิบายกลุ่ม

2. STC มีลักษณะการทำงานเร็ว

3. STC ให้ผลการจัดกลุ่มแบบ overlap cluster คือเอกสารหนึ่งๆสามารถปรากฏได้มากกว่าหนึ่งกลุ่ม เป็นการขยายขอบเขตและเพิ่มความหนาแน่นของข้อมูลที่ตรงกับความต้องการทำให้ผู้สืบค้นสามารถเข้าถึงข้อมูลที่ต้องการได้หลายช่องทอง

4. STC มีลักษณะการจัดกลุ่มแบบ incremental clustering คือ เมื่อมีเอกสารเข้ามาใหม่มันสามารถจัดกลุ่มต่อได้เลย โดยไม่ต้องเริ่มต้นกระบวนการทำงานใหม่

5. รายงานวิจัยฉบับนี้ใช้ snippet ที่ได้จากระบบสืบค้นแทนเอกสาร เพื่อให้จัดกลุ่มได้เร็วกว่าการใช้เอกสารต้นแบบทั้งฉบับ

จุดด้อยของรายงานวิจัยฉบับนี้คือ

1. STC ไม่สามารถแสดงผลการจัดกลุ่มได้ทั้งหมด เนื่องจากโครงสร้างการทำงานของ STC ทำให้ได้กลุ่มจำนวนมาก และต้องเลือกแสดงผลเฉพาะกลุ่มที่สนใจ ทำให้เกิดปัญหาการ Coverage ในช่วงกลุ่มที่สนใจและการกำหนดจำนวนกลุ่มที่แน่นอน (Fixed cluster)

2. การจัดกลุ่มโดยยอมให้มีการทับซ้อนของข้อมูลในแต่ละกลุ่มจะเป็นผลเสียมากกว่า ถ้าข้อมูลที่เกิดการทับซ้อนกันนั้นไม่ใช่ข้อมูลที่ตรงกับความต้องการ ซึ่งส่งผลให้ประสิทธิภาพของกลุ่มลดลง และผลการจัดกลุ่มมีค่าความถูกต้องน้อยลงด้วยเช่นกัน

2.3.2 รายงานวิจัยเรื่อง Grouper: A Dynamic Clustering Interface to Web Search Results

ระบบ Grouper จะมีความยาวของคำอธิบายกลุ่มที่ไม่คงที่แต่คำต้องอยู่ติดกัน เพื่อแก้ปัญหาผลการสืบค้นในลักษณะลำดับรายการยาวๆ ทำให้ผู้สืบค้นไม่ได้รับความสะดวกและเสียเวลาในการค้นหาเอกสารที่ต้องการ จึงสร้าง Grouper เพื่อเป็นหน้าจอที่หุ้มระบบสืบค้นที่มีลักษณะเป็น meta search engine ชื่อ Husky-Search โดยมีขั้นตอนการทำงานดังนี้ [3]

1. ใช้ Suffix Tree เพื่อค้นหากลุ่มเอกสาร เช่นเดียวกับ STC

2. การสร้างคำอธิบายกลุ่ม (description the cluster) เป็นการเลือกกลุ่ม โดยพิจารณาจาก ระดับของเอกสารที่ประกอบอยู่ในกลุ่ม (coverage) และความยาวของ phrase คือจำนวนของคำที่อยู่ใน phrase แต่ไม่นับ Stop word หรือ คำที่ปรากฏอยู่ใน query ซึ่งจะเลือกกลุ่มตามเงื่อนไขต่อไปนี้

2.1 Word Overlap คือถ้าคำใน phrase ซ้ำกันมากกว่า 60 % ให้เลือก phrase ที่มีค่า coverage มากกว่า เช่นจากตารางที่ 2.4 phrase ที่7 ไม่ถูกเลือกผลเพราะ 75%ของคำที่ปรากฏในคำอธิบายกลุ่มของมันไปปรากฏอยู่ใน phrase ที่ 6 ซึ่งมีค่า coverage สูงกว่า

2.2 Sub - and Super String คือถ้า phrase นั้นมี sub-phrase และ super-phrase คือ phrase ที่อยู่ตรงกลางประโยค จะไม่แสดงมันออกมา เช่น จากตารางที่ 2.4 phrase ที่ 3 และ 4 ไม่ถูกเลือกเพราะมันมี sub-phrase และ super-phrase

2.3 Most-General phrase and Low coverage คือ จะเลือก phrase ที่สั้นๆ เพราะมันเป็น Most-General phrase คือไม่มี sub-phrase ที่ต่อจากมัน แต่ต้องเปรียบเทียบกับ Most-specific phrase คือไม่มี super-phrase ก่อนมัน และดูค่า coverage ต้องต่างกันตั้งแต่ 20 % ขึ้นไป แล้วเลือกแสดงผลตัวที่เป็น super-phrase เช่น phrase ที่ 8 และ 6 มีค่า coverage ต่างกันเพียง 5 % จึงไม่เลือก phrase ที่ 8 ดังแสดงตัวอย่างการเลือกกลุ่มตามตารางที่ 2.5

ตารางที่ 2.5 แสดงตัวอย่างการเลือกกลุ่มเพื่อแสดงผล

Num.	Phrase	Coverage	Most-Spec.	Most-Gen.	Selected
1	earth summit	60%	+	-	✓
2	vice president of the united states of america	30%	+		✓
3	president of the united states of	40%			
4	united states of america	50%			
5	united states	65%		+	✓
6	greenhouse gas emissions forecast	40%	+		✓
7	reducing emissions of greenhouse gas	30%	+		
8	greenhouse gas	45%		+	

แนวทางในการสร้างความเร็วในการทำงานของ Grouper ใช้แนวทาง 3 แนวทาง คือ

1. สร้างขั้นตอนการทำงานให้เป็น Incremental Clustering Algorithm
2. ใช้ snippets แทนการใช้เอกสารแบบเต็มรูปแบบในการจัดกลุ่ม
3. ใช้ suffix tree เป็นลักษณะคำไม่ใช่ทีตัวอักษร และไม่สนใจ phrase ที่มีคำ stop word ต่อท้ายและนำหน้าประโยค ทำให้ลดภาระในการทำงานลงได้

ระบบของ Grouper ให้ความสำคัญกับการเลือกกลุ่มเอกสารซึ่งพิจารณาจากคำอธิบายกลุ่ม (description the cluster) ร่วมกับจำนวนเอกสารภายในกลุ่ม

2.3.3 รายงานวิจัยเรื่อง Clustering Web Documents : A Phrase-Based Method for Grouping Search Engine Results

พัฒนาระบบการจัดกลุ่มผลการสืบค้นโดยใช้ STC และมองว่า STC เป็นกระบวนการทำงานเรื่องการจัดกลุ่มแนวใหม่ ที่ออกแบบมาเพื่อตอบสนองความต้องการจัดกลุ่มผลการสืบค้น ซึ่งใช้การจัดกลุ่มผลการสืบค้นจากคำที่มีความหมาย ด้วยการนำประสิทธิภาพของโครงสร้างข้อมูลชื่อ suffix tree มาระบุกลุ่มของเอกสารโดยพิจารณาจากการใช้ common phrase ร่วมกันของเอกสาร และใช้ข้อมูลนั้นในการสร้างกลุ่ม และทำการรวมกลุ่มเข้าด้วยกันตามกระบวนการทำงานเพื่อแสดงผลต่อผู้สืบค้น ดังขั้นตอนการทำงานต่อไปนี้ [2]

1. Document Parsing ขั้นตอนแรกนี้ประกอบด้วยขั้นตอนดังต่อไปนี้
   1.1 การแปลงเอกสารให้อยู่ในรูปของลำดับของคำ
   1.2 กำจัดสิ่งที่ไม่ใช่คำออก เช่น สัญลักษณ์ ตัวเลข เป็นต้น
   1.3 การทำ stemming words ระบบนี้จะใช้ของ Light Stemming Algorithm
2. Phrase Cluster Identification คือการกำหนดกลุ่มพื้นฐานโดยพิจารณาจากการใช้ phrase ร่วมกันของกลุ่มเอกสารที่ได้จาก suffix tree ประกอบด้วย

2.1 สร้าง stop list ซึ่งประกอบด้วย คำ stop words ในภาษาอังกฤษ เช่น " the" , " there" , " their" , " thus" เป็นต้น คำที่ปรากฏน้อยมากเช่น 3 หรือน้อยกว่า หรือคำที่ปรากฏมากๆ เช่น ปรากฏมากกว่า 40 %ของเอกสาร หรือคำสืบค้น หรือคำที่ปรากฏมากกว่า 5 % ของเอกสาร เช่น "java" , " mail " , " netscape" เพื่อใช้ในการตรวจสอบคำที่ปรากฏใน phrase เมื่อพบคำที่อยู่ใน stop list ปรากฏอยู่ใน phrase ให้กลุ่มมีคะแนน (maximum phrase) เป็น 0

2.2 คำนวณคะแนนค่า maximum phrase ของแต่ละ Phrase ตามสูตร

$$S(m) = |m| * f(|m_p|) * \sum tfidf(w_i)$$

เมื่อ

s(m)	คือ คะแนน maximum phrase ของ cluster m กับ phrase m_p
$\|m\|$	คือ จำนวนของเอกสารใน phrase cluster m
w_i	คือ คำที่ปรากฏใน phrase m_p
tfidf(w_i)	คือ คะแนนความถี่ของคำปรากฏใน phrase m_p
($\|m_p\|$)	คือ จำนวนของเอกสารใน phrase cluster m ที่ไม่รวม stop word
f	คือ ฟังก์ชั่นของการเกิดเหตุการณ์ phrase มีความยาวเป็นหนึ่ง

สูตรในการคำนวณค่า tfidf(w_i) คือ

$$tfidf(w_i, d) = (1 + \log(tf(w_i, d))) * \log(1 + N / df(w_i))$$

เมื่อ

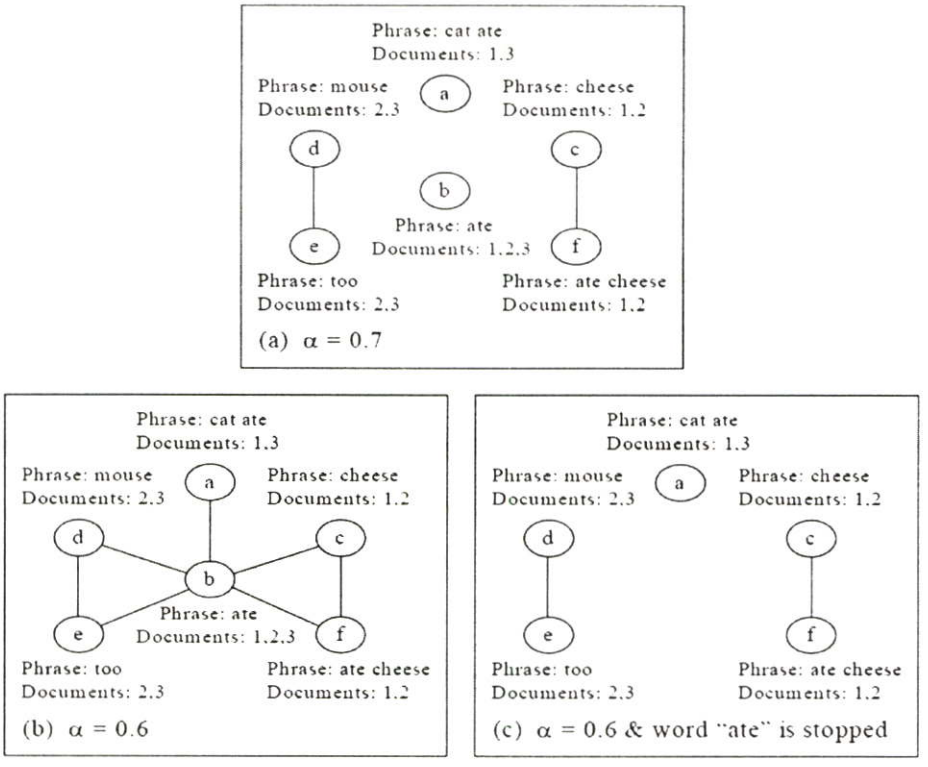
df	คือ จำนวนเอกสารที่มีคำ w_i ปรากฏ
tf(w_i ,d)	คือ จำนวนความถี่ของคำที่ปรากฏในเอกสาร d
N	คือ จำนวนเอกสารทั้งหมดในกลุ่มเอกสาร

3. Phrase Cluster Merging นำตัวแทนกลุ่มมารวมกัน เพื่อลดจำนวนของกลุ่มก่อนการแสดงผลให้กับผู้สืบค้น โดยพิจารณาจากเงื่อนไขการคิดค่า similarity ตามสูตร

$$sim(m_i, m_j) = 1 \; if \; |m_i \cap m_j| / |m_i| > \alpha \; and \; |m_i \cap m_j| / |m_j| > \alpha$$
$$sim(m_i, m_j) = 0 \; ortherwise$$

เมื่อ α คือค่าคงที่ระหว่าง 0 – 1 และ รายงานฉบับนี้ใช้ 0.6

ถ้าคำนวณค่า similarity แล้วมีผลเป็น 1 กลุ่มจะถูกรวมเข้าด้วยกันเป็นกลุ่มเดียว และเลือกกลุ่มที่มีค่า s(m) สูงกว่าเป็นตัวแทนกลุ่ม แต่ถ้าค่า similarity = 0 กลุ่มจะไม่ถูกรวม ดังตัวอย่างการรวมกลุ่มในรูปที่ 2.16



รูปที่ 2.16 แสดงตัวอย่างการรวมกลุ่มพื้นฐาน

จากรูปที่ 2.16 เป็นตัวอย่างกราฟการรวมกลุ่มพื้นฐานด้วยค่าพื้นฐานขั้นต่ำที่แตกต่างกัน ทำให้การได้รับกลุ่มแตกต่างกันด้วย เช่น รูปที่ (a) คือ ผลการรวมกลุ่มพื้นฐานถ้ากำหนดให้ค่าพื้นฐานขั้นต่ำ หรือ α มีค่าเท่ากับ 0.7 จะได้ผลการจัดกลุ่มเป็น 4 กลุ่มคือ กลุ่ม {a} , {b} , {d, e} และ {c, f} , รูปที่ (b) คือ ผลการรวมกลุ่มพื้นฐานที่กำหนดให้ค่าพื้นฐานขั้นต่ำมีค่าเท่ากับ 0.6 จะได้ผลการจัดกลุ่มเป็น 1 กลุ่ม คือกลุ่ม { a, b, c, d, e, f} และ รูปที่ (c) คือผลการรวมกลุ่มพื้นฐานที่กำหนดให้ค่าพื้นฐานขั้นต่ำมีค่าเท่ากับ 0.6 แต่ตัดกลุ่ม b ออก โดยกำหนดให้มันเป็นคำ stop word จะได้ผลการจัดกลุ่มเป็น 3 กลุ่มคือ กลุ่ม {a} , {d, e} และ {c, f}

4. การคำนวณคะแนน final score เป็นการรวมคะแนน s(m) ของแต่ละกลุ่มพื้นฐานที่ถูกรวมเข้าเป็นกลุ่มเดียวกันใน final cluster โดยกลุ่มใดมีคะแนน final score cluster สูงอยู่ในระดับต้นๆจะถูกนำไปเป็นป้ายชื่อกลุ่มและแสดงผลให้กับผู้สืบค้น ซึ่งโดยปกติแล้วจะแสดงผลเฉพาะ 10 -15 กลุ่มแรกเท่านั้น

2.3.4 รายงานวิจัยเรื่อง Web search results clustering in Polish : experimental evaluation of Carrot

เป็นการจัดกลุ่มผลการสืบค้นสำหรับภาษาโปแลนด์ โดยการใช้เทคนิคการจัดกลุ่มของ Suffix Tree Clustering (STC) ในการจัดกลุ่มผลการสืบค้น เพื่อปรับปรุงให้ STC เหมาะกับภาษาโปแลนด์ โดยปรับปรุงในส่วนของขั้นตอนการทำงาน pre-processing ที่ย่อมไม่เหมือนกัน เพราะภาษาอังกฤษไม่เหมือนภาษาโปแลนด์ มีขั้นตอนการทำงานดังนี้ [4]

1. pre-processing มีลักษณะการทำงานเช่นเดียวกับระบบอื่นๆ คือกำจัดสิ่งที่ไม่ใช่คำ หรือสิ่งที่ไม่มีความหมายในการจัดกลุ่ม พร้อมทั้งการทำงานอีก 2 กระบวนการทำงานที่ทำให้ลดเวลาในการประมวลของระบบจัดกลุ่มคือ กำจัด stop word ของภาษาโปแลนด์ และ ทำ stemming โดยใช้ Algorithm ของภาษาโปแลนด์

2. ค้นหากลุ่มพื้นฐาน (base cluster) ด้วยกระบวนการทำงานของ Suffix Tree Clustering (STC) เช่นเดียวกับระบบอื่นๆ และคำนวณคะแนนของกลุ่มด้วยสูตร

$$S(m) = |m| * f(|m_p|) * \sum tfidf(w_i)$$

3. ค้นหา merged cluster ตามลักษณะของ STC ด้วยสูตร

$$sim(m_i, m_j) = 1 \; if \; |m_i \cap m_j| / |m_i| > \alpha \; and \; |m_i \cap m_j| / |m_j| > \alpha$$
$$sim(m_i, m_j) = 0 \; ortherwise$$

4. จัดลำดับผลการสืบค้นโดยใช้คะแนน final cluster เช่นเดียวกับ STC ในภาษาอังกฤษ

ผลการทำงานของ STC กับภาษาโปแลนด์ ในระบบของ CARROT บ่งบอกว่า

1. minimum base cluster score threshold มีความสำคัญมาก เพราะมันเป็นตัวกำหนดการรวมกลุ่ม base cluster

2. STC มีความอ่อนไหวกับตำแหน่งของข้อมูลนำเข้า ซึ่งมีผลทั้งการสร้าง phrase และคะแนนของคำ ดังนั้นขั้นตอนในการทำ pre-processing จึงมีความสำคัญมากเช่นกัน เพราะมีผลกระทบต่อความมั่นคงและคุณภาพของกลุ่มเอกสาร

3. การใช้ phrase ในการจัดกลุ่ม บางครั้งก็อาจจะไม่ได้กลุ่มที่ดีที่สุด ดังเช่นในภาษาโปแลนด์ ตำแหน่งของคำอาจหลอกในด้านของความหมาย หรือคำไม่ชัดเจนด้านความหมาย จึงต้องให้ลำดับของคำที่ถูกต้องมาทำงานแทนในการจัดกลุ่ม

2.3.5 Learning to Cluster Web Search Results

พัฒนาระบบการจัดกลุ่มผลการสืบค้นโดยใช้ Suffix Tree Clustering (STC) และเปลี่ยนการทำงานจากแนวทางของการเรียนรู้แบบไม่มีการชี้นำ (Unsupervised learning) ให้มีลักษณะเป็นการเรียนรู้แบบมีการชี้นำ (Supervised learning) ในส่วนของการจัดลำดับความสำคัญ (ranking) ซึ่งขั้นตอนแรกจะระบุกลุ่มของเอกสารที่ใช้วลีพื้นฐาน (common phrase) ร่วมกัน โดยนำแนวทางของ n-gram มาใช้ร่วมกับ suffix tree เพื่อกำหนด candidate phrase และเพิ่มการคำนวณในหลายๆส่วนเพื่อกำหนด salient phrase และสร้างเครื่องมือในการเรียนรู้เพื่อจัดลำดับความสำคัญของ salient phrase แบบมีการชี้นำ (Supervised learning) โดยเพิ่ม training data ให้กับระบบ เพราะการจัดลำดับกลุ่มที่ผ่านมาไม่ได้ให้ความสำคัญกับชื่อกลุ่ม ประกอบด้วยขั้นตอนการทำงาน 4 ขั้นตอนดังนี้ [5]

1. Search result fetching นำ ผลการสืบค้น มาวิเคราะห์โดยใช้ HTML parser และ นำผลที่ได้เอาเฉพาะ title และ snippets ผ่านการกรองจาก HTML parser

2. Document parsing and phrase property calculation ประกอบด้วย

2.1 กำหนดค่าค่าน้ำหนักที่แตกต่างกันของ title และ snippets โดยพิจารณาตามหลักความน่าจะเป็นที่คะแนนค่าน้ำหนักของ title สูงกว่า snippet

2.2 ทำ stemming words โดยใช้ Porter's algorithm

2.3 สร้าง n-gram โดยให้มี Stop words ประกอบอยู่ใน n-gram ด้วย เพราะมันจะช่วยให้คำอธิบายกลุ่มมีความหมายมากยิ่งขึ้น แล้วกรอง candidate phrase ออกในขั้นตอน post-processing ถ้าคำอธิบายกลุ่มนั้นปรากฏเฉพาะคำ stop words หรือ คำสืบค้น

2.4 ค้นหา candidate phrase ตามขั้นตอนของ STC

2.5 นำ phrase ที่ได้คำนวณ 5 properties ซึ่งประกอบด้วยรายละเอียดดังต่อไปนี้

a. Phrase Frequency/Inverted Document Frequency คำนวณความถี่ของวลี (phrase)

$$fidf = f(w) * \log \frac{N}{|D(w)|}$$

b. Phrase Length คำนวณความยาวของวลี (phrase) คือ $LEN = n$

c. Intra-Cluster Similar คำนวณความเหมือนกันภายในกลุ่ม

$$o = \frac{1}{|D(w)|} \sum_{d_i \in D(w)} d_i$$

$$ICS = \frac{1}{|D(w)|} \sum_{d_i \in D(w)} \cos(d_i, o)$$

d. Cluster Entropy คำนวณค่าความทับซ้อนของข้อมูลในแต่ละกลุ่ม

$$CE = -\sum_{t} \frac{|D(w) \cap D(t)|}{|D(w)|} \log \frac{|D(w) \cap D(t)|}{|D(w)|}$$

e. Phrase Independence คำนวณค่าความเป็นอิสระของ phrase

$$IND_l = -\sum_{t=l(w)} \frac{f(t)}{TF} \log \frac{f(t)}{TF}$$

$$IND = \frac{IND_l + IND_r}{2}$$

3. Salient phrase ranking นำ 5 properties ที่คำนวณได้มาใช้ประโยชน์ใน regression model ที่เรียนรู้จาก training data แล้วรวมทั้ง 5 properties เป็นค่าเดียว คือ x = (TFIDF , LEN , ICS , CE , IND) เพื่อนำเข้าสู่กระบวนการของ regression model (ค่า y ที่นำมาใช้แทนใน regression model จะมาจากผู้เชี่ยวชาญ 3 คน ให้คะแนน phrase ตามเงื่อนไขที่กำหนด และเมื่อรวมคะแนนแล้ว phrase ที่มีคะแนนมากกว่า 100 คะแนน จะมีค่า y =1 นอกนั้นเป็น 0) ซึ่งประกอบด้วยรายละเอียดดังต่อไปนี้

a. Linear Regression คือ $y = b_0 + \sum_{j=1}^{p} b_j x_j + e$

b. Logistic Regression คือ $\log it(q) = \log \frac{q}{1-q} = b_0 + \sum_{j=1}^{p} b_j x_{j+e}$

c. Support Vector Regression คือ

$$L_\varepsilon(y, f(\mathrm{X}, \omega)) = \begin{cases} 0 & if\ |y - f(\mathrm{X}, \omega)| \le \varepsilon \\ |y - f(\mathrm{X}, \omega)| - \varepsilon & ortherwise \end{cases}$$

รูปแบบของ minimization คือ $\min \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right)$

พร้อมทั้งข้อจำกัดคือ $s.t. \begin{cases} y_i - f(\mathrm{X}_i, \omega) \le \varepsilon + \xi_i^* \\ f(\mathrm{X}_i, \omega) - y_i \le \varepsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0, i = 1, \ldots, n \end{cases}$

4. Post-processing ประกอบด้วยขั้นตอนดังต่อไปนี้

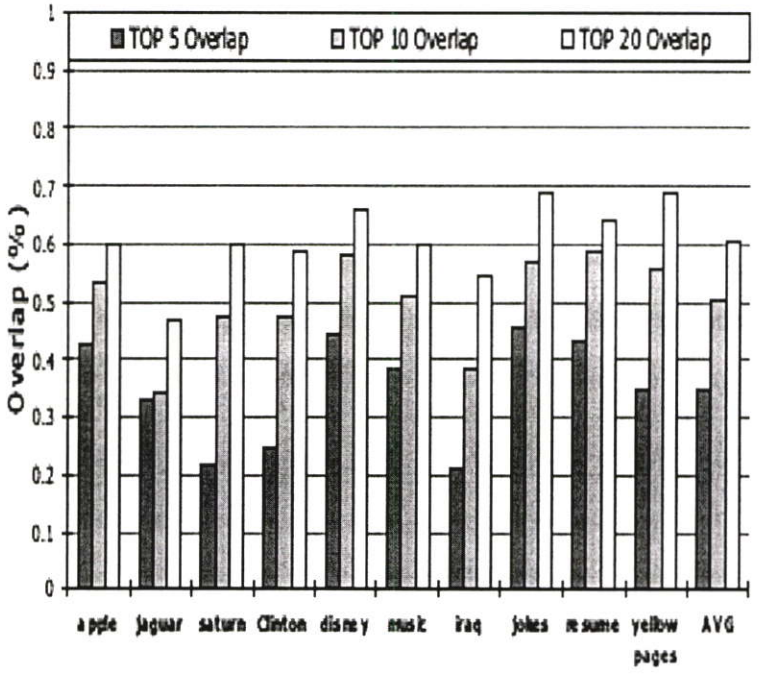
4.1 กำจัด phrase หรือกลุ่มที่มีป้ายชื่อปรากฏเฉพาะคำ Query และ stop words ออกจากระบบ

4.2 ค้นหา Merged phrase and cluster เพื่อลดจำนวนกลุ่มที่ทับซ้อนกัน คือ คิดค่า similarity โดยพิจารณาจากจำนวนเอกสารที่ทับซ้อนกันตั้งแต่ร้อยละ 75 แล้วรวม 2 กลุ่มให้เป็นกลุ่มเดียวและสร้างกลุ่มใหม่จากการรวม 2 กลุ่มเข้าด้วยกันนั้น

4.3 แสดงผลกลุ่มที่มีคะแนนสูงสุดให้กับผู้สืบค้น

รายงานวิจัยฉบับนี้มุ่งเน้นที่จะแก้ปัญหา เรื่องการจัดลำดับความสำคัญของผลการจัดกลุ่ม (ranking salient phrase) ทำให้ยังคงมีข้อด้อยเรื่องผลการจัดกลุ่ม ในด้านต่างๆดังนี้

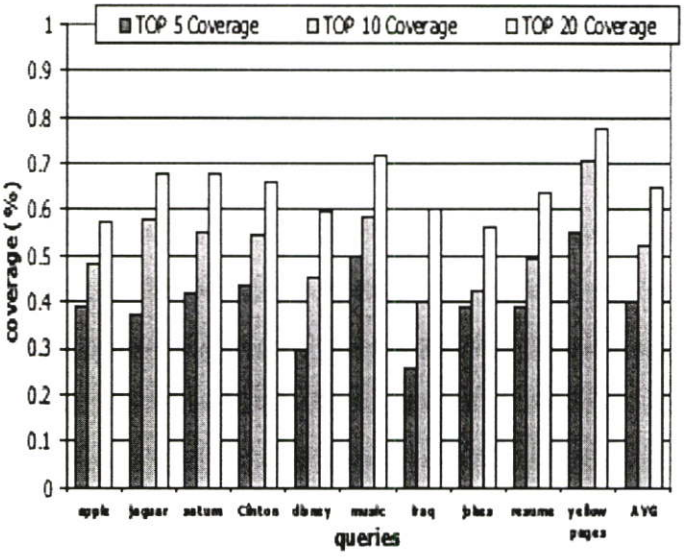
1. ระดับการทับซ้อน(overlap)ของข้อมูลในแต่ละกลุ่ม ดังผลการทดลองจากรูปที่ 2.17



รูปที่ 2.17 แสดงอัตราการทับซ้อนของข้อมูล

จากรูปที่ 2.17 พบว่าระบบจัดกลุ่มมีอัตราการทับซ้อนของข้อมูลเฉลี่ยร้อยละ 0.4

2. อัตราความคลอดคลุมในการจัดกลุ่มเอกสาร (coverage) ระดับเอกสารที่สามารถจัดกลุ่มและแสดงผลการสืบค้นให้กับผู้สืบค้นได้อยู่ในระดับไม่สูง ดังผลการทดลองจากรูปที่ 2.18



รูปที่ 2.18 แสดงระดับของอัตราการ coverage ของข้อมูลที่นำมาใช้ในการจัดกลุ่ม

จากรูปที่ 2.18 พบว่าระบบมีการจัดกลุ่มผลการสืบค้นแล้วมีอัตราการทับซ้อนของข้อมูลเฉลี่ยประมาณร้อยละ 0.5 การแก้ไขปัญหาด้านการ coverage ในอนาคตจะใช้การจัดกลุ่มผลการสืบค้นในลักษณะการรวมกลุ่ม (Cluster Merge Algorithm) แนวใหม่

3. การจัดกลุ่มผลการสืบค้นควรจะเป็นลำดับชั้น (Hierarchical Clustering) เพราะจะเป็นการอำนวยความสะดวกในการเข้าถึงข้อมูลให้กับผู้สืบค้นอย่างแท้จริง

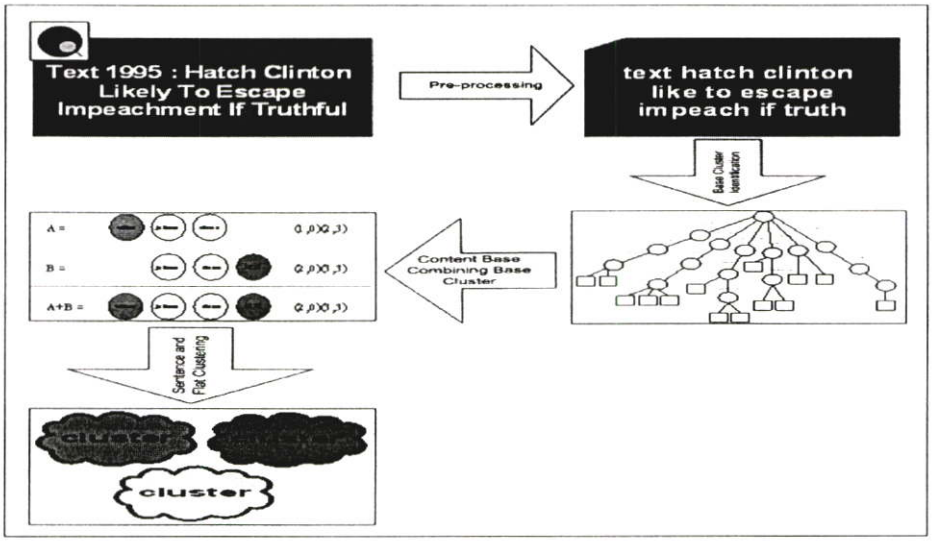
จากการศึกษารายงานผลการวิจัยและทฤษฎีต่างๆ ดังที่กล่าวไว้ในข้างต้นทำให้พบปัญหาของระบบการจัดกลุ่มผลการสืบค้นในลักษณะต่างๆ เช่นปัญหาของการจัดกลุ่มผลการสืบค้นด้วยการใช้คำเดี่ยวเป็นตัวกำหนดลักษณะหรือตัวแทนของเอกสาร ไม่มีอำนาจจำแนกดีพอ หรือปัญหาด้านการคำนวณที่ซับซ้อนของการจัดกลุ่มในลักษณะของลำดับชั้น (Hierarchical Clustering) ปัญหาระดับ coverage และ overlap ของเอกสาร ทำให้มีการพัฒนาการจัดกลุ่มในลักษณะการนำวลี (phrase) มาเป็นตัวกำหนดลักษณะ หรือตัวแทนของเอกสารที่มีอำนาจจำแนกดีกว่าคำเดี่ยว และการจัดกลุ่มในลักษณะของลำดับรายการ (Flat Clustering) ที่มีการคำนวณและการทำงานที่ซับซ้อนน้อยกว่าลักษณะของลำดับชั้น (Hierarchical Clustering) ดังจะกล่าวรายละเอียดในบทที่3

บทที่ 3
การจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งแนวใหม่

การจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งแนวใหม่ (A New Suffix Tree Clustering Algorithm) มีลักษณะเป็น Sentence and Flat Clustering ซึ่งนำคำอธิบายสั้นๆ (Snippets) ของผลการสืบค้น (Search Results) เป็นตัวแทนเอกสารในการจัดกลุ่ม โดยปรับปรุงการรวมกลุ่มพื้นฐาน (Merged Base Cluster) ของเทคนิค STC ด้วยการพิจารณาการมีอยู่ร่วมกันของคำ (Join Phrase) ที่ปรากฏในป้ายชื่อกลุ่ม ร่วมกับการพิจารณาตำแหน่งที่เชื่อมต่อกันของคำในเอกสารเดียวกันที่อยู่ต่างกลุ่มของคู่กลุ่มเอกสาร เพื่อนำมาสร้างกลุ่มใหม่ด้วยการเชื่อมป้ายชื่อกลุ่มที่ได้จากการทำงานร่วมกันของโครงสร้างข้อมูลชื่อ Suffix Tree และเทคนิค n-gram ให้เป็นป้ายชื่อกลุ่มใหม่ที่มีความสมบูรณ์ ดังมีรายละเอียดการทำงานต่อไปนี้

3.1 โครงสร้างการจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งแนวใหม่

การจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งแนวใหม่ (A New Suffix Tree Clustering Algorithm) เป็นการจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่ง (Suffix Tree Clustering) ร่วมกับเทคนิคเอ็นแกรม (n-gram) และ การเชื่อมวลี (Join Phrase) เพื่อค้นหาป้ายชื่อกลุ่ม (Cluster Label) ที่ขาดความสมบูรณ์ของกลุ่มเอกสาร ประกอบด้วย 4 ขั้นตอนการทำงาน ดังแสดงในรูปภาพที่ 3.1



รูปที่ 3.1 แสดงโครงสร้างการจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งแนวใหม่

จากรูปภาพที่ 3.1 สามารถนำมาเขียนเป็นกระบวนการทำงาน (Pseudo-code) ซึ่งเริ่มจากการแตกข้อความหรือคำอธิบายสั้นๆ (snippet) ให้อยู่ในรูปของประโยคที่ประกอบด้วยคำแต่ละคำตามลำดับของคำ ดังแสดงรายละเอียดในรูปที่ 3.2

```
Split text into sentences consisting of words
Phase 1: pre-processing
        Pre-processing of words is using stop-word and stemming-word
Phase 2: Base Cluster Identification
        2.1 Creation of a Generalized Suffix Tree with n-gram technique of all sentences
            for each document{
                split sentence into n-gram block
                  for each n-gram{
                   Insert word into node of suffix tree and number of current document
                   and word's  position into last node ;
                    }
                Slide n-gram until to last word of sentence into generalized suffix tree;
            }
        2.2 Update or compact internal node with the index to current document while
           rearranging  the tree
        2.3 Build a list of base cluster that number of document in node's sub-tree > 1
Phase 3: Content Based Combining Base Cluster
        for each base cluster{
        3.1 Delete cluster B if cluster B is subset of cluster A or delete document of
        cluster B that  overlap with cluster A if phrase of cluster B length equal 1
        3.2  Joint base cluster A and B if {a1 = b0 , a2 = b1 , …., an = bn-1 } and  number of
             word's position in some document of pair cluster > 1
        }
Phase 4: Ranking Content base cluster
```

$$s(m) = |d| * f(|m_p|) * \sum tfidf(p_i)$$

รูปที่ 3.2 แสดงอัลกอริทึมการจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งแนวใหม่

3.2 ขั้นตอนการทำงานของอัลกอริทึม

การจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งแนวใหม่ (A New Suffix Tree Clustering Algorithm) เป็นกระบวนการทำงานเพื่อปรับปรุงแก้ไขข้อด้อยของการจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่ง (Suffix Tree Clustering) ซึ่งประกอบด้วยอัตราการทับซ้อนที่มากเกินไป จนเป็นเหตุให้อัตราความถูกต้องหรือคุณภาพของกลุ่มลดลง และการนำเทคนิค n-gram มาช่วยให้การสร้าง suffix tree มีความสะดวกมากยิ่งขึ้นและลดปริมาณการใช้หน่วยความจำ แต่ทำให้ป้ายชื่อกลุ่มขาดความสมบูรณ์ เพราะถูกตัดขาดด้วยขนาดของ n-gram เช่นป้ายชื่อกลุ่มที่แท้จริงคือ " William Jefferson Clinton Documents" แต่เมื่อกำหนดให้ n-gram มีขนาดของ n<= 3 ป้ายชื่อกลุ่มจะถูกตัดตามขนาดของ n-gram ประกอบด้วยกลุ่ม " William Jefferson Clinton" และ "Jefferson Clinton Documents" ดังนั้นการจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งแนวใหม่ จึงใช้แนวทางการเชื่อมวลีที่ได้จากการจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งร่วมกับเทคนิค n-gram เพื่อลดอัตราการทับซ้อน (Overlap) และค้นหาป้ายชื่อกลุ่มที่ขาดไปเนื่องจากขนาดของ n-gram ดังแสดงรายละเอียดการทำงานดังต่อไปนี้

3.2.1 ขั้นตอนจัดเตรียมข้อมูลนำเข้า

การจัดเตรียมข้อมูลนำเข้า (Pre-processing) เพื่อช่วยลดสิ่งที่มีผลทำให้ผลการจัดกลุ่มอาจด้อยประสิทธิภาพ และประสิทธิผลในการจัดกลุ่มผลการสืบค้น ประกอบด้วย 3 กระบวนการทำงาน ดังต่อไปนี้

3.2.1.1 โทเคนไนเซชั่น (Tokenization) กำจัดสิ่งที่ไม่ใช่คำ พิจารณาเป็นสายอักขระ (String) อักขระที่ไม่ใช่ตัวอักษร เช่น ตัวเลข สัญลักษณ์ และเครื่องหมายวรรคตอนต่างๆ และส่วนที่เหลือจะเปลี่ยนให้เป็นตัวพิมพ์เล็ก (Lower Case) ทั้งหมด

3.2.1.2 การกำจัดคำหยุด (Stop word Eliminating) เป็นการกรองเอาคำที่ปรากฏเป็นจำนวนมากในเอกสาร ซึ่งมันมีค่าในการแบ่งแยกต่ำ ตัวอย่างคำหยุดประกอบด้วย คำนำหน้า (Articles) เช่น a, an, the คำบุพบท (Prepositions) เช่น in, on, of คำสันธาน (Conjunctions) เช่น and, or, but, if คำสรรพนาม (Pronouns) เช่น I, you , them, it และคำต่างๆ ที่เป็นไปได้ เช่น site, back , copyright , enter เป็นต้น

3.2.1.3 การสเต็มมิ่ง (Stemming word) เป็นการแทนที่รูปแปร (Variant) ต่างๆของคำด้วยรากของคำเพียงคำเดียว รูปแปรประกอบด้วยพหูพจน์ รูปแบบของคำกริยาที่เติมท้ายด้วย "ing" หน่วยคำเติมหลังบุคคลที่สาม หน่วยคำเติมหลังอดีตกาล (Past Tense) เป็นต้น ตัวอย่างเช่น คำว่า connect เป็นรากของคำต่อไปนี้ connects, connected, connecting, connection เป็นต้น ซึ่งคำต่างๆ เหล่านี้มีความหมายที่คล้ายคลึงกัน โดยมาตรฐานที่ใช้ส่วนใหญ่ คือ อัลกอริทึมการสเต็มมิ่งของพอร์ตเตอร์ (Porter's Stemming Algorithm) ซึ่งทำให้ลดจำนวนคำได้ถึง 40-50 %

3.2.2 ขั้นตอนการกำหนดกลุ่มพื้นฐาน

ขั้นตอนการกำหนดกลุ่มพื้นฐาน (Base Cluster Identification) คือการระบุกลุ่มเอกสารโดยการประยุกต์ใช้โครงสร้างข้อมูลชื่อ Suffix Tree ร่วมกับเทคนิคการแตกประโยคเพื่อสร้างลำดับรายการของคำ n คำ (n-gram) ประกอบด้วยรายละเอียดดังต่อไปนี้

3.2.2.1 การสร้าง suffix tree ร่วมกับการใช้เทคนิค n-gram จะทำการกำหนดขนาดของ n-gram เพื่อกำหนดขนาดความสูงของ suffix tree โดย n-gram จะเริ่มเลื่อนจากตำแหน่งที่ 0 ถึงตำแหน่งสุดท้ายของคำในเอกสาร และลดขนาดลงเมื่อคำในเอกสารไม่เพียงพอต่อขนาดของ n-gram ที่กำหนด ดังตัวอย่างการเลื่อนตำแหน่งคำตามขนาดของ n-gram <= 3 และขั้นตอนการสร้าง suffix tree ของแต่ละ n-gram ดังต่อไปนี้

1. แตก snippet ให้อยู่ในรูปของ sentences หรือ phrase แล้วนำมาแตกประโยคตามขนาดของ n-gram ตามลำดับของคำ

2. แตก phrase ของ n-gram ออกเป็นคำเดี่ยวตามลำดับของคำ เช่น William Jefferson Clinton แบ่งเป็นคำว่า “William” , “Jefferson” และ “Clinton”

3. สร้าง suffix tree ซึ่งมี root node เป็น directed tree ซึ่งพิจารณาจาก 2 เงื่อนไข คือ (2.1) ถ้าคำแรกของ n-gram ซ้ำกับคำใน node แรกของ sub tree ใด ให้นำ n-gram เข้าสร้าง suffix tree ใน sub tree เดียวกันนั้นและตรวจสอบความซ้ำซ้อนของคำในทุก node ที่ผ่านถ้าซ้ำให้ใช้ node เดิม ถ้าไม่ซ้ำให้สร้าง node ใหม่จนครบตามคำของขนาด n-gram และ (2.2) ถ้าคำแรกของ n-gram ไม่ซ้ำกับคำใน node แรกของ sub tree ใด ให้เริ่มสร้าง suffix tree ใน sub tree ใหม่ ด้วย n-gram นั้นจนครบตามขนาดของ n-gram

4. เก็บตำแหน่งคำพร้อมหมายเลขเอกสารในโหนด (node) สุดท้าย เพื่อให้ทราบจุดสิ้นสุดของแต่ละ n-gram และง่ายต่อการค้นหากลุ่มเอกสารที่ใช้คำหรือวลีร่วมกัน

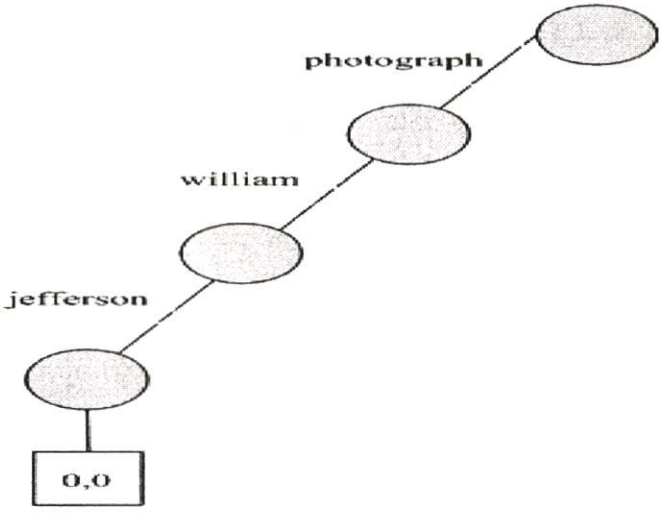
การสร้าง suffix tree จะกำหนดให้โหนด แทนกลุ่มเอกสาร และลิงก์ (Link) แทนป้ายชื่อกลุ่ม ดังตัวอย่าง จากเอกสาร 5 เอกสาร ดังแสดงตัวอย่างเอกสารในรูปที่ 3.3

เอกสารที่0 : photographs of William Jefferson Clinton
เอกสารที่1 : President William Jefferson Clinton !
เอกสารที่2 : President William Jefferson Clinton 324 Document
เอกสารที่3 : William Jefferson Clinton ?
เอกสารที่4 : Jefferson Clinton was Documents

รูปที่ 3.3 แสดงเอกสารตัวอย่างซึ่งผ่านขั้นตอน pre-processing

เอกสารที่0 : **photographs william jefferson** clinton

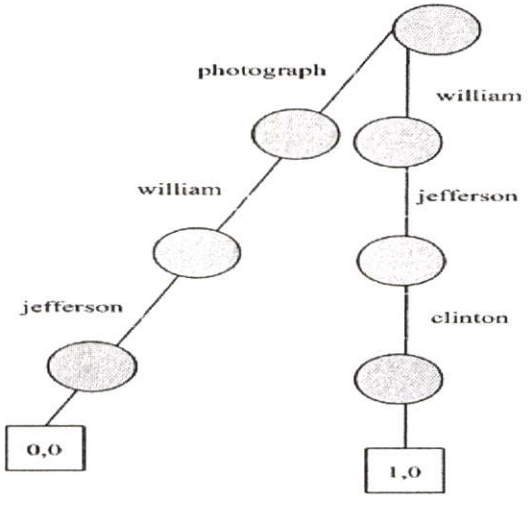
การสร้าง suffix tree เริ่มจากตำแหน่งที่ 0 ของเอกสารที่0 กำหนดขนาด n-gram <= 3 ผลคือ " photographs William Jefferson " และ node สุดท้ายคือ "Jefferson " จะเก็บตำแหน่ง 0 ของเอกสารที่0 และ edge แทนคำแต่ละคำใน n-gram ดังแสดงตัวอย่างในรูปที่ 3.4



รูปที่ 3.4 แสดงตัวอย่างการสร้าง suffix tree ด้วย n-gram <= 3 ของคำตำแหน่งที่ 0 ในเอกสารที่ 0

เอกสารที่ 0 : photographs **william jefferson clinton**

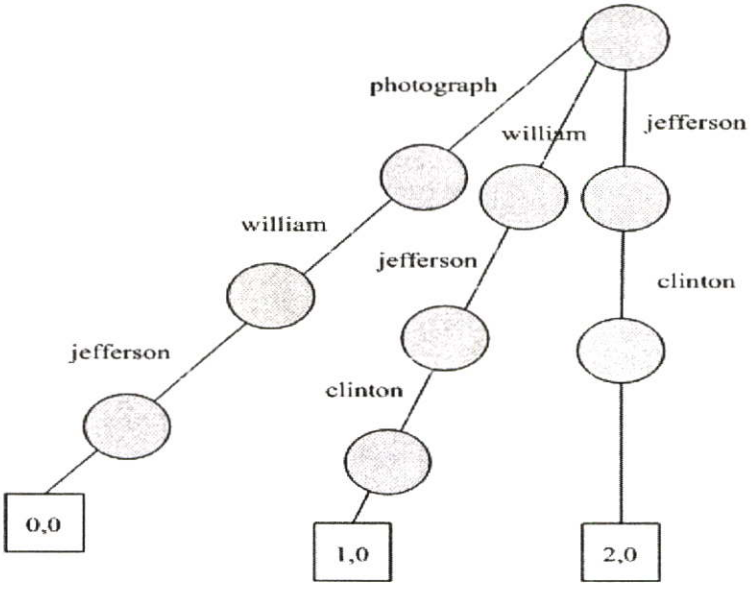
n-gram จะทำการเลื่อนตำแหน่งของคำในเอกสารเดิมครั้งละ 1 ตำแหน่ง ผลคือ "william jefferson clinton" และนำเข้าสู่ขั้นตอนของการสร้าง suffix tree ดังแสดงตัวอย่างในรูปที่ 3.5



รูปที่ 3.5 แสดงตัวอย่างการสร้าง suffix tree ด้วย n-gram <= 3 ของคำตำแหน่งที่1 ในเอกสารที่ 0

เอกสารที่ 0 : photographs william **jefferson clinton**

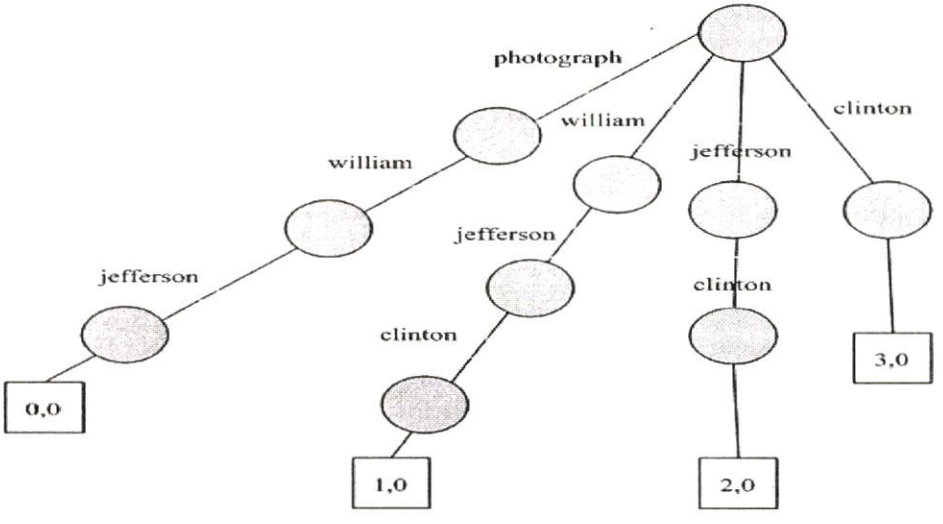
รอบนี้ n-gram จะลดขนาดเหลือ 2 คำเพราะคำในเอกสารไม่เพียงพอต่อขนาดสูงสุดที่กำหนด ผลคือ "jefferson clinton" และนำเข้าสู่ขั้นตอนของการสร้าง suffix tree ดังรูปที่ 3.6



รูปที่ 3.6 แสดงตัวอย่างการสร้าง suffix tree ด้วย n-gram <= 3 ของคำตำแหน่งที่ 2 ในเอกสารที่ 0

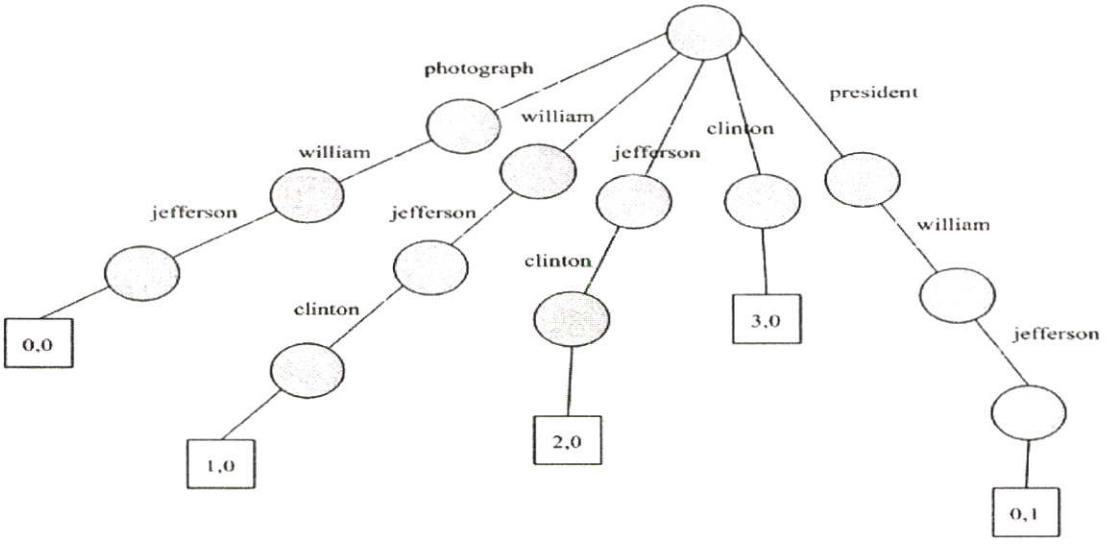
เอกสารที่ 0 : photographs william jefferson **clinton**

เมื่อ n-gram มีขนาดเท่ากับ 1 การทำงานจะเป็นรอบสุดท้ายเพราะเป็นคำนั่นหมายถึงทำงานมาถึงคำสุดท้ายของเอกสาร ดังรูปที่ 3.7



รูปที่ 3.7 แสดงตัวอย่างการสร้าง suffix tree ด้วย n-gram <= 3 ของคำตำแหน่งที่ 3 ในเอกสารที่ 0

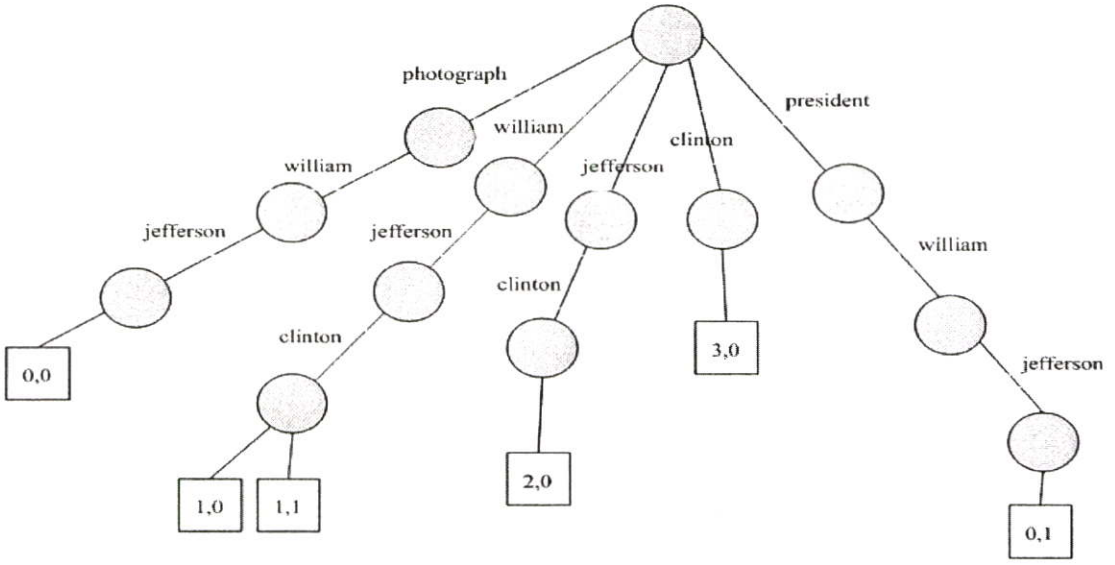
เอกสารที่1 : **president william jefferson** clinton เมื่อเริ่มเอกสารใหม่ n-gram จะมีขนาดเท่ากับ 3 คำ และนำเข้าสู่ขั้นตอนของการสร้าง suffix tree ดังรูปที่ 3.8



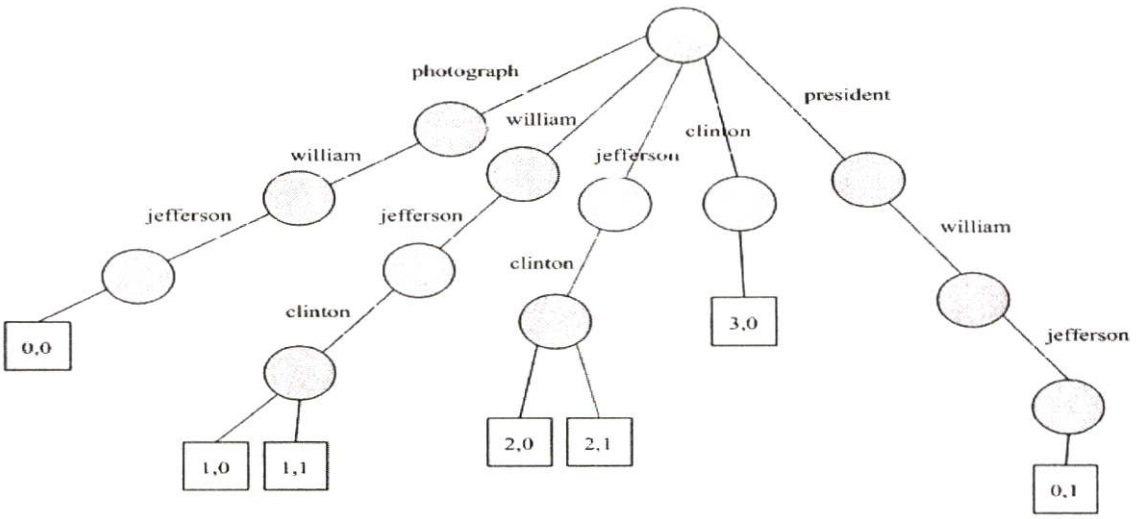
รูปที่ 3.8 แสดงตัวอย่างการสร้าง suffix tree ด้วย n-gram <= 3 ของคำตำแหน่งที่ 0 ในเอกสารที่ 1

เอกสารที่1 : president **william jefferson clinton**

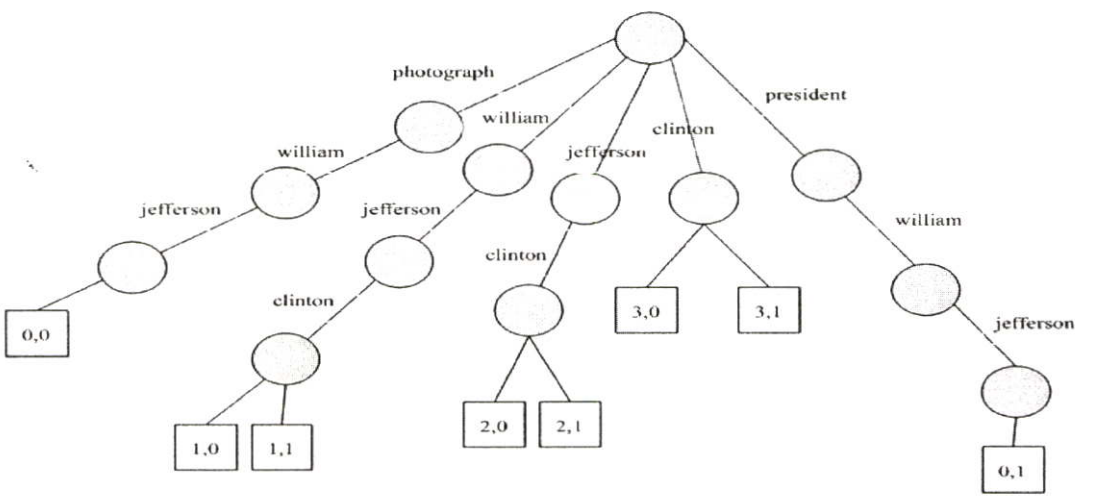
รอบนี้การสร้าง suffix tree คำแรกซ้ำกับ node แรก คือ " William " ต้องเริ่มสร้าง suffix tree ที่ sub tree ของ "William Jefferson Clinton " และเริ่มที่ node แรกเดียวกันจะต้องตรวจสอบว่า nod e ต่อไปซ้ำกันหรือไม่ ถ้าซ้ำก็ไม่ต้องเพิ่ม node ดังรูปที่ 3.9 , 3.10 และ 3.11



รูปที่ 3.9 แสดงตัวอย่างการสร้าง suffix tree ด้วย n-gram <= 3 ของคำตำแหน่งที่ 1 ในเอกสารที่ 1

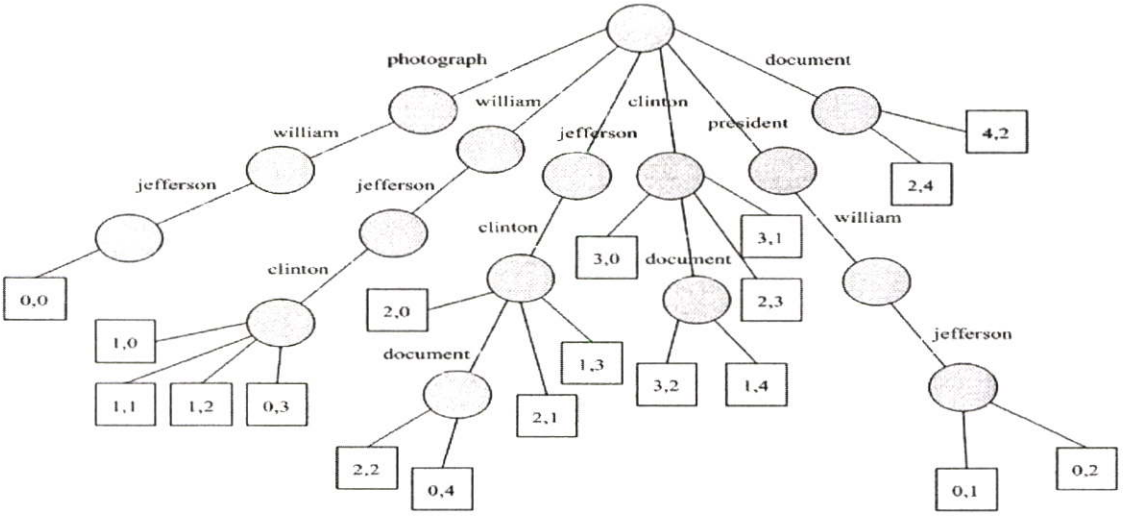


รูปที่ 3.10 แสดงตัวอย่างการสร้าง suffix tree ด้วย n-gram <= 3 ของคำตำแหน่งที่ 2 ในเอกสารที่ 1



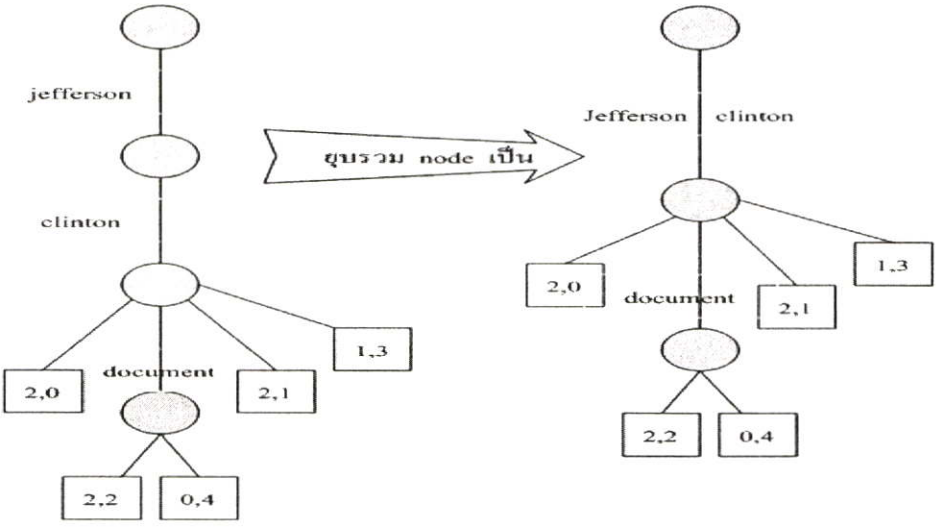
รูปที่ 3.11 แสดงตัวอย่างการสร้าง suffix tree ด้วย n-gram <= 3 ของคำตำแหน่งที่ 3 ในเอกสารที่ 1

การสร้าง suffix tree จะทำงานวนซ้ำไปเรื่อยๆ ตั้งแต่คำแรกของอกสารแรก จนถึงคำสุดท้ายของเอกสารสุดท้าย โดยสามารถจัดกลุ่มเอกสารเพิ่มขึ้นได้เรื่อยๆโดยไม่ต้องเริ่มต้นการจัดกลุ่มใหม่เมื่อมีเอกสารใหม่เข้ามาเราเรียกว่า Incremental Clustering ซึ่งการทำงานแต่ละครั้งเราจะพบกลุ่มเอกสารที่ปรากฏในแต่ละโหนดและป้ายชื่อกลุ่มในแต่ละลิงก์ ทำให้ระบบได้รับกลุ่มเอกสารพร้อมป้ายชื่อกลุ่มเราเรียกว่า Dynamic Clustering and Labeling การทำงานดังกล่าวจะมีผลกลุ่มและป้ายชื่อกลุ่มเปลี่ยนแปลงตามคำภายในผลการสืบค้นที่เข้ามาเราเรียกว่า on-the-fly web search results การทำงานในลักษณะนี้เราเรียกว่า Incremental and Dynamic or on-the-fly Web Search Results Clustering and Labeling ดังแสดงผลการจัดกลุ่มรูปที่ 3.12



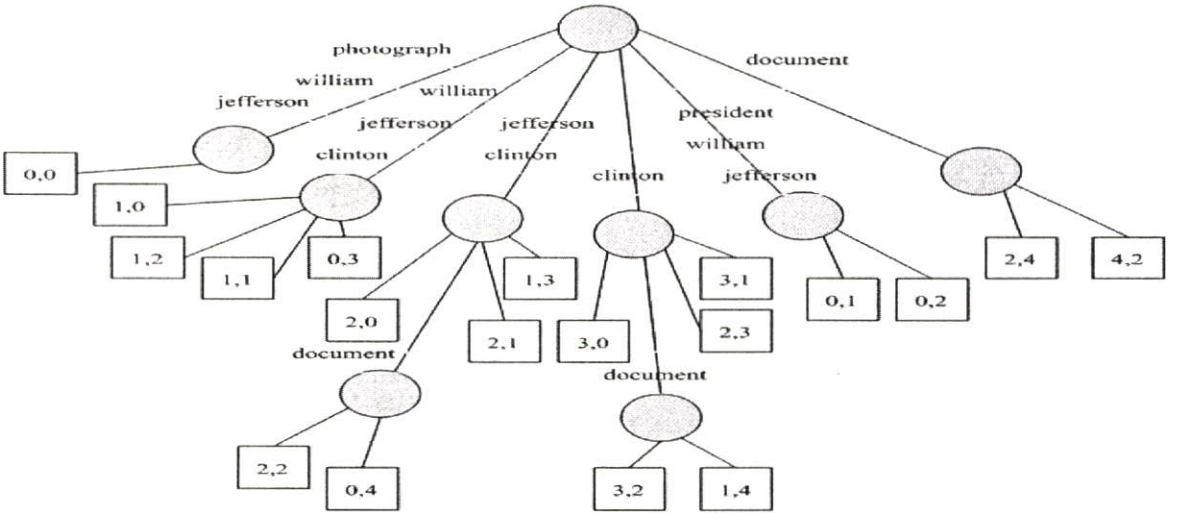
รูปที่ 3.12 แสดงตัวอย่างการสร้าง suffix tree ด้วย n-gram <= 3 ของเอกสารทั้ง 5 เอกสาร

3.2.2.2 การยุบรวม Internal node (Compact Internal Node) เพื่อลดจำนวน node และเวลาในการประมวลผลการสร้างและค้นหา common phrase ที่กลุ่มเอกสารใช้ร่วมกัน จึงทำการยุบรวม internal node ที่ไม่มีเอกสารประกอบอยู่ภายใน node และมีหนึ่ง link ออกจาก node โดยหลักการยุบรวม internal node คือการนำคำของแต่ละ node ที่ถูกยุบรวมจากบนลงล่างของ sub-tree มาต่อกันให้เป็น phrase ดังแสดงตัวอย่างในรูปที่ 3.13 คือ โหนด " Jefferson " ถูกยุบรวมกับ node " Clinton " และนำคำของทั้ง 2 โหนด มารวมกันให้เป็น phrase ผลคือ " Jefferson Clinton " และ node ใหม่คือ "Jefferson Clinton " ไม่ถูกยุบรวมกับ node " document " เพราะ node "Jefferson Clinton " มีเอกสารประกอบอยู่ภายใน node



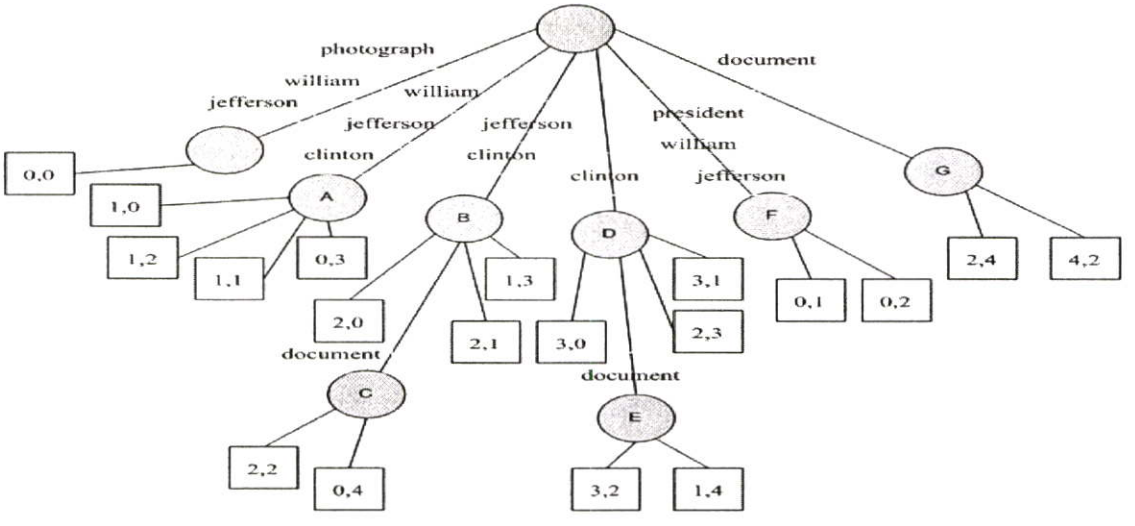
รูปที่ 3.13 แสดงตัวอย่างการยุบรวม internal node " Jefferson " รวมกับ node " Clinton "

จากตัวอย่างการสร้าง suffix tree ของเอกสารตัวอย่าง 5 เอกสาร เมื่อยุบรวม node แล้ว จะได้ผลการยุบรวม internal node เป็น phrase ทำให้ค้นหา phrase ได้ง่าย ดังรูปที่ 3.14



รูปที่ 3.14 แสดงผลการยุบรวม internal node ของเอกสารทั้ง 5 เอกสาร

3.2.2.3 สร้างลำดับรายการของกลุ่มพื้นฐาน จากการกำหนดกลุ่มพื้นฐานซึ่งพิจารณาจากการใช้ common phrase ร่วมกันของกลุ่มเอกสาร โดยพิจารณา node ที่มีเอกสารมากกว่า 1 เอกสารให้เป็นกลุ่มพื้นฐาน ดังแสดงตัวอย่างการกำหนดกลุ่มพื้นฐานในรูปที่ 3.15 และรายการลำดับชื่อกลุ่มพื้นฐาน (List of Base Cluster) ที่ต้องนำคำของแต่ละ node มาต่อกันดังแสดงผลในตารางที่ 3.1



รูปที่ 3.15 แสดงการกำหนดกลุ่มพื้นฐานจากการใช้ common phrase ร่วมกันของกลุ่มเอกสาร

ตารางที่ 3.1 แสดงรายชื่อกลุ่มพื้นฐานที่ได้จากการใช้โครงสร้าง suffix tree ร่วมกับเทคนิค n-gram

Node	Common Phrase	Document(position,document)	Number
A	william jefferson clinton	(1, 0) , (1, 1) , (1, 2) , (0, 3)	4
B	jefferson clinton	(2, 0) , (2, 1) , (1, 3) , (2, 2) , (0, 4)	5
C	jefferson clinton document	(2, 2) , (0, 4)	2
D	Clinton	(3, 0) , (3, 1) , (2, 3) , (3, 2) , (1, 4)	5
E	clinton document	(3, 2) , (1, 4)	2
F	president william jefferson	(0, 1) , (0, 2)	2
G	Document	(4, 2) , (2, 4)	2

จากขั้นตอนการกำหนดกลุ่มพื้นฐาน (Base Cluster Identification) จากการทำงานร่วมกันของโครงสร้างข้อมูลชื่อ suffix tree และเทคนิค n-gram ทำให้ป้ายชื่อกลุ่มพื้นฐานที่ได้รับขาดความสมบูรณ์ เพราะถูกตัดด้วยขนาดของ n-gram เมื่อเทียบกับคำที่ปรากฏจริงในเอกสาร เช่น กลุ่ม " Jefferson Clinton document " ประกอบด้วยเอกสารที่ 2 และ 4 เมื่อนำป้ายชื่อกลุ่มมาเปรียบเทียบกับคำที่ปรากฏจริงในเอกสาร common phrase ที่แท้จริงคือ " William Jefferson Clinton document" ดังนั้นเพื่อความสมบูรณ์ของป้ายชื่อกลุ่ม เราจึงทำการปรับแต่งกลุ่มและป้ายชื่อกลุ่มด้วยขั้นตอนการเชื่อมป้ายชื่อกลุ่มพื้นฐาน (Content Based Combining Base Cluster)

3.2.3 ขั้นตอนการเชื่อมป้ายชื่อกลุ่ม

การรวมกลุ่มพื้นฐานแบบเดิมของ STC จะพิจารณาอัตราความคล้ายคลึงของกลุ่มจากเอกสารที่เป็นสมาชิกภายในคู่กลุ่มเอกสารเท่านั้น ทำให้ได้กลุ่มลักษณะไม่เฉพาะเจาะจงและอัตราการทับซ้อนของเอกสารสูง เช่น A และ B ไม่ถูกรวมเข้าด้วยกันเพราะอัตราความคล้ายคลึงต่ำกว่าค่าความคล้ายคลึงขั้นต่ำที่กำหนดคือ 0.75 แต่ A และ B มีค่าความคล้ายคลึงเพียง 0.6

A = {data , mining} (2, 3, 5, 7, 10)

B = {data , mining , tool} (2,3,5)

นอกจากนี้การจัดกลุ่มผลการสืบค้นด้วยการสร้าง suffix tree โดยการใช้เทคนิค n-gram เข้ามามีส่วนสำคัญในการกำหนดขนาดความสูงของ suffix tree ทำให้มีความสะดวกต่อการพัฒนาและการจัดการกับ suffix tree และประหยัดพื้นที่ในหน่วยความจำ แต่ทำให้ป้ายชื่อกลุ่มขาดความสมบูรณ์ เนื่องจากมันถูกตัดขาดด้วยขนาดของ n-gram ดังนั้นเพื่อความสมบูรณ์ของป้ายชื่อกลุ่มที่ถูกตัดขาด จึงต้องนำป้ายชื่อกลุ่มที่ได้จากการทำงานร่วมกันของ suffix tree กับ n-gram มาเชื่อมต่อกัน โดยพิจาณาการเชื่อมป้ายชื่อกลุ่มหรือ phrase ระหว่างกลุ่มต้นแบบ (A Cluster) และกลุ่มย่อย (B Cluster) จากเงื่อนไขของกรณีการเชื่อมป้ายชื่อกลุ่มเพื่อค้นหากลุ่มใหม่ ดังแสดง

รายละเอียดในหัวข้อที่ 3.2.3.1 และ เพื่อเป็นการลดอัตราการทับซ้อนของเอกสารในของแต่ละกลุ่ม ซึ่งอาจส่งผลให้คุณภาพของกลุ่มลดลงเนื่องจากภายในกลุ่มมีเอกสารที่ไม่ตรงกับความต้องการหรือเอกสารปนเปื้อนภายในกลุ่มมีจำนวนมากเกินไป และทำให้กลุ่มมีขนาดใหญ่ ส่งผลให้ผู้สืบค้นต้องเสียเวลาในการค้นหาเอกสารที่ต้องการภายในกลุ่ม ดังนั้นเพื่อให้กลุ่มมีลักษณะเฉพาะเจาะจงมากขึ้น สมาชิกภายในกลุ่มน้อยลง ลดเอกสารปนเปื้อนภายในกลุ่ม เราจึงมีการลบกลุ่มย่อยและลบเอกสารภายใต้เงื่อนไขการลบในกรณีการลบกลุ่มย่อยและลบเอกสารภายในกลุ่ม ดังแสดงรายละเอียดในหัวข้อที่ 3.2.3.2

3.2.3.1 กรณีการเชื่อมป้ายชื่อกลุ่มเพื่อค้นหากลุ่มใหม่

การเชื่อมป้ายชื่อกลุ่มเพื่อค้นหากลุ่มใหม่ กระบวนการทำงานจะทำงานภายใต้เงื่อนไขการเชื่อมป้ายชื่อกลุ่ม ดังแสดงรายละเอียดดังต่อไปนี้

1. กลุ่มต้นแบบ (A Cluster) หรือกลุ่มย่อย (B Cluster) ต้องมีความยาวของป้ายชื่อกลุ่มมากกว่าหรือเท่ากับขนาดของ n-gram
2. พิจารณาคำที่ปรากฏภายในป้ายชื่อกลุ่มของคู่กลุ่ม คือพิจารณาการมีอยู่ร่วมกันหรือการใช้คำเดียวกันประกอบเป็นป้ายชื่อกลุ่มหรือ phrase ตามเงื่อนไขในสมการที่ 3.1

$$A \oplus B = \left\{ \begin{array}{ll} a_0 & \oplus \\ a_1 = b_0 & \\ a_2 = b_1 & \\ \vdots & \\ a_n = b_{n-1} & \\ \oplus & b_n \end{array} \right\} if \ \mathrm{A}_{(d)} \in \mathrm{B}_{(d)} > 1 \qquad (3.1)$$

เมื่อ a คือ เซตของคำที่ปรากฏใน base cluster A

b คือ เซตของคำที่ปรากฏใน base cluster B

$A_{(d)}$ คือ เซตของเอกสารที่มีตำแหน่งเชื่อมต่อกันในกลุ่ม A

$B_{(d)}$ คือ เซตของเอกสารที่มีตำแหน่งเชื่อมต่อกันในกลุ่ม B

3. พิจาณาตำแหน่งของคำในเอกสารเดียวกันของคู่กลุ่มเอกสารที่เชื่อมต่อกันมีมากกว่า 1 เอกสาร ตามเงื่อนไขในสมการที่ 3.1

ตัวอย่างการเชื่อมป้ายชื่อกลุ่มเพื่อค้นหาป้ายชื่อกลุ่มใหม่ จากป้ายชื่อกลุ่มพื้นฐานที่ถูกตัดขาดจากการใช้เทคนิค n-gram

A = { President , William , Jefferson } (0, 1) , (0, 2)

B = { William , Jefferson , Clinton } (1, 0) , (1, 1) , (1, 2) , (0, 3)

จากตัวอย่างกลุ่ม A และกลุ่ม B จะเป็นไปตามเงื่อนไขทั้ง 3 กรณีคือ

1. คำที่ปรากฏเป็นป้ายชื่อกลุ่มพื้นฐานของทั้ง 2 กลุ่มพื้นฐาน มีการทับซ้อนกันในลักษณะที่สามารถเชื่อมต่อกันได้และมีขนาดที่เท่ากัน ตามสมการที่ 3.1 ดังแสดงในรูปที่ 3.16

A = { president , william , jefferson } (0,1)(0,2)

B = { william , jefferson , clinton } (1,0)(1,1)(1,2)(0,3)

รูปที่ 3.16 ตัวอย่างการมีอยู่ร่วมกันของป้ายชื่อกลุ่มพื้นฐานที่สามารถนำมาเชื่อมต่อกันได้

2. เอกสารที่เป็นสมาชิกของกลุ่มย่อย (B Cluster) มากกว่า 1 เอกสาร เป็นสมาชิกของกลุ่มต้นแบบ (A Cluster) ดังแสดงในสมการที่ 2

3. ตำแหน่งของคำในเอกสารเดียวกันจะต้องเชื่อมต่อกัน และตำแหน่งของกลุ่ม A จะต้องน้อยกว่าหรือมากกว่าตำแหน่งของกลุ่ม B เท่ากับ 1 ตำแหน่ง จากตัวอย่างตำแหน่งของกลุ่ม A คือ ตำแหน่งที่ 0 ในเอกสารที่1 และตำแหน่งที่ 0 ในเอกสารที่ 2 ตำแหน่งของกลุ่ม B คือ ตำแหน่งที่ 1 ในเอกสารที่1 และตำแหน่งที่ 1 ในเอกสารที่ 2 ซึ่งเป็นไปตามเงื่อนไขการเชื่อมต่อกันของตำแหน่งคำในเอกสาร

จากตัวอย่าง กลุ่ม A และกลุ่ม B เป็นไปตามเงื่อนไขในสมการที่ 3.1 ดังนั้นต้องนำกลุ่มทั้งสองกลุ่มมาเชื่อมต่อกัน เพื่อให้ได้กลุ่มใหม่และป้ายชื่อกลุ่มใหม่ ดังแสดงผลลัพธ์การเชื่อมป้ายชื่อกลุ่มและกลุ่มใหม่ที่ได้รับคือ

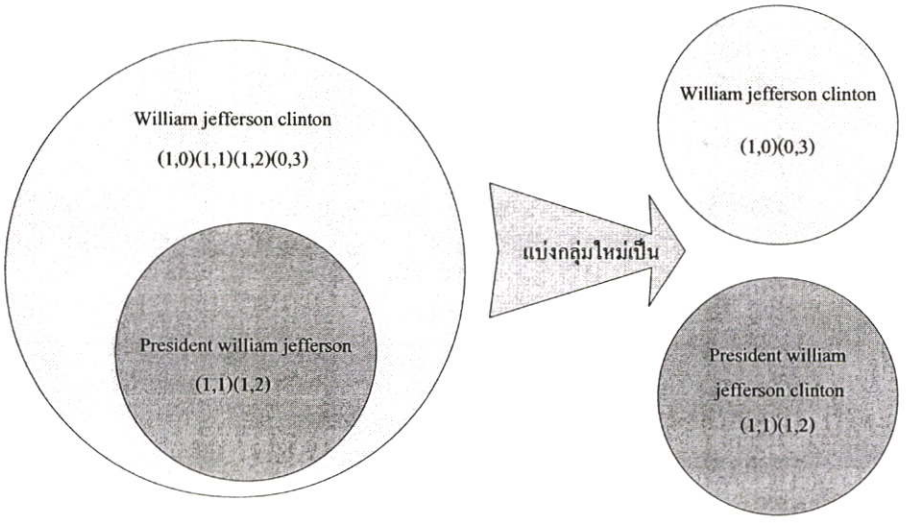
$A \oplus B$ = { President , William , Jefferson , Clinton } (1, 1) , (1, 2)

เมื่อได้กลุ่มใหม่ คือกลุ่มที่มีป้ายชื่อกลุ่มที่เกิดจากจากการเชื่อมป้ายชื่อกลุ่มหรือ phrase ของ 2 กลุ่มเอกสารเข้าด้วยกัน แล้วกลายเป็นป้ายชื่อกลุ่มหรือ phrase ใหม่ พร้อมกับการเปลี่ยนตำแหน่งของคำโดยการใช้ตำแหน่งที่สูงกว่าเข้าแทนที่ เพื่อเพื่อรองรับการเชื่อมป้ายชื่อกลุ่มที่อาจมีขึ้นอีก เพราะตำแหน่งของคำเป็นปัจจัยสำคัญต่อการตัดสินว่าระบบจะทำการเชื่อมป้ายชื่อกลุ่ม เช่น

{ William , Jefferson , Clinton } (1, 0) , (0, 3) และ

{ President , William , Jefferson , Clinton } (1, 1) , (1, 2)

ดังแสดงตัวอย่างการแบ่งกลุ่มใหม่จากการเชื่อมป้ายชื่อกลุ่ม ในรูปที่ 3.17



รูปที่ 3.17 แสดงตัวอย่างการแบ่งกลุ่มใหม่จากการเชื่อมป้ายชื่อกลุ่ม

เพื่อความเร็วในการเชื่อมป้ายชื่อกลุ่มเพื่อค้นหากลุ่มใหม่ กระบวนการทำงานในขั้นตอนนี้จะมีการแบ่งรายการ (List) ของกลุ่มออกเป็น 2 ส่วนคือ ลำดับรายการ (List) ของกลุ่มพื้นฐาน (List_OldSTC) และลำดับรายการ (List) ของกลุ่มที่ได้จากการเชื่อมป้ายชื่อกลุ่ม (List_NewSTC) โดยจะทำการนำกลุ่มใหม่ที่ได้จากการเชื่อมป้ายชื่อกลุ่มมาเก็บไว้ที่ลำดับรายการของ List_NewSTC และการเชื่อมป้ายชื่อกลุ่มในรอบที่ป้ายชื่อมีความยาวมากกว่าขนาดของ n-gram ที่กำหนดในเบื้องต้น การทำงานจะทำอยู่เฉพาะภายในลำดับรายการของ List_NewSTC เท่านั้น เพราะรายงานวิจัยฉบับนี้จะให้ความสำคัญต่อการเชื่อมป้ายชื่อกลุ่มเพื่อค้นหากลุ่มใหม่ เมื่อการเชื่อมป้ายชื่อกลุ่มเพื่อค้นหากลุ่มใหม่สิ้นสุดลง คือไม่มีกลุ่มใดสามารถเชื่อมต่อกันได้อีก กระบวนการทำงานต่อมาคือนำ 2 ลำดับรายการคือ List_NewSTC และ List_OldSTC มาเชื่อมต่อกันเป็นลำดับรายการเดียวกันคือ List_ClusterSTC เพื่อนำไปจัดลำดับความสำคัญของกลุ่มและแสดงผลต่อผู้สืบค้นต่อไป ดังแสดงกระบวนการทำงานในขั้นตอนการเชื่อมป้ายชื่อกลุ่ม (Content Based Combining Base Cluster) ในรูปที่ 3.18

ขั้นตอนการเชื่อมป้ายชื่อกลุ่ม (Content Based Combining Base Cluster)

Step 1: ทำงานกับ List_OldSTC

1.1 ค้นหากลุ่มต้นแบบ (A Cluster) ตามเงื่อนไขกลุ่มต้นแบบต้องมีความยาวของป้ายชื่อกลุ่มมากกว่า 1 คำ

1.2 เปรียบเทียบกลุ่มต้นแบบ (A Cluster) และกลุ่มย่อย (B Cluster) เพื่อตรวจสอบการทำงาน 2 กรณีคือ

1.2.1 กรณีเชื่อมป้ายชื่อกลุ่มเพื่อสร้างกลุ่มใหม่ คือการปรับแต่งกลุ่มใหม่ ถ้าป้ายชื่อกลุ่มต้นแบบ (A cluster) และป้ายชื่อกลุ่มย่อย (B Cluster) มีความยาวป้ายชื่อกลุ่มเท่ากับขนาดของ n-gram และป้ายชื่อกลุ่มเชื่อมกันในลักษณะ $a_1 = b_0$, $a_2 = b_1$, $a_n = b_{n-1}$ และมีจำนวนเอกสารภายในกลุ่มที่มีตำแหน่งคำเชื่อมต่อกันมากกว่า 1

1.2.1.1 เมื่อมีการเชื่อมกลุ่ม ให้ทำสัญลักษณ์ (mark) กับเอกสารที่เหมือนกันของทั้งสองกลุ่ม เพื่อลบเอกสารเหล่านี้ออกจากกลุ่มเดิม เมื่อทำงานจนครบทุกกลุ่มพื้นฐาน

1.2.1.2 นำกลุ่มใหม่ที่ได้จากการเชื่อมป้ายชื่อกลุ่มเก็บไว้ที่ List_NewSTC

1.2.2 กรณีลบกลุ่มย่อยหรือลบเฉพาะเอกสารที่ทับซ้อนกัน ถ้ากลุ่มย่อยมีป้ายชื่อกลุ่มเป็น subset ของกลุ่มต้นแบบ หรือลบเอกสารของกลุ่มย่อยที่ทับซ้อนกับกลุ่มต้นแบบ ถ้ากลุ่มย่อยมีป้ายชื่อกลุ่มเป็นคำเดี่ยว

1.3 กลับไปทำงานในขั้นตอนที่ 1.1 จนกว่าจะครบทุกกลุ่มใน List_OldSTC

1.4 ลบเอกสารที่มีเครื่องหมายสัญลักษณ์การลบออกจากกลุ่มพื้นฐาน

Step 2: ทำงานกับ List_NewSTC

2.1 ค้นหากลุ่มต้นแบบ (A Cluster) ตามเงื่อนไขความยาวของป้ายชื่อกลุ่มมีค่ามากกว่า n-gram และขนาดจะเพิ่มขึ้นครั้งละ 1 คำ

2.2 เปรียบเทียบกลุ่มต้นแบบ (A Cluster) และกลุ่มย่อย (B Cluster) ที่มีขนาดความยาวของป้ายชื่อกลุ่มเท่ากัน เพื่อตรวจสอบการเชื่อมป้ายชื่อกลุ่มเพื่อสร้างกลุ่มใหม่เช่นเดียวกับ 1.2.1

2.3 กลับไปที่ 2.1 จนกว่าจะหยุดการทำงานลงเมื่อไม่มีกลุ่มที่สามารถเชื่อมกันได้อีก

2.4 ลบเอกสารที่มีเครื่องหมายสัญลักษณ์การลบออกจากกลุ่มพื้นฐาน

Step 3: นำ List_OldSTC และ List_NewSTC มารวมกัน

รูปที่ 3.18 แสดงกระบวนการทำงานของการเชื่อมป้ายชื่อกลุ่ม

จากขั้นตอนของการเชื่อมป้ายชื่อกลุ่มพื้นฐาน (Content Based Combining Base Cluster) เพื่อค้นหาป้ายชื่อกลุ่มที่ขาดความสมบูรณ์ เนื่องจากถูกตัดด้วยขนาดของ n-gram และทำการปรับแต่งกลุ่มใหม่ จะช่วยให้ขนาดของกลุ่มเล็กลงเพราะในขณะที่เชื่อมป้ายชื่อกลุ่มเพื่อค้นหากลุ่มใหม่นั้น ระบบจะทำการลบเอกสารของกลุ่มเล็กที่ทับซ้อนในกลุ่มใหญ่ออกด้วย การลบสมาชิกภายในกลุ่มต้นแบบ (A Cluster) และกลุ่มย่อย (B Cluster) ระบบจะไม่ลบเอกสารออกจากกลุ่มในทันที แต่ใช้วิธีการทำเครื่องหมายหรือสัญลักษณ์การลบไว้ก่อน เพราะอาจมี Phrase อื่นๆที่สามารถเชื่อมต่อได้อีก ซึ่งจะช่วยให้อัตราการทับซ้อนของเอกสารลดลงและเมื่ออัตราการทับซ้อนของเอกสารลดลง ส่งผลให้ระดับของจำนวนเอกสารที่ไม่ตรงกับความต้องการ (Irrelevant Document) ภายในกลุ่มลดลงตามไปด้วยเช่นกัน ดังนั้นคุณภาพของกลุ่มจะสูงขึ้น เพราะมีเอกสารที่ตรงกับความต้องการสูงกว่าเอกสารที่ไม่ตรงกับความต้องการ และสามารถจัดกลุ่มได้ถูกต้อง (Precision) กว่าระบบการจัดกลุ่มผลการสืบค้นด้วย Suffix Tree Clustering (STC) ในรูปแบบเดิม

การพิจารณาตำแหน่งของคำในเอกสารเดียวกัน ดังแสดงในสมการที่ 3.1 เพื่อป้องกันความผิดพลาดในการเชื่อมป้ายชื่อกลุ่ม เพราะบางครั้งเอกสารหนึ่งๆ อาจมีคำที่สามารถเชื่อมต่อกันได้แต่ความเป็นจริงแล้วไม่ใช่กลุ่มที่สามารถเชื่อมกันเพื่อให้มันกลายเป็นกลุ่มเอกสารใหม่ เช่น

A = {transport , property} (1,34) , (3,53) และ

B = {property , pure} (5,34) , (4,53)

กรณีเช่นนี้ 2 กลุ่มพื้นฐานมีป้ายชื่อกลุ่มที่สามารถเชื่อมกันได้ และมีเอกสารที่เป็นสมาชิกในกลุ่มเป็นเอกสารเดียวกัน แต่เมื่อดูตำแหน่งการเชื่อมต่อกันของ phrase 2 กลุ่มพื้นฐานนี้จะไม่ถูกเชื่อมเข้าเป็น phrase เดียวกัน เพราะเอกสารที่ 53 และ 34 ไม่ได้ใช้ phrase " transport property pure " ร่วมกันจริง มันปรากฏเฉพาะในเอกสารที่ 53 เท่านั้น แต่ในเอกสารที่ 34 จะปรากฏดังนี้ คือ " ..ตำแหน่งที่ 0.. , transport , property , ...ตำแหน่งที่ 3.. , ...ตำแหน่งที่ 4... , property , pure " จะเห็นว่าตำแหน่งของคำไม่ต่อกัน ดังนั้นทั้ง 2 กลุ่มไม่ควรถูกรวมเข้าเป็นกลุ่มเดียวกัน

3.2.3.1 กรณีการลบกลุ่มย่อยและลบเอกสารภายในกลุ่ม

เพื่อลดเวลาในการทำงานและลดอัตราการทับซ้อนของเอกสาร เพราะเอกสารหนึ่งๆอาจปรากฏได้มากกว่า 1 กลุ่มเอกสาร หรืออาจกล่าวได้ว่าเอกสารหนึ่งอาจมีเนื้อหาสอดคล้องกับหลายเอกสาร แต่การทับซ้อนของเอกสารจำนวนมากจะทำให้คุณภาพหรือค่าความถูกต้องของกลุ่มลดลง เพื่อลดจำนวนกลุ่มและจำนวนเอกสารที่ทับซ้อนกัน และลดเวลาในการประมวลผล การจัดกลุ่มผลการสืบค้นด้วย Suffix Tree Clustering แนวใหม่ จึงมีการลบกลุ่มหรือลบเอกสารของกลุ่มย่อย โดยกลุ่มย่อยคือกลุ่มพื้นฐานที่ไม่ใช่กลุ่มต้นแบบและมีป้ายชื่อกลุ่มเป็น subset ของกลุ่ม

ต้นแบบหรือป้ายชื่อกลุ่มเป็นคำเดี่ยว การลบกลุ่มหรือลบเอกสารของกลุ่มย่อยที่เข้าเงื่อนไขการลบกลุ่มออกจากลำดับรายการของกลุ่มพื้นฐาน หรือ จาก 2 กรณีคือ

1. ลบเอกสารของกลุ่มย่อยที่ทับซ้อนกับเอกสารของกลุ่มต้นแบบ เช่น

A = { William , Jefferson , Clinton } (1,0) (1,1) (1,2) (0,3)

B = { Jefferson , Clinton } (2,0) (1,3) (5,6) (7,8)

กลุ่ม B จะถูกลบสมาชิกที่เหมือนกับกลุ่ม A คือ (2,0) (1,3) กลุ่ม B จะเหลือสมาชิกเพียง 2 เอกสารคือ B = { Jefferson , Clinton } (5,6) (7,8)

2. ระบบจะทำการลบกลุ่มพื้นฐาน B เมื่อเปรียบเทียบกับกลุ่มพื้นฐาน A แล้วปรากฏว่ากลุ่ม B คือ sub-set ของกลุ่ม A โดยจะทำการลบส่วนประกอบทั้งหมดของกลุ่มพื้นฐาน B ซึ่งประกอบด้วย ป้ายชื่อกลุ่ม , ตำแหน่งของคำ และหมายเลขเอกสารที่เป็นสมาชิกภายในกลุ่มออกจากลำดับรายการของกลุ่มพื้นฐาน เมื่อพบว่า ป้ายชื่อกลุ่มของกลุ่มย่อย (B Cluster) ทั้งหมดทับซ้อนกับป้ายชื่อกลุ่มของกลุ่มต้นแบบ (A Cluster) และเอกสารที่เป็นสมาชิกของกลุ่ม B ทุกตัวเป็นสมาชิกของกลุ่ม A

ตัวอย่างกลุ่มพื้นฐาน B ที่จะต้องถูกลบทิ้งในกรณีเป็นไปตามเงื่อนไขของสมการที่ 3.2

A = { William , Jefferson , Clinton } (1,0) (1,1) (1,2) (0,3)

B = { Jefferson , Clinton } (2,0) (2,1) (2,2) (1,3)

จากตัวอย่างคำที่ปรากฏเป็นป้ายชื่อของกลุ่มพื้นฐาน B ตั้งแต่ตำแหน่งแรกจนถึงตำแหน่งสุดท้ายซ้ำกับคำที่ปรากฏเป็นป้ายชื่อของกลุ่มพื้นฐาน A และสมาชิกทุกตัวของกลุ่มพื้นฐาน B เป็นสมาชิกของกลุ่มพื้นฐาน A ดังนั้นกลุ่มพื้นฐาน B จะถูกลบออกจากลำดับรายการของกลุ่มพื้นฐาน เพราะถือว่ากลุ่ม B เป็นกลุ่มย่อยของกลุ่ม A และกลุ่มเอกสารเหล่านั้นได้มีการปรากฏเป็นกลุ่มแล้วในกลุ่ม A จึงไม่จำเป็นต้องแสดงหรือมีกลุ่มอีก ซึ่งเป็นการลดความทับซ้อนของกลุ่มเอกสาร

การเชื่อมป้ายชื่อกลุ่มเพื่อค้นหากลุ่มใหม่ในขั้นตอนการเชื่อมป้ายชื่อกลุ่มพื้นฐาน (Content Based Combining Base Cluster) จะช่วยให้ค้นพบป้ายชื่อกลุ่มที่แท้จริงหรือป้ายชื่อกลุ่มที่สมบูรณ์ของกลุ่มเอกสาร ซึ่งจากเดิมถูกตัดความสมบูรณ์ของป้ายชื่อกลุ่มด้วยขนาดของ n-gram และการลบกลุ่มย่อยหรือลบเอกสารของกลุ่มย่อยที่ทับซ้อนกับกลุ่มต้นแบบ จะช่วยลดจำนวนกลุ่มและขนาดของกลุ่ม พร้อมทั้งลดอัตราการทับซ้อนของเอกสารที่มีมากเกินไป ทำให้ค่าความถูกต้อง (Precision) สูงขึ้น คุณภาพของกลุ่มย่อมสูงขึ้นตามไปด้วยเพราะอัตราของเอกสารที่ไม่ตรงกับความต้องการลดต่ำลง ดังตัวอย่างการทำงานดังต่อไปนี้

ตัวอย่างจากเอกสาร 5 เอกสาร ซึ่งเป็นผลการสืบค้นจากคำว่า " Jefferson Clinton " ดังมีกลุ่มพื้นฐานแสดงในตารางที่3.1 ระบบจะทำการค้นหากลุ่มต้นแบบ (A Cluster) ซึ่งต้องมีความยาวป้ายชื่อกลุ่มมากกว่า 1 คำ จากตารางที่ 3.1 กลุ่มต้นแบบกลุ่มแรกคือกลุ่ม william jefferson clinton

และกลุ่มที่เหลือเป็นกลุ่มย่อย (B Cluster) แล้วทำการเปรียบเทียบตามเงื่อนไขของสมการที่ 3.1 และการลบกลุ่มและการเชื่อมป้ายชื่อกลุ่มเพื่อให้ได้กลุ่มใหม่ ดังแสดงผลการทำงานในตารางที่ 3.2

ตารางที่ 3.2 แสดงผลการทำงานและผลการทำงานของกลุ่ม A = william jefferson clinton

Num.	Common Phrase (Cluster Label)	Document(position,doc)	หมายเหตุ
1	william jefferson clinton	(1,0)(1,1)(1,2)(0,3)	
2	jefferson clinton	(0,4)	ถูกลบเอกสารที่ทับซ้อนกับกลุ่ม A เพราะป้ายชื่อกลุ่ม B เป็น sub-set ของกลุ่ม A
3	jefferson clinton document	(2,2)(0,4)	ไม่เชื่อมเพราะสมาชิกของ B cluster เป็นสมาชิกของ A cluster ไม่มากกว่า 1 เอกสาร
4	clinton	(1,4)	ถูกลบเอกสารที่ทับซ้อนกับกลุ่ม A เพราะป้ายชื่อกลุ่ม B เป็น sub-set ของกลุ่ม A
5	clinton document	(3,2)(1,4)	
6	president william jefferson	(0,1)(0,2)	
7	document	(4,2)(2,4)	

จากตารางที่ 3.2 จะพบว่ากลุ่มย่อย (B cluster) ที่มีป้ายชื่อกลุ่มทับซ้อนกับกลุ่มต้นแบบ (A cluster) ทั้งหมดหรืออาจกล่าวได้ว่ากลุ่มย่อยมีป้ายชื่อกลุ่มเป็นsub-set ของกลุ่มต้นแบบแต่สมาชิกของกลุ่มไม่เป็น sub-set ของกลุ่มต้นแบบ ระบบจะทำการลบเอกสารที่ทับซ้อนกับกลุ่มต้นแบบ รวมทั้งตำแหน่งของคำออกจากกลุ่มย่อย ทำให้อัตราการทับซ้อนของเอกสารลดลง

เมื่อการทำงานของกลุ่ม A สิ้นสุดลง กลุ่ม A จะถูกเปลี่ยนเป็นกลุ่มถัดมาที่มีความยาวป้ายชื่อกลุ่มมากกว่า 1 คำ และกลุ่มที่เหลือตั้งแต่กลุ่มที่ 1 ถึงกลุ่มสุดท้ายจะเป็นกลุ่ม B จากตารางที่ 3.2 กลุ่มต้นแบบคือกลุ่ม " jefferson clinton" และ กลุ่มย่อย "clinton" ถูกลบกลุ่มทิ้งไป เพราะป้ายชื่อกลุ่มทับซ้อนกับป้ายชื่อกลุ่ม A ทั้งหมด และสมาชิกภายในกลุ่มทุกตัวเป็นสมาชิกของกลุ่ม A ดังแสดงผลการทำงานในตารางที่ 3.3

ตารางที่ 3.3 แสดงผลการทำงานของกลุ่ม A = jefferson clinton

Num.	Common Phrase (Cluster Label)	Document(position,doc)	หมายเหตุ
1	william jefferson clinton	(1,0)(1,1)(1,2)(0,3)	
2	jefferson clinton	(0,4)	
3	jefferson clinton document	(2,2)(0,4)	
4	clinton document	(3,2)(1,4)	
5	president william jefferson	(0,1)(0,2)	
6	document	(4,2)(2,4)	

ผลจากตารางที่ 3.3 กลุ่มต้นแบบกลุ่มต่อมาคือกลุ่ม " jefferson clinton document" และกลุ่มย่อยที่ถูกลบทิ้งคือกลุ่ม "jefferson clinton", "clinton document" และกลุ่ม "document" เพราะป้ายชื่อกลุ่มทับซ้อนกับป้ายชื่อกลุ่ม A ทั้งหมด และสมาชิกภายในกลุ่มทุกตัวเป็นสมาชิกของกลุ่ม A ทำให้จำนวนกลุ่มภายในลำดับรายการของ List_OldSTC คงเหลือเท่ากับ 3 กลุ่ม ดังแสดงผลการทำงานในตารางที่ 3.4

ตารางที่ 3.4 แสดงผลการทำงานของกลุ่ม A = jefferson clinton document

Num.	Common Phrase (Cluster Label)	Document(position,doc)	หมายเหตุ
1	william jefferson clinton	(1,0)(1,1)(1,2)(0,3)	
2	jefferson clinton document	(2,2)(0,4)	
3	president william jefferson	(0,1)(0,2)	

จากตารางที่ 3.4 กลุ่มต้นแบบ (A Cluster) ต่อมาคือกลุ่ม "president william jefferson" ซึ่งมีป้ายชื่อกลุ่มเชื่อมต่อกับกลุ่มย่อย (B Cluster) คือกลุ่ม "william jefferson clinton" ในลักษณะ $a_1 = b_0$ ถึง $a_n = b_{n-1}$ และมีตำแหน่งของคำเชื่อมต่อกันในเอกสารเดียวกันของทั้งคู่กลุ่มมากกว่า 1 เอกสาร คือ กลุ่มต้นแบบ (A Cluster) = (0,1)(0,2) และกลุ่มย่อย (B Cluster) = (1,0)(1,1) ดังนั้น กลุ่มต้นแบบและกลุ่มย่อยดังกล่าวต้องถูกเชื่อมกลุ่มเข้าด้วยกัน และนำไปเก็บใส่ไว้ในลำดับรายการของ List_NewSTC ดังแสดงผลการเชื่อมป้ายชื่อกลุ่มเพื่อค้นหากลุ่มใหม่ในตารางที่ 3.5

ตารางที่ 3.5 แสดงผลการเชื่อมป้ายชื่อกลุ่มภายในลำดับรายการของ List_NewSTC

Num.	Common Phrase (Cluster Label)	Document(position,doc)	หมายเหตุ
1	president william jefferson clinton	(1, 1) , (1, 2)	

หมายเหตุ จากรูปที่ 3.18 กระบวนการทำงานของการเชื่อมป้ายชื่อกลุ่มพื้นฐานการเชื่อมป้ายชื่อกลุ่มเพื่อค้นหากลุ่มใหม่ หลังจากตรวจสอบเงื่อนไข เพื่อเข้าสู่กระบวนการเชื่อมป้ายชื่อกลุ่มเพื่อค้นหากลุ่มใหม่ ระบบจะทำการเชื่อมป้ายชื่อกลุ่มและสร้างกลุ่มใหม่โดยการนำสมาชิกของกลุ่มที่ทับซ้อนกันพร้อมกับตำแหน่งของคำที่มากกว่ามาเป็นสมาชิกของกลุ่มใหม่ พร้อมทั้งลบสมาชิกที่เหมือนกันของกลุ่มต้นแบบและกลุ่มย่อยออกจากกลุ่มทั้งสองกลุ่ม ซึ่งการลบสมาชิกภายในกลุ่มต้นแบบ (A Cluster) และกลุ่มย่อย (B Cluster) ใช้วิธีการทำเครื่องหมายหรือสัญลักษณ์ (Mark) การลบไว้ก่อน โดยจะไม่ลบในทันทีเพราะอาจมีป้ายชื่อกลุ่มอื่นๆที่สามารถเชื่อมต่อได้อีก และจะลบเอกสารออกจากกลุ่มเมื่อกลุ่มต้นแบบทำงานมาถึงกลุ่มพื้นฐานสุดท้าย ดังแสดงตัวอย่างในตารางที่ 3.5 และ ผลการทำงานในตารางที่ 3.6

ตารางที่ 3.6 แสดงตัวอย่างการ mark เอกสารที่ต้องถูกลบทิ้งเมื่อมีการเชื่อมป้ายชื่อกลุ่ม

Num.	Common Phrase (Cluster Label)	Document(position,doc)	หมายเหตุ
1	william jefferson clinton	(1,0)(1,1*)(1,2*)(0,3)	
2	jefferson clinton document	(2,2)(0,4)	
3	president william jefferson	(0,1*)(0,2*)	

การทำงานในขั้นตอนการรวมกลุ่มพื้นฐาน (Content Based Combining Base Cluster) จะสิ้นสุดเมื่อไม่มีการลบกลุ่มหรือลบเอกสาร และไม่มีการรวมกลุ่มพื้นฐานใดๆในรอบการทำงาน เอกสารที่ถูกทำเครื่องหมายเพื่อลบทิ้งจะถูกลบทิ้งไป และกลุ่มที่ไม่มีเอกสารภายในกลุ่มจะถูกลบทิ้งไปด้วยเช่นกัน ดังแสดงผลการรวมกลุ่มพื้นฐานในตารางที่ 3.7

ตารางที่ 3.7 แสดงกลุ่มคงเหลือภายในลำดับรายการของ List_OldSTC

Num.	Common Phrase(Cluster Label)	Document(position,document)	จำนวน
0	william jefferson clinton	(1, 0) , (0, 3)	2
1	jefferson clinton document	(2, 2) , (0, 4)	2

จากตารางที่ 3.5 แสดงผลการเชื่อมป้ายชื่อกลุ่มภายในลำดับรายการของ List_NewSTC และ 3.7 แสดงกลุ่มคงเหลือภายในลำดับรายการของ List_OldSTC เมื่อในลำดับรายการของ List_OldSTC และ List_NewSTC ไม่มีการกระทำใดๆเกิดขึ้น เนื่องจาก List_NewSTC มีเพียงกลุ่มเดียว ต่อมาจึงนำทั้ง 2 ลำดับรายการมารวมกันเป็นลำดับรายการเดียวคือ List_ClusterSTC ดังแสดงผลในตารางที่ 3.8

ตารางที่ 3.8 แสดงกลุ่มภายในลำดับรายการของ List_ClusterSTC

Num.	Common Phrase(Cluster Label)	Document(position,document)	จำนวน
0	william jefferson clinton	(1, 0) , (0, 3)	2
1	jefferson clinton document	(2, 2) , (0, 4)	2
2	president william jefferson clinton	(1, 1) , (1, 2)	2

จากตัวอย่างเอกสารทั้ง 5 เอกสารจะเห็นว่าการรวมกลุ่มพื้นฐานด้วย content based combining base cluster สามารถค้นพบป้ายชื่อกลุ่มที่ขาดหายไปจากการใช้เทคนิค n-gram ทำให้ป้ายชื่อกลุ่มสมบูรณ์ตามที่ปรากฏในเอกสาร ขนาดและจำนวนกลุ่มเอกสารลดลง ต่อมาระบบจะทำการจัดลำดับความสำคัญของกลุ่มตามขั้นตอนการทำงานในส่วนของการจัดลำดับความสำคัญของกลุ่ม (Ranking content based cluster)

3.2.4 ขั้นตอนการจัดลำดับความสำคัญของกลุ่ม

การจัดลำดับความสำคัญของกลุ่มเอกสาร (Ranking content based cluster) ที่ได้จากการรวมกลุ่มพื้นฐาน เป็นการแสดงผลการจัดกลุ่มตามลำดับคะแนนความสำคัญของกลุ่มเอกสาร ซึ่งรายงานวิจัยฉบับนี้จะให้ลำดับความสำคัญของกลุ่มเอกสาร จากการพิจารณา 3 ส่วนประกอบตามแนวความคิดของงานวิจัยเรื่อง “Learning to Cluster Web Search Results” [5] ประกอบด้วย

1. ป้ายชื่อกลุ่ม (Cluster Label) การจัดลำดับความสำคัญกลุ่มจะให้ความสำคัญกับป้ายชื่อที่มีลักษณะเป็น phrase ที่มีขนาดไม่ยาวเกินไปมากกว่าคำเดี่ยวหรือป้ายชื่อกลุ่มที่มีความยาวมากจนเกิดไป เพราะ phrase มีอำนาจจำแนกมากกว่าคำ และบางครั้งป้ายชื่อที่มีลักษณะเป็นคำเดี่ยว อาจเกิดข้อผิดพลาดของการจัดกลุ่มในเรื่องความหมาย เช่นกลุ่มเอกสารชื่อ new อาจเกิดมาจากเอกสารที่มีคำว่า “news” ปรากฏอยู่ในเอกสาร และ เอกสารที่มีคำว่า “New York” ปรากฏอยู่ในเอกสาร เพราะระบบไม่สามารถแยกได้ว่า “new” มาจากคำที่มีความหมายใด ดังนั้นระบบจึงลดความสำคัญของกลุ่มที่มีป้ายชื่อกลุ่มเป็นคำเดี่ยว โดยการให้คะแนนความสำคัญเป็น 0 คะแนน และป้ายชื่อกลุ่มที่ยาวมากๆ อาจเกิดจากเอกสารทั้งสองเอกสารหรือเอกสารทั้งกลุ่มคือเอกสารที่มีเนื้อหาเดียวกัน หรือเป็นเอกสารเดียวกัน ซึ่งไม่มีประโยชน์ที่แสดงผลในลำดับต้นๆ

2. จำนวนเอกสารที่เป็นสมาชิกภายในกลุ่มเอกสาร ลำดับความสำคัญของกลุ่มเอกสารจะขึ้นอยู่กับจำนวนสมาชิกภายในกลุ่มเอกสาร เพื่อให้ค่าน้ำหนักตกไปที่กลุ่มขนาดใหญ่ในลักษณะของ phrase มากกว่ากลุ่มขนาดใหญ่แต่มีป้ายชื่อเป็นคำเดี่ยว

3. เนื่องจากระบบให้ความสำคัญกับ phrase มากกว่าคำเดี่ยว เมื่อต้องจัดลำดับความสำคัญของกลุ่ม ระบบจึงใช้ความถี่ของ phrase ที่ปรากฏในเอกสารที่เป็นสมาชิกของกลุ่มเป็นตัวกำหนดความสำคัญของกลุ่ม โดยการทำ invert ความถี่ของ phrase คือการทำให้ค่าคะแนนความถี่ของ phrase ลดลงก่อนนำมาคำนวณรวมกับการคิดคะแนนความสำคัญของกลุ่ม เพื่อป้องกัน phrase ที่มีความถี่ในเอกสารจำนวนมาก แต่อาจเป็น phrase ที่ไม่มีความหมายในการจำแนกกลุ่ม เช่น ในรายการเอกสารมีจำนวนเอกสารทั้งหมด 100 เอกสาร และมีเอกสารที่มี phrase p_i ปรากฏอยู่จำนวน 90 เอกสาร นั่นคือ phrase p_i จะไม่มีอำนาจจำแนกในการจัดกลุ่ม ดังนั้นค่าคะแนน $(tf(p_i,d))$ ของ phrase จะลดต่ำลง ดังสมการที่ 3.3

$$tfidf(p_i,d) = (1+\log(tf(p_i,d))) * \log(1+N/df(p_i)) \qquad (3.3)$$

เมื่อ $(tf(p_i,d))$ คือ จำนวนความถี่ของ phrase p_i ในเอกสาร d

$df(p_i)$ คือ จำนวนความถี่ของเอกสารที่มี phrase p_i ปรากฏอยู่

N คือ จำนวนเอกสารทั้งหมด

จากองค์ประกอบ 3 องค์ประกอบ ในการจัดลำดับความสำคัญของกลุ่ม ทำให้กลุ่มที่ได้แสดงผลในลำดับต้นๆ คือกลุ่มขนาดใหญ่ ที่มีป้ายชื่อเป็น phrase และ phrase จะต้องปรากฏอยู่ภายในเอกสารที่เป็นสมาชิกของกลุ่มในจำนวนที่เหมาะสม ตามหลักการคิดคะแนนความสำคัญของกลุ่ม ดังสมการที่ 3.4

$$s(m) = |d| * f|m_p| * \sum tfidf(p_i) \qquad (3.4)$$

$$f|m_p| = \begin{cases} 0, if\ |p| = 1 \\ |p|, if\ 2 \le |m_p| \le 8 \\ \alpha, if\ |p| > 8 \end{cases}$$

เมื่อ $|d|$ คือ จำนวนเอกสารภายในกลุ่ม

$|m_p|$ คือ จำนวนคำที่ปรากฏในป้ายชื่อกลุ่ม

$tfidf(p_i)$ คือ ค่า invert ความถี่ของ phrase ที่ปรากฏอยู่ในกลุ่ม

ตัวอย่างการคำนวณค่า $tfidf(p_i)$ ของกลุ่ม jefferson clinton document ในเอกสารที่ 2 ปรากฏคำว่า "jefferson clinton document" 1 ครั้ง ค่า tf(p_i ,d) = 1 และจำนวนเอกสารที่ปรากฏคำว่า "jefferson clinton document" คือเอกสาร 2,4 ค่า df = 2 จากสมการที่5 แทนค่าในสมการดังนี้

$$tfidf(p_i,d) = (1+\log(tf(p_i,d))) * \log(1+N/df(p_i))$$

$$= (1+\log(1)) * \log(1+ (5/2))$$

$$= (1+0) * 1.098$$

$$= 1.098$$

จากแนวทางการจัดลำดับความสำคัญด้วยการใช้ *phrase* ของสมการที่ 6 สามารถจัดลำดับความสำคัญของกลุ่มของเอกสารตัวอย่างทั้ง 5 เอกสาร ดังแสดงผลการจัดลำดับในตารางที่ 3.10

เอกสารที่0 : Photographs William Jefferson Clinton

เอกสารที่1 : President William Jefferson Clinton

เอกสารที่2 : President William Jefferson Clinton Document

เอกสารที่3 : William Jefferson Clinton

เอกสารที่4 : Jefferson Clinton Documents

ตารางที่ 3.9 แสดงผลการจัดลำดับความสำคัญของกลุ่ม

Num.	Common Phrase	Document	$s(m)$
1	President william jefferson clinton	1,2	2*4*2.196 = 17.568
2	william jefferson clinton	0,3	2*3*2.196 = 13.176
3	jefferson clinton document	2,4	2*3*2.196 = 13.176

3.3 ตัวอย่างการทำงานของอัลกอริทึม

จากกระบวนการทำงานของการจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งแนวใหม่ (A New Suffix Tree Clustering Algorithm) ซึ่งเป็นกระบวนการทำงานในแนวทางการรวมกลุ่ม (Merged Clusters) แบบใหม่ ของการจัดกลุ่มผลการสืบค้น (Web Search Results Clustering) เพื่อให้กลุ่มเอกสารมีลักษณะเฉพาะเจาะจงมากขึ้น ดังแสดงตัวอย่างการทำงานของอังกอริทึมหรือกระบวนการทำงาน ด้วยตัวอย่างเอกสารซึ่งเป็นผลการสืบค้นจากคำว่า " Clinton " จากระบบสืบค้นชื่อ " google.com " จำนวน 19 ผลการสืบค้นดังนี้

1.Hillary Rodham Clinton, Senator from New York Official site, including information about New York, biography, constituent services, contact details, committees and legislation, issues, news and speeches, links, and latest events.

2.Biography of William J. Clinton Short biography from the official White House site.

3.Welcome to Clinton County, NY! Links to government agencies, schedule of events, weather forecasts, and a photo gallery of the area.

4.Bill Clinton - Wikipedia, the free encyclopedia Encyclopedia article on Clinton's life and Presidential administration. Includes hypertext links to related articles.

5.William J. Clinton Foundation Charitable foundation organized by President Clinton, focussing on global issues of health security, economic empowerment, leadership and citizenship, and racial, ethnic and religious reconciliation ...

6.Village of Clinton, British Columbia, Canada - Village of Clinton: P.O. Box 309, 1423 Cariboo Hwy, Clinton, B.C. Canada, V0K 1K0 E-mail: admin@village.clinton.bc.ca Phone 1-250-459-2261, Fax 1-250-459-2227

7.Clinton County welcomes you! — Clinton County, Ohio ... A growing community answering the challenges of merging rural and urban ideals. Includes list of county agencies.

8.Clinton Community College 11.18.05 CLINTON COMMUNITY COLLEGE BOARD OF TRUSTEES MEET Read more Click here for more campus announcements Prospective Students About our area Admissions Continuing Education College Catalog ...

9.Village Of Clinton: [Digital Towpath] The Village of Clinton is 218 years old! The picture below is the Gazebo in the Village Park News: 2nd Annual Sotrmwater Management Report May 31, 2005 The 2005 Annual Stormwater Management Report is ...

10.The Clinton Herald, Clinton, Iowa--Homepage Local and sports news, classifieds, business directory, obituaries, weather, community information and advertising details included.

11.William J. Clinton Foundation President Clinton in the News: 1/01: 60 Minutes Transcript: Bill Clinton's Efforts to Combat AIDS Worldwide 12/13: Announcement of Partnership with Viet Nam to Fight Against HIV/AIDS 12/07: Announcement of ...

12.Welcome to Clinton, Massachusetts Includes local headlines, town calendar, facts, forum, history, and images.

13.Home Page Includes downloadable accident reports, citation payment, bike laws, and local advisories.

14.Clinton County, Illinois--Home Page Government site offering information on the county board, offices, meeting calendar, local municipalities and townships, historical society and museum, war memorial, county fair and 4H activities.

15.Clinton County Government Includes information on departments and services.

16.Clinton Presidential Materials Project Clinton Presidential Materials Project introduction page. The Clinton Presidential Materials Project has become the William J. Clinton Presidential Library and Museum. Visit the new web site, http ...

17.Link to Article Iraq repeatedly blocked UNSCOM from inspecting suspect sites. Iraq repeatedly restricted UNSCOM's ability to obtain necessary evidence. Iraq tried to stop an UNSCOM biological weapons team from ...

18.The Official Site of Clinton, Mississippi Official city site. Features information on the history, government, landmarks, schools, and attractions.

19.Clinton County Iowa Home Page The official Web Site for Clinton County Government. A portal to information on elected officials, departments, and services provided by Clinton County.

จาก 19 ผลการสืบค้นที่ได้จาก google.com เมื่อต้องการนำมาผ่านกระบวนการทำงานของการจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งแนวใหม่ (A New Suffix Tree Clustering Algorithm) จะต้องผ่านขั้นตอนการทำงานดังต่อไปนี้

1. ขั้นตอนการทำ Pre-processing ซึ่งเป็นการกำจัดสิ่งที่ไม่มีประโยชน์ต่อการจัดกลุ่มผลการสืบค้น หรือกำจัดสิ่งปนเปื้อนต่างๆ ที่ทำให้การทำงานต้องสิ้นเปลืองเวลาในการประมวลผลกับบางสิ่งที่สามารถลบออกไปได้ และหน่วยความจำในการทำงาน เช่น การกำจัดตัวเลข สัญลักษณ์ต่างๆ , การทำ Stop-words และการทำ Stemming Words ดังแสดงผลการทำงานของขั้นตอนการทำ Pre-processing ต่อไปนี้

0 [hillari, rodham, clinton, senat, new, york, offici, includ, inform, new, york, biographi, constitu, servic, contact, detail, committe, legisl, issu, new, speech, latest, event]

1 [biographi, william, j, clinton, short, biographi, offici, white, hous]

2 [welcom, clinton, counti, ny, link, govern, agenc, schedul, event, weather, forecast, photo, galleri, area]

3 [bill, clinton, wikipedia, free, encyclopedia, encyclopedia, articl, clinton, life, presidenti, administr, includ, hypertext, relat, articl]

4 [william, j, clinton, foundat, charit, foundat, organ, presid, clinton, focuss, global, issu, health, secur, econom, empower, leadership, citizenship, racial, ethnic, religi, reconcili]

5 [villag, clinton, british, columbia, canada, villag, clinton, p, o, box, cariboo, hwy, clinton, b, c, canada, v-k, k, e-mail, admin-villag, clinton, bc, ca, phone, fax]

6 [clinton, counti, welcom, you, clinton, counti, ohio, a, grow, commun, answer, challeng, merg, rural, urban, ideal, includ, list, counti, agenc]

7 [clinton, commun, colleg, clinton, commun, colleg, board, of, truste, meet, read, click, campu, announc, prospect, student, about, area, admiss, continu, educ, colleg, catalog]

8 [villag, clinton, digit, towpath, villag, clinton, year, old, pictur, gazebo, villag, park, new, nd, annual, sotrmwat, manag, report, mai, annual, stormwat, manag, report]

9 [clinton, herald, clinton, iowa-homepag, local, sport, new, classifi, busi, directori, obituari, weather, commun, inform, advertis, detail, includ]

10 [william, j, clinton, foundat, presid, clinton, new, minut, transcript, bill, clinton, effort, combat, aid, worldwid, announc, partnership, viet, nam, fight, against, hiv-aid, announc]

11 [welcom, clinton, massachusett, includ, local, headlin, town, calendar, fact, forum, histori, imag]

12 [home, page, includ, download, accid, report, citat, payment, bike, law, local, advisori]

13 [clinton, counti, illinoi-home, page, govern, offer, inform, counti, board, offic, meet, calendar, local, municip, township, histor, societi, museum, war, memori, counti, fair, h, activ]

14 [clinton, counti, govern, includ, inform, depart, servic]

15 [clinton, presidenti, materi, project, clinton, presidenti, materi, project, introduct, clinton, presidenti, materi, project, william, j, clinton, presidenti, librari, museum, visit, http]

16 [link, articl, iraq, repeatedli, block, unscom, inspect, suspect, iraq, repeatedli, restrict, unscom, abil, obtain, evid, iraq, stop, unscom, biolog, weapon, team]

17 [offici, site, clinton, mississippi, offici, citi, featur, inform, histori, govern, landmark, school, attract]

18 [clinton, counti, iowa, home, page, offici, web, site, clinton, counti, govern, a, portal, inform, elect, offici, depart, servic, provid, clinton, counti]

2. ขั้นตอนการกำหนดกลุ่มพื้นฐาน ดังแสดงผลการทำงานในตารางที่ 3.10

ตารางที่ 3.10 แสดงผลการกำหนดกลุ่มพื้นฐาน

No.	กลุ่มพื้นฐาน	ตำแหน่งคำและเอกสารภายในกลุ่ม
0	clinton	(2, 0) , (3, 1) , (19, 2) , (1, 3) , (0, 4) , (4, 5) , (0, 6) , (0, 7) , (8, 8) , (0, 9) , (1, 10) , (7, 11) , (2, 13) , (2, 14) , (8, 15) , (1, 17) , (6, 18)
1	clinton counti	(19, 2) , (1, 6) , (0, 13) , (4, 14) , (0, 18)
2	clinton counti govern	(0, 14) , (8, 18)
3	clinton foundat	(2, 4) , (2, 10)
4	new	(4, 0) , (9, 8) , (19, 9) , (12, 10)
5	offici	(6, 0) , (6, 1) , (0, 17) , (4, 18)
6	includ	(16, 0) , (7, 3) , (3, 6) , (11, 9) , (16, 11) , (3, 12) , (2, 14)
7	includ inform	(7, 0) , (3, 14)
8	inform	(8, 0) , (13, 9) , (6, 13) , (4, 14) , (7, 17) , (13, 18)
9	biographi	(11, 0) , (0, 1)
10	servic	(6, 0) , (13, 14) , (17, 18)
11	detail	(15, 0) , (15, 9)
12	issu	(18, 0) , (11, 4)
13	event	(22, 0) , (8, 2)
14	william j clinton	(1, 1) , (0, 4) , (0, 10) , (13, 15)
15	j clinton	(2, 1) , (1, 4) , (1, 10) , (14, 15)
16	j clinton foundat	(1, 4) , (1, 10)
17	welcom	(0, 2) , (0, 6) , (2, 11)
18	welcom clinton	(0, 2) , (0, 11)
19	counti	(20, 2) , (2, 6) , (1, 13) , (5, 14) , (18, 18)
20	counti govern	(1, 14) , (9, 18)
21	link	(4, 2) , (0, 16)
22	govern	(5, 2) , (4, 13) , (2, 14) , (9, 17) , (10, 18)
23	agenc	(19, 2) , (6, 6)
24	weather	(9, 2) , (11, 9)

ตารางที่ 3.10 (ต่อ)

No.	กลุ่มพื้นฐาน	ตำแหน่งคำและเอกสารภายในกลุ่ม
25	area	(13, 2) , (17, 7)
26	bill clinton	(0, 3) , (9, 10)
27	articl	(14, 3) , (6, 16)
28	presidenti	(9, 3) , (1, 15)
29	foundat	(3, 4) , (5, 10)
30	presid clinton	(7, 4) , (4, 10)
31	villag	(0, 5) , (5, 8)
32	villag clinton	(0, 5) , (5, 8)
33	commun	(9, 6) , (1, 7) , (4, 9)
34	board	(6, 7) , (8, 13)
35	meet	(9, 7) , (10, 13)
36	announc	(22, 7) , (13, 10)
37	report	(22, 8) , (17, 12)
38	local	(4, 9) , (4, 11) , (10, 12) , (12, 13)
39	calendar	(7, 11) , (11, 13)
40	histori	(10, 11) , (8, 17)
41	home page	(0, 12) , (3, 18)
42	page	(1, 12) , (3, 13) , (4, 18)
43	museum	(17, 13) , (18, 15)
44	depart servic	(5, 14) , (16, 18)
45	site clinton	(1, 17) , (7, 18)

3. ขั้นตอนการเชื่อมป้ายชื่อกลุ่ม เพื่อค้นหากลุ่มใหม่ที่ป้ายชื่อมีความสมบูรณ์ ดังแสดงผลการทำงานในตารางที่ 3.11

ตารางที่ 3.11 แสดงผลการเชื่อมป้ายชื่อกลุ่ม

No.	ป้ายชื่อกลุ่ม	เอกสารภายในกลุ่ม
0	clinton	[7, 9]
1	clinton counti	[2, 6, 13]
2	clinton counti govern	[14, 18]
3	includ inform	[0, 14]
4	william j clinton	[1, 15]
5	welcom clinton	[2, 11]
6	link	[16]
7	bill clinton	[3, 10]
8	presid clinton	[4, 10]
9	villag clinton	[5, 8]
10	home page	[12, 18]
11	depart servic	[14, 18]
12	site clinton	[17, 18]
13	william j clinton foundat	[4, 10]

4. ขั้นตอนการจัดลำดับความสำคัญของกลุ่ม ดังแสดงผลการทำงานในตารางที่ 3.12 โดยมีผลการจัดลำดับความสำคัญดังแสดงในตารางที่ 3.12

ตารางที่ 3.12 แสดงผลการจัดลำดับความสำคัญของกลุ่ม

No.	ป้ายชื่อกลุ่ม	เอกสารภายในกลุ่ม
0	clinton counti	[2, 6, 13]
1	william j clinton foundat	[4, 10]
2	villag clinton	[5, 8]
3	clinton counti govern	[14, 18]
4	william j clinton	[1, 15]
5	includ inform	[0, 14]
6	welcom clinton	[2, 11]
7	bill clinton	[3, 10]
8	presid clinton	[4, 10]
9	home page	[12, 18]
10	depart servic	[14, 18]
11	site clinton	[17, 18]
12	clinton	[7, 9]
13	link	[16]

ในบทต่อไปจะกล่าวถึง การทดลองและผลการทดลองกับชุดข้อมูลที่นำมาทดลองกับกระบวนการทำงาน (Algorithm) ของงานวิจัยนี้ และงานวิจัยที่เกี่ยวข้อง

บทที่ 4

การทดลองและผลการทดลอง

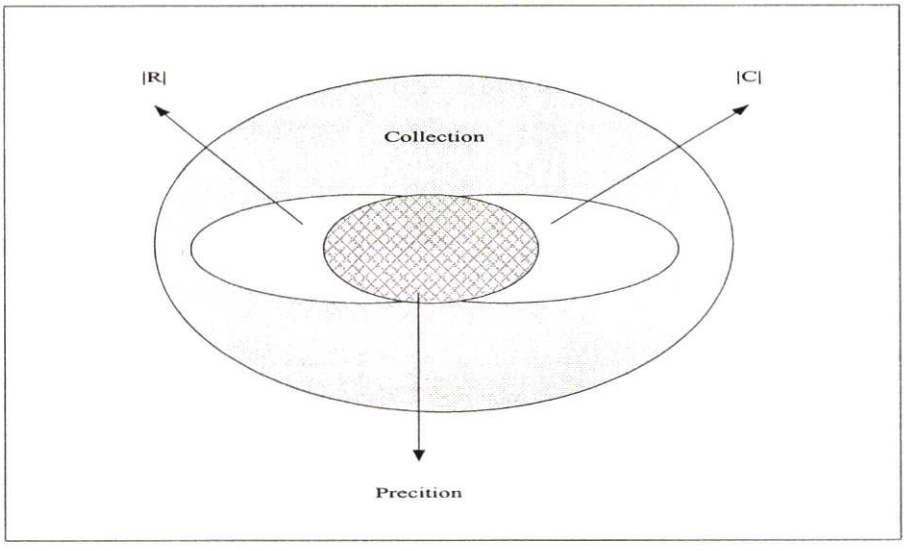
ในบทนี้จะกล่าวถึงชุดข้อมูลที่ใช้ในการทดลอง การทดลอง และผลการทดลอง ซึ่งเป็นการเปรียบเทียบประสิทธิภาพการจัดกลุ่มผลการสืบค้นระหว่างเทคนิค Suffix Tree Clustering (STC_Original), Suffix Tree Clustering ร่วมกับเทคนิค n-gram (STC with n-gram) และ A New Suffix Tree Clustering (New STC) โดยกระบวนการทำงานของ A New Suffix Tree Clustering (New STC) เป็นผลงานวิจัยฉบับนี้

4.1 การวัดประสิทธิภาพการทำงานของโมเดล

การวัดประสิทธิภาพการทำงานของโมเดล ได้ใช้ตัววัดประสิทธิภาพที่ใช้ในงานวิจัยทั่วไป โดยมีรายละเอียดดังต่อไปนี้

4.1.1 ตัววัดประสิทธิภาพความถูกต้อง

ตัววัดประสิทธิภาพความถูกต้อง (Precision) หรืออาจเรียกว่าเอกสารที่ตรงกับความต้องการ (Relevance) คือการประเมินความถูกต้องในการจัดกลุ่มผลการสืบค้น รายงานวิจัยฉบับนี้ซึ่งใช้มาตรฐานความถูกต้องจากการเปรียบเทียบผลการจัดกลุ่มกับระบบสืบค้นลักษณะ Directories Search Engine ที่มีการจัดกลุ่มผลการสืบค้นโดยมนุษย์คือ Dmoz.com ดังแสดงรายละเอียดส่วนที่ใช้วัดความถูกต้องในรูปที่ 4.1 และสมการที่ 4.1



รูปที่ 4.1 แสดงส่วนที่ใช้วัดค่าความถูกต้องในการจัดกลุ่มผลการสืบค้น

$$P@N = \frac{|C \cap R|}{|R|} \tag{4.1}$$

โดยที่ C คือกลุ่มของเอกสารที่จัดโดยมนุษย์

R คือกลุ่มของเอกสารที่จัดโดยระบบของ A New STC Algorithm

N คือจำนวนกลุ่มเอกสาร

4.1.2 **ตัววัดอัตราความทับซ้อนของเอกสาร**

ตัววัดอัตราการทับซ้อนของเอกสาร (Overlap) คือการวัดจำนวนเอกสารที่ปรากฏมากกว่าหนึ่งกลุ่มเอกสาร ดังแสดงรายละเอียดในสมการที่ 4.2

$$Overlap = (c / s) - 1 \tag{4.2}$$

โดยที่ c คือ ผลรวมของจำนวน snippets ในแต่ละกลุ่ม

s คือ จำนวนของ snippets ทั้งหมด

4.1.3 ตัววัดประสิทธิภาพด้านความคลอบคลุมการจัดกลุ่ม

ตัววัดประสิทธิภาพด้านความคลอบคลุมการจัดกลุ่ม (Coverage) คือการคำนวณจำนวนผลการสืบค้นที่สามารถจัดกลุ่มได้จากผลการสืบค้นทั้งหมด ดังแสดงรายละเอียดในสมการที่ 4.3

$$Coverage = a / s \tag{4.3}$$

โดยที่ a คือ จำนวนของ snippets ที่ถูกจัดกลุ่ม

s คือ จำนวนของ snippets ทั้งหมด

4.1.4 ตัววัดอัตราเอกสารไม่ตรงกับความต้องการภายในกลุ่ม

ตัววัดอัตราเอกสารไม่ตรงกับความต้องการภายในกลุ่ม (Irrelevant Documents) คือจำนวนเอกสารปนเปื้อนภายในกลุ่ม ที่เหลือจากเอกสารที่ตรงกับความต้องการ (Relevant Documents) ดังแสดงรายละเอียดในสมการที่ 4.4

$$\mathbf{Id = 1 - precision} \tag{4.4}$$

โดยที่ precision คือ อัตราความถูกต้องในการจัดกลุ่มของกลุ่ม

4.1.5 ตัววัดระยะห่างระหว่างเอกสารที่ตรงกับความต้องการและไม่ตรงกับความต้องการภายในกลุ่มเอกสาร

Above คือช่วงระยะห่างระหว่างค่าของอัตราจำนวนเอกสารที่ตรงกับความต้องการและไม่ตรงกับความต้องการภายในกลุ่มเอกสาร ถ้าค่ายิ่งสูงแสดงว่าจำนวนของเอกสารที่ตรงกับความต้องการอยู่ห่างจากจำนวนเอกสารที่ไม่ตรงกับความต้องการมีมากขึ้น ดังแสดงรายละเอียดในสมการที่ 4.5

$$\textbf{Above} = \textbf{precision} / \textbf{Id} \tag{4.5}$$

โดยที่ precision คือ อัตราของเอกสารที่จัดกลุ่มได้ถูกต้อง

Id คือ อัตราของเอกสารที่ไม่ตรงกับความต้องการ

4.1.6 ตัววัดขนาดกลุ่ม

ตัววัดขนาดกลุ่ม (Size of Cluster) คือการคำนวณค่าเฉลี่ยของขนาดกลุ่มผลการสืบค้น เนื่องจากวัตถุประสงค์หลักของการจัดกลุ่มผลการสืบค้น คือการอำนวยความสะดวกให้กับผู้สืบค้นในด้านการเข้าถึงผลการสืบค้นที่ต้องการได้สะดวกและรวดเร็ว ดังนั้นเพื่อให้สอดคล้องกับวัตถุประสงค์หลักในการจัดกลุ่มผลการสืบค้น กลุ่มผลการสืบค้นต้องมีขนาดกะทัดรัดเพื่อให้ผู้สืบค้นสามารถเข้าถึงผลการสืบค้นได้ง่ายขึ้น การคำนวณค่าเฉลี่ยของขนาดกลุ่มแสดงรายละเอียดในสมการที่ 4.6

$$size\_cluster = \frac{\sum_{i=0}^{n} c_i}{n+1} \tag{4.6}$$

โดยที่ C_i คือ จำนวนเอกสารภายในกลุ่ม C

n คือ จำนวนกลุ่ม

การวัดประสิทธิภาพกระบวนการทำงานของ A New Suffix Tree Clustering (STC) นอกจากตัวชี้วัดประสิทธิภาพที่ได้อธิบายในข้างต้นแล้ว เรายังใช้จำนวนโหนด (Node) หรือจำนวนคำที่สร้างอยู่ภายในซัฟฟิกทรี (Suffix Tree) ในการชี้วัดเรื่องจำนวนหน่วยความจำที่ใช้ในการทำงาน และระยะเวลา (Time) ที่ใช้ในการทำงาน เพราะการทำงานในลักษณะออนไลน์ (Online) ต้องสามารถให้ผลการทำงานในเวลาที่รวดเร็ว

4.2 ชุดข้อมูลที่ใช้ในการทดลอง

รายงานฉบับนี้ใช้ข้อมูลในการทดสอบโมเดลจำนวน 3 ชุดข้อมูล คือ ชุดข้อมูลของผลการสืบค้นจำนวน 2 ชุดข้อมูล และ ชุดข้อมูลที่มีลักษณะเป็นข้อความ 1 ชุดข้อมูล ในการทดลองหาประสิทธิภาพของโมเดล New STC โดยการเปรียบเทียบกับโมเดลของ STC_Original และ STC with n-gram ประกอบด้วยรายละเอียดดังต่อไปนี้

4.2.1 ชุดข้อมูลผลการสืบค้นภายในกลุ่มของ Dmoz.com

เป็นชุดข้อมูลผลการสืบค้นที่สร้างจากคำอธิบายสั้นๆ (Snippets) ซึ่งใช้เป็นตัวแทนเอกสารหรือผลการสืบค้นภายในกลุ่มของ Directories Search Engine ที่จัดกลุ่มผลการสืบค้นโดยมนุษย์ชื่อ Dmoz.com และ ทำการทดลองผลการจัดกลุ่มในลักษณะเดียวกับงานวิจัยเรื่อง A Concept-Driven Algorithm for Clustering Search Results [19] โดยทำการแบ่งลักษณะของชุดข้อมูลออกเป็น 3 ลักษณะตามเนื้อหาเอกสาร ประกอบด้วย

1. ลักษณะของกลุ่มเอกสารที่มีเนื้อหาเป็นเรื่องเดียวกัน เพื่อใช้ในการทดสอบความถูกต้องในการจัดกลุ่มเอกสารในลักษณะที่เอกสารของแต่ละกลุ่มมีเนื้อหาเกี่ยวข้องกันหรือเนื้อหาเป็นเรื่องเดียวกัน เช่น นำผลการสืบค้นเรื่อง SQL Database จาก 3 กลุ่มเอกสาร มาทดสอบด้วยโมเดลเพื่อทดสอบว่าเมื่อเอกสารเป็นเอกสารเรื่องเดียวกันโมเดลจะสามารถจัดกลุ่มได้ถูกต้องอย่างไรบ้าง

2. ลักษณะของกลุ่มเอกสารที่มีเนื้อหาแตกต่างกัน เพื่อใช้ในการทดสอบความถูกต้องในการจัดกลุ่มเอกสารในลักษณะที่เอกสารของแต่ละกลุ่มมีเนื้อหาแตกต่างกัน เช่น นำผลการสืบค้นที่เกี่ยวข้องกับเรื่องคอมพิวเตอร์ , สุขภาพ และ ภาพยนตร์ ซึ่งทั้ง 3 กลุ่มต้องไม่มีการซ้ำซ้อนของข้อมูลเช่นกลุ่มเอกสารที่มีเนื้อหาเกี่ยวกับคอมพิวเตอร์ประกอบด้วย 5 กลุ่มย่อย คือ กลุ่ม SQL Database , Data Warehouse , XML , Java และ Text และ Text Editor การทดลองจะเลือกมาเฉพาะกลุ่มใดกลุ่มหนึ่งเท่านั้น เพื่อนำเอกสารของทั้ง 3 เนื้อหามารวมกันแล้วทดสอบด้วยโมเดลเพื่อทดสอบว่า เมื่อเอกสารเป็นเอกสารที่กล่าวถึงเรื่องราวที่แตกต่างกัน โมเดลจะสามารถจัดกลุ่มได้ถูกต้องอยู่ในระดับใด

3. ลักษณะของกลุ่มเอกสารที่มีเนื้อหาผสม เพื่อใช้ในการทดสอบความถูกต้องในการจัดกลุ่มเอกสารในลักษณะที่เอกสารของแต่ละกลุ่มมีเนื้อหาผสม เช่น นำนำผลการสืบค้นที่เกี่ยวข้องกับเรื่องคอมพิวเตอร์ , สุขภาพ และ ภาพยนตร์ ซึ่งทั้ง 3 กลุ่มโดยไม่คำนึงถึงการซ้ำซ้อนของข้อมูลเช่นกลุ่มเอกสารที่มีเนื้อหาเกี่ยวกับคอมพิวเตอร์ประกอบด้วย 5 กลุ่มย่อย คือ กลุ่ม SQL Database , Data Warehouse , XML , Java และ Text และ Text Editor การทดลองอาจจะเลือกมาทุกกลุ่มหรือบางกลุ่ม เพื่อนำเอกสารของทั้ง 3 เนื้อหามารวมกันแล้วทดสอบด้วยโมเดลเพื่อทดสอบว่า เมื่อเอกสารเป็นเอกสารที่กล่าวถึงเรื่องราวที่ผสมผสานกันหลายๆเนื้อหาหรือบางกลุ่มมีเนื้อหาเดียวกัน โมเดลจะสามารถจัดกลุ่มได้ถูกต้องอยู่ในระดับใด

4.2.2 ชุดข้อมูลผลการสืบค้นด้วยคำสืบค้นภายใน Dmoz.com

เป็นชุดข้อมูลผลการสืบค้น (Web Search Results) ที่สร้างจากการสืบค้นด้วยคำสืบค้นภายใน Dmoz.com (Query Word) ซึ่งเป็น Directories Search Engine ที่มีการนำผลการสืบค้นมาจัดกลุ่มโดยมนุษย์ และใช้คำอธิบายสั้นๆ (Snippets) เป็นตัวแทนเอกสารเพื่อใช้ในการทดลองที่มีลักษณะการทดลองเช่นเดียวกับหลายๆระบบที่เป็นงานวิจัยด้านการจัดกลุ่มผลการสืบค้น เช่น รายงานวิจัยเรื่อง Learning to Clustering Web Search Results [5] และ A Method of web search result clustering based on rough sets [20] โดยแบ่งลักษณะของคำสืบค้น (Query Words) เป็น 3 ลักษณะ ประกอบด้วย

1. คำสืบค้นที่มีลักษณะเป็นคำทั่วๆไป (General Concepts) ซึ่งคำเหล่านี้จะมีความหมายไปในทิศทางเดียว ประกอบด้วย " Search " , " Computer " , " Music" เพื่อทดสอบว่าโมเดลสามารถจัดกลุ่มผลการสืบค้นได้ถูกต้องอยู่ในระดับใด เมื่อคำสืบค้นเป็นคำทั่วๆไป

2. คำสืบค้นที่มีลักษณะเป็นชื่อเฉพาะ (Entity Name) ซึ่งคำจะมีลักษณะเป็นชื่อเฉพาะประกอบด้วย " Thailand " , " Iraq " , " Disney " เพื่อทดสอบว่าโมเดลสามารถจัดกลุ่มผลการสืบค้นได้ถูกต้องอยู่ในระดับใด เมื่อคำสืบค้นเป็นคำชื่อเฉพาะ

3. คำสืบค้นที่มีลักษณะเป็นคำกำกวม (Ambiguous Queries) หรือคำที่มีความหมายหลากหลาย ประกอบด้วย " Matrix " ซึ่งหมายถึง เมตริกซ์ทางคณิตศาสตร์ หรือบางครั้งอาจหมายถึงแม่พิมพ์ แหล่งกำเนิดหรือ เนื้อเยื่อเสริมสร้างและผลิตเซลล์ที่โคนเล็บ , " Apple " คำนี้อาจหมายถึง ผลไม้ชื่อแอปเปิ้ล หรืออาจหมายถึงชื่อของผลิตภัณฑ์ทางคอมพิวเตอร์ และ " Jaguar " ซึ่งหมายถึง สัตว์ตระกูลเสือชื่อจากัวร์ หรืออาจหมายถึงชื่อของผลิตภัณฑ์ทางรถยนต์ เพื่อทดสอบว่าโมเดลสามารถจัดกลุ่มผลการสืบค้นได้ถูกต้องอยู่ในระดับใด เมื่อคำสืบค้นเป็นคำกำกวม

4.2.3 ชุดข้อมูลจาก Ohsumed Collection

เป็นชุดข้อมูลที่มีลักษณะเป็นข้อความจาก TEXT CATEGORIZATION CORPORA [23] ซึ่งนิยมนำมาใช้ในการทดลองในงานการจัดกลุ่มข้อความ (Text Categorization) ประกอบด้วยบทคัดย่อ (Abstracts) ทางการแพทย์จาก MeSH Categories ในปี 1991 ชื่อ Ohsumed Collection [21] และการทดลองในรายงานฉบับนี้จะทำการสุ่มเลือกเอกสารภายในกลุ่ม โดยแต่ละกลุ่มจะต้องไม่มีเอกสารซ้ำซ้อนกัน ผลการสุ่มเลือกเอกสารจากกลุ่มของ TEXT CATEGORIZATION CORPORA เราใช้กลุ่มจำนวน 22 กลุ่มเอกสาร รวม 4,380 เอกสาร เพื่อทดสอบว่าโมเดลสามารถจัดกลุ่มผลการสืบค้นได้ถูกต้องอยู่ในระดับใด เมื่อเอกสารไม่ใช่ผลการสืบค้น หรือคำอธิบายสั้นๆ (Snippets) แต่เป็นเอกสารที่มีลักษณะเป็นข้อความจำนวนมากกว่า 50 คำ

4.3 ตัวอย่างชุดข้อมูลที่ใช้ในการทดลองและผลการทดลอง

4.3.1 ชุดข้อมูลเอกสารภายในกลุ่มเอกสารของ Dmoz.com

ประกอบด้วยผลการสืบค้นจำนวน 9 กลุ่มผลการสืบค้น รวมผลการสืบค้นจำนวน 479 ผลการสืบค้น ที่มีความเกี่ยวข้องกับ 4 เนื้อหา ประกอบด้วย คอมพิวเตอร์ (Computer) จำนวน 5 กลุ่ม , สุขภาพ (Health Care), รูปภาพ (Photography) และ ภาพยนต์ (Movies) จำนวน 2 กลุ่ม ดังแสดงรายละเอียดในตารางที่ 4.1

ตารางที่ 4.1 แสดงกลุ่มผลการสืบค้นของข้อมูลจาก Dmoz.com

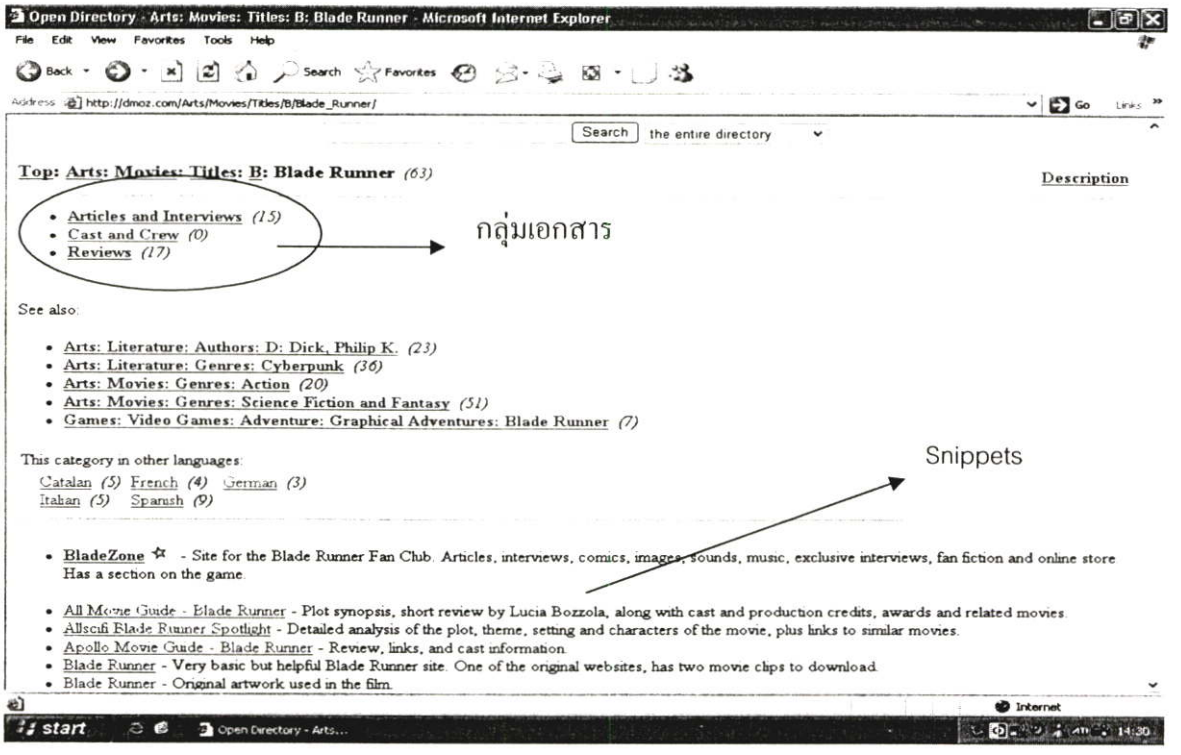
ชื่อกลุ่ม	จำนวนเอกสาร	คำอธิบายกลุ่ม
SQL Database	103	กลุ่มเอกสารระบบฐานข้อมูล ที่เกี่ยวกับ SQL
Data Warehouse	33	กลุ่มเอกสารที่กล่าวถึง Data Warehouse
XML	14	กลุ่มเอกสารที่กล่าวถึง XML
Java	48	กลุ่มเอกสารที่กล่าวถึงโปรแกรมจาวา
Text Editor	41	กลุ่มเอกสารที่กล่าวถึง Text Editor
Health Care	79	กลุ่มเอกสารที่กล่าวถึงเรื่องสุขภาพร่างกายมนุษย์
Photography	20	กลุ่มเอกสารที่กล่าวถึงรูปภาพ
Lord Of The Ring	80	กลุ่มเอกสารที่กล่าวถึงภาพยนตร์เรื่อง Lord Of The Ring
Blade Runner	61	กลุ่มเอกสารที่กล่าวถึงถาพยนต์เรื่อง Brad Runner
รวม	479	

การทดลองและผลการทดลอง แบ่งข้อมูลออกเป็น 3 ลักษณะ คือ

ขั้นตอนการทดลอง แบ่งออกเป็น 4 ขั้นตอนคือ

1. นำคำอธิบายสั้นๆ (Snippets) จากแต่ละกลุ่มของ Dmoz.com ที่ถูกจัดกลุ่มโดยมนุษย์จากกลุ่มข้อมูลที่ต้องการ โดยไม่ใช้คำสืบค้น มาเป็นตัวแทนเอกสารของแต่ละกลุ่ม โดยการแบ่งแยกตามกลุ่มทั้งกลุ่มย่อยและกลุ่มใหญ่ เช่น กลุ่มข้อมูลเกี่ยวกับภาพยนตร์เรื่อง Blade Runner ประกอบด้วย 3 กลุ่มย่อย คือ กลุ่ม Blade Runner มีเอกสาร 31 เอกสาร , Articles and Interviews มีเอกสารจำนวน 15 เอกสาร และกลุ่ม Reviews มีเอกสารจำนวน 15 เอกสาร ดังแสดงตัวอย่างกลุ่มและคำอธิบายสั้นๆ (Snippets) ซึ่งใช้เป็นตัวแทนเอกสารภายในกลุ่มในรูปที่ 4.2

เพื่อนำมาใช้ในการทดลองตามชุดข้อมูลที่ต้องการ ซึ่งประกอบด้วย ชุดข้อมูลที่มีเนื้อหาเป็นเรื่องเดียวกัน , ชุดข้อมูลที่มีเนื้อหาแตกต่างกัน และชุดข้อมูลที่มีเนื้อหาผสม



รูปที่ 4.2 ตัวอย่างกลุ่ม และSnippets ภายในกลุ่มที่มีเนื้อหาเป็นเรื่องเดียวกัน

2. นำคำอธิบายสั้นๆ (Snippets) ซึ่งใช้เป็นตัวแทนเอกสาร มาผ่านกระบวนการ pre-processing คือทำ Stop-words โดยการกำจัดคำที่ไม่มีความหมาย เพื่อให้จำนวนคำลดลง และการทำ Stemming words โดยการทำให้คำอยู่ในรากศัพท์เดิม เพื่อลดความหลากหลายของคำที่จะใช้ในการสร้างซัฟฟิกทรี (Suffix Tree)

3. นำคำที่เหลือจากการทำงานในขั้นตอนการทำ pre-processing เข้าสู่กระบวนการทำงานของโมเดล และคำนวณหาค่าต่างๆของตัวชี้วัดประสิทธิภาพของโมเดล

4. ศึกษาค่าต่างๆของตัวชี้วัด โดยการแบ่งการศึกษาค่ามาตรฐานต่างๆออกเป็นช่วงๆของผลการจัดกลุ่ม ประกอบด้วย ค่ามาตรฐานในช่วง 5 กลุ่มแรกของผลการจัดกลุ่ม , ช่วง 10 กลุ่มแรกของผลการจัดกลุ่ม , ช่วง 15 กลุ่มแรกของผลการจัดกลุ่ม , ช่วง 20 กลุ่มแรกของผลการจัดกลุ่ม และทุกกลุ่มซึ่งจะเป็นการแสดงภาพรวมทั้งหมดของการจัดกลุ่มในแต่ละชุดข้อมูล ดังแสดงในตารางผลการทดลอง

การทดลองด้วยชุดข้อมูลที่มีลักษณะเนื้อหาเป็นเรื่องเดียวกัน เราจะทำการเลือกเอกสารจากกลุ่มเอกสารที่มีเนื้อหาเป็นเรื่องเดียวกันจากกลุ่มย่อยต่างๆ ดังแสดงตัวอย่างกลุ่ม พร้อมกับคำอธิบายสั้นๆ (Snippets) ของกลุ่มในรูปที่ 4.2 ซึ่งประกอบด้วยรายละเอียดดังต่อไปนี้

1 ชุดข้อมูลที่มีลักษณะเนื้อหาเป็นเรื่องเดียวกัน ประกอบด้วย

1.1 ชุด Lord Of The Ring ประกอบด้วยเนื้อหาของภาพยนตร์เรื่อง Lord Of The Ring ประกอบด้วยกลุ่ม Image_Galleries จำนวน 9 เอกสาร , กลุ่ม Fellowship_of_the_Ring,_The จำนวน 32 เอกสาร , กลุ่ม Return_of_the_King,_The_-_2003 จำนวน 9 เอกสาร และ กลุ่ม Two_Towers,_The จำนวน 30 เอกสาร รวม 80 เอกสาร ซึ่งแต่ละกลุ่มจะเป็นกลุ่มย่อยของกลุ่ม Lord_of_the_Rings_Series (http://dmoz.com/Arts/Movies/Titles/L/Lord_of_the_Rings_Series) ดังแสดงผลการทดลองในตารางที่ 4.2

1.2 ชุด Brad Runner ประกอบด้วยเนื้อหาของภาพยนตร์เรื่อง Brad Runner จำนวน 3 กลุ่ม ประกอบด้วย กลุ่ม Blade_Runner จำนวน 29 เอกสาร , กลุ่ม Articles_and_Interviews จำนวน เอกสาร 15 และกลุ่ม Reviews จำนวน 17 เอกสาร รวม 61 เอกสาร ซึ่งแต่ละกลุ่มจะเป็นกลุ่มย่อยของกลุ่ม Blade_Runner (http://dmoz.com/Arts/Movies/Titles/B/Blade_Runner) ดังแสดงผลการทดลองในตารางที่ 4.3

1.3 ชุด SQL Database ประกอบด้วยเนื้อหาของระบบฐานข้อมูลที่เกี่ยวข้องกับ SQL จำนวน 3 กลุ่ม ประกอบด้วย กลุ่ม PostgreSQL จำนวน 45 เอกสาร , กลุ่ม MySQL จำนวน 38 เอกสาร และกลุ่ม Microsoft_SQL_Server จำนวน 20 เอกสาร รวม 103 เอกสาร ซึ่งแต่ละกลุ่มจะเป็นกลุ่มย่อยของกลุ่ม Databases (http://dmoz.com/Computers/Software/Databases) ดังแสดงผลการทดลองในตารางที่ 4.4 และแสดงผลค่าเฉลี่ยในตารางที่ 4.5 และ 4.6 ตามลำดับ

ตารางที่ 4.2 แสดงผลการทดลองข้อมูลชุด LORD OF THE RING

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	115	5	10	15	20	115	5	10	15	20	56
จำนวนกลุ่ม BASE CLUSTER	148					146					146				
จำนวนกลุ่ม MERGED CLUSTER	115					115					56				
จำนวน NODE	5,184					1,691					1,691				
เวลาที่ใช้ (TIME)	1 SEC.					1 SEC.					0 SEC.				
ขนาดกลุ่ม	19.6	12.3	9.07	7.45	4.93	19.6	12.3	9.07	7.45	4.93	7.8	5.6	4.6	4.1	2.66
OVERLAP	0.81	1.08	1.27	1.29	6.09	0.81	1.08	1.27	1.29	6.09	0.05	0.19	0.28	0.41	0.86
COVERAGE	0.68	0.74	0.75	0.81	1	0.68	0.74	0.75	0.81	1	0.46	0.59	0.68	0.73	1
PRECISION	0.84	0.78	0.78	0.79	0.72	0.84	0.78	0.78	0.79	0.72	0.68	0.7	0.73	0.74	0.83
IRRELEVANT	0.16	0.22	0.22	0.21	0.28	0.16	0.22	0.22	0.21	0.28	0.32	0.3	0.27	0.26	0.17
ABOVE	5.25	3.55	3.55	3.76	2.57	5.25	3.55	3.55	3.76	2.57	2.13	2.33	2.70	2.85	4.88

ตารางที่ 4.3 แสดงผลการทดลองข้อมูลชุด BLADE RUNNER

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	112	5	10	15	20	112	5	10	15	20	35
จำนวนกลุ่ม BASE CLUSTER	132					133					133				
จำนวนกลุ่ม MERGED CLUSTER	112					112					35				
จำนวน NODE	5,582					1,612					1,612				
เวลาที่ใช้ (TIME)	1 SEC.					1 SEC.					0 SEC.				
ขนาดกลุ่ม	10.4	6.9	5.3	4.5	3.79	10.8	6.9	5.3	4.5	3.79	8.4	5.3	4.2	3.7	2.8
OVERLAP	0.21	0.5	0.67	0.88	5.97	0.23	0.5	0.67	0.89	5.97	0.08	0.18	0.31	0.46	0.61
COVERAGE	0.7	0.75	0.79	0.79	1	0.72	0.75	0.79	0.79	1	0.64	0.74	0.79	0.82	1
PRECISION	0.79	0.77	0.78	0.79	0.79	0.79	0.77	0.78	0.79	0.79	0.65	0.74	0.76	0.8	0.85
IRRELEVANT	0.21	0.23	0.22	0.21	0.21	0.21	0.23	0.22	0.21	0.21	0.35	0.26	0.24	0.2	0.15
ABOVE	3.76	3.35	3.55	3.76	3.76	3.76	3.35	3.55	3.76	3.76	1.86	2.85	3.17	4.00	5.67

ตารางที่ 4.4 แสดงผลการทดลองข้อมูลชุด SQL Database

ค่ามาตรฐาน	STC Original					STC with N-Gram					New STC				
	5	10	15	20	191	5	10	15	20	119	5	10	15	20	76
จำนวนกลุ่ม Base Cluster	218					218					218				
จำนวนกลุ่ม Merged Cluster	191					119					76				
จำนวน Node	7,173					2,371					2,371				
เวลาที่ใช้ (Time)	1 sec.					1 sec.					1 sec.				
ขนาดกลุ่ม	9.8	6.2	5.53	4.65	5.51	9.8	6.2	5.53	4.65	5.51	5.4	4.2	3.5	3.1	2.3
Overlap	0.2	0.38	0.48	0.6	7.29	0.2	0.38	0.48	0.6	7.29	0	0.14	0.16	0.27	0.65
Coverage	0.39	0.43	0.54	0.56	1	0.39	0.43	0.54	0.56	1	0.26	0.36	0.43	0.47	1
Precision	0.84	0.79	0.83	0.87	0.75	0.84	0.79	0.83	0.87	0.75	0.97	0.92	0.88	0.91	0.88
Irrelevant	0.16	0.21	0.17	0.13	0.25	0.16	0.21	0.17	0.13	0.25	0.03	0.08	0.12	0.09	0.12
Above	5.25	3.76	4.88	6.69	3.00	5.25	3.76	4.88	6.69	3.00	32.33	11.50	7.33	10.11	7.33

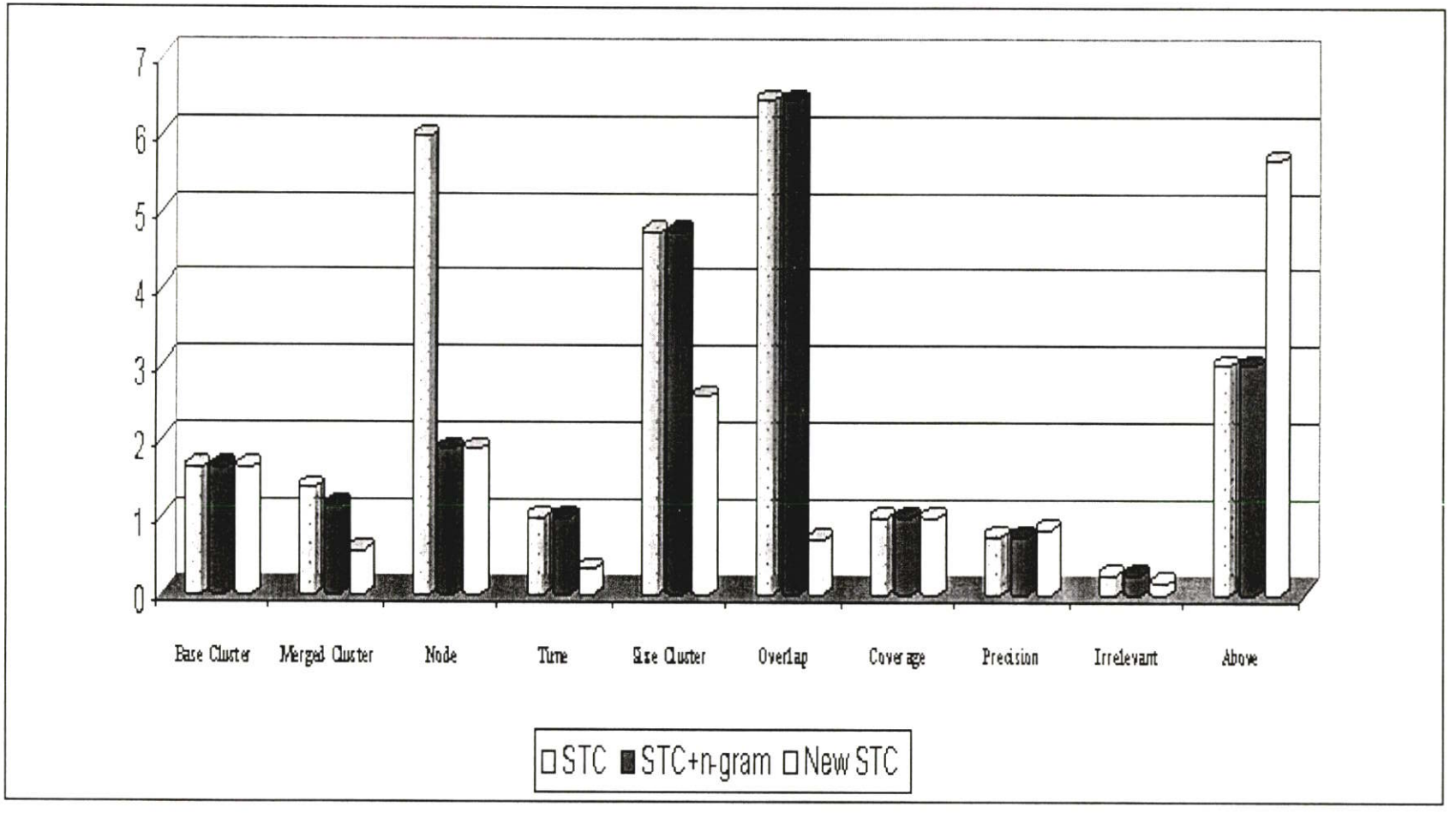
ตารางที่ 4.5 แสดงค่าเฉลี่ยผลการทดลองชุดข้อมูลเนื้อหาเดียวกัน

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	ALL	5	10	15	20	ALL	5	10	15	20	ALL
จำนวนกลุ่ม BASE CLUSTER	166					166					166				
จำนวนกลุ่ม MERGED CLUSTER	139					115					56				
จำนวน NODE	5,980					1,891					1,891				
เวลาที่ใช้ (TIME)	1 SEC.					1 SEC.					0.33 SEC.				
ขนาดกลุ่ม	13.27	8.47	6.63	5.53	4.74	13.4	8.47	6.63	5.53	4.74	7.20	5.03	4.10	3.63	2.59
OVERLAP	0.41	0.65	0.81	0.92	6.45	0.41	0.65	0.81	0.93	6.45	0.04	0.17	0.25	0.38	0.71
COVERAGE	0.59	0.64	0.69	0.72	1	0.6	0.64	0.69	0.72	1	0.45	0.56	0.63	0.67	1.00
PRECISION	0.82	0.78	0.8	0.82	0.75	0.72	0.78	0.8	0.82	0.75	0.77	0.79	0.79	0.82	0.85
IRRELEVANT	0.18	0.22	0.2	0.18	0.25	0.28	0.22	0.2	0.18	0.25	0.23	0.21	0.21	0.18	0.15
ABOVE	4.56	3.55	4.00	4.56	3.00	2.57	3.55	4.00	4.56	3.00	3.35	3.76	3.76	4.56	5.67

ตารางที่ 4.6 แสดงค่า NORMALIZE ค่าเฉลี่ยผล ของชุดข้อมูลที่มีเนื้อหาเป็นเรื่องเดียวกัน

ค่ามาตรฐาน	STC ORGINAL	STC WITH N-GRAM	NEW STC	หมายเหตุ
จำนวนกลุ่ม BASE CLUSTER	1.66	1.66	1.66	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกนำมาทำให้อยู่ในช่วง 1-10 ด้วยการหารด้วย 100
จำนวนกลุ่ม MERGED CLUSTER	1.39	1.15	0.56	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกนำมาทำให้อยู่ในช่วง 1-10 ด้วยการหารด้วย 100
จำนวน NODE	5.98	1.89	1.89	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกนำมาทำให้อยู่ในช่วง 1-10 ด้วยการหารด้วย 1,000
เวลาที่ใช้ (TIME)	1 SEC.	1 SEC.	0 SEC.	
ขนาดกลุ่ม (SIZE CLUSTER)	4.74	4.74	2.59	
OVERLAP	6.45	6.45	0.71	
COVERAGE	1	1	1.00	
PRECISION	0.75	0.75	0.85	
IRRELEVANT	0.25	0.25	0.15	
ABOVE	3.00	3.00	5.67	

จากตารางที่ 4.6 ค่า Normalize ค่าเฉลี่ยผลการทำงาน 3 โมเดล ของชุดข้อมูลที่มีเนื้อหาเป็นเรื่องเดียวกัน สามารถสร้างกราฟได้ดังแสดงในรูปภาพที่ 4.3



รูปที่ 4.3 กราฟแสดงผลการทำงานของ 3 โมเดล ในชุดข้อมูลที่มีเนื้อหาเป็นเรื่องเดียวกัน

จากรูปที่ 4.3 คือกราฟสรุปผลการทำงานของทั้ง 3 โมเดล คือ STC_Original , STC with n-gram และ New STC ของชุดข้อมูลที่มีเนื้อหาใกล้เคียงกัน เพื่อศึกษาความสามารถด้านต่างๆ ของโมเดล ด้วยตัวชี้วัดมาตรฐานต่างๆที่ใช้วัดประสิทธิภาพของโมเดลดังที่กล่าวในข้างต้น และศึกษาความสามารถในการจัดกลุ่มเอกสารหรืออำนาจจำแนกกลุ่มเอกสาร ที่เอกสารภายในกลุ่มมีเนื้อหาใกล้เคียงกัน ดังแสดงรายละเอียดของผลการทำงานดังต่อไปนี้

1. จำนวนกลุ่มพื้นฐาน (Base Cluster) เท่ากันทั้ง 3 โมเดล คือ 166 กลุ่มพื้นฐาน เนื่องจากไม่มีเอกสารใดที่วลี (Phrase) เหมือนกันยาวมากกว่า 3 คำ

2. จำนวนกลุ่ม Merged Cluster คือกลุ่มที่ได้รับการปรับแต่งตามรูปแบบการรวมกลุ่มพื้นฐานของแต่ละโมเดล ผลการทดลองพบว่าโมเดลของ New STC มีจำนวนกลุ่ม Merged Cluster น้อยที่สุด คือ 56 กลุ่ม การลดลงของจำนวนกลุ่มคิดเป็นร้อยละ 51.30 เมื่อเปรียบเทียบกับ STC_Original ซึ่งมีจำนวน 139 กลุ่ม และการลดลงของจำนวนกลุ่มคิดเป็นร้อยละ 42.45 เมื่อเปรียบเทียบกับ STC with n-gram ซึ่งมีจำนวน 115 กลุ่ม

3. จำนวนโหนด (Node) ภายในซาฟฟิกทรี (Suffix Tree) ผลการทดลองพบว่าโมเดลของ New STC และ STC with n-gram มีจำนวน 1,891 โหนด ลดลงจากของโมเดลของ STC_Original ซึ่งมีจำนวน 5,980 โหนด และอัตราการลดลงของโหนดคิดเป็นร้อยละ 68.38 แสดงให้เห็นว่าความต้องการในการใช้งานหน่วยความจำลดลง เมื่อนำแนวทางการทำงานของเอ็นแกรม (n-gram) มาช่วยในการสร้างซาฟฟิกทรี (Suffix Tree)

4. เวลา (Time) ที่ใช้ในการทำงาน ผลการทดลองพบว่าโมเดลของ New STC ใช้เวลาในการทำงานน้อยที่สุด คือเฉลี่ยเท่ากับ 0.33 วินาที เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ที่ใช้เวลาในการทำงานเฉลี่ยเท่ากับ 1 วินาที

5. ขนาดกลุ่ม (Size Cluster) คือจำนวนเอกสารภายในกลุ่ม ผลการทดลองพบว่าจำนวนเอกสารภายในกลุ่มของโมเดล New STC เฉลี่ยกลุ่มละ 2.59 เอกสาร ซึ่งมีขนาดเล็กที่สุด เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram เฉลี่ยกลุ่มละ 4.74 เอกสาร ทำให้จำนวนเอกสารภายในกลุ่มหรือขนาดของกลุ่มลดลงคิดเป็นร้อยละ 45.36

6. อัตราการทับซ้อน(Overlap) ผลการทดลองพบว่าอัตราการทับซ้อนของกลุ่มเอกสารที่ถูกปรับแต่งด้วยโมเดลของ New STC มีค่า เฉลี่ยเท่ากับ 0.71 ซึ่งมีค่าน้อยที่สุด เมื่อนำมาเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยอัตราการทับซ้อนเท่ากับ 6.45 ทำให้อัตราการทับซ้อนของเอกสารลดลงคิดเป็นร้อยละ 98.45

7. อัตราการครอบคลุม (Coverage) ในการจัดกลุ่มเอกสารของทั้ง 3 โมเดล มีค่าเป็น 1 เท่ากัน แสดงให้เห็นว่าโมเดลของ STC_Original , STC with n-gram และ New STC สามารถจัดกลุ่มเอกสารได้ทุกเอกสาร

8. อัตราความถูกต้อง (Precision) ผลการทดลองพบว่าโมเดลของ New STC มีอัตราความถูกต้องในการจัดกลุ่มสูงที่สุดคือ 0.85 เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยอัตราการความถูกต้องเท่ากับ 0.75

9. อัตราเอกสารที่ไม่ตรงกับความต้องการ (Irrelevant Documents) ภายในกลุ่ม ผลการทดลองพบว่าโมเดลของ New STC มีจำนวนเอกสารที่ไม่ตรงกับความต้องการหรือเอกสารปนเปื้อนน้อยที่สุดคือ 0.15 เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยอัตราเอกสารที่ไม่ตรงกับความต้องการหรือเอกสารปนเปื้อนเท่ากับ 0.25

10. ระยะห่างระหว่างเอกสารที่ตรงกับความต้องการและไม่ตรงกับความต้องการ (Above) ผลการทดลองพบว่าโมเดลของ New STC มีระยะห่างสูงที่สุดคือ 5.67 เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยระยะห่างระหว่างเอกสารที่ตรงกับความต้องการและไม่ตรงกับความต้องการเท่ากับ 3

2 ชุดข้อมูลที่มีลักษณะเนื้อหาแตกต่างกัน

การทดลองด้วยชุดข้อมูลที่มีเนื้อหาแตกต่างกัน จะใช้จำนวนเอกสารภายในกลุ่มเป็นตัวบ่งชี้ในการเลือกกลุ่มมาเป็นตัวแทน เพื่อรวมกับกลุ่มเอกสารอื่นที่มีเนื้อหาแตกต่างกัน เช่นกลุ่มเอกสารในลักษณะเนื้อหาเกี่ยวข้องกับคอมพิวเตอร์ มีจำนวน 5 กลุ่มเอกสาร เราจะเลือกเฉพาะกลุ่มที่มีเอกสารมากที่สุดมาเป็นตัวแทนกลุ่มเอกสารที่มีเนื้อหาเกี่ยวข้องกับคอมพิวเตอร์ คือ กลุ่มเอกสารที่เกี่ยวกับ SQL Database จำนวน 104 เอกสาร , กลุ่มเอกสารที่เกี่ยวกับ Java จำนวน 48 เอกสาร และ กลุ่มเอกสารที่เกี่ยวกับ Text Editor จำนวน 41 เอกสาร ดังแสดงรายละเอียดต่อไปนี้

2.1 ชุด Distinct_1 เป็นชุดเอกสารที่เอกสารมีเนื้อหาแตกต่างกันชุดที่ 1 ประกอบด้วย 4 กลุ่มเอกสารที่มีเนื้อหาแตกต่างกัน คือ Lord Of The Ring จำนวน 80 เอกสาร , SQL Database จำนวน 104 เอกสาร , Photography จำนวน 20 เอกสาร และ Health Care จำนวน 79 เอกสาร รวมเอกสารจำนวน 283 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.7

2.2 ชุด Distinct_2 เป็นชุดเอกสารที่เอกสารมีเนื้อหาแตกต่างกันชุดที่ 2 ประกอบด้วย 4 กลุ่มเอกสารที่มีเนื้อหาแตกต่างกัน คือ Java จำนวน 48 เอกสาร , Health Care จำนวน 79 เอกสาร , Photography จำนวน 20 เอกสาร และ Brad Runner จำนวน 61 เอกสาร รวมเอกสารจำนวน 209 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.8

2.3 ชุด Distinct_3 ประกอบด้วย 4 กลุ่มเอกสารที่มีเนื้อหาแตกต่างกัน คือ Text Editor จำนวน 41 เอกสาร, Health Care จำนวน 79 เอกสาร , Photography จำนวน 20 เอกสาร และ Lord Of The Ring จำนวน 80 เอกสาร รวมเอกสารจำนวน 221 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.9 และแสดงผลค่าเฉลี่ยในตารางที่ 4.10 และ 4.11 ตามลำดับ

ตารางที่ 4.7 แสดงผลการทดลองข้อมูลชุด DISTINT_1

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	540	5	10	15	20	539	5	10	15	20	226
จำนวนกลุ่ม BASE CLUSTER	626					622					622				
จำนวนกลุ่ม MERGED CLUSTER	540					539					226				
จำนวน NODE	22,479					6,853					6,853				
เวลาที่ใช้ (TIME)	5 SEC.					4 SEC.					2 SEC				
ขนาดกลุ่ม	20.2	14.5	12.07	10.45	4.65	21.1	15.2	12.2	10.6	4.66	9.2	6.8	5.7	5.25	2.2
OVERLAP	0.49	0.58	0.55	0.61	7.88	0.34	0.5	0.56	0.59	7.87	0	0.05	0.06	0.14	0.76
COVERAGE	0.24	0.33	0.41	0.46	1	0.28	0.36	0.41	0.48	1	0.16	0.023	0.29	0.33	1
PRECISION	1	0.98	0.98	0.98	0.9	1	1	0.98	0.98	0.89	1	1	1	1	0.99
IRRELEVANT	0	0.02	0.02	0.02	0.1	0	0	0.02	0.02	0.11	0	0	0	0	0.01
ABOVE	-	49.00	49.00	49.00	9.00	-	-	49.00	49.00	8.09	-	-	-	-	99.00

ตารางที่ 4.8 แสดงผลการทดลองข้อมูลชุด DISTINT_2

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	473	5	10	15	20	472	5	10	15	20	159
จำนวนกลุ่ม BASE CLUSTER	533					531					531				
จำนวนกลุ่ม MERGED CLUSTER	473					472					159				
จำนวน NODE	24,140					6,104					6,104				
เวลาที่ใช้ (TIME)	3 SEC.					3 SEC.					1 SEC.				
ขนาดกลุ่ม	15.6	9.9	6.67	6.7	3.99	15.6	9.9	6.67	6.7	3.99	9.4	6.6	5.53	4.9	2.23
OVERLAP	0.04	0.24	0.35	0.37	8.04	0.04	0.24	0.33	0.38	8.03	0.04	0.06	0.14	0.17	0.7
COVERAGE	0.36	0.38	0.41	0.47	1	0.36	0.38	0.42	0.48	1	0.22	0.3	0.35	0.4	1
PRECISION	1	1	0.97	0.97	0.88	1	1	0.97	0.97	0.89	0.92	0.96	0.97	0.98	0.99
IRRELEVANT	0	0	0.03	0.03	0.12	0	0	0.03	0.03	0.11	0.08	0.04	0.03	0.02	0.01
ABOVE	-	-	32.33	32.33	7.33	-	-	32.33	32.33	8.09	11.50	24.00	32.33	49.00	99.00

ตารางที่ 4.9 แสดงผลการทดลองข้อมูลชุด DISTINT_3

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	464	5	10	15	20	463	5	10	15	20	179
จำนวนกลุ่ม BASE CLUSTER	540					536					536				
จำนวนกลุ่ม MERGED CLUSTER	464					463					179				
จำนวน NODE	21,078					5,956					5,956				
เวลาที่ใช้ (TIME)	3 SEC.					3 SEC.					1 SEC.				
ขนาดกลุ่ม	20	14.1	10.73	9.25	4.36	20	14.7	11.2	9.3	4.36	8.6	6.2	5.47	4.85	2.21
OVERLAP	0.49	0.6	0.68	0.75	8.14	0.49	0.56	0.58	0.75	8.13	0.02	0.07	0.12	0.18	0.79
COVERAGE	0.3	0.4	0.43	0.48	1	0.3	0.43	0.48	0.48	1	0.19	0.26	0.33	0.37	1
PRECISION	1	1	1	1	0.9	1	1	1	1	0.9	1	1	1	1	0.99
IRRELEVANT	0	0	0	0	0.1	0	0	0	0	0.1	0	0	0	0	0.01
ABOVE	-	-	-	-	9.00	-	-	-	-	9.00	-	-	-	-	99.00

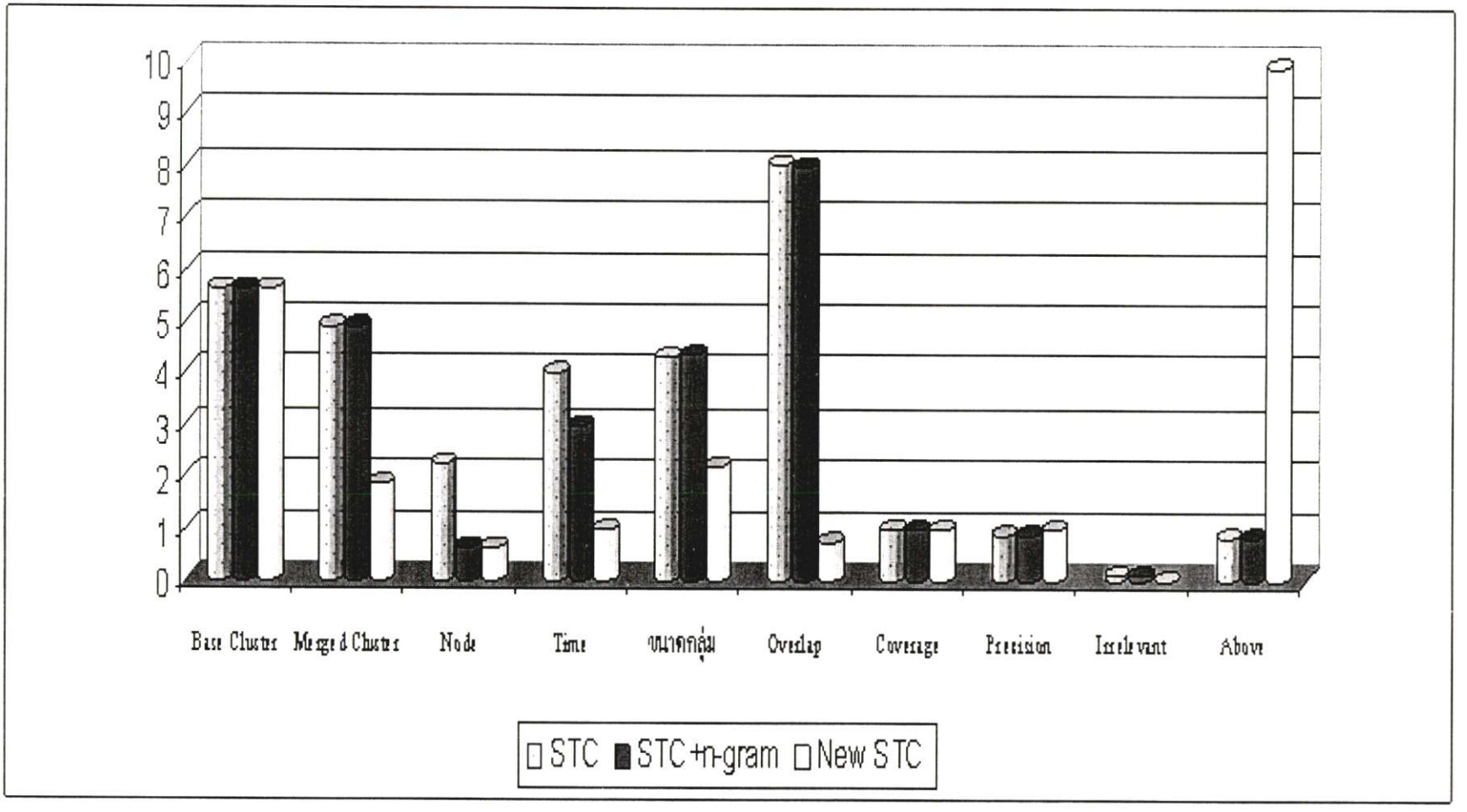
ตารางที่ 4.10 แสดงค่าเฉลี่ยผลการทดลองชุดข้อมูลเนื้อหาแตกต่างกัน

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	ALL	5	10	15	20	ALL	5	10	15	20	ALL
จำนวนกลุ่ม BASE CLUSTER	566					563					563				
จำนวนกลุ่ม MERGED CLUSTER	492					491					188				
จำนวน NODE	22,566					6,304					6,304				
เวลาที่ใช้ (TIME)	4					3					1				
ขนาดกลุ่ม	18.6	12.38	9.82	8.8	4.33	18.9	13.27	10.02	8.87	4.34	9.07	6.53	5.57	5.00	2.21
OVERLAP	0.34	0.47	0.53	0.58	8.02	0.29	0.43	0.49	0.57	8.01	0.02	0.06	0.11	0.16	0.75
COVERAGE	0.3	0.37	0.42	0.47	1	0.31	0.39	0.44	0.48	1	0.19	0.19	0.32	0.37	1.00
PRECISION	1	0.99	0.98	0.98	0.89	1	1	0.98	0.98	0.89	0.97	0.99	0.99	0.99	0.99
IRRELEVANT	0	0.01	0.02	0.02	0.11	0	0	0.02	0.02	0.11	0.03	0.01	0.01	0.01	0.01
ABOVE	0.00	99.00	49.00	49.00	8.09	-	-	49.00	49.00	8.09	32.33	99.00	99.00	99.00	99.00

ตารางที่ 4.11 แสดงค่า NORMALIZE ค่าเฉลี่ยผลการทำงาน 3 โมเดล ของชุดข้อมูลที่มีเนื้อหาแตกต่างกัน

ค่ามาตรฐาน	STC ORIGINAL	STC WITH N-GRAM	NEW STC	หมายเหตุ
จำนวนกลุ่ม BASE CLUSTER	5.66	5.63	5.63	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกนำมาทำให้อยู่ในช่วง 0-10 ด้วยการหารด้วย 100
จำนวนกลุ่ม MERGED CLUSTER	4.92	4.91	1.88	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกนำมาทำให้อยู่ในช่วง 0-10 ด้วยการหารด้วย 100
จำนวน NODE	2.26	0.63	0.63	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกนำมาทำให้อยู่ในช่วง 0-10 ด้วยการหารด้วย 10,000
เวลาที่ใช้ (TIME)	4	3	1	
ขนาดกลุ่ม	4.33	4.34	2.21	
OVERLAP	8.02	8.01	0.75	
COVERAGE	1	1	1.00	
PRECISION	0.89	0.89	0.99	
IRRELEVANT	0.11	0.11	0.01	
ABOVE	0.81	0.81	9.9	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกนำมาทำให้อยู่ในช่วง 0-10 ด้วยการหารด้วย 10

จากตารางที่ 4.11 ค่า Normalize ค่าเฉลี่ยผลการทำงาน 3 โมเดล ของชุดข้อมูลที่มีเนื้อหาแตกต่างกัน สามารถสร้างกราฟได้ดังแสดงในรูปภาพที่ 4.4



รูปที่ 4.4 กราฟแสดงผลการทำงานของ 3 โมเดล ของชุดข้อมูลที่มีเนื้อหาแตกต่างกัน

จากรูปที่ 4.4 คือกราฟสรุปผลการทำงานของทั้ง 3 โมเดล คือ STC_Original , STC with n-gram และ New STC ของชุดข้อมูลที่มีเนื้อหาแตกต่างกัน เพื่อศึกษาความสามารถด้านต่างๆ ของโมเดล ด้วยตัวชี้วัดมาตรฐานต่างๆที่ใช้วัดประสิทธิภาพของโมเดลดังที่กล่าวในข้างต้น และศึกษาความสามารถในการจัดกลุ่มเอกสารหรืออำนาจจำแนกกลุ่มเอกสาร ที่เอกสารภายในกลุ่มมีเนื้อหาแตกต่างกัน ดังแสดงรายละเอียดของผลการทำงานดังต่อไปนี้

1. จำนวนกลุ่มพื้นฐาน (Base Cluster) ผลการทดลองพบว่า จำนวนกลุ่มพื้นฐานของโมเดลที่ใช้เทคนิค n-gram มาช่วยในการสร้าง Suffix Tree มีจำนวน 563 กลุ่ม ซึ่งต่ำกว่าโมเดล STC_Original ซึ่งมีจำนวนกลุ่มเท่ากับ 566 กลุ่ม

2. จำนวนกลุ่ม Merged Cluster คือกลุ่มที่ได้รับการปรับแต่งตามรูปแบบการรวมกลุ่มพื้นฐานของแต่ละโมเดล ผลการทดลองพบว่าโมเดลของ New STC มีจำนวนกลุ่ม Merged Cluster น้อยที่สุด คือ 188 กลุ่ม และการลดลงของจำนวนกลุ่มคิดเป็นร้อยละ 61.79 เมื่อเปรียบเทียบกับ STC_Original ซึ่งมี 492 กลุ่ม และการลดลงของจำนวนกลุ่มคิดเป็นร้อยละ 61.71 เมื่อเปรียบเทียบกับ STC with n-gram ซึ่งมี 491 กลุ่ม

3. จำนวนโหนด (Node) ภายในซาฟฟิกทรี (Suffix Tree) ของโมเดล New STC มีจำนวน 6,304 โหนด ลดลงจากของโมเดล STC_Original ซึ่งมีจำนวน 22,566 โหนด อัตราการลดลงของโหนดคิดเป็นร้อยละ 72.06 ทำให้ความต้องการในการใช้งานหน่วยความจำลดลง เมื่อนำแนวทางการทำงานของเอ็นแกรม (n-gram) มาช่วยในการสร้างซาฟฟิกทรี (Suffix Tree)

4. เวลา (Time) ที่ใช้ในการทำงาน ผลการทดลองพบว่าโมเดลของ New STC ใช้เวลาในการทำงานน้อยที่สุด คือเฉลี่ยเท่ากับ 1 วินาที เมื่อเปรียบเทียบกับ STC_Original ที่ใช้เวลาในการทำงานเฉลี่ยเท่ากับ 4 วินาที และ STC with n-gram ที่ใช้เวลาในการทำงานเฉลี่ยเท่ากับ 3 วินาที 5. ขนาดกลุ่ม (Size Cluster) คือจำนวนเอกสารภายในกลุ่ม ผลการทดลองพบว่าจำนวนเอกสารภายในกลุ่มของโมเดล New STC เฉลี่ยกลุ่มละ 2.21 เอกสาร ซึ่งมีขนาดเล็กที่สุดเมื่อเปรียบเทียบกับ STC_Original ที่มีขนาดกลุ่มเฉลี่ยกลุ่มละ 4.33 เอกสาร ซึ่งการลดลงของขนาดกลุ่มเอกสารคิดเป็นร้อยละ 48.96 และ STC with n-gram มีขนาดกลุ่มเฉลี่ยกลุ่มละ 4.34 เอกสาร ซึ่งการลดลงของขนาดกลุ่มเอกสารคิดเป็นร้อยละ 49.08

6. อัตราการทับซ้อน(Overlap) ของกลุ่มเอกสารที่ถูกปรับแต่งด้วยโมเดลของ New STC มี เฉลี่ยเท่ากับ 0.75 ซึ่งมีอัตราการทับซ้อนน้อยที่สุด เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยอัตราการทับซ้อนเท่ากับ 8.02 และ 8.01 ตามลำดับ การลดลงของอัตราการทับซ้อนของเอกสารคิดเป็นร้อยละ 90.65 และ 90.64 ตามลำดับ

7. อัตราการครอบคลุม (Coverage) ในการจัดกลุ่มเอกสารของทั้ง 3 โมเดล มีค่าเป็น 1 เท่ากัน แสดงให้เห็นว่าโมเดลของ STC_Original , STC with n-gram และ New STC สามารถจัดกลุ่มเอกสารได้ทุกเอกสาร

8. อัตราความถูกต้อง (Precision) ผลการทดลองพบว่าโมเดลของ New STC มีอัตราความถูกต้องในการจัดกลุ่มสูงที่สุดคือ 0.99 เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยอัตราการความถูกต้องเท่ากับ 0.89

9. อัตราเอกสารที่ไม่ตรงกับความต้องการ (Irrelevant Documents) ภายในกลุ่ม ผลการทดลองพบว่าโมเดลของ New STC มีจำนวนเอกสารที่ไม่ตรงกับความต้องการหรือเอกสารปนเปื้อนน้อยที่สุดคือ 0.01 เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยอัตราเอกสารที่ไม่ตรงกับความต้องการเท่ากับ 0.11

10. ระยะห่างระหว่างเอกสารที่ตรงกับความต้องการและไม่ตรงกับความต้องการ (Above) ผลการทดลองพบว่าโมเดลของ New STC มีระยะห่างสูงที่สุดคือ 9.9 เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยระยะห่างระหว่างเอกสารที่ตรงกับความต้องการและไม่ตรงกับความต้องการเท่ากับ 0.81

3 ชุดข้อมูลที่มีลักษณะผสมทุกๆเนื้อหาเข้าด้วยกัน ประกอบด้วย

เป็นชุดข้อมูลที่นำทุกกลุ่มของแต่ละเรื่องราวมารวมกัน เพื่อทดสอบประสิทธิภาพในการจัดกลุ่มที่แต่ละเอกสารภายของชุดข้อมูลมีทั้งแตกต่างกันและเหมือนกัน โดยแบ่งออกเป็น 3 ชุดข้อมูล ดังแสดงรายละเอียดต่อไปนี้

3.1 ชุด Mixed_1 เป็นการนำเอกสารของทุกกลุ่มมารวมอยู่ของชุดข้อมูลเดียว ประกอบด้วย 9 กลุ่มเอกสารคือ SQL Database , Data Warehouse , XML , Java , Text Editor , Health Care , Photography , Lord Of The Ring และ Brad Runner จำนวน 479 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.12

3.2 ชุด Mixed_2 เป็นการนำเอกสารของกลุ่มที่มีเนื้อหาเกี่ยวกับคอมพิวเตอร์มารวมอยู่ของชุดข้อมูลเดียวกัน ประกอบด้วย 5 กลุ่มเอกสารที่เกี่ยวข้องกับคอมพิวเตอร์ คือ SQL Database , Data Warehouse , XML และ Java , Text Editor จำนวน 240 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.13

3.3 ชุด Mixed_3 เป็นการสุ่มเลือกกลุ่มเอกสารมารวมเป็นชุดเดียวกันโดยไม่แบ่งแยกเนื้อหา ประกอบด้วย 5 กลุ่มเอกสารคือ Data Warehouse , Health Care , Photography , Lord Of The Ring และ Brad Runner จำนวน 273 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.14 และแสดงผลค่าเฉลี่ยในตารางที่ 4.15 และ 4.16 ตามลำดับ

ตารางที่ 4.12 แสดงผลการทดลองข้อมูลชุด MIXED_1

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	997	5	10	15	20	996	5	10	15	20	389
จำนวนกลุ่ม BASE CLUSTER	1,155					1,143					1,143				
จำนวนกลุ่ม MERGED CLUSTER	997					996					389				
จำนวน NODE	47,790					12,304					12,304				
เวลาที่ใช้ (TIME)	4 SEC.					4 SEC.					2 SEC.				
ขนาดกลุ่ม	28.8	20.2	16.07	14.7	4.93	28.8	20.8	16.53	14.95	4.93	14.4	10.7	8.73	7.6	2.37
OVERLAP	0.23	0.26	0.31	0.4	9.24	0.23	0.22	0.36	0.39	9.23	0.01	0.01	0.04	0.1	0.92
COVERAGE	0.24	0.33	0.38	0.44	1	0.24	0.35	0.38	0.45	1	0.15	0.22	0.26	0.29	1
PRECISION	0.99	0.92	0.94	0.93	0.76	0.99	0.92	0.94	0.93	0.76	0.95	0.91	0.94	0.92	0.92
IRRELEVANT	0.01	0.08	0.06	0.07	0.24	0.01	0.08	0.06	0.07	0.24	0.05	0.09	0.06	0.08	0.08
ABOVE	99.00	11.50	15.67	13.29	3.17	99.00	11.50	15.67	13.29	3.17	19.00	10.11	15.67	11.50	11.50

ตารางที่ 4.13 แสดงผลการทดลองข้อมูลชุด MIXED_2

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	521	5	10	15	20	521	5	10	15	20	196
จำนวนกลุ่ม BASE CLUSTER	601					594					594				
จำนวนกลุ่ม MERGED CLUSTER	521					521					196				
จำนวน NODE	27,189					6,536					6,536				
เวลาที่ใช้ (TIME)	2 SEC.					2 SEC.					1 SEC.				
ขนาดกลุ่ม	12.6	11.4	9.93	8.55	4.78	14.2	11.4	9.93	8.7	4.78	7.8	6.5	5.53	4.95	2.31
OVERLAP	0.26	0.28	0.57	0.57	9.38	0.16	0.28	0.57	0.6	9.38	0.03	0.07	0.14	0.16	0.89
COVERAGE	0.21	0.37	0.4	0.45	1	0.25	0.37	0.4	0.45	1	0.16	0.25	0.3	0.35	1
PRECISION	0.88	0.94	0.95	0.94	0.78	0.88	0.94	0.95	0.92	0.78	0.83	0.91	0.94	0.91	0.93
IRRELEVANT	0.12	0.06	0.05	0.06	0.22	0.12	0.06	0.05	0.08	0.22	0.17	0.09	0.06	0.09	0.07
ABOVE	7.33	15.67	19.00	15.67	3.55	7.33	15.67	19.00	11.50	3.55	4.88	10.11	15.67	10.11	13.29

ตารางที่ 4.14 แสดงผลการทดลองข้อมูลชุด MIXED_3

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	571	5	10	15	20	570	5	10	15	20	224
จำนวนกลุ่ม BASE CLUSTER	698					686					686				
จำนวนกลุ่ม MERGED CLUSTER	571					570					224				
จำนวน NODE	26,141					7,099					7,099				
เวลาที่ใช้ (TIME)	2 SEC.					2 SEC.					1 SEC.				
ขนาดกลุ่ม	29.8	18.1	15.67	13.5	4.44	29.8	18.8	16	13.65	4.45	14	9.3	7.27	6.45	2.36
OVERLAP	0.22	0.38	0.56	0.69	8.29	0.22	0.4	0.54	0.71	8.29	0.01	0.07	0.09	0.12	0.93
COVERAGE	0.48	0.48	0.55	0.59	1	0.45	0.49	0.57	0.59	1	0.25	0.32	0.37	0.42	1
PRECISION	1	0.98	0.95	0.94	0.82	1	0.98	0.95	0.94	0.82	1	0.96	0.97	0.98	0.95
IRRELEVANT	0	0.02	0.05	0.06	0.18	0	0.02	0.05	0.06	0.18	0	0.04	0.03	0.02	0.05
ABOVE	-	49.00	19.00	15.67	4.56	-	49.00	19.00	15.67	4.56	-	24.00	32.33	49.00	19.00

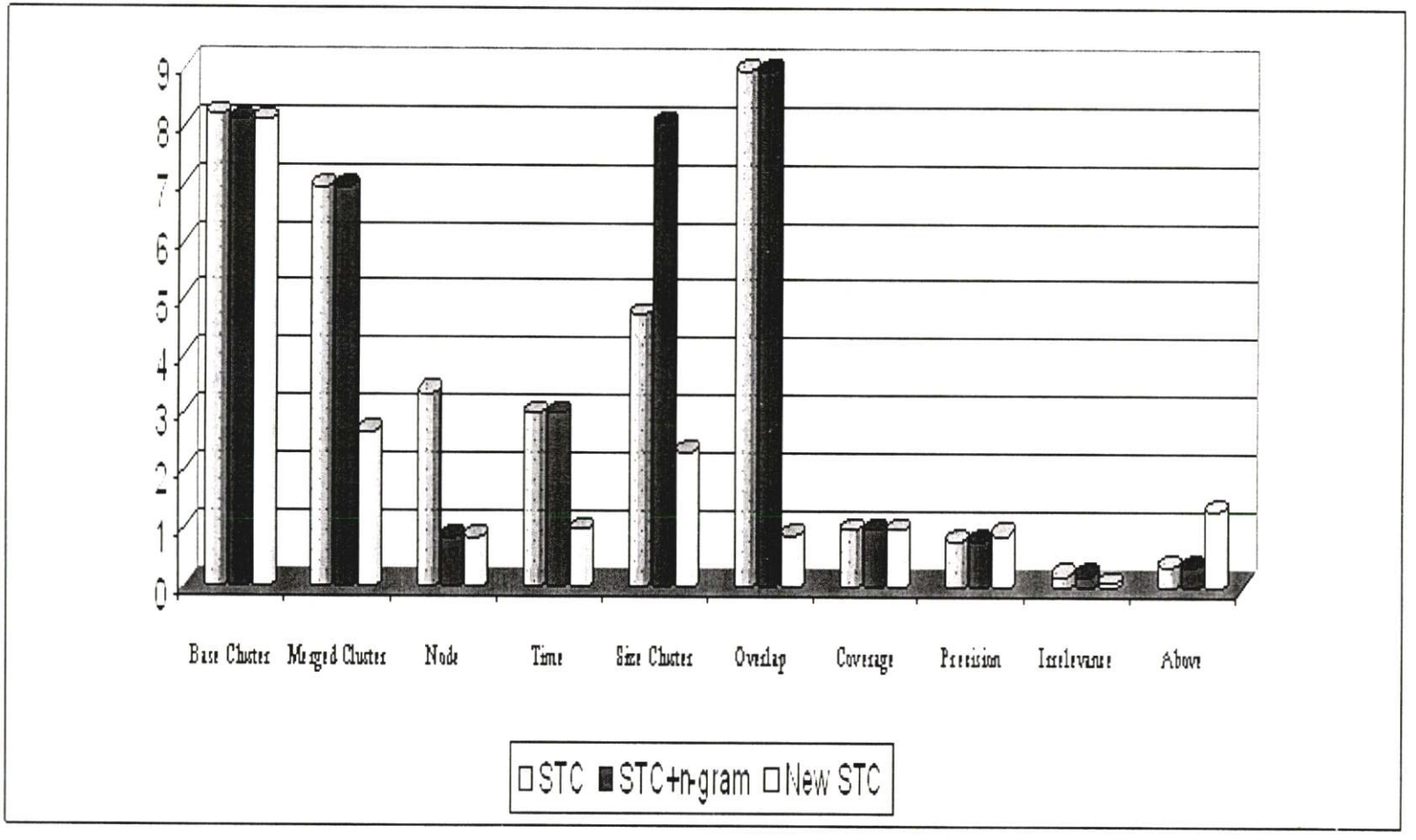
ตารางที่ 4.15 แสดงค่าเฉลี่ยผลการทดลองชุดข้อมูลเนื้อหาผสม

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	ALL	5	10	15	20	ALL	5	10	15	20	ALL
จำนวนกลุ่ม BASE CLUSTER	818					808					808				
จำนวนกลุ่ม MERGED CLUSTER	696					696					270				
จำนวน NODE	33,707					8,646					8,646				
เวลาที่ใช้ (TIME)	3 SEC.					3 SEC.					1 SEC.				
ขนาดกลุ่ม	23.73	16.57	13.89	12.25	4.72	24.27	17	14.15	12.43	4.72	102.07	8.83	7.18	6.33	2.35
OVERLAP	0.24	0.31	0.48	0.55	8.97	0.2	0.3	0.49	0.57	8.97	0.02	0.05	0.09	0.13	0.91
COVERAGE	0.31	0.39	0.44	0.49	1	0.31	0.4	0.45	0.05	1	0.19	0.26	0.31	0.35	1.00
PRECISION	0.96	0.95	0.95	0.94	0.79	0.96	0.95	0.95	0.93	0.79	0.93	0.93	0.95	0.94	0.93
IRRELEVANT	0.04	0.05	0.05	0.06	0.21	0.04	0.05	0.05	0.07	0.21	0.07	0.07	0.05	0.06	0.07
ABOVE	24.00	19.00	19.00	15.67	3.76	24.00	19.00	19.00	13.29	3.76	13.29	13.29	19.00	15.67	13.29

ตารางที่ 4.16 แสดงค่า NORMALIZE ของค่าเฉลี่ยผลการทำงาน 3 โมเดล ของชุดข้อมูลที่มีเนื้อหาผสม

ค่ามาตรฐาน	STC ORGINAL	STC WITH N-GRAM	NEW STC	หมายเหตุ
จำนวนกลุ่ม BASE CLUSTER	8.18	8.08	8.08	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10
จำนวนกลุ่ม MERGED CLUSTER	6.96	6.96	2.7	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร100
จำนวน NODE	3.37	0.86	0.86	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10,000
เวลาที่ใช้ (TIME)	3	3	1	
ขนาดกลุ่ม (SIZE CLUSTER)	4.72	8.05	2.35	
OVERLAP	8.97	8.97	0.91	
COVERAGE	1	1	1	
PRECISION	0.79	0.79	0.93	
IRRELEVANCE	0.21	0.21	0.07	
ABOVE	0.38	0.38	1.33	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10

จากตารางที่ 4.16 ค่า Normalize ค่าเฉลี่ยผลการทำงาน 3 โมเดล ของชุดข้อมูลที่มีเนื้อหาผสม สามารถสร้างกราฟได้ดังแสดงในรูปภาพที่ 4.5



รูปที่ 4.5 แสดงกราฟแสดงผลการทำงานของ 3 โมเดล ของชุดข้อมูลที่มีเนื้อหาผสม

จากรูปที่ 4.5 คือกราฟสรุปผลการทำงานของทั้ง 3 โมเดล คือ STC_Original , STC with n-gram และ New STC ของชุดข้อมูลที่มีเนื้อหาผสม เพื่อศึกษาความสามารถด้านต่างๆของโมเดล ด้วยตัวชี้วัดมาตรฐานต่างๆที่ใช้วัดประสิทธิภาพของโมเดลดังที่กล่าวในข้างต้น และศึกษาความสามารถในการจัดกลุ่มเอกสารหรืออำนาจจำแนกกลุ่มเอกสาร ที่เอกสารภายในกลุ่มมีเนื้อหาผสม ดังแสดงรายละเอียดของผลการทำงานดังต่อไปนี้

1. จำนวนกลุ่มพื้นฐาน (Base Cluster) ผลการทดลองพบว่า จำนวนกลุ่มพื้นฐานของโมเดล STC with n-gram และ New STC ซึ่งใช้เทคนิค n-gram มาช่วยในการสร้าง Suffix Tree ทำให้มีจำนวนกลุ่มพื้นฐานเท่ากับ 808 กลุ่ม ซึ่งต่ำกว่าโมเดล STC_Original ซึ่งมีจำนวนกลุ่มพื้นฐานเท่ากับ 818 กลุ่ม

2. จำนวนกลุ่ม Merged Cluster คือกลุ่มที่ได้รับการปรับแต่งตามรูปแบบการรวมกลุ่มพื้นฐานของแต่ละโมเดล ผลการทดลองพบว่าโมเดลของ New STC มีจำนวนกลุ่ม Merged Cluster น้อยที่สุด คือ 270 กลุ่ม และการลดลงของจำนวนกลุ่มคิดเป็นร้อยละ 61.21 เมื่อเปรียบเทียบกับ STC_Original ซึ่งมี 696 กลุ่ม และการลดลงของจำนวนกลุ่มคิดเป็นร้อยละ 61.21 เมื่อเปรียบเทียบกับ STC with n-gram ซึ่งมี 696 กลุ่ม

3. จำนวนโหนด (Node) ภายในซาฟฟิกทรี (Suffix Tree) ผลการทดลองพบว่าโมเดลของ STC with n-gram และ New STC มีจำนวนโหนดเท่ากับ 8,646 โหนด ซึ่งลดลงจากจำนวนโหนดของโมเดล STC_Original ซึ่งมีจำนวน 33,707 โหนด และอัตราการลดลงของโหนดคิดเป็นร้อยละ 74.35 ทำให้ความต้องการในการใช้งานหน่วยความจำลดลง

4. เวลา (Time) ที่ใช้ในการทำงาน ผลการทดลองพบว่าโมเดลของ New STC ใช้เวลาในการทำงานน้อยที่สุด คือเฉลี่ยเท่ากับ 1 วินาที เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งใช้เวลาในการทำงานเฉลี่ยเท่ากับ 3 วินาที

5. ขนาดกลุ่ม (Size Cluster) คือจำนวนเอกสารภายในกลุ่ม ผลการทดลองพบว่าจำนวนเอกสารภายในกลุ่มของโมเดล New STC เฉลี่ยกลุ่มละ 2.35 เอกสาร ซึ่งมีขนาดเล็กที่สุดเมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ที่มีขนาดกลุ่มเฉลี่ยกลุ่มละ 4.72 เอกสาร ซึ่งการลดลงของขนาดกลุ่มเอกสารคิดเป็นร้อยละ 50.21

6. อัตราการทับซ้อน(Overlap) ของกลุ่มเอกสารที่ถูกปรับแต่งด้วยโมเดลของ New STC มี เฉลี่ยเท่ากับ 0.91 ซึ่งมีอัตราการทับซ้อนน้อยที่สุด เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยอัตราการทับซ้อนเท่ากับ 8.97 และการลดลงของอัตราการทับซ้อนของเอกสารคิดเป็นร้อยละ 89.86

7. อัตราการครอบคลุม (Coverage) ในการจัดกลุ่มเอกสารของทั้ง 3 โมเดล มีค่าเป็น 1 เท่ากัน แสดงให้เห็นว่าโมเดลของ STC_Original , STC with n-gram และ New STC สามารถจัดกลุ่มเอกสารได้ทุกเอกสาร

8. อัตราความถูกต้อง (Precision) ผลการทดลองพบว่าโมเดลของ New STC มีอัตราความถูกต้องในการจัดกลุ่มสูงที่สุดคือ 0.93 เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยอัตราการความถูกต้องเท่ากับ 0.79

9. อัตราเอกสารที่ไม่ตรงกับความต้องการ (Irrelevant Documents) ภายในกลุ่ม ผลการทดลองพบว่าโมเดลของ New STC มีจำนวนเอกสารที่ไม่ตรงกับความต้องการหรือเอกสารปนเปื้อนน้อยที่สุดคือ 0.07 เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยอัตราเอกสารที่ไม่ตรงกับความต้องการเท่ากับ 0.21

10. ระยะห่างระหว่างเอกสารที่ตรงกับความต้องการและไม่ตรงกับความต้องการ (Above) ผลการทดลองพบว่าโมเดลของ New STC มีระยะห่างสูงที่สุดคือ 13.29 เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยระยะห่างระหว่างเอกสารที่ตรงกับความต้องการและไม่ตรงกับความต้องการเท่ากับ 3.76

จากผลการทดลองของข้อมูลชุดผลการสืบค้นภายในกลุ่มของ Dmoz.com ทั้ง 3 ชุดข้อมูล คือ ชุดข้อมูลที่เอกสารมีเนื้อหาใกล้เคียงกัน , ชุดข้อมูลที่เอกสารมีเนื้อหาแตกต่างกัน และชุดข้อมูลที่มีเนื้อหาผสม ผลการทดลองพบว่า

1. โมเดลของ New STC สามารถจัดกลุ่มได้ถูกต้องมากกว่าโมเดลของ STC_Original และ STC with n-gram ส่งผลให้จำนวนเอกสารที่ไม่ตรงกับความต้องการภายในกลุ่มลดลง และช่วงระยะห่างของเอกสารที่ตรงกับความต้องการและไม่ตรงกับความต้องการสูงขึ้น

2. โมเดล New STC ใช้เวลาน้อยกว่าโมเดล STC_Original และ STC with n-gram เพราะโมเดลของ New STC ไม่ต้องเสียเวลาในการคำนวณคะแนนกลุ่ม (คะแนน Sb เพื่อใช้ในการวัดความน่าจะเป็นในการเป็นป้ายชื่อกลุ่มของกลุ่มเอกสาร) และ กระบวนการทำงานของโมเดล New STC มีการลบเอกสารที่ซ้ำซ้อนภายในกลุ่มเอกสาร ทำให้จำนวนเอกสารภายในกลุ่มลดลง และลบกลุ่มที่ป้ายชื่อกลุ่ม และสมาชิกภายในกลุ่มทับซ้อนกับกลุ่มอื่นทั้งหมด ทำให้จำนวนกลุ่มลดลง และการเชื่อมป้ายชื่อกลุ่มจะพิจารณาเฉพาะกลุ่มที่มีจำนวนคำภายป้ายชื่อกลุ่มเท่ากันและจำนวนคำมีค่ามากกว่าหรือเท่ากับขนาดของ n-gram เท่านั้น ทำให้ลดจำนวนกลุ่มในการพิจารณาลงทุกรอบการทำงาน

3. โมเดล New STC ใช้หน่วยความจำ (Space) น้อยกว่าโมเดล STC_Original และ STC with n-gram เนื่องจากการใช้เทคนิคของ n-gram มาช่วยในการสร้าง Suffix Tree ให้มีจำนวนโหนด (Node) หรือจำนวนของคำภายใน Suffix Tree

4. โมเดล New STC มีอัตราการทับซ้อน (Overlap) ของเอกสารน้อยกว่าโมเดล STC_Original และ STC with n-gram

5. อัตราการครอบคลุม (Coverage) ในการจัดกลุ่มเอกสารของทั้ง 3 โมเดล มีค่าเป็น 1 เท่ากัน แสดงให้เห็นว่าโมเดลของ STC_Original , STC with n-gram และ New STC สามารถจัดกลุ่มเอกสารได้ทุกเอกสาร

6. ขนาดของกลุ่มที่จัดกลุ่มด้วยโมเดลของ New STC มีขนาดเล็กกว่าขนาดกลุ่มของโมเดล STC_Original และ STC with n-gram

4.3.2 ชุดข้อมูลผลการสืบค้นด้วยคำสืบค้นภายใน Dmoz.com

ขั้นตอนการทดลอง แบ่งออกเป็น 4 ขั้นตอนคือ

1. ทำการสืบค้นภายใน Dmoz.com ที่ถูกจัดกลุ่มโดยมนุษย์ ด้วยคำสืบค้น (Query Word) จากกลุ่มข้อมูลทั้ง 16 กลุ่ม ประกอบด้วย กลุ่ม Arts , กลุ่ม Business , กลุ่ม Computers , กลุ่ม Games , กลุ่ม Health , กลุ่ม Home , กลุ่ม Kids and Teens , กลุ่ม News , กลุ่ม Recreation , กลุ่ม Reference , กลุ่ม Regional , กลุ่ม Science , กลุ่ม Shopping , กลุ่ม Society , กลุ่ม Sports และกลุ่ม World เพื่อทดสอบชุดข้อมูลในลักษณะผลการสืบค้นจากคำสืบค้นใน 3 ลักษณะ ประกอบด้วย

1.1 คำสืบค้นที่มีลักษณะเป็นคำทั่วๆไป (General Concepts) ประกอบด้วย "Search", "Computer", "Music" เพื่อทดสอบว่าโมเดลสามารถจัดกลุ่มผลการสืบค้นได้ถูกต้องอยู่ในระดับใด เมื่อคำสืบค้นเป็นคำทั่วๆไป

1.2 คำสืบค้นที่มีลักษณะเป็นชื่อ (Entity Name) ซึ่งคำจะมีลักษณะเป็นชื่อเฉพาะ ประกอบด้วย "Thailand", "Iraq", "Disney" เพื่อทดสอบว่าโมเดลสามารถจัดกลุ่มผลการสืบค้นได้ถูกต้องอยู่ในระดับใด เมื่อคำสืบค้นเป็นคำชื่อเฉพาะ

1.3 คำสืบค้นที่มีลักษณะเป็นคำกำกวม (Ambiguous Queries) หรือคำที่มีความหมายหลากหลาย ประกอบด้วย "Matrix", "Apple" และ "Jaguar" เพื่อทดสอบว่าโมเดลสามารถจัดกลุ่มผลการสืบค้นได้ถูกต้องอยู่ในระดับใด เมื่อคำสืบค้นเป็นคำกำกวม

จากชุดข้อมูลผลสืบค้นประกอบด้วย 9 คำสืบค้น เราจะนำคำอธิบายสั้นๆ (Snippets) มาเป็นตัวแทนเอกสารของแต่ละกลุ่ม เพื่อนำมาใช้ในการทดลอง

2. นำคำอธิบายสั้นๆ (Snippets) ซึ่งใช้เป็นตัวแทนเอกสาร มาผ่านกระบวนการทำงานด้าน pre-processing ด้วยการทำ Stop-words โดยการกำจัดคำที่ไม่มีความหมาย เพื่อให้จำนวนคำลดลง และ การทำ Stemming words โดยการทำให้คำอยู่ในรากศัพท์เดิม เพื่อลดความหลากหลายของคำที่จะใช้ในการสร้างซัฟฟิกทรี (Suffix Tree)

3. นำคำที่เหลือจากการทำงานในขั้นตอนการทำ pre-processing เข้าสู่กระบวนการทำงานของโมเดล และคำนวณหาค่าต่างๆของตัวชี้วัดประสิทธิภาพของโมเดล

4. ศึกษาค่าต่างๆของตัวชี้วัด โดยการแบ่งการศึกษาค่ามาตรฐานต่างๆออกเป็นช่วงๆของผลการจัดกลุ่ม ประกอบด้วย ค่ามาตรฐานในช่วง 5 กลุ่มแรกของผลการจัดกลุ่ม , ช่วง 10 กลุ่มแรกของผลการจัดกลุ่ม , ช่วง 15 กลุ่มแรกของผลการจัดกลุ่ม , ช่วง 20 กลุ่มแรกของผลการจัดกลุ่ม และภาพรวมทั้งหมดของผลการจัดกลุ่มในแต่ละชุดข้อมูล ดังแสดงในตารางผลการทดลอง

การทดลองในชุดข้อมูลนี้ จะทำการแบ่งชุดข้อมูล และแสดงผลการทดลองตามลักษณะของคำสืบค้น เพื่อสรุปผลการทำงานในแต่ละลักษณะคำสืบค้น และสรุปรวมค่าเฉลี่ยผลการทดลองของทั้ง 9 คำสืบค้น ซึ่งการทดลองด้วยชุดข้อมูลซึ่งได้จากคำสืบค้น และผลการทดลองประกอบด้วยรายละเอียดดังต่อไปนี้

1. คำสืบค้นที่มีลักษณะเป็นคำทั่วๆไป (General Concepts) ประกอบด้วย

1.1 ชุดข้อมูลผลการสืบค้นของคำว่า "Search" เป็นคำสืบค้นในลักษณะทั่วไป (General Concepts) จาก 10 กลุ่มเอกสาร ประกอบด้วยกลุ่ม Computers จำนวน 1,299 เอกสาร , กลุ่ม Health จำนวน 720 เอกสาร , กลุ่ม Recreation จำนวน 576 เอกสาร , กลุ่ม Reference จำนวน 727 เอกสาร , กลุ่ม Regional จำนวน 1,222 เอกสาร , กลุ่ม Science จำนวน 756 เอกสาร , กลุ่ม Shopping จำนวน 410 เอกสาร , กลุ่ม Society จำนวน 1,570 เอกสาร , กลุ่ม Sports จำนวน 197 เอกสาร และ กลุ่ม World จำนวน 490 เอกสาร รวมทั้งหมด 7,967 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.17

1.2 ชุดข้อมูลผลการสืบค้นของคำว่า "Computer" เป็นคำสืบค้นในลักษณะทั่วไป (General Concepts) จาก 7 กลุ่มเอกสาร ประกอบด้วย กลุ่ม Business จำนวน 2,255 เอกสาร , กลุ่ม Games จำนวน 937 เอกสาร , กลุ่ม Health จำนวน 154 เอกสาร , กลุ่ม Kids and Teens จำนวน 267 เอกสาร , กลุ่ม Reference จำนวน 1,329 เอกสาร , กลุ่ม Science จำนวน 1,191 เอกสาร และ กลุ่ม Society จำนวน 941 เอกสาร รวมทั้งหมด 7,074 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.18

1.3 ชุดข้อมูลผลการสืบค้นของคำว่า "Music" เป็นคำสืบค้นในลักษณะทั่วไป (General Concepts) จาก 9 กลุ่มเอกสาร ประกอบด้วย กลุ่ม Arts จำนวน 2,195 เอกสาร , กลุ่ม Business จำนวน 2,298 เอกสาร , กลุ่ม Computers จำนวน 1,209 เอกสาร , กลุ่ม Games จำนวน 669 เอกสาร , กลุ่ม Health จำนวน 116 เอกสาร , กลุ่ม Home จำนวน 68 เอกสาร , กลุ่ม Kids and Teens จำนวน 608 เอกสาร , กลุ่ม Recreation จำนวน 456 เอกสาร และ กลุ่ม Society จำนวน 1,515 เอกสาร รวมทั้งหมด 9,134 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.19

จากชุดข้อมูลคำสืบค้นที่มีลักษณะเป็นคำทั่วๆไป (General Concepts) ของทั้ง 3 คำสืบค้น นำมาคำนวณค่าเฉลี่ยได้ดังแสดงผลการทดลองในตารางที่ 4.20 , 4.21 และกราฟแสดงผลการทดลองเฉลี่ยในรูปที่ 4.6

ตารางที่ 4.17 แสดงผลการจัดกลุ่มชุดข้อมูลผลการสืบค้นของคำว่า "SEARCH"

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	16,182	5	10	15	20	15,744	5	10	15	20	11,662
จำนวนกลุ่ม BASE CLUSTER	20,375					19,125					19,125				
จำนวนกลุ่ม MERGED CLUSTER	16,182					15,744					11,662				
จำนวน NODE	897,009					177,360					177,360				
เวลาที่ใช้ (TIME)	29 MIN 39 SEC					28 MIN 6 SEC					5 MIN 48 SEC				
ขนาดกลุ่ม (SIZE CLUSTER)	577.2	363.6	271.67	230.65	8.69	577.2	363.6	276.47	233.5	8.84	82.4	63.7	56.87	50.75	2.79
OVERLAP	0.19	0.3	0.36	0.41	16.67	0.19	0.3	0.4	0.43	16.49	0	0.2	0.3	0.3	3.08
COVERAGE	0.3	0.35	0.37	0.41	1	0.3	0.35	0.37	0.41	1	0.05	0.08	0.1	0.12	1
PRECISION	0.66	0.63	0.68	0.63	0.7	0.66	0.63	0.68	0.63	0.69	0.36	0.38	0.36	0.38	0.77
IRRELEVANT	0.34	0.37	0.32	0.37	0.3	0.34	0.37	0.32	0.37	0.31	0.64	0.62	0.64	0.62	0.23
ABOVE	1.94	1.70	2.13	1.70	2.33	1.94	1.70	2.13	1.70	2.23	0.56	0.61	0.56	0.61	3.35

ตารางที่ 4.18 แสดงผลการจัดกลุ่มชุดข้อมูลผลการสืบค้นของคำว่า "COMPUTER"

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	15,748	5	10	15	20	15,575	5	10	15	20	11,465
จำนวนกลุ่ม BASE CLUSTER	19,722					18,860					18,860				
จำนวนกลุ่ม MERGED CLUSTER	15,748					15,575					11,465				
จำนวน NODE	938,434					169,464					169,464				
เวลาที่ใช้ (TIME)	26 MIN 46 SEC.					26 MIN 26 SEC.					6 MIN 9 SEC.				
ขนาดกลุ่ม (SIZE CLUSTER)	387.4	250.3	216	187.85	8.71	387.4	263.6	221.93	187.85	8.78	79.2	64.9	55.93	51	2.84
OVERLAP	0.2	0.29	0.29	0.33	18.4	0.2	0.28	0.29	0.33	18.34	0	0	0	0.01	3.61
COVERAGE	0.23	0.27	0.36	0.4	1	0.23	0.29	0.37	0.4	1	0.06	0.09	0.12	0.14	1
PRECISION	0.62	0.68	0.61	0.59	0.72	0.62	0.62	0.57	0.59	0.72	0.52	0.51	0.68	0.58	0.79
IRRELEVANT	0.38	0.32	0.39	0.41	0.28	0.38	0.38	0.43	0.41	0.28	0.48	0.49	0.32	0.42	0.21
ABOVE	1.63	2.13	1.56	1.44	2.57	1.63	1.63	1.33	1.44	2.57	1.08	1.04	2.13	1.38	3.76

ตารางที่ 4.19 แสดงผลการจัดกลุ่มชุดข้อมูลผลการสืบค้นของคำว่า "MUSIC"

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	19,151	5	10	15	20	18,924	5	10	15	20	14,140
จำนวนกลุ่ม BASE CLUSTER	22,034					21,265					21,265				
จำนวนกลุ่ม MERGED CLUSTER	19,151					18,924					14,140				
จำนวน NODE	984,464					197,222					197,222				
เวลาที่ใช้ (TIME)	41 MIN. 28 SEC.					39 MIN. 45 SEC.					5 MIN 38 SEC				
ขนาดกลุ่ม (SIZE CLUSTER)	224	197.7	172	156.85	8.98	196.6	197.7	175.2	160.85	9.06	88.4	68	59.47	54.85	3.03
OVERLAP	0.18	0.23	0.39	0.38	17.84	0.05	0.23	0.33	0.31	17.77	0.14	0.09	0.08	0.1	3.68
COVERAGE	0.1	0.18	0.2	0.25	1	0.1	0.18	0.22	0.27	1	0.04	0.07	0.09	0.11	1
PRECISION	0.77	0.83	0.79	0.81	0.7	0.76	0.7	0.72	0.67	0.69	0.88	0.67	0.63	0.62	0.75
IRRELEVANT	0.23	0.17	0.21	0.19	0.3	0.24	0.3	0.28	0.33	0.31	0.12	0.33	0.37	0.38	0.25
ABOVE	3.35	4.88	3.76	4.26	2.33	3.17	2.33	2.57	2.03	2.23	7.33	2.03	1.70	1.63	3.00

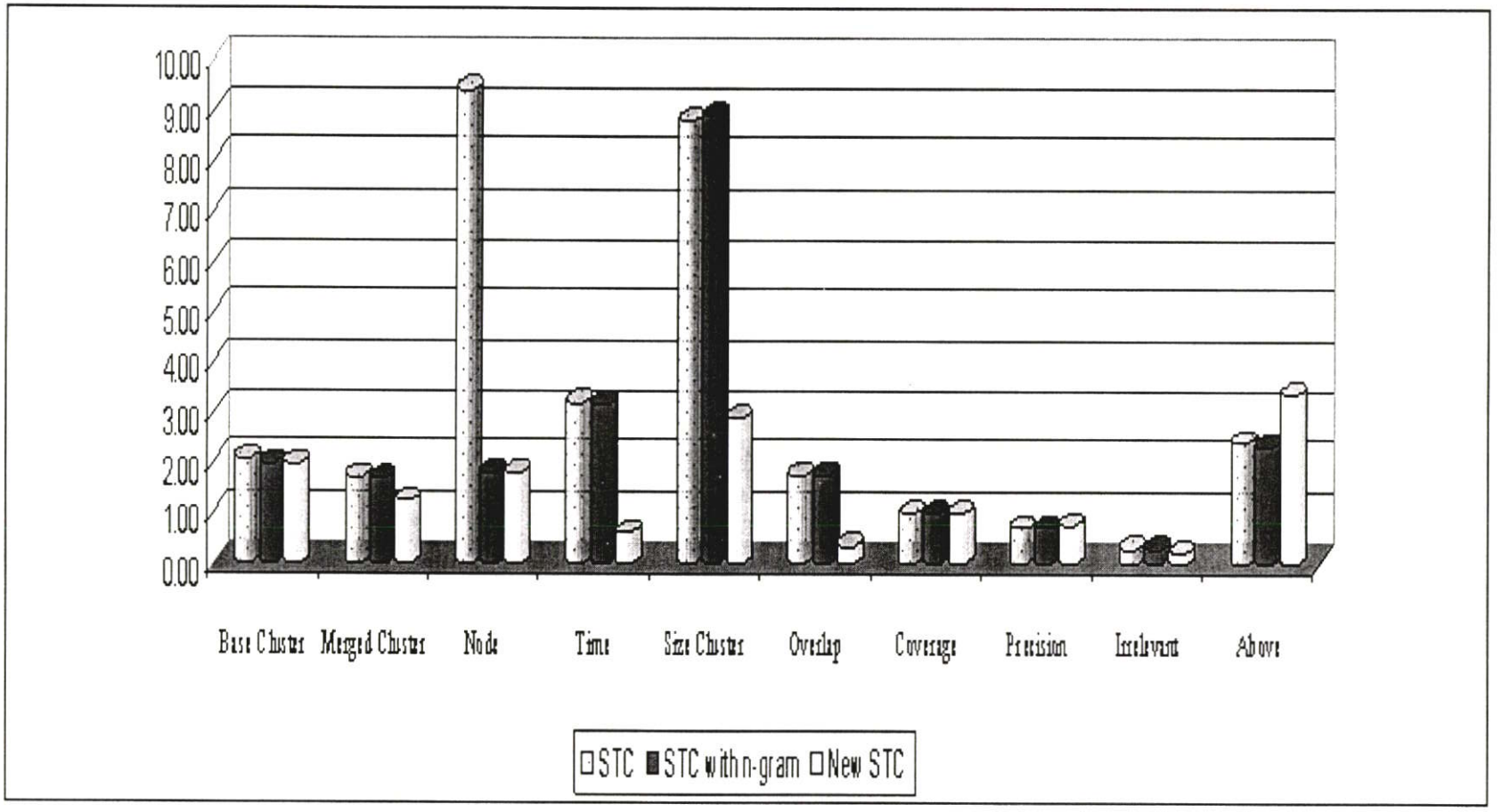
ตารางที่ 4.20 แสดงค่าเฉลี่ยผลการจัดกลุ่มของคำสืบค้นที่มีลักษณะเป็นคำทั่วๆไป

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	ALL	5	10	15	20	ALL	5	10	15	20	ALL
จำนวนกลุ่ม BASE CLUSTER	20,710.33					19,750.00					19,750.00				
จำนวนกลุ่ม MERGED CLUSTER	17,027.00					16,747.67					12,422.33				
จำนวน NODE	939,969.00					181,348.67					181,348.67				
เวลาที่ใช้ (TIME)	97 MIN. 53 SEC.					94 MIN. 17 SEC.					17 MIN. 35 SEC.				
ขนาดกลุ่ม (SIZE CLUSTER)	396.20	270.53	219.89	191.78	8.79	387.07	274.97	224.53	194.07	8.89	83.33	65.53	57.42	52.20	2.89
OVERLAP	0.19	0.27	0.35	0.37	17.64	0.15	0.27	0.34	0.36	17.53	0.05	0.10	0.13	0.14	3.46
COVERAGE	0.21	0.27	0.31	0.35	1.00	0.21	0.27	0.32	0.36	1.00	0.05	0.08	0.10	0.12	1.00
PRECISION	0.68	0.71	0.69	0.68	0.71	0.68	0.65	0.66	0.63	0.70	0.59	0.52	0.56	0.53	0.77
IRRELEVANT	0.32	0.29	0.31	0.32	0.29	0.32	0.35	0.34	0.37	0.30	0.41	0.48	0.44	0.47	0.23
ABOVE	2.31	2.90	2.48	2.47	2.41	2.25	1.89	2.01	1.72	2.34	2.99	1.23	1.46	1.21	3.37

ตารางที่ 4.21 แสดงค่า NORMALIZE ของค่าเฉลี่ยผลการจัดกลุ่มของคำสืบค้นที่มีลักษณะเป็นคำทั่วๆไป

ค่ามาตรฐาน	STC ORIGINAL	STC WITH N-GRAM	NEW STC	หมายเหตุ
จำนวนกลุ่มBASE CLUSTER	2.07	1.98	1.98	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10,000
จำนวนกลุ่ม MERGED CLUSTER	1.70	1.67	1.24	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10,000
จำนวนNODE	9.40	1.81	1.81	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 100,000
เวลาที่ใช้ (TIME)	3.22	3.16	0.61	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10
ขนาดกลุ่ม (SIZE CLUSTER)	8.79	8.89	2.89	
OVERLAP	1.76	1.75	0.35	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10
COVERAGE	1.00	1.00	1.00	
PRECISION	0.71	0.70	0.77	
IRRELEVANT	0.29	0.30	0.23	
ABOVE	2.41	2.34	3.37	

จากตารางที่ 4.21 ค่า Normalize ของค่าเฉลี่ยผลการจัดกลุ่มของคำสืบค้นที่มีลักษณะเป็นคำทั่วๆไป (General Concepts) สามารถสร้างกราฟได้ดังแสดงในรูปภาพที่ 4.6



รูปที่ 4.6 กราฟแสดงผลของชุดข้อมูลที่มีคำสืบค้นที่มีลักษณะเป็นคำทั่วๆไป

จากรูปที่ 4.6 อัตราความถูกต้องในการจัดกลุ่มของโมเดล New STC เท่ากับ 0.77 ซึ่งสูงกว่า STC_Original ที่มีค่าเท่ากับ 0.71 และ STC with n-gram ที่มีค่าเท่ากับ 0.70 ในลักษณะที่คำสืบค้นที่มีลักษณะเป็นคำทั่วๆไป (General Concepts)

2. คำสืบค้นที่มีลักษณะเป็นชื่อ (Entity Name) ประกอบด้วย

2.1 ชุดข้อมูลผลการสืบค้นของคำว่า "Thailand" เป็นคำสืบค้นในลักษณะเป็นชื่อ (Entity Name) จาก 6 กลุ่มเอกสาร ประกอบด้วย กลุ่ม Arts จำนวน 103 เอกสาร , กลุ่ม Business จำนวน 180 เอกสาร , กลุ่ม News จำนวน 40 เอกสาร , กลุ่ม Regional จำนวน 40 เอกสาร , กลุ่ม Society จำนวน 100 เอกสาร และกลุ่ม World จำนวน 120 เอกสาร รวมทั้งหมด 583 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.22

2.2 ชุดข้อมูลผลการสืบค้นของคำว่า "Iraq" เป็นคำสืบค้นในลักษณะเป็นชื่อ (Entity Name) จาก 4 กลุ่มเอกสาร ประกอบด้วย กลุ่ม Regional จำนวน 298 เอกสาร , กลุ่ม Science จำนวน 21 เอกสาร , กลุ่ม Society bจำนวน 255 เอกสาร และกลุ่ม World จำนวน 89 เอกสาร รวมทั้งหมด 663 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.23

2.3 ชุดข้อมูลผลการสืบค้นของคำว่า "Disney" เป็นคำสืบค้นในลักษณะเป็นชื่อ (Entity Name) จาก 8 กลุ่มเอกสาร ประกอบด้วย กลุ่ม Arts จำนวน 343 เอกสาร , กลุ่ม Business จำนวน 63 เอกสาร , กลุ่ม Kids and Teens จำนวน 161 เอกสาร , กลุ่ม News จำนวน 146 เอกสาร , กลุ่ม Regional จำนวน 200 เอกสาร , กลุ่ม Shopping จำนวน 106 เอกสาร , กลุ่ม Society จำนวน 106 เอกสาร และกลุ่ม World จำนวน 245 เอกสาร รวมทั้งหมด 1,318 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.24

จากชุดข้อมูลคำสืบค้นที่มีลักษณะเป็นชื่อ (Entity Name) ของทั้ง 3 คำสืบค้น นำมาคำนวณค่าเฉลี่ยได้ดังแสดงผลการทดลองในตารางที่ 4.25 , 4.26 และกราฟแสดงผลการทดลองเฉลี่ยในรูปที่ 4.7

ตารางที่ 4.22 ผลการจัดกลุ่มชุดข้อมูลผลการสืบค้นของคำว่า "THAILAND"

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	1,326	5	10	15	20	1,319	5	10	15	20	595
จำนวนกลุ่มBASE CLUSTER	1,676					1,626					1,626				
จำนวนกลุ่มMERGED CLUSTER	1,326					1,319					595				
จำนวน NODE	69,564					16,107					16,107				
เวลาที่ใช้ (TIME)	7 SEC.					7 SEC.					3 SEC.				
ขนาดกลุ่ม (SIZE CLUSTER)	26.2	18.7	15.8	14.7	5.22	28.6	20.7	16.93	14.7	5.23	12.6	9.9	8.73	7.9	2.45
OVERLAP	0.34	0.4	0.53	0.64	10.86	0.35	0.47	0.51	0.57	10.33	0.03	0.04	0.04	0.05	1.5
COVERAGE	0.17	0.23	0.27	0.31	1	0.18	0.24	0.29	0.32	1	0.1	0.16	0.21	0.26	1
PRECISION	0.89	0.9	0.88	0.86	0.8	0.89	0.89	0.85	0.81	0.8	0.77	0.72	0.7	0.69	0.9
IRRELEVANT	0.11	0.1	0.12	0.14	0.2	0.11	0.11	0.15	0.19	0.2	0.23	0.28	0.3	0.31	0.1
ABOVE	8.09	9.00	7.33	6.14	4.00	8.09	8.09	5.67	4.26	4.00	3.35	2.57	2.33	2.23	9.00

ตารางที่ 4.23 ผลการจัดกลุ่มชุดข้อมูลผลการสืบค้น ของคำว่า "IRAQ"

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	1,733	5	10	15	20	1,730	5	10	15	20	779
จำนวนกลุ่ม BASE CLUSTER	2,292					2,249					2,249				
จำนวนกลุ่ม MERGED CLUSTER	1,733					1,730					779				
จำนวน NODE	104,071					20,606					20,606				
เวลาที่ใช้ (TIME)	23 SEC.					22 SEC.					9 SEC.				
ขนาดกลุ่ม (SIZE CLUSTER)	33.8	25.8	22.93	19.55	5.77	33.8	27	23	19.2	5.77	15.8	13.4	11.73	10.35	2.54
OVERLAP	0.25	0.29	0.36	0.53	14.27	0.25	0.27	0.36	0.51	14.25	0.05	0.06	0.11	0.13	2.02
COVERAGE	0.2	0.3	0.38	0.39	0.99	0.2	0.32	0.38	0.39	0.99	0.11	0.19	0.28	0.28	0.99
PRECISION	0.92	0.81	0.76	0.8	0.72	0.92	0.78	0.73	0.77	0.72	0.81	0.73	0.71	0.69	0.79
IRRELEVANT	0.08	0.19	0.24	0.2	0.28	0.08	0.22	0.27	0.23	0.28	0.19	0.27	0.29	0.31	0.21
ABOVE	11.50	4.26	3.17	4.00	2.57	11.50	3.55	2.70	3.35	2.57	4.26	2.70	2.45	2.23	3.76

ตารางที่ 4.24 ผลการจัดกลุ่มชุดข้อมูลผลการสืบค้นของคำว่า "DISNEY"

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	2,813	5	10	15	20	2,790	5	10	15	20	1,495
จำนวนกลุ่ม BASE CLUSTER	3,853					3,711					3,711				
จำนวนกลุ่ม MERGED CLUSTER	2,813					2,790					1,495				
จำนวน NODE	135,569					31,400					31,400				
เวลาที่ใช้ (TIME)	34 SEC.					33 SEC.					15 SEC.				
ขนาดกลุ่ม (SIZE CLUSTER)	94.4	60.9	46.73	38.2	5.95	94.4	60.9	46.73	38.3	5.99	21.6	16.9	14.4	12.75	2.41
OVERLAP	0.38	0.42	0.41	0.47	11.69	0.38	0.42	0.41	0.48	11.65	0	0.03	0.06	0.09	1.73
COVERAGE	0.26	0.33	0.38	0.4	1	0.26	0.33	0.38	0.39	1	0.08	0.12	0.15	0.18	1
PRECISION	0.54	0.56	0.63	0.72	0.74	0.54	0.53	0.63	0.7	0.73	0.43	0.47	0.54	0.6	0.84
IRRELEVANT	0.46	0.44	0.37	0.28	0.26	0.46	0.47	0.37	0.3	0.27	0.57	0.53	0.46	0.4	0.16
ABOVE	1.17	1.27	1.70	2.57	2.85	1.17	1.13	1.70	2.33	2.70	0.75	0.89	1.17	1.50	5.25

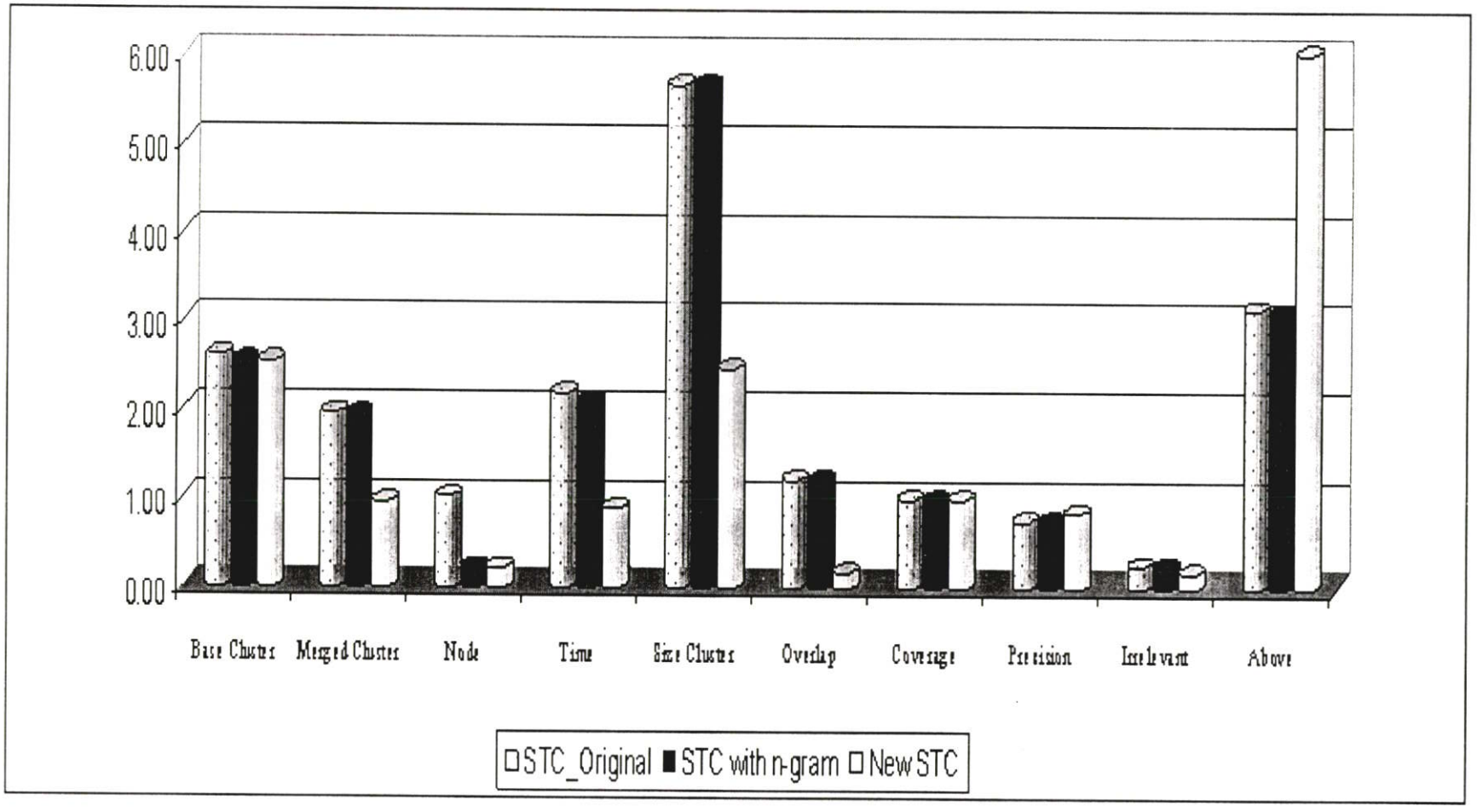
ตารางที่ 4.25 ค่าเฉลี่ยผลการจัดกลุ่มของคำสืบค้นที่มีลักษณะเป็นชื่อ (ENTITY NAME)

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	ALL	5	10	15	20	ALL	5	10	15	20	ALL
จำนวนกลุ่ม BASE CLUSTER	2,607.00					2,528.67					2,528.67				
จำนวนกลุ่ม MERGED CLUSTER	1,957.33					1,946.33					956.33				
จำนวน NODE	103,068.00					22,704.33					22,704.33				
เวลาที่ใช้ (TIME)	22 SEC.					21 SEC.					9 SEC.				
ขนาดกลุ่ม (SIZE CLUSTER)	51.47	35.13	28.49	24.15	5.65	52.27	36.20	28.89	24.07	5.66	16.67	13.40	11.62	10.33	2.47
OVERLAP	0.32	0.37	0.43	0.55	12.27	0.33	0.39	0.43	0.52	12.08	0.03	0.04	0.07	0.09	1.75
COVERAGE	0.21	0.29	0.34	0.37	1.00	0.21	0.30	0.35	0.37	1.00	0.10	0.16	0.21	0.24	1.00
PRECISION	0.78	0.76	0.76	0.79	0.75	0.78	0.73	0.74	0.76	0.75	0.67	0.64	0.65	0.66	0.84
IRRELEVANT	0.22	0.24	0.24	0.21	0.25	0.22	0.27	0.26	0.24	0.25	0.33	0.36	0.35	0.34	0.16
ABOVE	6.92	4.84	4.07	4.24	3.14	6.92	4.26	3.36	3.31	3.09	2.79	2.05	1.98	1.99	6.00

ตารางที่ 4.26 ค่า NORMALIZE ของค่าเฉลี่ยผลการจัดกลุ่มของคำสืบค้นที่มีลักษณะเป็นชื่อ

ค่ามาตรฐาน	STC ORIGINAL	STC WITH N-GRAM	NEW STC	หมายเหตุ
จำนวนกลุ่ม BASE CLUSTER	2.61	2.53	2.53	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 1,000
จำนวนกลุ่ม MERGED CLUSTER	1.96	1.95	0.96	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 1,000
จำนวน NODE	1.03	0.23	0.23	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 100,000
เวลาที่ใช้ (TIME)	2.2	2.1	0.9	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10
ขนาดกลุ่ม (SIZE CLUSTER)	5.65	5.66	2.47	
OVERLAP	1.23	1.21	0.18	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10
COVERAGE	1.00	1.00	1.00	
PRECISION	0.75	0.75	0.84	
IRRELEVANT	0.25	0.25	0.16	
ABOVE	3.14	3.09	6.00	

จากตารางที่ 4.26 ค่า Normalize ของค่าเฉลี่ยผลการจัดกลุ่มของคำสืบค้น ที่มีลักษณะเป็นชื่อ (Entity Name) สามารถสร้างกราฟได้ดังแสดงในรูปภาพที่ 4.7



รูปที่ 4.7 กราฟแสดงผลของชุดข้อมูลที่มีคำสืบค้นที่มีลักษณะเป็นชื่อ

จากรูปที่ 4.7 อัตราความถูกต้องในการจัดกลุ่มของโมเดล New STC เท่ากับ 0.84 ซึ่งสูงกว่า STC_Original และ STC with n-gram ที่มีค่าเท่ากับ 0.75 ในลักษณะที่คำสืบค้นที่มีลักษณะเป็นชื่อ (Entity Name)

3. คำสืบค้นที่มีลักษณะเป็นคำกำกวม (Ambiguous Queries) ประกอบด้วย

3.1 ชุดข้อมูลผลการสืบค้นของคำว่า "Matrix" เป็นคำสืบค้นในลักษณะกำกวม (Ambiguous Queries) จาก 7 กลุ่มเอกสาร ประกอบด้วย กลุ่ม Arts จำนวน 499 เอกสาร , กลุ่ม Business จำนวน 138 เอกสาร , กลุ่ม Computers จำนวน 174 เอกสาร , กลุ่ม Games จำนวน 54 เอกสาร , กลุ่ม Regional จำนวน 137 เอกสาร , กลุ่ม Science จำนวน 114 เอกสาร , และกลุ่ม World จำนวน 249 เอกสาร รวมทั้งหมด 1,365 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.27

3.2 ชุดข้อมูลผลการสืบค้นของคำว่า "Apple" เป็นคำสืบค้นในลักษณะกำกวม (Ambiguous Queries) จาก 8 กลุ่มเอกสาร ประกอบด้วย กลุ่ม Arts จำนวน 121 เอกสาร , กลุ่ม Business จำนวน 105 เอกสาร , กลุ่ม Computers จำนวน 774 เอกสาร , กลุ่ม Home จำนวน 340 เอกสาร , กลุ่ม News จำนวน 208 เอกสาร , กลุ่ม Recreation จำนวน 71 เอกสาร , กลุ่ม Shopping จำนวน 146 เอกสาร และกลุ่ม World จำนวน 553 เอกสาร รวมทั้งหมด 2,316 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.28

3.3 ชุดข้อมูลผลการสืบค้นของคำว่า "Jaguar" เป็นคำสืบค้นในลักษณะกำกวม (Ambiguous Queries) จาก 5 กลุ่มเอกสาร ประกอบด้วย กลุ่ม Games จำนวน 88 เอกสาร , กลุ่ม Recreation จำนวน 69 เอกสาร , กลุ่ม Regional จำนวน 292 เอกสาร , กลุ่ม Shopping จำนวน 61 เอกสาร และกลุ่ม World จำนวน 124 เอกสาร รวมทั้งหมด 633 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.29

จากชุดข้อมูลของทั้ง 3 คำสืบค้น นำมาคำนวณค่าเฉลี่ยได้ดังแสดงผลการทดลองในตารางที่ 4.30, 4.31 และกราฟแสดงผลการทดลองเฉลี่ยในรูปที่ 4.8

ตารางที่ 4.27 ผลการจัดกลุ่มของชุดข้อมูลสืบค้นคำว่า "MATRIX"

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	2,723	5	10	15	20	2,715	5	10	15	20	1,262
จำนวนกลุ่ม BASE CLUSTER	3,537					3,469					3,469				
จำนวนกลุ่ม MERGED CLUSTER	2,723					2,715					1,262				
จำนวน NODE	147,434					34,529					34,529				
เวลาที่ใช้ (TIME)	32					32					12				
ขนาดกลุ่ม (SIZE CLUSTER)	80.8	53.9	39.53	34.05	6.04	81.8	53.9	40.73	33.3	6.05	27.6	21.4	17.73	15.8	2.58
OVERLAP	0.19	0.54	0.64	0.64	11.05	0.19	0.54	0.57	0.62	11.03	0.06	0.06	0.09	0.11	1.39
COVERAGE	0.25	0.26	0.26	0.3	1	0.25	0.26	0.29	0.3	1	0.1	0.15	0.18	0.21	1
PRECISION	0.79	0.87	0.89	0.9	0.74	0.79	0.87	0.87	0.9	0.74	0.89	0.84	0.85	0.82	0.85
IRRELEVANT	0.21	0.13	0.11	0.1	0.26	0.21	0.13	0.13	0.1	0.26	0.11	0.16	0.15	0.18	0.15
ABOVE	3.76	6.69	8.09	9.00	2.85	3.76	6.69	6.69	9.00	2.85	8.09	5.25	5.67	4.56	5.67

ตารางที่ 4.28 ผลการจัดกลุ่มของชุดข้อมูลสืบค้นคำว่า "APPLE"

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	4,554	5	10	15	20	4,534	5	10	15	20	2,295
จำนวนกลุ่ม BASE CLUSTER	5,662					5,589					5,589				
จำนวนกลุ่มMERGED CLUSTER	4,554					4,534					2,295				
จำนวน NODE	231,813					54,369					54,369				
เวลาที่ใช้ (TIME)	1 MIN 26 SEC.					1 MIN 27 SEC.					26 SEC.				
ขนาดกลุ่ม (SIZE CLUSTER)	114	83.5	66.4	58.2	6.45	114	83.5	67.2	58.25	6.47	42.4	30.9	26.07	22.85	2.67
OVERLAP	0.27	0.3	0.4	0.4	11.68	0.27	0.3	0.4	0.39	11.67	0.01	0.03	0.03	0.03	1.64
COVERAGE	0.19	0.28	0.31	0.36	1	0.19	0.28	0.31	0.36	1	0.09	0.13	0.16	0.19	1
PRECISION	0.61	0.71	0.67	0.66	0.79	0.61	0.71	0.69	0.67	0.79	0.58	0.65	0.64	0.65	0.88
IRRELEVANT	0.39	0.29	0.33	0.34	0.21	0.39	0.29	0.31	0.33	0.21	0.42	0.35	0.36	0.35	0.12
ABOVE	1.56	2.45	2.03	1.94	3.76	1.56	2.45	2.23	2.03	3.76	1.38	1.86	1.78	1.86	7.33

ตารางที่ 4.29 ผลการจัดกลุ่มของชุดข้อมูลสืบค้นคำว่า "JAGUAR"

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	1,350	5	10	15	20	1,343	5	10	15	20	724
จำนวนกลุ่ม BASE CLUSTER	1,749					1,684					1,684				
จำนวนกลุ่ม MERGED CLUSTER	1,350					1,343					724				
จำนวน NODE	65,893					15,443					15,443				
เวลาที่ใช้ (TIME)	11					9					3				
ขนาดกลุ่ม (SIZE CLUSTER)	45.6	28.8	25.13	21.95	6.01	45.6	32.2	26.33	22.2	6.02	23	17.4	14.2	12.25	2.54
OVERLAP	0.21	0.39	0.42	0.51	11.81	0.21	0.33	0.43	0.5	11.78	0.49	0.33	0.35	0.35	1.91
COVERAGE	0.3	0.33	0.42	0.46	1	0.3	0.38	0.44	0.47	1	0.12	0.21	0.25	0.29	1
PRECISION	0.86	0.87	0.85	0.83	0.79	0.86	0.81	0.83	0.81	0.79	0.91	0.83	0.8	0.79	0.85
IRRELEVANT	0.14	0.13	0.15	0.17	0.21	0.14	0.19	0.17	0.19	0.21	0.09	0.17	0.2	0.21	0.15
ABOVE	6.14	6.69	5.67	4.88	3.76	6.14	4.26	4.88	4.26	3.76	10.11	4.88	4.00	3.76	5.67

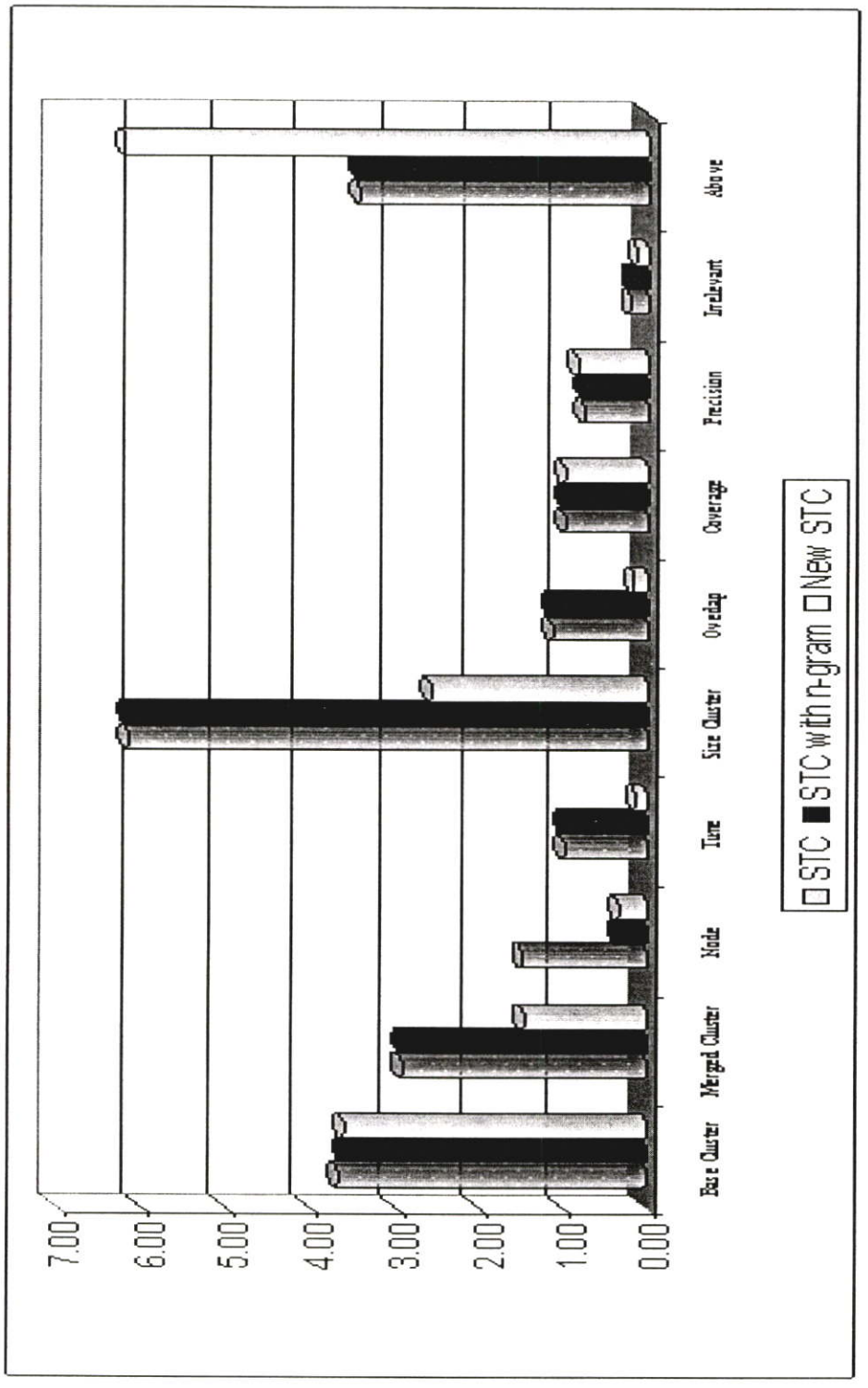
ตารางที่ 4.30 ค่าเฉลี่ยผลการจัดกลุ่มของคำสืบค้นที่มีลักษณะเป็นคำกำกวม

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	ALL	5	10	15	20	ALL	5	10	15	20	ALL
จำนวนกลุ่ม BASE CLUSTER	3,649.33					3,580.67					3,580.67				
จำนวนกลุ่ม MERGED CLUSTER	2,875.67					2,864.00					1,427.00				
จำนวน NODE	148,380.00					34,780.33					34,780.33				
เวลาที่ใช้ (TIME)	1 MIN.					1 MIN.					14 SEC.				
ขนาดกลุ่ม (SIZE CLUSTER)	80.13	55.40	43.69	38.07	6.17	80.47	56.53	44.75	37.92	6.18	31.00	23.23	19.33	16.97	2.60
OVERLAP	0.22	0.41	0.49	0.52	11.51	0.22	0.39	0.47	0.50	11.49	0.19	0.14	0.16	0.16	1.65
COVERAGE	0.25	0.29	0.33	0.37	1.00	0.25	0.31	0.35	0.38	1.00	0.10	0.16	0.20	0.23	1.00
PRECISION	0.75	0.82	0.80	0.80	0.77	0.75	0.80	0.80	0.79	0.77	0.79	0.77	0.76	0.75	0.86
IRRELEVANT	0.25	0.18	0.20	0.20	0.23	0.25	0.20	0.20	0.21	0.23	0.21	0.23	0.24	0.25	0.14
ABOVE	3.82	5.28	5.26	5.27	3.46	3.82	4.47	4.60	5.10	3.46	6.53	4.00	3.82	3.39	6.22

ตารางที่ 4.31 ค่า NORMALIZE ของค่าเฉลี่ยผลการจัดกลุ่มของคำสืบค้นที่มีลักษณะเป็นคำกำกวม

ค่ามาตรฐาน	STC ORIGINAL	STC WITH N-GRAM	NEW STC	หมายเหตุ
จำนวนกลุ่ม BASE CLUSTER	3.65	3.58	3.58	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 1,000
จำนวนกลุ่ม MERGED CLUSTER	2.88	2.86	1.43	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 1,000
จำนวน NODE	1.48	0.35	0.35	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 100,000
เวลาที่ใช้ (TIME)	1.00	1.00	0.14	
ขนาดกลุ่ม (SIZE CLUSTER)	6.17	6.18	2.60	
OVERLAP	1.15	1.15	0.17	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลจึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10
COVERAGE	1.00	1.00	1.00	
PRECISION	0.77	0.77	0.86	
IRRELEVANT	0.23	0.23	0.14	
ABOVE	3.46	3.46	6.22	

จากตารางที่ 4.31 ค่า Normalize ของค่าเฉลี่ยผลการจัดกลุ่มของคำสืบค้น ที่มีลักษณะเป็นชื่อ (Entity Name) สามารถสร้างกราฟได้ดังแสดงในรูปภาพที่ 4.8



รูปที่ 4.8 กราฟแสดงผลของชุดข้อมูลทดสอบสี่ชุดในหัวข้อลักษณะเป็นคำกำกวม

จากรูปที่ 4.8 อัตราความถูกต้องของ New STC สูงกว่า STC_Original และ STC with n-gram และจากผลการจัดกลุ่มผลการสืบค้นของทั้ง 3 ลักษณะของคำสืบค้น ทั้ง 9 คำสืบค้น สามารถสรุปผลการจัดกลุ่มผลการสืบค้นได้ดังแสดงผลค่าเฉลี่ยในตารางที่ 4.32 และ Normalize ค่าเฉลี่ย ดังแสดงผลในตารางที่ 4.33 เพื่อให้ง่ายต่อการสร้างกราฟซึ่งแสดงผลในรูปภาพที่4.9

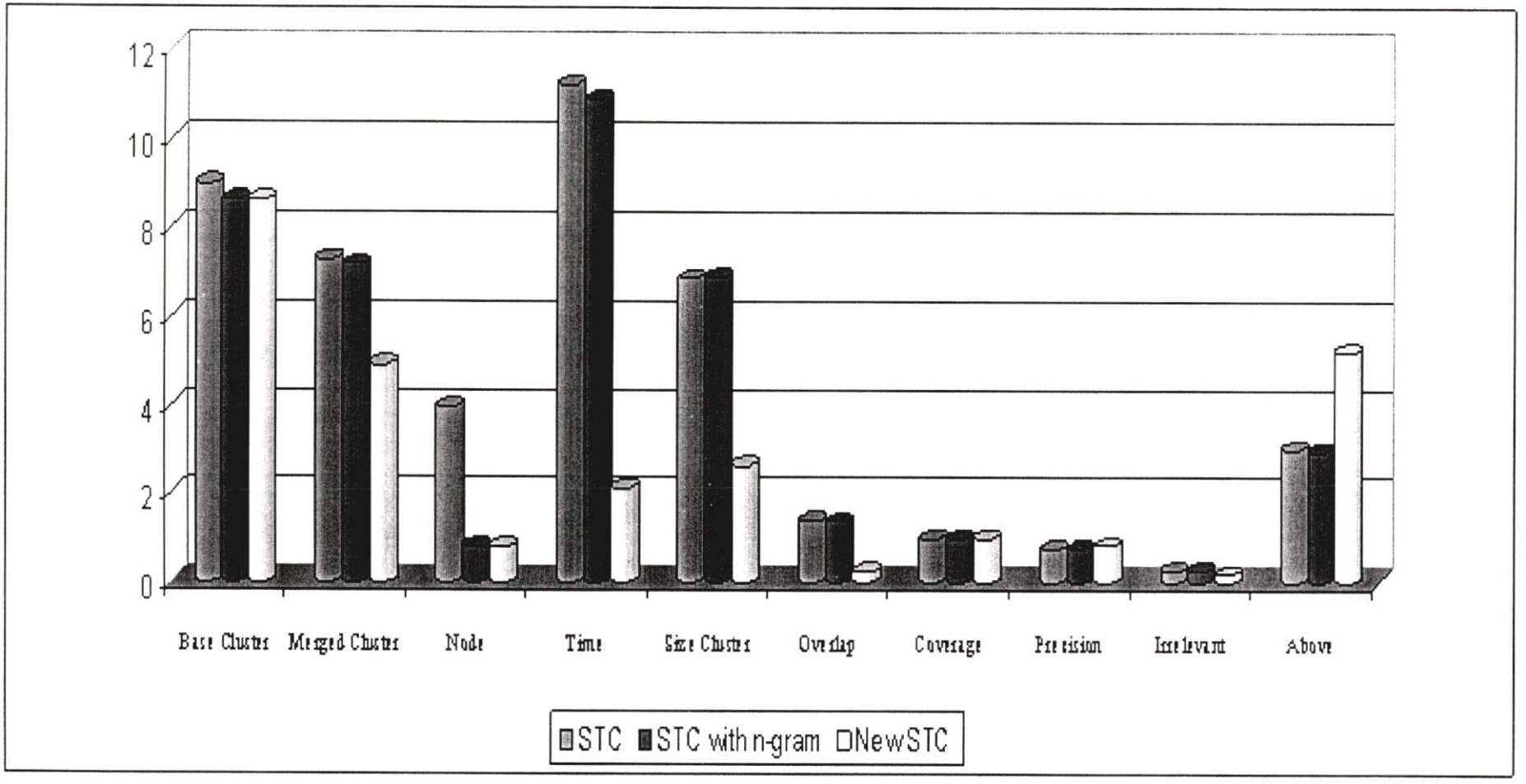
ตารางที่ 4.32 ค่าเฉลี่ยผลการทดลองชุดผลการสืบค้นด้วยคำสืบค้นภายใน DMOZ.COM

ค่ามาตรฐาน	AVERAGE														
	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	ALL	5	10	15	20	ALL	5	10	15	20	ALL
จำนวนกลุ่ม BASE CLUSTER	8,989					8,620					8,620				
จำนวนกลุ่ม MERGED CLUSTER	7,287					7,186					4,935				
จำนวน NODE	397,139					79,611					79,611				
เวลาที่ใช้ (TIME)	(101/9) 11.22					(98/9) 10.88					(19/9) 2.11				
ขนาดกลุ่ม (SIZE CLUSTER)	175.93	120.36	97.35	84.67	6.87	173.27	122.57	99.39	85.35	6.91	43.67	34.06	29.46	26.50	2.65
OVERLAP	0.25	0.35	0.42	0.48	13.81	0.23	0.35	0.41	0.46	13.70	0.09	0.09	0.12	0.13	2.28
COVERAGE	0.22	0.28	0.33	0.36	1.00	0.22	0.29	0.34	0.37	1.00	0.08	0.13	0.17	0.20	1.00
PRECISION	0.74	0.76	0.75	0.76	0.74	0.74	0.73	0.73	0.73	0.74	0.68	0.64	0.66	0.65	0.82
IRRELEVANT	0.26	0.24	0.25	0.24	0.26	0.26	0.27	0.27	0.27	0.26	0.32	0.36	0.34	0.35	0.18
ABOVE	4.35	4.34	3.94	3.99	3.00	4.33	3.54	3.32	3.38	2.96	4.10	2.43	2.42	2.19	5.20

จากตารางที่ 4.33 การ Normalize ค่าเฉลี่ยผลการทำงานของ 3 โมเดล ของชุดข้อมูลผล

ตารางที่ 4.33 ค่า NORMALIZE ค่าเฉลี่ยของชุดข้อมูลผลการสืบค้นด้วยคำสืบค้นภายใน DMOZ.COM

ค่ามาตรฐาน	STC ORIGINAL	STC WITH N-GRAM	NEW STC	หมายเหตุ
จำนวนกลุ่ม BASE CLUSTER	8.99	8.62	8.62	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 1,000
จำนวนกลุ่ม MERGED CLUSTER	7.29	7.19	4.94	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 1,000
จำนวน NODE	3.97	0.80	0.80	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 100,000
เวลาที่ใช้ (TIME)	11.22	10.88	2.11	
ขนาดกลุ่ม (SIZE CLUSTER)	6.87	6.91	2.65	
OVERLAP	1.38	1.37	0.23	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10
COVERAGE	1	1	1	
PRECISION	0.74	0.74	0.82	
IRRELEVANT	0.26	0.26	0.18	
ABOVE	3.00	2.96	5.20	



รูปที่ 4.9 กราฟแสดงผลของชุดข้อมูลผลการสืบค้นด้วยคำสืบค้นภายใน DMOZ.COM

จากรูปที่ 4.9 คือกราฟสรุปผลการทำงานของทั้ง 3 โมเดล คือ STC_Original , STC with n-gram และ New STC ของชุดข้อมูลผลการสืบค้นด้วยคำสืบค้นภายใน Dmoz.com เพื่อศึกษาความสามารถด้านต่างๆของโมเดล ด้วยตัวชี้วัดมาตรฐานต่างๆที่ใช้วัดประสิทธิภาพของโมเดลดังที่กล่าวในข้างต้น และศึกษาความสามารถในการจัดกลุ่มเอกสารหรืออำนาจจำแนกกลุ่มเอกสาร ที่เอกสารภายในกลุ่มผลการสืบค้นด้วยคำสืบค้นภายใน Dmoz.com ดังแสดงรายละเอียดของผลการทำงานดังต่อไปนี้

1. จำนวนกลุ่มพื้นฐาน (Base Cluster) ผลการทดลองพบว่า จำนวนกลุ่มพื้นฐานของโมเดล STC with n-gram และ New STC ซึ่งใช้เทคนิค n-gram มาช่วยในการสร้าง Suffix Tree ทำให้มีจำนวนกลุ่มพื้นฐานเท่ากับ 8,620 กลุ่ม ซึ่งต่ำกว่าโมเดล STC_Original ซึ่งมีจำนวนกลุ่มพื้นฐานเท่ากับ 8,989 กลุ่ม

2. จำนวนกลุ่ม Merged Cluster คือกลุ่มที่ได้รับการปรับแต่งตามรูปแบบการรวมกลุ่มพื้นฐานของแต่ละโมเดล ผลการทดลองพบว่าโมเดลของ New STC มีจำนวนกลุ่ม Merged Cluster น้อยที่สุด คือ 4,935 กลุ่ม และการลดลงของจำนวนกลุ่มคิดเป็นร้อยละ 32.28 เมื่อเปรียบเทียบกับ STC_Original ซึ่งมี 7,287 กลุ่ม และการลดลงของจำนวนกลุ่มคิดเป็นร้อยละ 31.33 เมื่อเปรียบเทียบกับ STC with n-gram ซึ่งมี 7,186 กลุ่ม

3. จำนวนโหนด (Node) ภายในซาฟฟิกทรี (Suffix Tree) ผลการทดลองพบว่าโมเดลของ STC with n-gram และ New STC มีจำนวนโหนดเท่ากับ 79,611 โหนด ซึ่งลดลงจากจำนวนโหนดของโมเดล STC_Original ซึ่งมีจำนวน 397,139 โหนด และอัตราการลดลงของโหนดคิดเป็นร้อยละ 79.95 ทำให้ความต้องการในการใช้งานหน่วยความจำลดลง

4. เวลา (Time) ที่ใช้ในการทำงาน ผลการทดลองพบว่าโมเดลของ New STC ใช้เวลาในการทำงานน้อยที่สุด คือเฉลี่ยเท่ากับ 2 นาที 22 วินาที เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งใช้เวลาในการทำงานเฉลี่ยเท่ากับ 11 นาที 3 วินาที และ 11 นาที ตามลำดับ ขนาดกลุ่ม (Size Cluster) คือจำนวนเอกสารภายในกลุ่ม ผลการทดลองพบว่าจำนวนเอกสารภายในกลุ่มของโมเดล New STC เฉลี่ยกลุ่มละ 2.65 เอกสาร ซึ่งมีขนาดเล็กที่สุด เมื่อเปรียบเทียบกับ STC_Original ที่มีขนาดกลุ่มเฉลี่ยกลุ่มละ 6.87 ซึ่งการลดลงของขนาดกลุ่มเอกสารคิดเป็นร้อยละ 61.43 และ STC with n-gram ที่มีขนาดกลุ่มเฉลี่ยกลุ่มละ 6.91 เอกสาร ซึ่งการลดลงของขนาดกลุ่มเอกสารคิดเป็นร้อยละ 61.65

6. อัตราการทับซ้อน(Overlap) ของกลุ่มเอกสารที่ถูกปรับแต่งด้วยโมเดลของ New STC มีค่า เฉลี่ยเท่ากับ 2.28 ซึ่งมีอัตราการทับซ้อนน้อยที่สุด เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยอัตราการทับซ้อนเท่ากับ 13.81 และ 13.70 ตามลำดับ ซึ่งการลดลงของอัตราการทับซ้อนของเอกสารคิดเป็นร้อยละ 83.49 และ 83.36

7. อัตราการครอบคลุม (Coverage) ในการจัดกลุ่มเอกสารของทั้ง 3 โมเดล มีค่าเฉลี่ยเท่ากับ 1 เท่ากัน แสดงให้เห็นว่าโมเดลของ STC_Original , STC with n-gram และ New STC สามารถจัดกลุ่มเอกสารได้ทุกเอกสาร แต่ผลการทดลองในส่วนของ " Iraq " มีค่า Coverage เท่ากับ 0.99 สาเหตุเนื่องมาจากมีหนึ่งเอกสารที่ไม่มีคำซ้ำกับเอกสารใดเลย ทำให้ไม่สามารถจัดกลุ่มให้กับเอกสารดังกล่าวได้

8. อัตราความถูกต้อง (Precision) ผลการทดลองพบว่าโมเดลของ New STC มีอัตราความถูกต้องในการจัดกลุ่มสูงที่สุดคือ 0.82 เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยอัตราการความถูกต้องเท่ากับ 0.74

9. อัตราเอกสารที่ไม่ตรงกับความต้องการ (Irrelevant Documents) ภายในกลุ่ม ผลการทดลองพบว่าโมเดลของ New STC มีจำนวนเอกสารที่ไม่ตรงกับความต้องการหรือเอกสารปนเปื้อนน้อยที่สุดคือ 0.18 เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยอัตราเอกสารที่ไม่ตรงกับความต้องการเท่ากับ 0.26

10. ระยะห่างระหว่างเอกสารที่ตรงกับความต้องการและไม่ตรงกับความต้องการ (Above) ผลการทดลองพบว่าโมเดลของ New STC มีระยะห่างสูงที่สุดคือ 5.2 เมื่อเปรียบเทียบกับ STC_Original และ STC with n-gram ซึ่งมีค่าเฉลี่ยระยะห่างระหว่างเอกสารที่ตรงกับความต้องการและไม่ตรงกับความต้องการเท่ากับ 3 และ 2.96 ตามลำดับ

จากผลการทดลองของข้อมูลชุดผลการสืบค้นภายในกลุ่มของ Dmoz.com ทั้ง 9 คำสืบค้น ผลการทดลองพบว่า

1. โมเดลของ New STC สามารถจัดกลุ่มได้ถูกต้องมากกว่าโมเดลของ STC_Original และ STC with n-gram ส่งผลให้จำนวนเอกสารที่ไม่ตรงกับความต้องการภายในกลุ่มลดลง และช่วงระยะห่างของเอกสารที่ตรงกับความต้องการและไม่ตรงกับความต้องการสูงขึ้น

2. โมเดล New STC ใช้เวลาน้อยกว่าโมเดล STC_Original และ STC with n-gram

3. โมเดล New STC ใช้หน่วยความจำ (Space) น้อยกว่าโมเดล STC_Original และ STC with n-gram

4. โมเดล New STC มีอัตราการทับซ้อน (Overlap) ของเอกสารน้อยกว่าโมเดล STC_Original และ STC with n-gram

5. อัตราการครอบคลุม (Coverage) ในการจัดกลุ่มเอกสารของทั้ง 3 โมเดล มีค่าเป็น 1 เท่ากัน แสดงให้เห็นว่าโมเดลของ STC_Original , STC with n-gram และ New STC สามารถจัดกลุ่มเอกสารได้ทุกเอกสาร

6. ขนาดของกลุ่มที่จัดกลุ่มด้วยโมเดลของ New STC มีขนาดเล็กกว่าขนาดกลุ่มของโมเดล STC_Original และ STC with n-gram

4.3.3 ชุดข้อมูลจาก Ohsumed Collection

ขั้นตอนการทดลอง แบ่งออกเป็น 4 ขั้นตอนคือ

1. ทำการสุ่มเลือกเอกสารจากกลุ่มเอกสารภายใน Ohsumed Collection โดยกำหนดให้แต่ละกลุ่มต้องไม่มีเอกสารทับซ้อนกัน จำนวน 22 กลุ่ม รวม 4,380 เอกสาร เพื่อนำมาใช้ในการทดลองประสิทธิภาพด้านการจัดกลุ่มเอกสารที่คำมีจำนวนมากกว่า 50 คำของโมเดล New STC , STC_Original และ STC with n-gram

2. นำภายในบทคัดย่อของแต่ละเอกสาร ซึ่งใช้เป็นตัวแทนเอกสาร มาผ่านกระบวนการทำงานด้าน pre-processing ด้วยการทำ Stop-words โดยการกำจัดคำที่ไม่มีความหมาย เพื่อให้จำนวนคำลดลง และ การทำ Stemming words โดยการทำให้คำอยู่ในรากศัพท์เดิม เพื่อลดความหลากหลายของคำที่จะใช้ในการสร้างซัฟฟิกทรี (Suffix Tree)

3. นำคำที่เหลือจากการทำงานในขั้นตอนการทำ pre-processing เข้าสู่กระบวนการทำงานของโมเดล และคำนวณหาค่าต่างๆของตัวชี้วัดประสิทธิภาพของโมเดล

4. ศึกษาค่าต่างๆของตัวชี้วัด โดยการแบ่งการศึกษาค่ามาตรฐานต่างๆออกเป็นช่วงๆ ของผลการจัดกลุ่ม ประกอบด้วย ค่ามาตรฐานในช่วง 5 กลุ่มแรกของผลการจัดกลุ่ม , ช่วง 10 กลุ่มแรกของผลการจัดกลุ่ม , ช่วง 15 กลุ่มแรกของผลการจัดกลุ่ม , ช่วง 20 กลุ่มแรกของผลการจัดกลุ่ม และภาพรวมทั้งหมดของผลการจัดกลุ่มในแต่และชุดข้อมูล ดังแสดงในตารางผลการทดลอง

จากการสุ่มเลือกเอกสารจากกลุ่มเอกสารภายใน Ohsumed Collection โดยกำหนดให้แต่ละกลุ่มต้องไม่มีเอกสารทับซ้อนกัน จำนวน 22 กลุ่ม รวม 4,380 เอกสาร ดังแสดงผลการทดลองในตารางที่ 4.34 , 4.35 และรูปที่ 4.10

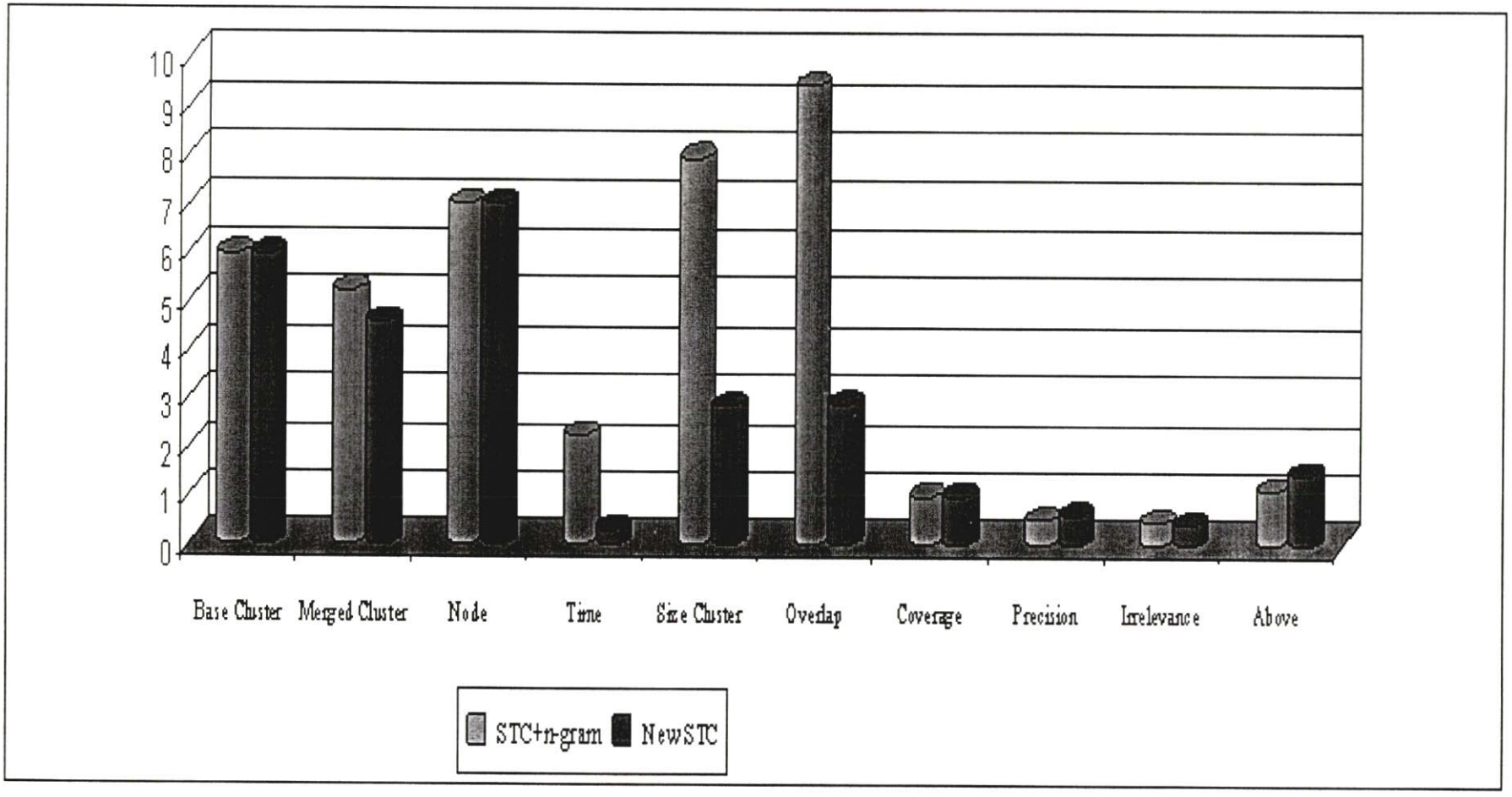
ตารางที่ 4.34 แสดงผลการจัดกลุ่มของชุดข้อมูลจาก OHSUMED COLLECTION

ค่ามาตรฐาน	STC ORIGINAL					STC WITH N-GRAM					NEW STC				
	5	10	15	20	ALL	5	10	15	20	ALL	5	10	15	20	ALL
จำนวนกลุ่ม BASE CLUSTER	ระบบไม่สามารถสร้าง SUFFIX TREE ได้ เนื่องจากโมเดลของ STC ORIGINAL ใช้หน่วยความจำมากทำให้หน่วยความจำไม่เพียงพอต่อการทำงาน					57,996					57,996				
จำนวนกลุ่ม MERGED CLUSTER						52,430					45,824				
จำนวน NODE						689,286					689,286				
เวลาที่ใช้ (TIME)						224 MIN. 32 SEC.					30 MIN. 32 SEC.				
ขนาดกลุ่ม						227	195.5	182.33	170.5	7.97	91.40	74.70	67.27	61.80	2.87
OVERLAP						0.14	0.25	0.39	0.49	94.38	0.08	0.11	0.15	0.17	29.05
COVERAGE						0.23	0.36	0.45	0.52	1	0.10	0.15	0.20	0.24	1.00
PRECISION						0.35	0.39	0.36	0.32	0.53	0.24	0.21	0.21	0.21	0.59
IRRELEVANCE						0.65	0.61	0.64	0.68	0.47	0.76	0.79	0.79	0.79	0.41
ABOVE						0.54	0.64	0.56	0.47	1.13	0.32	0.27	0.27	0.27	1.44

ตารางที่ 4.35 ค่า NORMALIZE ผลการทำงานของชุดข้อมูลจาก OHSUMED COLLECTION

ค่ามาตรฐาน	STC ORIGINAL	STC WITH N-GRAM	NEW STC	NORMALIZE STC WITH N-GRAM	NORMALIZE NEW STC	หมายเหตุ
จำนวนกลุ่ม BASE CLUSTER	-	57,996	57,996	6	6	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10,000
จำนวนกลุ่มMERGED CLUSTER	-	52,430	45,824	5.24	4.58	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10,000
จำนวน NODE	-	689,286	689,286	7	7	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10,000
เวลาที่ใช้ (TIME)	-	224 MIN. 32 SEC.	30 MIN. 32 SEC.	2.24	0.30	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 100
ขนาดกลุ่ม	-	7.97	2.87	7.97	2.87	
OVERLAP	-	94.38	29.05	9.44	2.91	เพื่อความสะดวกในการสร้างกราฟ ข้อมูลนี้จึงถูกสร้างให้อยู่ในช่วง 0-10 ด้วยการหาร 10
COVERAGE	-	1	1.00	1	1	
PRECISION	-	0.53	0.59	0.53	0.59	
IRRELEVANCE	-	0.47	0.41	0.47	0.41	
ABOVE	-	1.13	1.44	1.13	1.44	

จากตารางที่ 4.35 การ Normalize ค่าเฉลี่ยผลการทำงานของ 3 โมเดล ของชุดข้อมูลจาก Ohsumed Collection สามารถสร้างกราฟได้ดังรูปภาพที่ 4.10



รูปที่ 4.10 กราฟแสดงผลการทำงาน 3 โมเดล ของชุดข้อมูลจาก OHSUMED COLLECTION

จากรูปที่ 4.10 คือกราฟสรุปผลการทำงานของทั้ง 3 โมเดล คือ STC_Original , STC with n-gram และ New STC ของชุดข้อมูลจาก Ohsumed Collection เพื่อศึกษาความสามารถด้านต่างๆของโมเดล ด้วยตัวชี้วัดมาตรฐานต่างๆที่ใช้วัดประสิทธิภาพของโมเดลดังที่กล่าวในข้างต้น และศึกษาความสามารถในการจัดกลุ่มเอกสารหรืออำนาจจำแนกกลุ่มเอกสาร ที่เอกสารภายในกลุ่มประกอบด้วยคำมากกว่า 50 คำ จากผลการทดลองพบว่า โมเดลของ STC_Original ไม่สามารถสร้างกลุ่มพื้นฐานได้ เพราะไม่สามารถสร้าง Suffix Tree ได้สำเร็จ เนื่องจากโมเดลของ STC_Original ต้องใช้หน่วยความจำ (Space) จำนวนมาก เมื่อคำภายในเอกสารมีจำนวนมาก เพราะการสร้าง Suffix Tree หนึ่งคำเท่ากับหนึ่งโหนด (Node) ใน Suffix Tree ผลการทดลองในชุดข้อมูลนี้จึงเปรียบเทียบกับโมเดล New STC และ STC with n-gram ผลทดลองพบว่าโมเดล New STC ได้รับกลุ่มพื้นฐานเท่ากับ STC with n-gram เพราะมีจำนวนโหนดภายใน Suffix Tree เท่ากัน แต่ New STC ได้รับกลุ่ม Merged Cluster น้อยกว่า ในเวลาที่เร็วกว่า เพราะขนาดของกลุ่ม คือจำนวนเอกสารภายในกลุ่มมีจำนวนน้อยกว่า ทำให้อัตราการทับซ้อนของเอกสารน้อยกว่าด้วยเช่นกัน ความสามารถในการจัดกลุ่มเอกสารได้ทุกเอกสารหรือความคลอบคลุมในการจัดกลุ่มของทั้งสองโมเดลมีค่าเท่ากันคือสามารถจัดกลุ่มเอกสารได้ทุกเอกสาร แต่อัตราความถูกต้องในการจัดกลุ่มของโมเดล New STC สูงกว่า STC with n-gram ส่งผลให้จำนวนเอกสารที่ไม่ตรงกับความต้องการหรือเอกสารปนเปื้อนต่ำกว่า และระยะห่างระหว่างเอกสารที่ตรงกับความต้องการกับเอกสารที่ไม่ตรงกับความต้องการของโมเดล New STC สูงกว่า STC with n-gram

4.3 สรุปผลการทดลอง

จากการทดลองทดสอบการทำงานของโมเดล New STC และ เปรียบเทียบผลการทำงานกับ 2 โมเดลคือ STC_Original และ STC with n-gram ด้วยชุดข้อมูล 3 ชุดข้อมูล ผลการทดลองพบว่า โมเดล New STC ทำงานได้จำนวนกลุ่ม (Merged Cluster) น้อยกว่า ในเวลาที่เร็ว (Time) กว่า และใช้หน่วยความจำ (Space) น้อยกว่า ส่งผลให้ผลความถูกต้องในการจัดกลุ่มสูงกว่า แต่เมื่อเปรียบเทียบในกลุ่มลำดับต้นๆ จะพบว่าอัตราการความคลอบคลุม (Coverage) ในการจัดกลุ่มเอกสารของโมเดล New STC มีค่าน้อยกว่า STC_Original และ STC with n-gram เพราะขนาดของกลุ่มเอกสารภายในโมเดล New STC มีค่าน้อยกว่า คือมีจำนวนสมาชิกภายในกลุ่มน้อยกว่า เนื่องจากกลุ่มของโมเดล New STC มีลักษณะเป็นกลุ่มที่เฉพาะเจาะจงมากกว่ากลุ่มของโมเดล STC_Original และ STC with n-gram เพราะต้องการให้เอกสารปรากฏในกลุ่มที่เหมาะสมที่สุด โดยพิจารณาจากคำที่ปรากฏภายในเอกสาร

ในบทที่ 5 จะกล่าวถึงผลสรุปของงานวิจัยและข้อเสนอแนะ ซึ่งจะเป็นประโยชน์ในการพัฒนาการทำวิจัยต่อไป

บทที่ 5
สรุปผลการวิจัย และข้อเสนอแนะ

เนื่องจากการทดลองของงานวิจัยนี้มุ่งเน้นที่จะสร้างกลุ่มที่มีขนาดเล็ก ซึ่งมีลักษณะเฉพาะเจาะจง ตามป้ายชื่อกลุ่ม เพื่ออำนวยความสะดวกในการมองเห็นและเข้าถึงผลการสืบค้นที่ตรงกับความต้องการของผู้สืบค้นอย่างแท้จริง ซึ่งจะส่งผลให้อัตราการทับซ้อนของเอกสาร (Overlap) ลดลง และเพิ่มอัตราความถูกต้อง (Precision) ของการจัดกลุ่มผลการสืบค้นให้สูงขึ้น แต่ยังคงประสิทธิภาพด้านความคลอบคลุมการจัดกลุ่ม (Coverage) ในการทดลองใช้ชุดข้อมูลจำนวน 3 ชุด และเปรียบเทียบผลการทำงานของทั้ง 3 อัลกอริทึม คือ STC _Original , STC with n-gram และ New STC ซึ่งนำมาสรุปผลการวิจัยประกอบด้วยรายละเอียดดังต่อไปนี้

5.1 สรุปผลการวิจัย

เนื่องจากสมมุติฐานของงานวิจัยนี้ เริ่มต้นจากการมองข้อด้อยของอัลกอริทึม STC ซึ่งก็คือกลุ่มที่ได้จากการทำงานของ STC มีลักษณะไม่เฉพาะเจาะจง และ มีขนาดใหญ่ ทำให้อัตราการทับซ้อนของเอกสารสูงมากเกินไป ส่งผลให้อัตราความถูกต้องในการจัดกลุ่มลดลง สิ้นเปลืองเวลา (Time) และหน่วยความจำ (Space) ในการประมวลผล และป้ายชื่อกลุ่มขาดความสมบูรณ์เนื่องจากถูกตัดด้วยขนาดของ n-gram ดังนั้นงานวิจัยนี้จึงปรับปรุงอัลกอริทึม STC เพื่อลดข้อด้อยดังที่กล่าวมา และสามารถสรุปแต่ละประเด็น ดังรายละเอียดจากผลการทดลองที่ได้จากชุดข้อมูลทั้ง 3 ชุดข้อมูล โดยเปรียบเทียบกับอัลกอริทึมของ STC รูปแบบเดิม (STC_Original) และ STC ที่ทำงานร่วมกับเทคนิค n-gram (STC with n-gram) ดังต่อไปนี้

1. ประสิทธิภาพในการจัดกลุ่มผลการสืบค้น

จากผลการทดลองพบว่า การจัดกลุ่มด้วยแนวทางของ A New Suffix Tree Clustering มีอัตราความถูกต้องสูงกว่า STC_Original และ STC with n-gram เนื่องจากกลุ่มของ New STC คือกลุ่มเอกสารที่มีป้ายชื่อกลุ่มปรากฏในทุกเอกสารที่เป็นสมาชิกภายในกลุ่ม แต่ STC_Original และ STC with n-gram เอกสารที่เป็นสมาชิกภายในกลุ่มบางครั้งอาจไม่มีป้ายชื่อกลุ่มปรากฏในเอกสาร และขนาดกลุ่มของอัลกอริทึม New STC เล็กกว่าขนาดกลุ่มของ STC _Original และ STC with n-gram ทั้งนี้เพราะอัลกอริทึม New STC มีการปรับแต่งกลุ่มในลักษณะที่ต้องการให้กลุ่มมีขนาดเล็ก มีความเฉพาะเจาะจงมากกว่ากลุ่มของ STC _Original โดยการเลือกให้เอกสารปรากฏภายในกลุ่มที่มีความเหมาะสมมากที่สุด ซึ่งพิจารณาความเหมาะสมจากความยาวของป้ายชื่อกลุ่มเป็นสำคัญ เพราะป้ายชื่อกลุ่มที่มีความยาวมากกว่าย่อมให้ข้อมูลที่มากกว่ากลุ่มที่มีป้ายชื่อกลุ่มเป็นคำเดี่ยว ส่งผลให้ในภาพรวมของอัตราการทับซ้อน (Overlap) ของเอกสารลดลง ค่าความถูกต้อง

(Precision) ในการจัดกลุ่มสูงขึ้น ทำให้เอกสารที่ไม่ตรงกับความต้องการภายในกลุ่มลดลง และคงประสิทธิภาพด้านความคลอบคลุมเอกสาร (Coverage) ในการจัดกลุ่มของอัลกอริทึม STC _Original แต่เนื่องจากกลุ่มของ New STC มีลักษณะเฉพาะเจาะจงมากจนเกินไป ทำให้อัตราความถูกต้องในการจัดกลุ่มและอัตราความคลอบคลุมเอกสารในการจัดกลุ่มภายในกลุ่มลำดับต้นๆ คือ 5, 10, 15 และ 20 กลุ่มแรก มีค่าต่ำกว่าของอัลกอริทึม STC _Original และ STC with n-gram ดังแสดงผลการทดลองในบทที่ 4

2. ประสิทธิภาพด้านเวลาที่ใช้ในการประมวลผล

จากผลการทดลอง พบว่าการทำงานของอัลกอริทึม New STC เร็วกว่าการทำงานของอัลกอริทึม STC _Original และ STC with n-gram ทั้งนี้เนื่องจากอัลกอริทึม New STC มีการปรับแต่งกลุ่มในลักษณะที่ต้องการให้กลุ่มมีขนาดเล็ก กะทัดรัด มีความเฉพาะเจาะจงมากกว่ากลุ่มของ STC _Original โดยการลบเอกสารที่ทับซ้อนกันในแต่ละกลุ่มทำให้สมาชิกภายในกลุ่มลดลง และลบกลุ่มเอกสารที่ทับซ้อนกันทำให้จำนวนกลุ่มลดลง ซึ่งเป็นจุดเด่นในการลดข้อมูลที่ต้องเข้าไปตรวจสอบทุกครั้งในการทำงาน ทำให้ประหยัดเวลาในการทำงาน และการเชื่อมป้ายชื่อกลุ่มจะพิจารณาเฉพาะกลุ่มที่สามาถนำมาเชื่อมกันได้เท่านั้น ทำให้จำนวนกลุ่มที่จะนำมาใช้ในการพิจารณาการเชื่อมกลุ่มลดลงทุกครั้ง ตามขนาดความยาวของป้ายชื่อกลุ่มที่เพิ่มขึ้น

3. ประสิทธิภาพด้านการใช้งานพื้นที่หน่วยความจำ (Space)

จากผลการทดลอง พบว่าการทำงานของอัลกอริทึม New STC ใช้หน่วยความจำน้อยกว่า STC _Original และ STC with n-gram โดยพิจารณาจากจำนวนโหนดที่ใช้ในการสร้างซัฟฟิกทรี ทั้งนี้เนื่องจากอัลกอริทึม New STC มีการนำเทคนิค n-gram มาใช้งานร่วมกับโครงสร้างข้อมูลซื่อซัฟฟิกทรี ทำให้สามารถลดจำนวนโหนดหรือจำนวนคำที่ปรากฏภายในซัฟฟิกทรีได้เป็นจำนวนมาก เพราะการสร้างซัฟฟิกทรีไม่จำเป็นต้องนำคำเอกสารทั้งหมดมาสร้างซัฟฟิกทรีในแต่ละครั้งของซับทรี (sub-tree) ดังเช่นอัลกอริทึมของ STC _Original แต่จะนำมาตามขนาดของ n-gram ที่กำหนด และบางครั้งคำภายในผลการสืบค้นซ้ำกันทำให้จำนวนโหนดไม่เพิ่มขึ้น ดังตัวอย่างผลการทดลองด้วยชุดข้อมูลจาก Ohsumed Collection ในหัวข้อที่ 4.3.3 อัลกอริทึมของ STC_Original ไม่สามารถประมวลผลได้ เนื่องจากพื้นทีหน่วยความจำไม่เพียงพอ

5.2 ปัญหาและข้อเสนอแนะในการทำวิจัยต่อไป

เนื่องจากการลบเอกสารที่ทับซ้อนกันของคู่กลุ่มเอกสารที่ถูกเชื่อมป้ายชื่อกลุ่มเพื่อค้นหากลุ่มใหม่ ในขั้นตอนของการเชื่อมป้ายชื่อกลุ่ม ส่งผลให้บางกลุ่ม (Cluster) คงเหลือเอกสารเพียงหนึ่งเอกสารเท่านั้น ดังตัวอย่างเช่น

A = { Thailand , Manufactures , Exports} (2 , 262)(6 , 268)(8 , 254) และ

B = { Topic , Thailand , Manufactures} (1 , 262)(5 , 268)

ผลการเชื่อมป้ายชื่อกลุ่มพื้นฐาน คือ

$A \oplus B$ = { Topic , Thailand , Manufactures Exports} (2 , 262)(6 , 268) และ

A = { Thailand , Manufactures , Exports} (8 , 254)

เพื่อเป็นการรักษาระดับอัตราความคลอบคลุม (Coverage) หรือไม่ให้สารสนเทศ (Information) บางอย่างต้องเสียไป เราจึงไม่ควรลบกลุ่มเหล่านี้ทิ้งไป แต่ควรจัดลำดับให้กลุ่มเหล่านี้มีความสำคัญน้อยที่สุด คือให้ปรากฏอยู่ในลำดับท้ายๆของผลการจัดกลุ่มผลการสืบค้น

5.3 แนวทางการพัฒนาในอนาคต

1. พัฒนาให้ซัฟฟิกทรีคลัสเตอร์ริ่ง (Suffix Tree Clustering) เป็นการจัดกลุ่มผลการสืบค้นในลักษณะลำดับชั้น (Hierarchical Clustering) เนื่องจากการจัดกลุ่มผลการสืบค้น เป็นกระบวนการทำงานซึ่งมีวัตถุประสงค์เพื่ออำนวยความสะดวกในการมองเห็นและเข้าถึงผลการสืบค้นให้กับผู้สืบค้น ดังนั้นเพื่อให้การทำงานเป็นไปตามวัตถุประสงค์อย่างแท้จริง การจัดกลุ่มผลการสืบค้นควรมีลักษณะของผลการทำงานเป็นลำดับชั้น (Hierarchical Clustering) และจากการศึกษาพบว่าซัฟฟิกทรีคลัสเตอร์ริ่ง (Suffix Tree Clustering) มีลักษณะการทำงานที่เป็นลำดับชั้น แต่ผลการจัดกลุ่มมีลักษณะเป็นลำดับรายการ (Flat Clustering)

2. พัฒนาการจัดลำดับความสำคัญของกลุ่ม (Ranking) เนื่องจากการลบเอกสารที่ทับซ้อนของกลุ่มเพื่อให้กลุ่มมีลักษณะเฉพาะเจาะจง ทำให้อัตราการทับซ้อนของเอกสารลดลง ส่งผลให้อัตราความถูกต้อง (Precision) และอัตราความคลอบคลุมการจัดกลุ่ม (Coverage) สูงกว่าหรือเท่ากันกับเทคนิคการจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งรูปแบบเดิมในภาพรวม แต่ความถูกต้องของกลุ่มในลำดับที่ 5, 10, 15 และ 20 กลุ่มแรกต่ำกว่าการจัดกลุ่มผลการสืบค้นด้วยซัฟฟิกทรีคลัสเตอร์ริ่งรูปแบบเดิม ดังแสดงผลการทดลองในแต่ละตารางการทดลองในบทที่4 เพื่อให้อัตราความถูกต้องในลำดับต้นๆสูงขึ้นควรมีการจัดลำดับความสำคัญ (Ranking) แนวใหม่ที่จะส่งผลให้กลุ่มที่มีอัตราความถูกต้องสูงที่สุดมาแสดงผลก่อนในลำดับต้นๆ และจากการศึกษาพบว่ากลุ่มที่มีอัตราความถูกต้องในระดับสูงๆ คือกลุ่มที่ป้ายชื่อกลุ่มยาวมากกว่า 2 คำ

3. พัฒนาซัฟฟิกทรีคลัสเตอร์ริ่ง (Suffix Tree Clustering) ให้สามารถจัดกลุ่มผลการสืบค้นกับผลการสืบค้นที่เป็นภาษาไทยได้

บรรณานุกรม

[1] Oren Zamir and Oren Etzioni. **Document Clustering : A Feasibility Demonstation**. Proceedings of the 19th International ACM SIGIR Conference on Research and Development of Information Retrieval, pp 46-54, 1998.

[2] Oren E. Zamir. **Clustering Web Document : A Phrase-Based Method for Grouping Search Engine Results.** Doctoral Dissertation, University of Washington, 1999.

[3] Oren Zamir and Oren Etzioni. **Grouper : A Dynamic Clustering Interface to Web Search Results**. WWW8/Computer Networks, Amsterdam, Netherland, 1999.

[4] Jerzy Stefanowski and Dawid Weiss . **Web Search Results clustering in Polish: experimental evaluation of Carrot.** Advances in Soft Computing, Intelligent Information Processing and Web mining, Proceedings of the International IIS: IIPWM'03 Conference, Zakopane, Poland, vol. 579 (XIV), 2003, pp. 209-22.

[5] Hua-Jun Zeng and et.at. **Learning to Cluster Web Search Results**. SIGIR'04 , Peking University, 2004

[6] Soumen Chakrabarti. **Mining The Web: Discovering Knowledge From Hypertext Data.** San Francisco: Morgan Kaufmann Publishers., 2003.

[7] Paolo, Ferragina. and Antonio Gulli. **A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering.** IW3C2, Chiba, Japan, 2005.

[8] [Online]. Available : **http://ercolino.isti.cnr.it/webcat/**

[9] F. Giannotti , M. Nanni and D. Pedreschi. **WEBCAT: Automatic Categorization of Web Search Results**. In SEBD03.

[10] Jerzy Stefanowski and Dawid Weiss. **Carrot^2 and Language Properties in Web Search Results Clustering.** Proceeding of the First International Atlantic Web Intelligence Conference, Madrit, Spain, vol.2663, 2003, pp. 240-249.

[11] [Online]. Available : **http://carrot.cs.put.poznan.pl/carrot2-remote-controller**

[12] Jerzy Stefanowski and Stanislaw Osinski. **An Algorithm for Clustering of Web search Results Clustering**. Master thesis of Science, Poznan University, Poland, June 2003.

[13] [Online]. Available. http://credo.fub.it

[14] B. Fung, K. Wang, and M. Ester. **Large Hierarchical Document Clustering using Frequent Itemsets**. Simon Fraser University, BC, Canada, May 2003.

[15] [Online]. Available. http://snaket.di.unipi.it

[16] Dell Zhang and Yisheng Dong. **Semantic, Hierarchical, Online Clustering of Web Search Results**. Proceeding of the 6th of Asia Pacific Web Conference (APWEB), Hangzhou, China, April 2004.

[17] [Online].Available. http://www.dcs.bbk.ac.uk/~dell/publications/dellzhang_unpub_shoc.html

[18] N. Chambers, J. Tetreault and J. Allen. **Approaches for Automatically Tagging Affect.** American Association for Artificial Intelligence(www.aaai.org). 2004.

[19] Stanislaw Osinski and Dawid weiss. **A Concept-Driven Algorithm for Clustering Search Results.** IEEE Computer Society. May-June 2005.

[20] Chi Lang Ngo and Hung Son Nguyen. **A method of web search results clustering based on rough sets**. Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05). 2005.

[21] [Online]. Available. ftp://medir.ohsu.edu/pub/ohsumed

[22] [Online]. Available. http://ai-nlp.info.uniroma2.it/moschitti/corpora.htm

ภาคผนวก

ผลงานวิจัยที่ได้รับการตีพิมพ์เผยแพร่

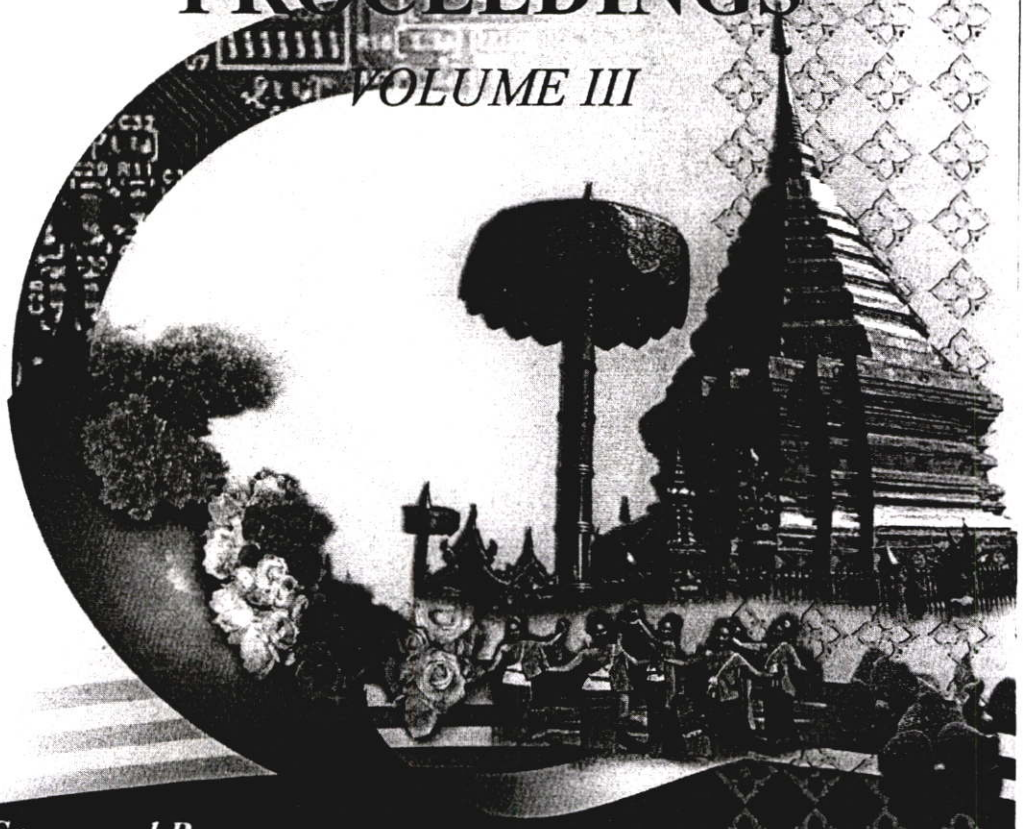
Worapoj Kreesuradej and Jongkol Janruang, "**A New Suffix Tree Clustering Algorithm**", Proceedings of 3th International Technical Conference On Circuits/Systems And Communications (ITC-CSCC 2006), Chiang Mai, 10–13 July 2006, pp. III101 - III103.

ITC-CSCC 2006

The International Technical Conference on Circuits/Systems, Computers and Communications

PROCEEDINGS

VOLUME III



Sponsored By

THE ELECTRICAL ENGINEERING/ELECTRONICS, COMPUTER, TELECOMMUNICATIONS AND INFORMATION ASSOCIATION (ECTI), THAILAND

THE INSTITUTE OF ELECTRONICS ENGINEERS OF KOREA (IEEK), KOREA

THE INSTITUTE OF ELECTRONICS, INFORMATION AND COMMUNICATION ENGINEERS (IEICE), JAPAN

NATIONAL ELECTRONICS AND COMPUTER TECHNOLOGY CENTER, THAILAND

In association with IEEE THAILAND SECTION

A New Suffix Tree Clustering Algorithm

Jongkol Janruang* and Worapoj Kreesuradej**

Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, 15320 Thailand
Email: tawan48@gmail.com* and worapoj@it.kmitl.ac.th**

ABSTRACT

Suffix Tree Clustering (STC) is incremental and linear time algorithm for on-the-fly web search results clustering. However, STC is inadequate since they generate interrupted cluster label due to using n-gram technique. The propose model provides more relevance and less overlap of web document clusters than conventional Suffix Tree Clustering algorithms. Experiments also show that the proposed algorithm has better performance than the conventional STC.

Keyword: web search results clustering, incremental clustering, content based combining.

1. INTRODUCTION

Traditional search engines such as Google, Yahoo and MSN often return a long list of search results. For example, if a user searches for word "data mining", Google returns a list of 54,800,000 web pages, Yahoo returns 34,900,000 web pages and MSN returns 2,794,005 web pages. Web users must go through the long list sequentially to find their required results. As a solution to this problem, clustering web search results into different group is proposed.

Several researches on clustering web search results are conducted to improve efficiency and effectiveness of clustering results. Basically, the approach of clustering web search results can be classified into four categories as follows [1].

- Single word and flat clustering, such as WABCAT [2], is a clustering which uses a single word and flat clustering techniques.
- Sentences and flat clustering, such as Lingo [3], is a clustering which uses phrase or sentences and flat clustering techniques.
- Single word and hierarchical clustering, such as FIHC [4], is clustering which uses a single word and hierarchical clustering techniques.
- Sentences and hierarchical clustering, such as SHOC [5], is a clustering which uses phrase or Sentences and hierarchical clustering technique.

Here, the method in this paper is classified as sentences and flat clustering. In addition, this paper proposes a new Suffix Tree Clustering algorithm that provides more efficient and effective than conventional Suffix Tree Clustering algorithms. The proposed algorithm provides smaller clusters and more readable cluster label than conventional Suffix Tree Clustering algorithms.

2. RELATE WORK

Suffix Tree Clustering (STC) is introduced for web search results clustering by Oren Zamir and Oren Etzioni since 1998. Oren Zamir and Oren Etzioni proposed Suffix Tree model to identify sets of documents that share common phases and then create clustering according to these phase [6][7][8]. In 2003, Hau-Jun Zeng and etc. [9] introduced an improved Suffix Tree Clustering by ranking salient phases as candidate cluster names based on a regression model learned from humane labeled training data. Several works on Suffix Tree Clustering have tried to improve relevance and overlap of web document clusters. Our base cluster label extraction process is similar to STC but we new merge base cluster and identify cluster label.

This paper also introduces a new Suffix Tree Clustering algorithm that content based combining base cluster to new cluster of web document and provides more relevance and less overlap of web document clusters than conventional Suffix Tree Clustering algorithms. Furthermore, the proposed STC algorithm also provides more readable cluster label than conventional Suffix Tree Clustering algorithms.

3. A NEW SUFFIX TREE CLUSTERING ALGORITHM

Unlike conventional Suffix Tree Clustering algorithms, the new Suffix Tree Clustering algorithm introduces a new way to merge base cluster and a new method to identify cluster label. As results, the proposed algorithm provides more relevance and less overlap of web document clusters than conventional Suffix Tree Clustering algorithms. The proposed algorithm composes of four steps which are presented as following.

3.1 Pre-processing

Pre-processing is selecting the most suitable terms that describing better document content. The term are transformed using stop words (such as Articles, pronouns, prepositions, etc.) are eliminate and stemming techniques [10].

ISBN 974-94418-6-9 ©2006 ECTI

3.2 Base Cluster Identification

Base Cluster Identification is identification base clusters using suffix tree with n-gram. The major steps to identify base clusters are shown as bellows.

- Building suffix tree with n-gram. As an example, we extract all n-gram from the snippets where n <= 3 to build suffix tree. As shown in figure 1. The suffix tree of 5 documents:
  D_1: photographs William Jefferson Clinton
  D_2: president William Jefferson Clinton
  D_3: president William Jefferson Clinton document
  D_4: William Jefferson Clinton
  D_5: Jefferson Clinton document

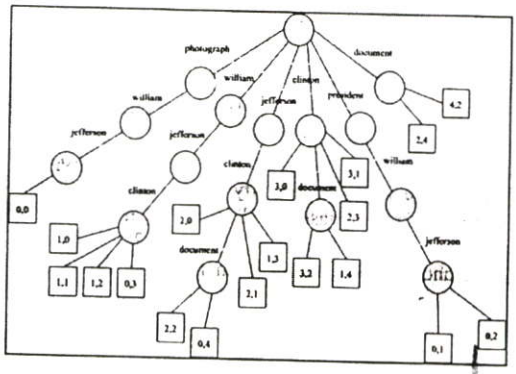


Fig. 1: A Suffix Tree with n-gram (where n<=3)

- As shown in Figure 2, each edge is labeled with a non-empty sub-string of S (hence it is a trie) . The label of a node in defined to be the concatenation of the edge-labels on the path from the root to that node

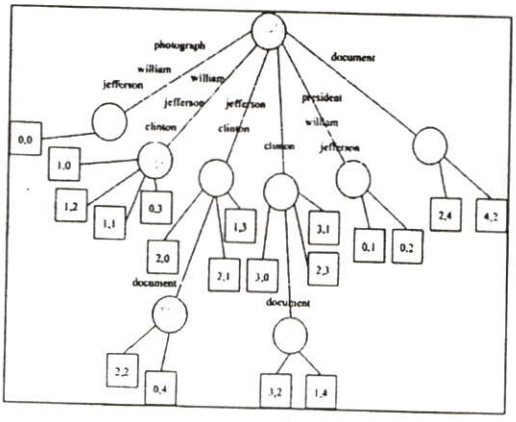


Fig. 2: Seven Base Cluster of Suffix Tree

- Base clusters identification is each node of the suffix tree. A base cluster represent a group document and a phrase that is cluster label. The set of documents tagging the suffix-nodes that are descendants of the node make up the document group. According to Fig. 2 , base clusters of the suffix tree are shown in Table 1.

Table 1: Seven nodes corresponding base clusters.

Node	Common Phrase	Document
A	William Jefferson Clinton	(1,0)(1,1)(1,2)(0,3)
B	Jefferson Clinton	(2,0)(2,1)(1,3)(2,2)(0,4)
C	Jefferson Clinton document	(2,2)(0,4)
D	Clinton	(3,0)(3,1)(2,3)(3,2)(1,4)
E	Clinton document	(3,2)(1,4)
F	president William Jefferson	(0,1)(0,2)
G	Document	(4,2)(2,4)

3.3 Content Base Combining Base Clusters

Document set of distinct base clusters may overlap and may be identical. For example, document set from A base cluster and that from B base cluster in table 1 are the same. Thus, combining base cluster is necessary to reduce number of clusters and overlap. Oren Zamir and Oren Etzioni introduced a binary similarity measure as criteria for based combining base cluster. Then, longer phase between those two similar base clusters is kept as common phase of the new cluster. This may lose information of the base cluster that has shorter phase. Unlike the previous works, a new joint base cluster from condition shown in Eq.1.

$$A \oplus B = \begin{cases} a_0 \oplus \\ a_1 = b_0 \\ a_2 = b_1 \\ \vdots \\ a_n = b_{n-1} \\ \oplus\ b_n \end{cases} \text{if } (A_{(d)} \subseteq B_{(d)} \text{ or } B_{(d)} \subseteq A_{(d)}) \text{ and } (A_{d \in B} d_{(pos)} - B_{d \in A} d_{(pos)} = 1 \text{ or } -1) \quad (1)$$

where A and B are base clusters that using compare similarity of cluster, $A_{(d)}$ is document in A cluster, $B_{(d)}$ is document in B cluster, $A_{d \in B} d_{(pos)}$ is phrase of origin in document A cluster, $B_{d \in A} d_{(pos)}$ is phrase of origin in document B cluster, $\{a_0, a_1, ..., a_n\}$ is set of term that appear in phrase of A cluster and $\{b_0, b_1, ..., b_n\}$ is set of term that appear in phrase of B cluster.

Content based combining base cluster uses a new joint operator to define a new common phase of a new cluster when merging two similar base clusters .

As an example, given that
A = { president, William, Jefferson } (0,1)(0,2) and
B = { William, Jefferson, Clinton } (1,0)(1,1)(1,2)(0,3)
a new common phase is defined as
$A \oplus B$ ={president, william, jefferson,clinton} (1,1)(1,2)

after combining pair similarity cluster, we delete member documents of the new cluster from the old cluster out

more document. Therefore, results of content based combining base cluster is

B = { William, Jefferson, Clinton } (1,0)(0,3)

$A \oplus B$ ={president,william,jefferson,clinton} (1,1)(1,2)

Example: According to *table:1* phrases that contain only the query words are filtered out. Here the query words is "Jefferson Clinton". Therefore, node B and D are filtered out. We then merge cluster and phrases, to reduce duplicated cluster. Then, we merge base clusters from remain clusters. The results are shows in *table 2.*

Table 2: Results of Content Based Combining Base Cluster of 5 document that had base clusters in *Table 1.*

Node	Common Phrase	Document
1	William Jefferson Clinton	(1,0) (0,3)
2	Jefferson Clinton document	(2,2)(0,4)
3	president William Jefferson Clinton	(1,1)(1,2)

3.4 Ranking Base Clusters

Ranking base clusters is reordering base clusters according to their score obtained from the Eq.3.

$$s(m) = |d| \cdot |m_p| \cdot \Sigma \text{ tfidf}(p_i,d) \qquad (3)$$

where |d| is the number of document in m cluster, $|m_p|$ is the number of word in m phrase and tfidf(p_i ,d) is a score that is calculated from the Eq.4. tfidf(p_i ,d) is an inverse phrase frequency and is defined as

$$\text{tfidf}(p_i,d) = (1+ \log(\text{tf}(p_i,d))) * \log(1+N/\text{df}(p_i)), \qquad (4)$$

where df(p_i) is number of document that phrase p_i appear in, tf(p_i ,d) is the number of occurrence of phrase p_i in document d and N is the total number of document in our document set, that show in *Table 3.*

Table 3: Results of Ranking

Num	cluster	document	S(m)
1	President William Jefferson Clinton	1,2	2*4*2.196=17.568
2	William Jefferson Clinton	0,3	2*3*2.196=13.176
3	Jefferson Clinton document	2,4	2*3*2.196=13.176

4 EXPERIMENTS

We use test data from the Open Directories Project (http://dmoz.org) [11] which is a human-collected directory of Web page links and descriptions. We obtain 1548 search results when submitting a query "computer" and "apple". Then, we compare the results of the proposed algorithm to the conventional STC. The results are shown in Table 4, 5, 6 and 7. From the results, a new Suffix Tree Clustering algorithm show better performance than the conventional STC.

Table 4: Results of Average Precision

Cluster	Conventional STC	New STC
5	0.55	0.53
10	0.59	0.54
15	0.59	0.58
20	0.59	0.59
All cluster	0.67	0.75

Table 5: Results of Average Overlap

Cluster	Conventional STC	New STC
5	0.05	0.03
10	0.25	0.12
15	0.56	0.24
20	1.25	0.47
All cluster	7.09	5.53

Table 6: Results of Average Coverage

cluster	Conventional STC	New STC
5	0.45	0.35
10	0.46	0.42
15	0.55	0.53
20	0.64	0.61
All cluster	1	1

Table 7: Results of Average Relevant and Irrelevant

	Conventional STC	New STC
Ratio number of cluster: Relevant document	0.67	0.75
Ratio number of cluster: Irrelevant document	0.33	0.25
Ratio of the above	2.03	3

5. CONCLUSION

This paper proposes a new Suffix Tree Clustering algorithm to overcome average irrelevant document in multiple clusters can hurt cluster quality problem of the conventional STC. The proposed algorithm provides smaller clusters and more readable cluster label that the preliminary experiment results also show that a new Suffix Tree Clustering algorithm has better performance than the conventional STC.

6 REFERENCES

[1] Paolo, Ferragina. and Antonio Gulli. **A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering.** IW3C2, Chiba, Japan, 2005.

[2] F. Giannotti , M. Nanni and D. Pedreschi. **WEBCAT: Automatic Categorization of Web Search Results.** In SEBD03.

[3] Jerzy Stefanowski and Stanislaw Osinski. **An Algorithm for Clustering of Web search Results**

Clustering. Master thesis of Science, Poznan University of Technology, Poland, 2003.

[4] B. Fung, K. Wang, and M. Ester. **Large Hierarchical Document Clustering using Frequent Itemsets**. Simon Fraser University, BC, Canada, May 2003

[5] Dell Zhang and Yisheng Dong. **Semantic, Hierarchical, Online Clustering of Web Search Results**. International Workshop on Web information and data management, Atlanta, Georgia.

[6] Oren Zamir and Oren Etzioni. **Document Clustering : A Feasibility Demonstation**. International ACM SIGIR, 1998, pp 46-54.

[7] Oren Zamir and Oren Etzioni. **Grouper : A Dynamic Clustering Interface to Web Search Results**. WWW8/Computer Networks, Amsterdam, Netherland, 1999.

[8] Oren E. Zamir. **Clustering Web Document : A Phrase-Based Method for Grouping Search Engine Results.** Doctoral Dissertation, University of Washington, 1999.

[9] Hua-Jun Zeng and et.at. **Learning to Cluster Web Search Results**. SIGIR'04 , Peking University, 2004.

[10] M. F. Porter. **An algorithm for suffix stripping.** Program, 14(3), pp.130-137, 1980.

[11] Open Directory Project. http://dmoz.org.

ประวัติผู้เขียน

ชื่อ	นางสาวจงกล จันทร์เรือง
วัน เดือน ปีเกิด	23 พฤศจิกายน 2517
ที่อยู่	เลขที่ 219 ตรอกสำโรงจันทร์ ต.ในเมือง อ .เมือง จ. นครราชสีมา 30000
ประวัติการศึกษา	2540 บริหารธุรกิจบัณฑิต สาขาคอมพิวเตอร์ธุรกิจ มหาวิทยาลัยวงษ์ชวลิตกุล
ประวัติการทำงาน	2540-2542 ครูผู้สอนโรงเรียนเบญจเทคโนบุรีรัมย์ 2542-2546 ครูผู้สอนโรงเรียนมารีย์บริหารธุรกิจ