บทที่ 2

ความรู้พื้นฐานและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและหาแนวทางในการพัฒนาระบบการวิเคราะห์ข้อคิดเห็นของผู้เชี่ยวชาญด้วยการวิเคราะห์ความหมายแอบแฝงและการจัดกลุ่มข้อความ สำหรับการวิจัยด้วยเทคนิคเดลฟาย ผู้วิจัยได้ศึกษาความรู้พื้นฐานเกี่ยวกับการวิจัยด้วยเทคนิคเดลฟาย ทฤษฎีทางศาสตร์คอมพิวเตอร์ที่นำมาประยุกต์ใช้ และงานวิจัยต่าง ๆ ที่เกี่ยวข้อง เพื่อนำความรู้ในทางทฤษฎีและงานวิจัยที่เกี่ยวข้องมาเป็นแนวทางในการพัฒนาระบบ ซึ่งในบทนี้จะกล่าวถึง การวิจัยด้วยเทคนิคเดลฟาย เทคนิคการหาความคล้ายคลึงระหว่างข้อความ เทคนิคการจัดกลุ่มข้อความ และงานวิจัยที่เกี่ยวข้อง โดยมีรายละเอียดดังต่อไปนี้

2.1 การวิจัยด้วยเทคนิคเดลฟาย

เทคนิคเดลฟาย (Delphi Technique) เป็นวิธีการทางวิทยาศาสตร์เพื่อรวบรวมข้อมูลจากผู้เชี่ยวชาญหลาย ๆ คน เพื่อมุ่งศึกษาและวิเคราะห์เกี่ยวกับองค์ความรู้ในอนาคตของศาสตร์ด้านต่าง ๆ โดยเฉพาะทางด้านวิทยาศาสตร์และเทคโนโลยีซึ่งเป็นศาสตร์ที่มีการเปลี่ยนแปลงอย่างต่อเนื่องตลอดเวลา (มนต์ชัย เทียนทอง, 2548) การวิจัยด้วยเทคนิคเดลฟายถูกจัดว่าเป็นการวิจัยเพื่ออนาคต หรืออนาคตศาสตร์ (Futurism) คือมุ่งเน้นวิจัยในเชิงลึกเพื่อให้ได้ความรู้ความเข้าใจเกี่ยวกับอนาคตได้ดียิ่งขึ้น สำหรับวัตถุประสงค์ที่สำคัญคือ การพยากรณ์ภาพในอนาคต การแสวงหาทางเลือกที่จะดำเนินการ การเตรียมการและการกระตุ้นเตือนให้ตระหนักถึงสิ่งต่าง ๆ ที่จะเกิดขึ้นในอนาคต อันจะนำไปสู่การจัดเตรียม การควบคุม การแก้ไข และการบริหารจัดการในอนาคตให้เป็นไปตามความต้องการ เช่น การพยากรณ์เกี่ยวกับเทคโนโลยีเครือข่ายไร้สายในอีก 10 ปีข้างหน้า หรือแนวทางการพัฒนาการเรียนการสอนออนไลน์ด้วยบทเรียน e-learning ในมหาวิทยาลัย เป็นต้น

กระบวนการของเทคนิคเดลฟายจะเก็บรวบรวมความคิดเห็นหรือการตัดสินใจจากกลุ่มผู้เชี่ยวชาญที่มีความชำนาญพิเศษในสาขาวิชาที่เกี่ยวข้อง เพื่อสรุปมติจากข้อค้นพบที่ได้ให้เป็นอันหนึ่งอันเดียวกันและมีความถูกต้อง โดยไม่ต้องนัดหมายกลุ่มผู้เชี่ยวชาญให้มาประชุมกัน แต่จะให้กลุ่มผู้เชี่ยวชาญแต่ละท่านแสดงความคิดเห็นจากการตอบแบบสอบถาม ซึ่งวิธีการนี้จะทำให้สามารถระดมความคิดเห็นจากผู้เชี่ยวชาญที่อยู่ในสถานที่และเวลาแตกต่างกันได้อย่างไม่มี

ข้อจำกัด อีกทั้งผู้เชี่ยวชาญแต่ละคนสามารถแสดงความคิดเห็นได้อย่างเต็มที่และอิสระ สามารถกลั่นกรองความคิดเห็นของตนเองได้อย่างรอบคอบปราศจากการชี้นำจากกลุ่ม ไม่ตกอยู่ภายใต้อิทธิพลทางความคิดของผู้อื่น ด้วยเหตุผลดังกล่าว เทคนิคเดลฟายจึงเป็นที่ยอมรับและได้รับความนิยม มีการใช้กันอย่างแพร่หลาย ไม่เฉพาะการวิจัยทางด้านวิทยาศาสตร์และเทคโนโลยี แต่ยังรวมไปถึงด้านธุรกิจ สังคม การเมือง เศรษฐกิจ และการศึกษา เนื่องจากทำให้ได้ข้อมูลของภาพในอนาคตที่น่าเชื่อถือสามารถนำไปใช้ประกอบการตัดสินใจได้ดี

2.1.1 กระบวนการวิจัยด้วยเทคนิคเดลฟาย

การดำเนินการวิจัยด้วยเทคนิคเดลฟายมีกระบวนการขั้นตอน การกำหนดปัญหาที่จะศึกษา การเลือกกลุ่มผู้เชี่ยวชาญ การสร้างเครื่องมือที่ใช้ในการวิจัยและเก็บรวบรวมข้อมูล โดยมีรายละเอียดดังต่อไปนี้

2.1.1.1 กำหนดปัญหาที่จะศึกษา

Judd (1971) อ้างถึงใน ชนิตา รักษ์พลเมือง (2535) กล่าวถึงการกำหนดปัญหาที่จะศึกษาในการวิจัยด้วยเทคนิคเดลฟายว่า เมื่อใดก็ตามที่ต้องการคาดการณ์สิ่งที่จะเกิดขึ้นในอนาคตหรือเมื่อใดก็ตามที่เห็นว่าความสอดคล้องต่อเนื่องกันระหว่างเป้าหมาย และวัตถุประสงค์เป็นสิ่งสำคัญแล้ว เมื่อนั้นควรใช้เทคนิคเดลฟาย และในด้านการศึกษานั้นเทคนิคเดลฟายยังอาจใช้ประโยชน์ในการหาค่านิยมที่สอดคล้องต้องกันในการประเมินผลสิ่งใด ๆ

จากคำกล่าวข้างต้นจะเห็นได้ว่าปัญหาที่จะศึกษาด้วยเทคนิคเดลฟายควรที่จะเป็นประเด็นปัญหาอันจะนำไปสู่การวางนโยบายหรือคาดการณ์อนาคตรวมทั้งการกำหนดทางเลือกต่าง ๆ หรือเป็นประเด็นปัญหาที่มุ่งหาความเห็นสอดคล้องต้องกันเพื่อแก้ปัญหาที่สลับซับซ้อนทั้งในเชิงโครงสร้างและการปฏิบัติงานหรือเพื่อสรุปเป็นหลักการแนวคิดร่วมกัน ปัญหาที่ศึกษาในการวิจัยด้วยเทคนิคเดลฟายจึงเป็นปัญหาในเชิงคุณลักษณะซึ่งไม่อาจตอบได้โดยอาศัยการศึกษาด้วยวิธีการเชิงสถิติ

2.1.1.2 เลือกกลุ่มผู้เชี่ยวชาญ

เนื่องจากคุณลักษณะเฉพาะของการวิจัยด้วยเทคนิคเดลฟาย คือ การอาศัยข้อคิดเห็นจากการตอบของผู้เชี่ยวชาญ ผลการวิจัยจะน่าเชื่อถือหรือไม่ขึ้นอยู่กับว่ากลุ่มผู้เชี่ยวชาญที่เลือกสรรมานั้น สามารถให้ข้อมูลที่น่าเชื่อถือได้เพียงใด ดังนั้นสิ่งที่ผู้วิจัยจะต้องคำนึงถึงในการเลือกกลุ่มผู้เชี่ยวชาญ ได้แก่ ความสามารถของกลุ่มผู้เชี่ยวชาญ ความร่วมมือของผู้เชี่ยวชาญจำนวนผู้เชี่ยวชาญและวิธีการเลือกสรรผู้เชี่ยวชาญ เป็นต้น สำหรับจำนวนผู้เชี่ยวชาญ

ที่ตอบแบบสอบถาม จะไม่มีข้อกำหนดตายตัวว่ามีจำนวนเท่าใด แต่จากผลการประชุมประจำปีของ California Jonior Colleges Association เมื่อปี พ.ศ 2514 (ชนิตา รักษ์พลเมือง, 2535) ได้ข้อสรุปเกี่ยวกับจำนวนผู้เชี่ยวชาญที่ใช้ในการวิจัยด้วยเทคนิคเดลฟายว่า ถ้าใช้ผู้เชี่ยวชาญจำนวน 17 คนขึ้นไป อัตราการลดลงของความคลาดเคลื่อน (Error) จะน้อยมาก การวิจัยด้วยเทคนิคเดลฟายจึงใช้ผู้เชี่ยวชาญจำนวน 17 คนเป็นส่วนใหญ่ อย่างไรก็ตามจำนวนของผู้เชี่ยวชาญสามารถน้อยลงกว่านี้ได้ แต่อัตราการลดลงของความคลาดเคลื่อนจะสูงขึ้นดังตารางที่ 2.1

ตารางที่ 2.1

จำนวนผู้เชี่ยวชาญที่ใช้ในการวิจัยด้วยเทคนิคเดลฟายที่มีผลต่อค่าความคลาดเคลื่อน

จำนวนผู้เชี่ยวชาญ	ช่วงความคลาดเคลื่อน	ความคลาดเคลื่อนลดลง
1-5	1.02-7.0	.50
5-9	.70-.58	.12
9-13	.58-.54	.04
13-17	.54-.50	.04
17-21	.50-.48	.02
21-25	.48-.46	.02
25-29	.46-.44	.02

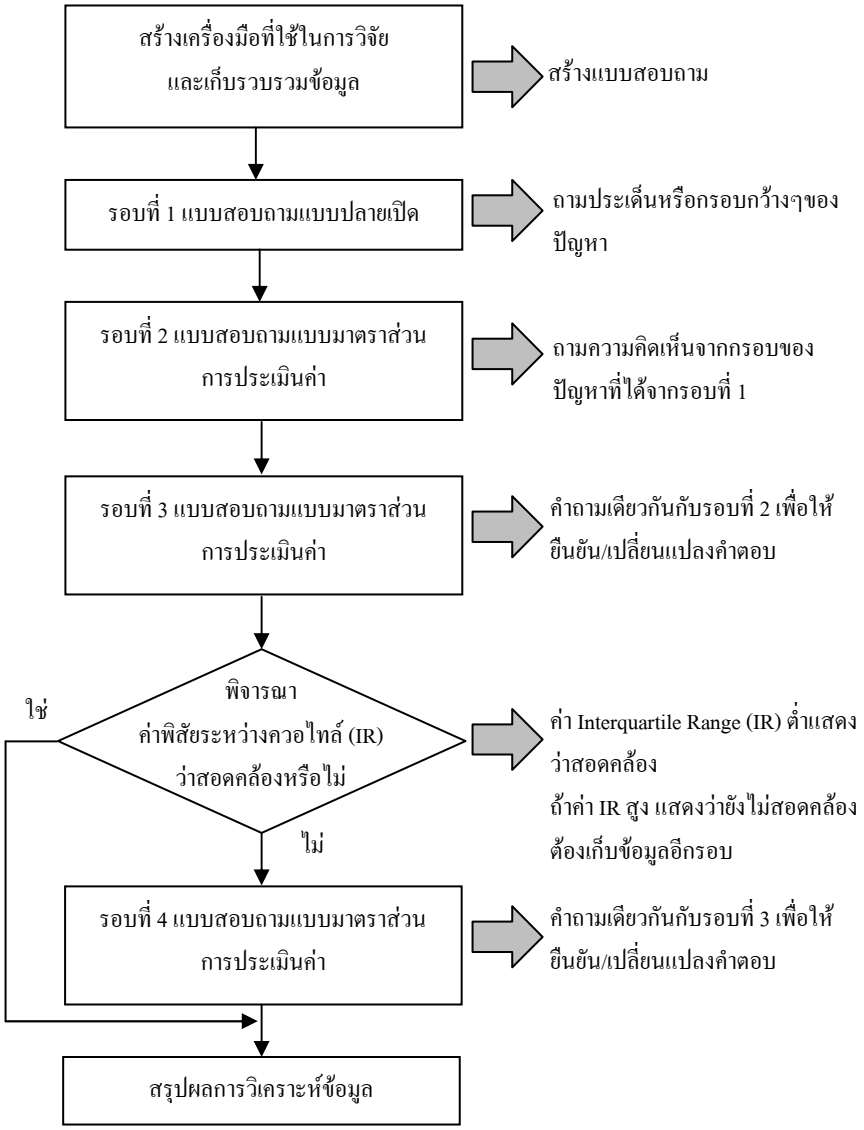
ที่มา: Macmillan, T.T., 1971:52 อ้างถึงใน ชนิตา รักษ์พลเมือง. (2535). การวิจัยแบบเทคนิคเดลฟาย. ใน *เทคนิควิธีการวิเคราะห์นโยบาย* (59-73). กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย. 2535:63

2.1.1.3 สร้างเครื่องมือที่ใช้ในการวิจัยและเก็บรวบรวมข้อมูล

องค์ประกอบที่สำคัญในการวิจัยด้วยเทคนิคเดลฟาย คือ แบบสอบถามซึ่งเป็นเครื่องมือในการวิจัย การเก็บข้อมูลโดยทั่วไปแบ่งออกเป็นสามถึงสี่รอบ ขึ้นอยู่กับผลการวิจัยในแต่ละรอบ โดยมีการระดมความคิดเห็นของกลุ่มผู้เชี่ยวชาญและให้ผู้เชี่ยวชาญได้กลั่นกรองความคิดเห็นของตนอย่างรอบคอบ การส่งและเก็บแบบสอบถามจึงมักดำเนินการทางไปรษณีย์ ซึ่งควรมีการสอดซองที่จ่าหน้าซองกลับมาหาผู้วิจัยเองพร้อมปิดตราไปรษณีย์เพื่ออำนวยความสะดวกให้กับผู้เชี่ยวชาญในการส่งกลับของแต่ละรอบ โดยกระบวนการสร้างเครื่องมือที่ใช้ในการวิจัยในแต่ละรอบแสดงในภาพที่ 2.1 มีรายละเอียดดังนี้

ภาพที่ 2.1

กระบวนการสร้างเครื่องมือที่ใช้ในการวิจัยด้วยเทคนิคเดลฟาย

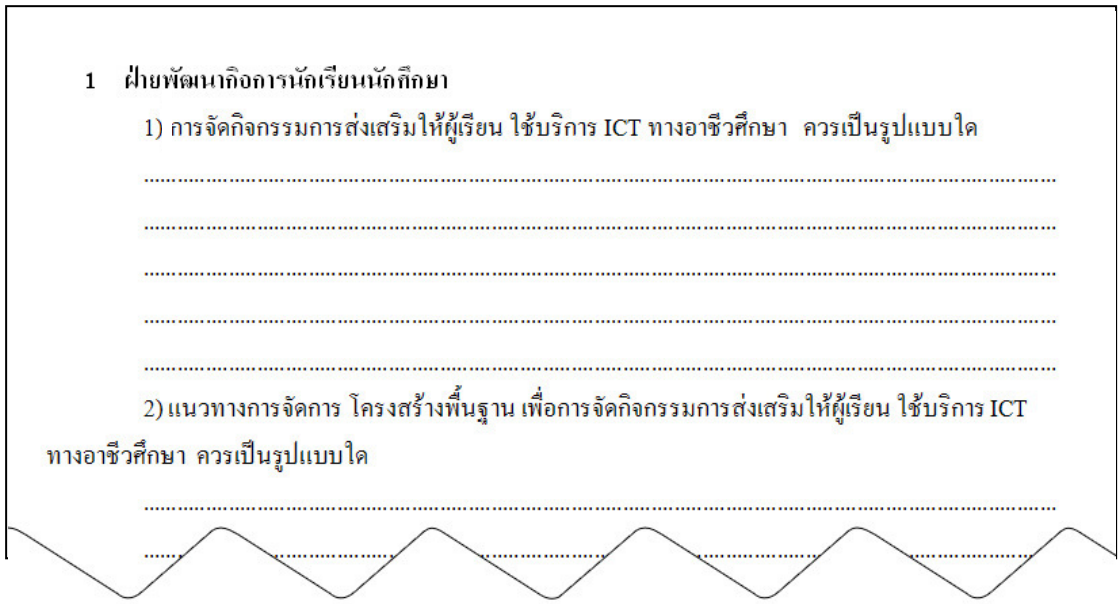


ที่มา: “สถิติและวิธีการวิจัยทางเทคโนโลยีสารสนเทศ,” โดย มนต์ชัย เทียนทอง, 2548, สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, น.172

รอบที่หนึ่งต้องกำหนดกรอบของการวิจัยเพื่อทำให้เห็นภาพของการวิจัยที่ชัดเจนขึ้น แล้วนำมาสร้างแบบสอบถามฉบับแรก ซึ่งเป็นคำถามปลายเปิด (Opened End) ดังตัวอย่างในภาพที่ 2.2 ผู้วิจัยจะต้องรวบรวมความคิดเห็นและวิเคราะห์คำตอบโดยละเอียดแล้วนำมาสังเคราะห์เป็นประเด็นต่าง ๆ เพื่อกำหนดกรอบของปัญหาในรอบต่อไป

ภาพที่ 2.2

ตัวอย่างแบบสอบถามรอบที่หนึ่งของการวิจัยด้วยเทคนิคเดลฟาย


1 ฝ่ายพัฒนากิจการนักเรียนนักศึกษา

1) การจัดกิจกรรมการส่งเสริมให้ผู้เรียน ใช้บริการ ICT ทางอาชีวศึกษา ควรเป็นรูปแบบใด

.....

.....

.....

.....

.....

2) แนวทางการจัดการ โครงสร้างพื้นฐาน เพื่อการจัดกิจกรรมการส่งเสริมให้ผู้เรียน ใช้บริการ ICT ทางอาชีวศึกษา ควรเป็นรูปแบบใด

.....

.....

รอบที่สอง เป็นขั้นตอนที่ยากและสำคัญ (ชนิตา รักษ์พลเมือง, 2535) โดยจะนำข้อคำตอบหรือข้อคิดเห็นของแต่ละข้อคำถามจากผู้เชี่ยวชาญที่ได้มาจากแบบสอบถามรอบแรกของแต่ละคนมาทำการวิเคราะห์เปรียบเทียบ โดยตัดข้อความซ้ำซ้อนออก และข้อคำตอบที่เหมือนกันจะรวมเข้าเป็นกลุ่มเดียวกันเพื่อตั้งเป็นข้อคำถาม ส่วนข้อคำตอบที่ไม่เหมือนกันก็แยกออกนำไปตั้งเป็นคำถามทีละข้อ หลังจากนั้นจึงสร้างแบบสอบถามรอบที่สองแล้ว ก็จะส่งกลับไปยังผู้เชี่ยวชาญกลุ่มเดิมอีกครั้งหนึ่ง ซึ่งรอบนี้จะเป็นแบบสอบถามแบบมาตราส่วนประเมินค่า (Rating Scale) เพื่อให้ผู้เชี่ยวชาญออกความคิดเห็นในลักษณะของการจัดระดับความสำคัญในคำถามแต่ละข้อ รวมทั้งระบุเหตุผลที่เห็นด้วยหรือไม่เห็นด้วยลงในช่องว่างท้ายข้อความ นอกจากนี้ยังสามารถเขียนคำแนะนำเพิ่มเติมได้อีกด้วยสำหรับการวิเคราะห์ข้อมูลของแบบสอบถาม ดังแสดงในภาพที่ 2.3

ภาพที่ 2.3

ตัวอย่างแบบสอบถามรอบที่สองของเทคนิคเดลฟาย

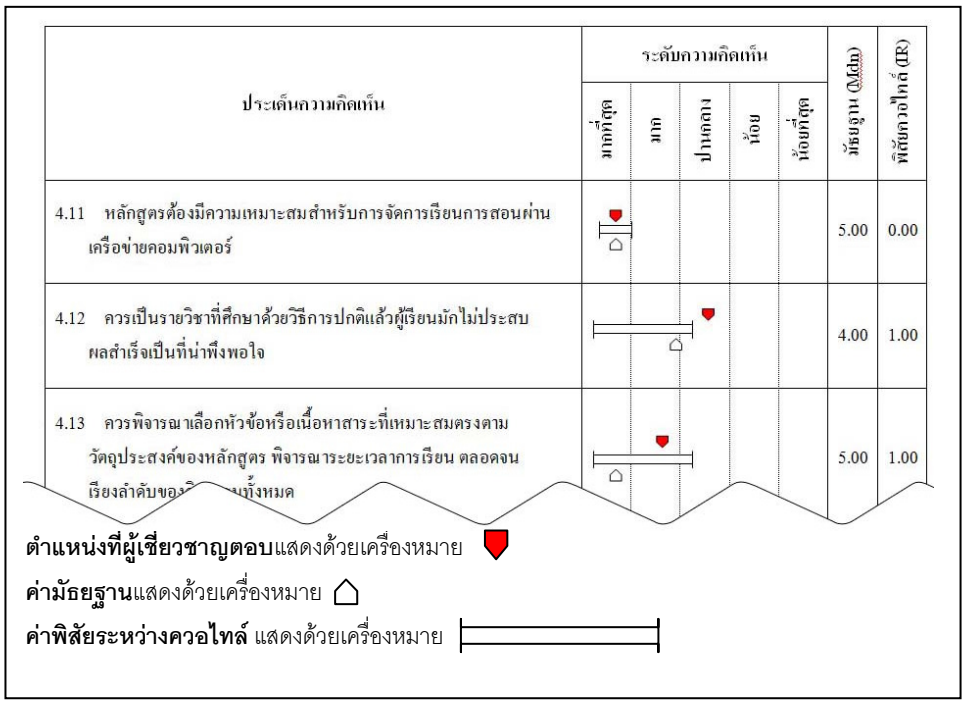
ประเด็นความคิดเห็น	ระดับความคิดเห็น				
	มากที่สุด	มาก	ปานกลาง	น้อย	น้อยที่สุด
11. หลักสูตรต้องมีความเหมาะสมสำหรับการจัดการเรียนการสอนผ่านเครือข่ายคอมพิวเตอร์	☐	☐	☐	☐	☐
12. ควรเป็นรายวิชาที่ศึกษาด้วยวิธีการปกติแล้วผู้เรียนมักไม่ประสบผลสำเร็จที่น่าพึงพอใจ	☐	☐	☐	☐	☐
13. ควรพิจารณาเลือกหัวข้อหรือเนื้อหาสาระที่เหมาะสมตรงตามวัตถุประสงค์ของหลักสูตร พิจารณาระยะเวลาการเรียน ตลอดจนเรียงลำดับของกิจกรรมทั้งหมด	☐	☐	☐	☐	☐
14. ออกแบบกิจกรรมการเรียนให้เป็นขั้นตอนที่ง่ายต่อการทำความเข้าใจและ	☐	☐	☐	☐	☐

รอบที่สาม หลังจากได้รับแบบสอบถามรอบที่สองจากผู้เชี่ยวชาญคืนแล้วจะวิเคราะห์ โดยการนำคำตอบแต่ละข้อมาหาค่ามัธยฐาน และค่าพิสัยระหว่างควอไทล์หรือค่า IR (Interquartile Range) โดยมีการกำหนดเครื่องหมายแสดงตำแหน่งที่ผู้เชี่ยวชาญท่านนั้นได้ตอบในแบบสอบถามฉบับรอบที่สองแล้วส่งกลับไปให้กับผู้เชี่ยวชาญกลุ่มเดิมอีกครั้ง เพื่อให้ผู้เชี่ยวชาญแต่ละคนได้เห็นความแตกต่างระหว่างคำตอบของตนเองเทียบกับ ค่ามัธยฐานและค่า IR ของคำตอบที่ได้จากกลุ่มผู้เชี่ยวชาญทั้งหมด เป็นเสมือนช่องทางที่จะให้ผู้เชี่ยวชาญได้พิจารณาทบทวนอีกครั้งว่าต้องการยืนยันคำตอบเดิมหรือต้องการเปลี่ยนคำตอบใหม่ แสดงตัวอย่างได้ดังภาพที่ 2.4

รอบที่สี่ กระทำตามขั้นตอนเดียวกันกับรอบที่สาม กรณีที่ผลการวิเคราะห์ข้อมูลที่ได้ในรอบนี้ได้คำตอบที่สอดคล้องกัน กล่าวคือ ถ้าค่าพิสัยระหว่างควอไทล์น้อยแสดงว่าความคิดเห็นที่ได้จากกลุ่มผู้เชี่ยวชาญมีความสอดคล้องเป็นไปในทิศทางเดียวกัน ก็สามารถยุติกระบวนการวิจัยและสรุปผลการวิจัยได้

ภาพที่ 2.4

ตัวอย่างแบบสอบถามรอบที่สาม ของเทคนิคเดลฟาย



2.1.2 ข้อดีและข้อเสียของการวิจัยด้วยเทคนิคเดลฟาย

กระบวนการวิจัยด้วยเทคนิคเดลฟายมีทั้งข้อดีและข้อเสีย (สุวรรณา เชื้อรัตนพงษ์, 2528; ชนิตา รักษ์พลเมือง, 2535; มนต์ชัย เทียนทอง, 2548; ศักดิ์ศรี ปาณะกุล, 2550) ดังนี้

2.1.2.1 ข้อดีของการวิจัยด้วยเทคนิคเดลฟาย

1. การวิจัยด้วยเทคนิคเดลฟายให้ผลที่น่าเชื่อถือและสามารถนำไปใช้ประกอบการตัดสินใจได้ดี เนื่องจากคำตอบที่ได้จากความคิดเห็นของกลุ่มผู้เชี่ยวชาญที่มีความชำนาญพิเศษในสาขาวิชานั้นอย่างแท้จริงอีกทั้งผลการวิจัยได้ผ่านกระบวนการพิจารณาจากการย้ำถามหลายรอบ จึงเป็นคำตอบที่กลั่นกรองมาอย่างรอบคอบช่วยให้เกิดความเชื่อมั่นของผลการวิจัยสูง และผู้เชี่ยวชาญแต่ละคนสามารถแสดงความคิดเห็นได้อย่างอิสระ ไม่ตกอยู่ภายใต้อิทธิพลทางความคิดของกลุ่ม เนื่องจากไม่มีการแจ้งผู้เชี่ยวชาญในกลุ่มให้ทราบว่าแต่ละคนเสนอความคิดเห็นอย่างไร ผู้เชี่ยวชาญแต่ละคนจึงมีโอกาสแสดงความคิดเห็นได้อย่างเท่าเทียมกัน และได้ตอบแบบสอบถามฉบับเดียวกันทุกรอบ รวมทั้งมีโอกาสปรับเปลี่ยนหรือยืนยันความคิดเห็นของตนจนเกิดความมั่นใจในคำตอบที่ได้

2. ใช้งบประมาณไม่มาก เนื่องจากไม่ต้องมีการพบปะโดยตรงของกลุ่มผู้เชี่ยวชาญ แต่ใช้แบบสอบถามเก็บข้อมูลแต่ละรอบ ทำให้ประหยัดค่าใช้จ่ายลงไปได้มาก

3. ทำการวิจัยได้ทุกสถานการณ์ สามารถเก็บข้อมูลจากผู้เชี่ยวชาญ ที่อยู่ในสถานที่แตกต่างกันได้ทั้งทางด้านสภาพภูมิศาสตร์และเวลา

4. เป็นวิธีวิจัยที่มีขั้นตอนการดำเนินการไม่ซับซ้อน รวมทั้งผู้วิจัยสามารถทราบลำดับความสำคัญของข้อมูลและเหตุผลในการตอบ รวมทั้งความสอดคล้องของความคิดเห็นในประเด็นต่าง ๆ

2.1.2.2 ข้อเสียของการวิจัยด้วยเทคนิคเดลฟาย

1. การคัดเลือกกลุ่มผู้เชี่ยวชาญที่เป็นผู้ตอบแบบสอบถาม หากไม่ใช่เป็นผู้ที่มีความเชี่ยวชาญในสาขาอย่างแท้จริง จะทำให้ผลการวิจัยเกิดความคลาดเคลื่อน

2. การเก็บรวบรวมข้อมูลอาจไม่ได้รับความร่วมมือจากผู้เชี่ยวชาญในการตอบแบบสอบถามโดยตลอด รวมทั้งเกิดความเบื่อหน่ายในการตอบแบบสอบถามหลายรอบ อันเนื่องมาจากสาเหตุต่าง ๆ เช่น เป็นเรื่องที่ไม่น่าสนใจ มีภารกิจมาก หรือปัญหาอื่น ส่งผลให้กระบวนการวิจัยล่าช้า อีกทั้งการตอบแบบสอบถามของผู้เชี่ยวชาญ หากผู้เชี่ยวชาญขาดความรอบคอบหรือเกิดความลำเอียงในการพิจารณาคำตอบ หรือผู้เชี่ยวชาญไม่ได้เป็นผู้ตอบแบบสอบถามด้วยตนเองอย่างแท้จริง ส่งผลให้ข้อมูลเกิดความคลาดเคลื่อนขาดความน่าเชื่อถือ

3. การสูญหายของแบบสอบถาม หรืออาจไม่ได้รับแบบสอบถามกลับคืนมาจากผู้เชี่ยวชาญ นอกจากนี้ยังพบว่าผู้เชี่ยวชาญซึ่งปกติจะมีภารกิจมากมักเดินทางไปต่างประเทศบ่อยครั้ง ทำให้ระยะเวลาการวิจัยล่าช้าไปจากกำหนดการ

4. การวิเคราะห์คำตอบในแต่ละรอบหากผู้วิจัยมีอคติหรือขาดความรอบครอบ ทำให้ผลการวิจัยคลาดเคลื่อนและขาดความน่าเชื่อถือ

การวิจัยด้วยเทคนิคเดลฟายเป็นการศึกษาอนาคต ซึ่งให้ประโยชน์ต่อการตัดสินใจอย่างมาก โดยการเก็บรวบรวมข้อคิดเห็นจากกลุ่มผู้เชี่ยวชาญที่มีความชำนาญพิเศษในสาขาวิชาที่เกี่ยวข้อง เพื่อสรุปเป็นมติจากข้อค้นพบที่ได้ให้เป็นอันหนึ่งอันเดียวกันและมีความถูกต้อง โดยที่ผู้วิจัยไม่ต้องนัดหมายกลุ่มผู้เชี่ยวชาญให้มาประชุมกันเหมือนกับการระดมสมอง ถึงแม้จะมีข้อจำกัดบางประการ แต่เทคนิคเดลฟายก็ได้ผลที่แน่นอนน่าเชื่อถือได้มาก และเป็นการให้ข้อมูลที่เป็นประโยชน์ในการตัดสินใจสำหรับอนาคต

2.2 ทฤษฎีที่เกี่ยวข้อง

การวิเคราะห์ข้อคิดเห็นของผู้เชี่ยวชาญเพื่อนำไปสู่การพัฒนาแบบสอบถามรอบที่สองสำหรับการวิจัยด้วยเทคนิคเดลฟายต้องอาศัยการประมวลผลข้อความในการหาความหมายที่คล้ายคลึงกันและการจัดกลุ่มข้อความ เพื่อให้สามารถพัฒนาระบบการวิเคราะห์ข้อคิดเห็นจากผู้เชี่ยวชาญได้ ดังนั้นทฤษฎีที่เกี่ยวข้องคือ เทคนิคการหาความคล้ายคลึงระหว่างข้อความ (Text Similarity) ซึ่งในงานวิจัยนี้เลือกใช้เทคนิคการวิเคราะห์ความหมายแอบแฝง (Latent Sementic Analysis) ที่มีความสามารถในการหาความคล้ายคลึงระหว่างข้อความที่มีความหมายในเชิงเนื้อหาหรือการตีความที่เป็นไปในทิศทางเดียวกันได้ และเทคนิคการจัดกลุ่มข้อความ (Text Clustering) เพื่อนำข้อความที่มีความคล้ายคลึงกันมาจัดไว้ในกลุ่มเดียวกัน โดยมีรายละเอียดดังต่อไปนี้

2.2.1 เทคนิคการหาความคล้ายคลึงระหว่างข้อความ (Text Similarity)

การหาความคล้ายคลึงระหว่างข้อความเป็นการประมวลผลทางภาษาธรรมชาติ (Natural Language Processing) เพื่อให้คอมพิวเตอร์สามารถประมวลผลทางภาษาในการหาความคล้ายคลึงระหว่างข้อความได้จะต้องศึกษาหลักการที่เกี่ยวข้องดังต่อไปนี้

2.2.1.1 การตัดคำ (Word Segmentation)

การตัดคำเป็นปัญหาที่สำคัญสำหรับการประมวลผลข้อความภาษาไทย เนื่องจากการเขียนข้อความภาษาไทยมีลักษณะการเขียนที่ต่อเนื่องกันไป ไม่สามารถหาขอบเขตของคำได้โดยตรงซึ่งแตกต่างกับภาษาอังกฤษที่มีช่องว่างในการแบ่งคำแต่ละคำอย่างชัดเจน มีงานวิจัยที่คิดค้นหลักการต่าง ๆ เพื่อหาขอบเขตของคำ แต่ผลที่ได้ของแต่ละวิธีการจะแตกต่างกันที่ ความถูกต้องของการตัดคำ ความเร็ว รวมถึงการใช้ทรัพยากรของระบบ ดังตารางที่ 2.2 แสดงรูปแบบผลลัพธ์ของการตัดคำที่ขึ้นอยู่กับหลักการที่นำมาประยุกต์ใช้

ตารางที่ 2.2
รูปแบบผลลัพธ์ของการตัดคำที่ใช้หลักการที่แตกต่างกัน

โปรแกรม	ผลลัพธ์ของการตัดคำ
A	ดวง|อาทิตย์|ขึ้น|ทาง|ทิศ|ตะวัน|ออก
B	ดวง|อาทิตย์|ขึ้น|ทาง|ทิศตะวันออก
C	ดวงอาทิตย์|ขึ้น|ทาง|ทิศ|ตะวัน|ออก
D	ดวงอาทิตย์|ขึ้น|ทาง|ทิศตะวันออก
....	.....

ที่มา: “Thai automatic indexing using Latent Semantic Indexing,” โดย Theeradej Tunpaiboon, 2000, Mahidol University, p.19

หลักการที่ใช้ในการตัดคำภาษาไทยมีดังนี้

1. หลักการตัดคำโดยใช้กฎการตัดพยางค์ (Word segmentation with rule based approach) เป็นหลักการแรกที่ถูกนำมาใช้สำหรับการตัดคำภาษาไทย โดย ยุพิน ไทยรัตนานนท์ (2524) สร้างกฎสำหรับการตัดคำโดยพิจารณาจากอักขระที่ปรากฏในพยางค์หรือคำ แบ่งออกเป็น 5 กลุ่ม คือ กลุ่มพยัญชนะ กลุ่มสระ กลุ่มวรรณยุกต์ กลุ่มตัวเลข และกลุ่มอักขระพิเศษ โดยกฎที่สร้างขึ้นนั้นจัดเก็บไว้ภายในรหัสต้นฉบับ (Source Code) ทำให้ไม่สามารถเพิ่มหรือแก้ไขกฎได้ จากผลการทดลองพบว่า การตัดคำด้วยกฎการตัดพยางค์ในงานวิจัยนี้ให้ผลความถูกต้อง 85% ต่อมา สุรินทร์ จรรยาพรพงษ์ (2526) พัฒนาการตัดคำภาษาไทยแบบใช้กฎการตัดพยางค์ โดยสร้างกฎเป็นสองกลุ่ม คือ กฎการหาขอบเขตหน้า และกฎการหาขอบเขตหลัง โดยพิจารณาจากรูปแบบการใช้สระแต่ละตัว เช่น สระ อ อิ อึ อี ถ้ามีวรรณยุกต์ จะต้องมีตัวสะกดหนึ่งตัวเสมอ โดยจากผลการทดลองพบว่าสามารถตัดพยางค์ได้ถูกต้องถึง 96% ซึ่งหลักการตัดคำโดยใช้กฎการตัดพยางค์ช่วยให้การตัดคำภาษาไทยทำงานได้เร็วขึ้นอีกทั้งไม่สิ้นเปลืองทรัพยากรของระบบอีกด้วย อย่างไรก็ตามหลักการตัดคำดังกล่าวยังมีข้อเสียในด้านการหาขอบเขตของคำ (ยืน ภู่วรวรรณ, 2529)

2. หลักการตัดคำโดยใช้พจนานุกรม (Word segmentation with dictionary approach) การตัดคำด้วยหลักการนี้ได้ถูกพัฒนาขึ้นเนื่องจากข้อจำกัดของการตัดคำด้วยกฎการตัดพยางค์ที่ไม่สามารถหาขอบเขตของคำได้ ยืน ภู่วรวรรณ (2529) จึงพัฒนาหลักการตัดคำโดยใช้

พจนานุกรม หลักการทำงานคือ ข้อความที่ได้รับเข้ามาในระบบตัดคำ ระบบจะตรวจสอบข้อความจากซ้ายไปขวา โดยการเปรียบเทียบกับคำในพจนานุกรม กรณีไม่พบคำในพจนานุกรมก็จะลดความยาวของอักขระลงหนึ่งตัวอักษร จนกว่าจะพบคำนั้นในพจนานุกรม และทำเครื่องหมายเพื่อเป็นจุดย้อนกลับ พร้อมกับทำการตรวจสอบและตัดคำอีกครั้ง เช่น “แนวทางการแก้ปัญหาหนี้สินเกษตรกร” เมื่อตรวจสอบข้อความนี้ไม่พบในพจนานุกรม ก็ลดความยาวของอักขระลงหนึ่งตัวอักษร โดยลดอักขระที่อยู่ทางขวาสุดดังนั้นข้อความจะเหลือเป็น “แนวทางการแก้ปัญหาหนี้สินเกษตรก” แล้วตรวจสอบข้อความกับพจนานุกรมอีกครั้ง เมื่อไม่พบก็ลดความยาวลงหนึ่งตัวอักษร เหลือเป็น “แนวทางการแก้ปัญหาหนี้สินเกษตร” ทำซ้ำตามหลักการเดิมจนสุดกว่าจะได้ข้อความที่ตรวจสอบจะพบในพจนานุกรม หลักการนี้เป็นการตัดคำโดยใช้พจนานุกรมแบบเทียบคำที่ยาวที่สุด (Longest Matching) จากผลการทดลองพบว่า สามารถตัดคำได้ถูกต้อง 80% แต่จุดด้อยของวิธีการนี้คือ การเลือกคำที่มีความยาวเกินไปตั้งแต่ต้นส่งผลทำให้ผลลัพธ์ของการตัดคำที่ตามมานั้นเกิดความผิดพลาด เช่น “ไปหามเหสี” จะตัดได้เป็น “ไป_หาม_เห_สี” เนื่องจากพบคำว่า “หาม” ก่อน “หา” ทำให้คำ ถัดไปก็ตัดผิดไปด้วย

วิรัช ศรเลิศล้ำวานิช (2536) ได้พัฒนาการตัดคำโดยเลือกแบบเหมือนมากที่สุด (Maximal Matching) วัตถุประสงค์เพื่อแก้ไขความบกพร่องของการตัดคำแบบเทียบคำที่ยาวที่สุด โดยวิธีการนี้จะใช้การตัดคำแบบวิธีการเทียบคำที่ยาวที่สุดก่อนจากนั้นจะทำการย้อนกลับ (Backtracking) เพื่อทำการตัดคำที่สามารถเป็นไปได้อีกครั้ง โดยใช้วิธีพิจารณาทางเลือกของการตัดคำที่มีค่าจำนวนคำ (Cost) จากการตัดที่ได้น้อยสุดที่สุด และเกิดคำที่ไม่พบในพจนานุกรมให้น้อยที่สุด ตัวอย่างการตัดคำ “ไปหามเหสี” ขั้นแรกจะใช้วิธีการเทียบคำที่ยาวที่สุด จะได้ผลการตัดคำคือ “ไป_หาม_เห_สี” หลังจากนั้นจะทำการย้อนกลับ พบว่าคำที่สอง “หาม” สามารถแบ่งได้เป็น “หา_ม” ทำให้คำที่เหลือสามารถตัดคำได้ 2 แบบ “มเหสี” และ “ม_เห_สี” แล้วทำการคำนวณหาค่าจำนวนคำ ซึ่งผลการตัดคำและคำนวณหาค่าจำนวนคำที่ได้คือ

ทางเลือกที่ 1 ผลการตัดคำ “ไป_หาม_เห_สี” มีค่าจำนวนคำเท่ากับ 4

ทางเลือกที่ 2 ผลการตัดคำ “ไป_หา_มเหสี” มีค่าจำนวนคำเท่ากับ 3

ทางเลือกที่ 3 ผลการตัดคำ “ไป_หา_ม_เห_สี” มีค่าจำนวนคำเท่ากับ 5

ผลการตัดคำที่ถูกเลือกนำมาใช้คือ ทางเลือกที่ 2 “ไป_หา_มเหสี” เนื่องจากมีค่าจำนวนคำน้อยที่สุด จากผลการทดลองหลักการตัดคำแบบเหมือนมากที่สุด ให้ค่าความถูกต้องในการตัดคำสูงถึงร้อยละ 90%

การตัดคำไม่ว่าจะตัดด้วยหลักการใดก็ตามถ้าหากสามารถหาขอบเขตของคำได้ การจัดเก็บข้อความและลำดับการทำงานในส่วนถัดไปก็ดำเนินไปได้อย่างสะดวกและถูกต้อง แต่อย่างไรก็ตามงานวิจัยของ Jaruskulchai (1981) ได้พิสูจน์แล้วว่าหลักการที่ใช้ในการตัดคำไม่ใช่ปัญหาหลักของขั้นตอนการทำงานทั้งหมดในการค้นคืนสารสนเทศสำหรับข้อความภาษาไทย

2.2.1.2 **การกำจัดคำหยุด** (Stop-Word List Removal)

การกำจัดคำหยุดเนื่องจากคำหยุดเป็นกลุ่มคำที่ไม่สามารถนิยามความหมายได้ หรือไม่สามารถเป็นตัวแทนของเอกสารได้ เป็นคำที่มักพบในทุก ๆ เอกสารและมีความถี่มากในการปรากฏ (ผุสดี บุญรอด, 2552) เช่น ที่ นั่น ใน ตาม และ ก็ เพื่อ ไว้ เป็นต้น ดังนั้นในขั้นตอนการประมวลผลข้อความจะทำการตัดคำหยุดออก โดยไม่ทำให้ใจความของเอกสารนั้นเกิดความเปลี่ยนแปลง ประเภทของคำที่จัดว่าเป็นคำหยุดในเอกสารภาษาไทย (ณิชาพร สุระ, 2549) มีดังต่อไปนี้

1. คำบุพบท เป็นคำที่ใช้เชื่อมคำหรือกลุ่มคำให้สัมพันธ์กัน มักใช้คำนำหน้าคำนาม คำสรรพนาม คำกริยา หรือคำวิเศษณ์ เพื่อแสดงความสัมพันธ์ของคำหรือกลุ่มคำที่อยู่หลังคำบุพบท ว่ามีความสัมพันธ์กับคำหรือกลุ่มคำที่อยู่หน้าคำบุพบทอย่างไร ตัวอย่างของคำบุพบท เช่น ของ ใน แก่ แด่ ต่อ ตั้ง ตั้งแต่ โดย เมื่อ กว่า กับ เป็น ดูก่อน ซ้ำแต่ ทาง สู่ แค่

2. คำสันธาน เป็นคำที่ทำหน้าที่เชื่อมคำกับคำ ประโยคกับประโยค ข้อความกับข้อความ เพื่อให้ติดต่อเป็นเรื่องเดียวกัน โดยมักจะใช้เชื่อมใจความที่คล้อยตามกัน ใจความที่ขัดแย้งกัน ใจความที่เป็นเหตุเป็นผลกัน หรือใจความที่เลือกอย่างใดอย่างหนึ่ง ตัวอย่างของคำสันธาน เช่น เพราะ เพราะว่า และ หรือ จึง ดังนั้น มิฉะนั้น ทั้ง แต่ แต่ว่า ครั้น

3. คำสรรพนาม เป็นคำที่ใช้แทนชื่อผู้พูด ผู้ฟัง ผู้ถูกกล่าวถึง และแทนคำนามที่กล่าวถึงมาแล้วในประโยค เพื่อไม่ต้องกล่าวคำนามนั้นซ้ำอีก ตัวอย่างของคำสรรพนาม เช่น ฉัน เรา เขา ดิฉัน กระผม กู คุณ ท่าน เธอ ใต้เท้า มัน

4. คำวิเศษณ์ เป็นคำที่ใช้ขยายคำอื่น เช่น คำนาม คำสรรพนาม คำกริยา หรือคำวิเศษณ์ เพื่อให้มีความหมายชัดเจนและมีรายละเอียดมากยิ่งขึ้น ตัวอย่างของคำวิเศษณ์ เช่น มาก น้อย ใหญ่ เล็ก มหึมา มโหฬาร อ้วน โต สูง ไพเราะ เยอะ หลาย สวย หอม นุ่ม เผ็ด เช้า ค่ำ นี่ นั่น เอง ทั้งนี้ ค่ะ ครับ ขอรับ จ้า จ๊ะ วะ ไม่ หามิได้ ทำไม อย่างไร หมด อดีต ปัจจุบัน

5. คำอุทาน เป็นคำที่แสดงอารมณ์ อาการ หรือความรู้สึกของผู้พูด รวมทั้งเป็นคำที่ใช้เสริมคำพูดตัวอย่างของคำอุทาน เช่น เอ๊ะ อ๊ะ อ่อ เฮ้อ ว้าย โธ่ โถ่ อนิจจา สิ หนอย ซะ นะ น่า แหม ตายล่ะ คุณพระช่วย ชิชะ อุวะ ไม่น่าเลย โอ้โห

คำหยุด มักเป็นคำที่ปรากฏขึ้นบ่อยครั้งในเอกสาร และปรากฏในเอกสารเกือบทุกฉบับ จึงถือได้ว่าคำหยุดเป็น คำที่มิได้เป็นประโยชน์ต่อการนำมาเป็นตัวแทนของเอกสารเชิงความหมายหรือคำสำคัญที่จะนำมาใช้ในการค้นคืนเอกสาร การกำจัดคำหยุดจึงเป็นกระบวนการที่ควรทำก่อนนำเข้าสู่การประมวลข้อความ เพื่อกำจัดคุณลักษณะที่ไม่เป็นประโยชน์ และลดขนาดของเอกสารลง ซึ่งจะช่วยประหยัดทั้งพื้นที่และเวลาในการประมวลผล

2.2.1.3 **การให้น้ำหนักคำ** (Term Weights)

การให้น้ำหนักคำเป็นการแสดงถึงความสำคัญของคำในเอกสาร โดยผ่านการนับความถี่ และนำมาคำนวณน้ำหนักคำเพื่อลดคำที่ไม่มีความสำคัญ (Noise) หรือเพิ่มน้ำหนักให้กับคำที่คาดว่ามีความสำคัญที่เกิดจากข้อความ เนื่องจากความถี่ของคำในเอกสารที่ปรากฏ (Term Frequency : tf) บางครั้งค่าความถี่ของคำมีค่าสูงมากปรากฏบ่อยครั้งอยู่ในทุกเอกสาร โดยไม่สามารถที่จะแยกแยะหรือเป็นตัวแทนของเอกสารได้ เพื่อแก้ปัญหาคำที่ไม่สำคัญแต่มีความถี่สูง การให้ค่าน้ำหนักของคำในเอกสารจึงพิจารณาค่าความถี่แบบผกผัน (Inverse Document Frequent : idf) ร่วมกับค่าความถี่คำ สูตรการคำนวณหาน้ำหนักของคำที่ใช้ คือ Term Frequency Inverse Document Frequency (TF-IDF) (Luhn, 1958) ดังสมการที่ 2.1

$$W_{ij} = tf_{ij} \times idf_i \qquad (2.1)$$

โดยที่ $idf_i = \log \frac{N}{df_i}$ ค่าความถี่ผกผันคำ i ของเอกสารทั้งหมด

W_{ij} คือ น้ำหนักของคำ t_iในเอกสาร d_j

tf_{ij} คือ ความถี่ของคำในเอกสาร

N คือ จำนวนเอกสารทั้งหมดในระบบ

df_i คือ จำนวนเอกสารที่มีคำ t_i ปรากฏ

i คือ ลำดับที่ของคำ

j คือ ลำดับที่ของเอกสาร

ความถี่ของคำที่ปรากฏ (Term Frequency: tf) เป็นสิ่งที่บ่งบอกถึงความสำคัญของคำคำนั้นที่มีต่อเอกสาร ซึ่งเอกสารใดที่มีคำที่ผู้ใช้ต้องการจะเป็นเอกสารที่มีประโยชน์ต่อผู้ใช้ เช่น ถ้าผู้ใช้ต้องการคำว่า “คอมพิวเตอร์” เอกสารที่มีคำว่า “คอมพิวเตอร์” ปรากฏเป็นความถี่ 10 ครั้ง จะเป็นประโยชน์มากกว่า เอกสารที่มีคำว่า “คอมพิวเตอร์” เพียงครั้งเดียว เมื่อมีกลุ่มเอกสารที่เป็น

ประโยชน์ตรงกับความต้องการ เช่น เอกสารเกี่ยวกับ “คอมพิวเตอร์” ภายในเอกสารนั้นจะพบคำว่า “คอมพิวเตอร์” โดยมีความถี่ค่อนข้างสูงและปรากฏขึ้นทุกเอกสารที่มีภายในกลุ่ม ทำให้ไม่สามารถแยกความแตกต่างระหว่างเอกสารได้ ด้วยเหตุนี้คำว่า “คอมพิวเตอร์” จึงมิใช่คำสำคัญของเอกสาร ดังนั้นการประเมินความสำคัญของคำก็จะมีค่าความสำคัญน้อยกว่าคำที่ใช้น้อยหรือปรากฏบางเอกสาร โดยการหาค่าความสำคัญของคำในเอกสารจากกลุ่มเอกสารที่ตรงกับความต้องการ จึงใช้การคำนวณที่เรียกว่า ค่าความถี่แบบผกผัน (Inverse Document Frequent : idf)

จากเหตุผลดังกล่าว เพื่อให้สามารถแยกแยะความแตกต่างระหว่างเอกสารภายในกลุ่ม การให้ค่าน้ำหนักคำสำคัญที่อาจเป็นตัวแทนของเอกสาร (Keyword) ของแต่ละเอกสารจากกลุ่มเอกสารทั้งหมดที่ตรงกับความต้องการจึงไม่ควรพิจารณาเฉพาะค่าความถี่ของคำที่ปรากฏ เพื่อประกอบการอธิบายสามารถแสดงตัวอย่างการคำนวณค่าน้ำหนักคำดังตารางที่ 2.3

ตารางที่ 2.3
แสดงการคำนวณค่าน้ำหนักคำ

N1: “เทคโนโลยีคอมพิวเตอร์ ช่วยในด้านการค้นคว้าศึกษาข้อมูล ทำให้การศึกษาง่ายขึ้น ไร้ขีดจำกัด ส่งเสริมความก้าวหน้าให้กับการศึกษาและวิธีการเรียนการสอนในอนาคต”
N2: “เทคโนโลยีคอมพิวเตอร์มีบทบาทในด้านเศรษฐกิจทำให้เปลี่ยนไปในทิศทางที่ดีขึ้น”
N3: “การศึกษาวิจัยเกี่ยวกับเทคโนโลยีคอมพิวเตอร์มีเพิ่มมากขึ้นช่วยให้ระบบคอมพิวเตอร์มีประสิทธิภาพในการทำงานสูงขึ้น”
N4: “เทคโนโลยีคอมพิวเตอร์ทำให้ระบบเศรษฐกิจของชาติพัฒนาไปเป็นสู่เศรษฐกิจโลกได้รวดเร็วขึ้น”

N= 4; IDF= $\log(N/df_i)$

Keyword	tf_{ij}				df_i	N/df_i	IDF_i	$W_{ij} = tf_{ij} \times idf_i$			
	N1	N2	N3	N4				N1	N2	N3	N4
เทคโนโลยี	1	1	1	1	4	1	0	0	0	0	0
คอมพิวเตอร์	1	1	2	1	4	1	0	0	0	0	0
ศึกษา	3	0	1	0	2	2.0	0.30	0.90	0	0.30	0
เศรษฐกิจ	0	1	0	2	2	2.0	0.30	0	0.30	0	0.60
ระบบ	0	0	1	1	2	2.0	0.30	0	0	0.30	0.30

ขั้นตอนก่อนการคำนวณค่าน้ำหนักคำ เริ่มจากนำข้อความของแต่ละเอกสารผ่านกระบวนการตัดคำและกำจัดคำหยุด แล้วนับความถี่คำเพื่อคำนวณหาค่าน้ำหนักคำตามสมการที่ 2.1 ผลค่าน้ำหนักคำสามารถอธิบายได้ดังนี้

กรณีของคำที่คาดว่าจะเป็นตัวแทนของเอกสาร (Keyword) มีความถี่ไม่ว่าจะสูงหรือน้อยแต่มีปรากฏในทุกเอกสารของกลุ่มเอกสารทั้งหมด ค่าน้ำหนักคำจึงเป็นศูนย์ ดังตัวอย่างในตารางที่ 2.3 คำว่า “เทคโนโลยี” และ “คอมพิวเตอร์” มีความถี่ในเอกสารและปรากฏในทุกเอกสาร กล่าวได้ว่าเอกสารในกลุ่มทั้งหมดนั้นเป็นเรื่องเกี่ยวกับเทคโนโลยีคอมพิวเตอร์ แต่เนื่องจากมีปรากฏในทุกเอกสารทำให้ไม่สามารถแยกความแตกต่างระหว่างเอกสารได้ คำว่า “เทคโนโลยี” กับ “คอมพิวเตอร์” จึงไม่มีความสำคัญไม่สามารถเป็นตัวแทนของเอกสารได้ ค่าน้ำหนักคำที่ได้จากการคำนวณจึงเท่ากับ 0

กรณีของคำที่คาดว่าจะเป็นตัวแทนของเอกสาร (Keyword) มีปรากฏเพียงบางเอกสารจากกลุ่มเอกสารทั้งหมด จึงสามารถตัวแทนของเอกสารเพื่อบ่งบอกความแตกต่างระหว่างเอกสารได้ ค่าน้ำหนักของคำจึงมากกว่าศูนย์แต่จะมีค่าน้ำหนักสูงหรือน้อยขึ้นอยู่กับความถี่ที่ปรากฏภายในเอกสารนั้น ๆ ดังตัวอย่างในตารางที่ 2.3 คำว่า “ศึกษา” จากเอกสารทั้งสี่นั้นมีปรากฏเพียงสองเอกสาร ทำให้สามารถแยกแยะความแตกต่างได้ว่าจากกลุ่มเอกสารทั้งหมดมีเพียงสองเอกสารที่เกี่ยวข้องกับการศึกษาโดยเอกสาร N1 มีค่าความถี่คำว่า “ศึกษา” เท่ากับ 3 ค่าน้ำหนักคำเท่ากับ 0.90 ในขณะที่เอกสาร N3 มีความถี่เท่ากับ 1 ค่าน้ำหนักคำเท่ากับ 0.30 จากตัวอย่างแสดงให้เห็นว่าค่าน้ำหนักคำมีค่าสูงหรือน้อยนั้นจะสัมพันธ์กันกับค่าความถี่ ในขณะที่คำว่า “คอมพิวเตอร์” ของเอกสาร N3 มีความถี่เท่ากับ 2 แต่ค่าน้ำหนักคำเท่ากับ 0 ไม่สัมพันธ์กันกับค่าความถี่ ทั้งนี้เนื่องมาจาก “คอมพิวเตอร์” เป็นคำที่มีปรากฏในทุกเอกสารเพราะฉะนั้นไม่ว่าค่าความถี่สูงหรือน้อย ค่าน้ำหนักคำก็จะเท่ากับ 0

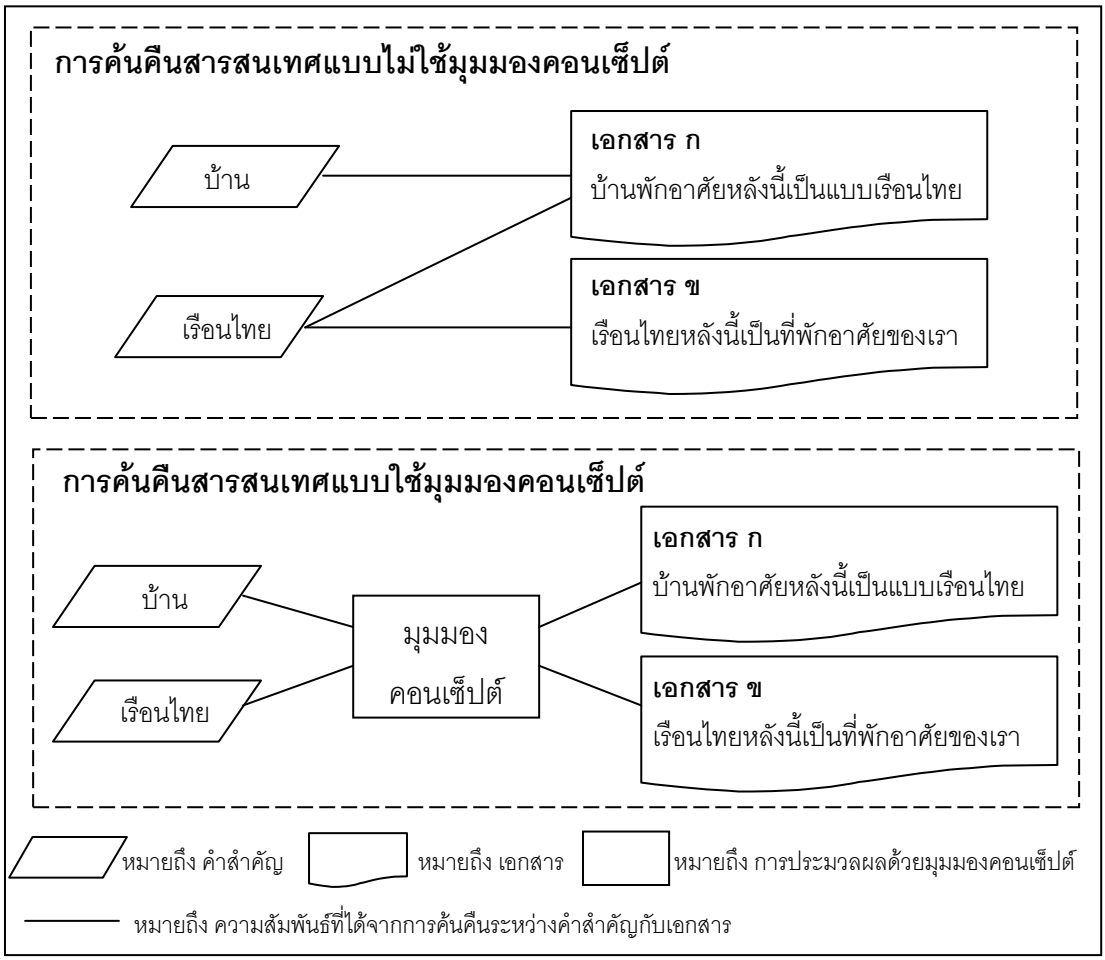
สรุปการคำนวณค่าน้ำหนักคำแบบ TF-IDF ให้ค่าน้ำหนักคำโดยพิจารณาจากความถี่ที่ปรากฏในแต่ละเอกสารร่วมกับความถี่ของคำคำนั้นที่ปรากฏในกลุ่มเอกสารทั้งหมด ทำให้สามารถแยกแยะความแตกต่างระหว่างเอกสารได้ อีกทั้งยังสามารถให้น้ำหนักคำตามความสำคัญของคำที่เป็นตัวแทนของเอกสารนั้นได้อีกด้วย

2.2.1.4 เทคนิคการวิเคราะห์ความหมายแอบแฝง (Latent Semantic Analysis: LSA)

การวิเคราะห์หาความหมายแฝงเป็นเทคนิคหนึ่งในการประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) ที่มีลักษณะเฉพาะในการหาความหมายด้วยเหตุนี้จึงมีการนำการวิเคราะห์ความหมายแอบแฝงหรือ LSA มาใช้กับการค้นคืนสารสนเทศ (Information Retrival) เนื่องจากในยุคแรกนั้นใช้คำสำคัญ (Keywords) ในการระบุสิ่งที่ต้องการและทำการค้นคืนจากคำสำคัญที่ระบุ แต่ปัญหาที่พบคือ บางเอกสารที่ไม่มีคำที่ผู้ใช้ระบุแต่เอกสารนั้นตรงกับความต้องการ (Relevant Document) ไม่ถูกค้นคืน ดังนั้นจึงมีการนำแนวคิดในการใช้มุมมองคอนเซ็ปต์ (Concept Space) หรือมุมมองความหมาย (Semantic Space) แทนคำสำคัญ (Keywords) (Deerwester et al., 1990) โดยแสดงการค้นคืนสารสนเทศแบบไม่ใช้มุมมองคอนเซ็ปต์และแบบใช้มุมมองคอนเซ็ปต์ ดังภาพที่ 2.5

ภาพที่ 2.5

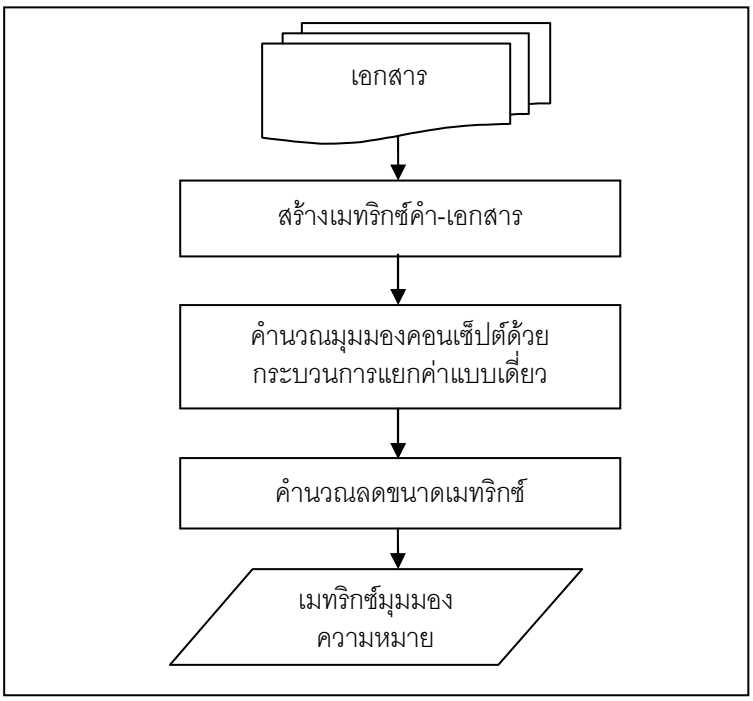
แสดงการค้นคืนสารสนเทศแบบไม่ใช้มุมมองคอนเซ็ปต์และแบบใช้มุมมองคอนเซ็ปต์



จากภาพที่ 2.5 กรณีใช้คำสำคัญ “บ้าน” ระบุเพื่อค้นคืนสารสนเทศ ถ้าเป็นการค้นคืนแบบไม่ใช้มุมมองคอนเซ็ปต์ จะไม่สามารถค้นคืนเอกสาร ข ได้ ทั้งที่คำสำคัญระหว่าง “บ้าน” กับ “เรือนไทย” เป็นคำที่มีความหมายคล้ายคลึงกัน กัน แต่เมื่อมีการใช้แนวคิดใหม่โดยการแทรกเลเยอร์มุมมองคอนเซ็ปต์ (Concept Space) หรือมุมมองความหมาย (Semantic Space) ระหว่างคำสำคัญกับเอกสาร ถ้าระบุคำสำคัญที่ต้องการค้นคืนด้วยคำว่า “บ้าน” ก็จะสามารถค้นคืนได้ทั้งเอกสาร ก และเอกสาร ข

ขั้นตอนการทำงานของการวิเคราะห์ความหมายแอบแฝง ข้อมูลอินพุตคือ ยูนิตหรือส่วนย่อยต่าง ๆ อาทิเช่น เอกสาร (Documents) ย่อหน้า (Paragraph) ประโยค (Sentence) หรือข้อความสั้น (Passage) โดยขั้นตอนแรกจะทำการสร้างเมทริกซ์คำ-เอกสาร (Terms-Documents Matrix) การทำงานต่อมาคือการคำนวณค่าหามุมมองคอนเซ็ปต์ (Concept Space) หรือมุมมองความหมาย (Semantic Space) ด้วยกระบวนการแยกค่าแบบเดี่ยว (Singular Value Decomposition : SVD) ซึ่งใช้เทคนิคทางคณิตศาสตร์และสถิติ จากนั้นลดขนาดของเมทริกซ์เพื่อพิจารณาความสัมพันธ์ทางด้านความหมายของเนื้อหาที่สำคัญ โดยลำดับขั้นตอนการทำงานของการวิเคราะห์ความหมายแอบแฝงสามารถแสดงได้ดังภาพที่ 2.6

ภาพที่ 2.6
แสดงขั้นตอนการทำงานของการวิเคราะห์ความหมายแอบแฝง



1. เมทริกซ์คำ-เอกสาร (Terms-Documents Matrix) เมทริกซ์คำ-เอกสารประกอบด้วย *m* แถวของคำ (Terms) และ *n* คอลัมน์ของเอกสาร (Documents) โดยที่ข้อมูลสมาชิกแต่ละตัวในเมทริกซ์ a_{ij} คือความถี่ของคำที่ *i* ภายในเอกสารที่ *j* หรือจะใช้เป็นระบบน้ำหนักของคำ

2. กระบวนการแยกค่าแบบเดี่ยว (Singular Value Decomposition) ระเบียบวิธีการแยกค่าแบบเดี่ยว (Singular Value Decomposition) หรือเขียนเป็นตัวย่อว่า SVD (Eckart & Young, 1936) เป็นวิธีการแก้ปัญหาโดยใช้ระบบสมการเชิงเส้น ซึ่งนิยมใช้ในการแก้ปัญหาทางด้านวิศวกรรมศาสตร์ และวิทยาศาสตร์ โดยอินพุตของ SVD เป็นเมทริกซ์คำ-เอกสาร (Term-Document Matrix) ผลลัพธ์ที่ได้จะเป็นความสัมพันธ์ของคำ-คำ (Word-Word) คำ-เอกสาร (Word-Document) และเอกสาร-เอกสาร (Document-Document) การคำนวณหาเมทริกซ์ A เป็นทฤษฎีของ Eckart และ Young กำหนด SVD (A) ดังแสดงในสมการที่ 2.2

$$SVD(\mathrm{A}) = U\Sigma V^T \qquad (2.2)$$

โดยที่ U คือ เมทริกซ์เชิงตั้งฉากแทนความสัมพันธ์ของคำ-คำ

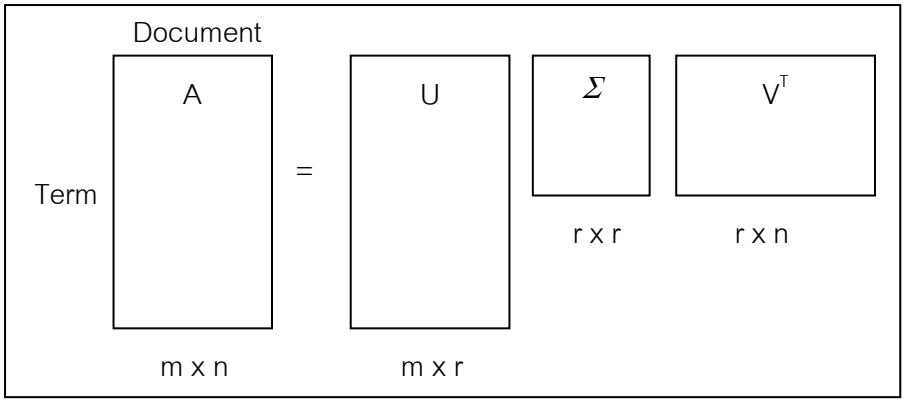
Σ คือ เมทริกซ์ทแยงมุมแทนความสัมพันธ์ของคำ-เอกสาร

V^T คือ เมทริกซ์เชิงตั้งฉากแทนความสัมพันธ์ของเอกสาร-เอกสาร และดำเนินการสับเปลี่ยนตำแหน่งตามหลักการ Transpose

การแยกค่าแบบเดี่ยวเป็นการแยกเมทริกซ์ A ซึ่งมีขนาด $m \geq n$ ออกเป็นเมทริกซ์สามเมทริกซ์ด้วยกัน ได้แก่ *U, Σ, V* ซึ่งเมทริกซ์ U มีขนาด *m x r* เมทริกซ์ Σ จะมีขนาด *r x r* และสุดท้ายเมทริกซ์ V^T มีขนาด *r x n* โดย *r* คือจำนวนลำดับ (Rank) ของเมทริกซ์ *A* ในกระบวนการแยกค่าแบบเดี่ยวของเมทริกซ์ A เขียนแทนด้วย SVD (A) ดังภาพที่ 2.7 แสดงถึงเมทริกซ์ที่ได้จากการแยกค่าแบบเดี่ยวของเมทริกซ์ A

ภาพที่ 2.7

การแทนเมทริกซ์ของการแยกค่าแบบเดี่ยวของ A ในรูปแบบภาพ



จากภาพที่ 2.7 เมทริกซ์ U คือเมทริกซ์เชิงตั้งฉาก (The Orthogonal Matrix) แทนความสัมพันธ์ของคำ-คำ ที่สามารถคำนวณค่าเวกเตอร์เจาะจง (Eigenvector) ของ AA^T เรียกว่า เวกเตอร์คำ (Term Vector) เป็นเวกเตอร์เชิงเดี่ยวทางซ้ายของเมทริกซ์ A (Left Singular Vector of A)

เมทริกซ์ Σ คือเมทริกซ์ทแยงมุม (The Diagonal Element) แทนความสัมพันธ์ของคำ-เอกสาร ที่สามารถคำนวณรากที่สองของค่าเจาะจง (Eigenvalue) ของ $A^T A$ โดยค่าที่นำมาคำนวณ คือ สมาชิกของเมทริกซ์หลักที่ r ทำการคำนวณเพื่อหาเส้นสมมติแสดงปริมาณและทิศทางแนวตั้งฉากที่มีความสัมพันธ์กันทั้งนี้สมาชิกแนวเส้นทแยงมุมของเมทริกซ์ตัวที่ 1 ถึงตัวที่ r มีค่ามากกว่าศูนย์ $\sigma_1 \geq \sigma_2 \ldots \geq \sigma_r > 0$ และสมาชิกแนวเส้นทแยงมุมของเมทริกซ์ตัวที่ r + 1 จนถึงสมาชิกของเมทริกซ์ตัวที่ n มีค่าเท่ากับศูนย์ $\sigma_{r+1} = \sigma_n = 0$ กล่าวได้ว่าเป็น ค่าแบบเดี่ยวของเมทริกซ์ A (Singular Value of A) สำหรับกระบวนการแยก (Decomposition) ของเมทริกซ์

เมทริกซ์ V คือเมทริกซ์เชิงตั้งฉาก (The Orthogonal Matrix) แทนความสัมพันธ์ของเอกสาร-เอกสาร ที่สามารถคำนวณค่าเวกเตอร์เจาะจง (Eigenvector) ของ $A^T A$ เรียกว่า เวกเตอร์เอกสาร (Document Vector) เป็นเวกเตอร์เชิงเดี่ยวทางขวาของเมทริกซ์ A (Left Singular Vector of A)

โดยค่าเมทริกซ์ V ที่ได้จากกระบวนการแยกค่าแบบเดี่ยวนั้น จะต้องทำการสับเปลี่ยนตำแหน่งตามหลักการ Transpose แล้วจึงนำเมทริกซ์ V^T ที่ได้ไปทำการคำนวณเพื่อหาค่าเมทริกซ์ A ตามสูตรดังกล่าวข้างต้น

3. **ลดขนาดของเมทริกซ์** (Reduced Rank of the Matrix) แนวคิดการแยกค่าแบบเดี่ยวนำมาใช้ในการวิเคราะห์ความหมายแอบแฝงเพื่อหามุมมองคอนเซ็ปต์หรือมุมมองความหมายจะต้องดำเนินการลดขนาดของเมทริกซ์ (k) โดย Deerwester et al. (1990) อธิบายว่าขั้นตอนการเลือกขนาดของเมทริกซ์ หรือการประมาณค่า k เพื่อลดขนาดของเมทริกซ์เป็นขั้นตอนที่สำคัญ โดยกำหนดค่า $k \leq r$ ถ้าประมาณค่า k น้อยเกินไปจะไม่สามารถพิจารณาความสัมพันธ์ทางด้านความหมายของเนื้อหาที่สำคัญ (Semantic Space) ได้อย่างชัดเจน แต่ถ้าหากทำการประมาณค่า k มากเกินไป ก็จะไม่สามารถกำจัดเนื้อหาที่ไม่สำคัญ (Noise) ออกไปได้ ดังนั้นการพิจารณาความสัมพันธ์ทางด้านความหมายก็จะไม่ชัดเจน

การลดขนาดเมทริกซ์สามารถทำได้โดยทฤษฎีของ Eckart และ Young (Eckart & Young, 1936 อ้างถึงใน Tunpaiboon, 2000, p.27) เพื่อการประมาณค่า k ที่ดี (A_k) โดยนำเมทริกซ์ U เมทริกซ์ Σ และเมทริกซ์ V^T มาสร้างเป็นเมทริกซ์ใหม่ คำนวณได้ดังสมการที่ 2.3 โดยเมทริกซ์ A_k เมทริกซ์ใหม่ที่เกิดจากการประมาณค่า k สามารถแสดงได้ดังภาพที่ 2.8

$$\{A_k\} = \{U_k\}\{\Sigma_k\}\{V_k\}^T \qquad (2.3)$$

โดยที่ U_k คือ เมทริกซ์เชิงตั้งฉากแทนความสัมพันธ์ของคำ-คำ หลังจากการลดขนาดของเมทริกซ์

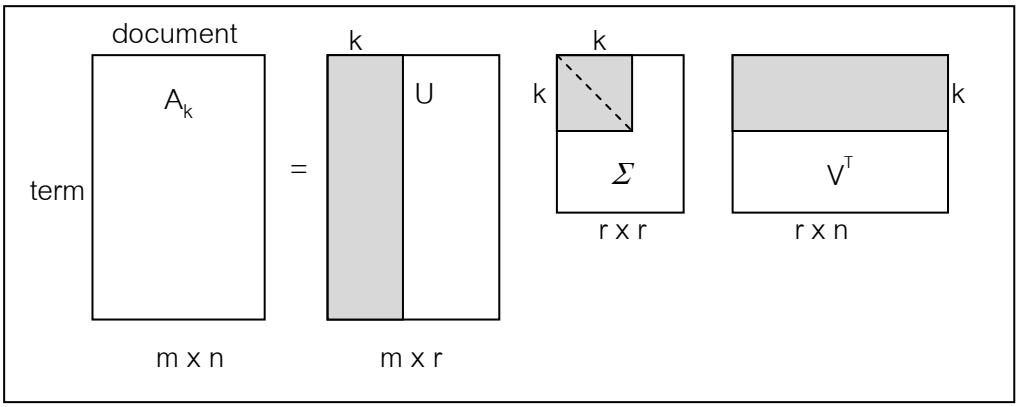
Σ_k คือ เมทริกซ์ทแยงมุมแทนความสัมพันธ์ของคำ-เอกสาร หลังจากการลดขนาดของเมทริกซ์

V_k^T คือ เมทริกซ์เชิงตั้งฉากแทนความสัมพันธ์ของคำ-เอกสาร หลังจากการลดขนาดของเมทริกซ์ ที่ดำเนินการสับเปลี่ยนตำแหน่งตามหลักการ Transpose แล้ว

ภาพที่ 2.8

แผนภาพอธิบายการสร้างเมทริกซ์ A_k จากกระบวนการแยกค่าแบบเดี่ยวหลัง

การเลือกขนาดเมทริกซ์



การประมาณค่า k เพื่อนำมาสร้างเมทริกซ์ A_k จึงจำเป็นต้องการเลือกค่า k ที่ดีและเหมาะสม จากการศึกษางานวิจัยที่เกี่ยวข้อง พบว่าการเลือกค่า k ที่เหมาะสมนั้นไม่สามารถกำหนดค่าที่แน่นอนได้ โดยค่า k จะเปลี่ยนแปลงไปตามแต่ละงานวิจัย ยกตัวอย่างเช่น งานวิจัยของ Landauer and Laham (1998) มีการกำหนดค่า k เท่ากับ 2 ส่วนงานวิจัยของ Berry and Brien (1994) มีการกำหนดค่า k เท่ากับ 2 ในขณะที่งานวิจัย ของ ศุภวรรณ เปรมกุศลชัย (2548) มีการกำหนดค่า k สำหรับบทคัดย่อวิทยานิพนธ์ฉบับที่ 1 เท่ากับ 69 บทคัดย่อวิทยานิพนธ์ฉบับที่ 2 และ 3 เท่ากับ 109

จากตัวอย่างดังกล่าว สังเกตเห็นได้ว่าขนาดเมทริกซ์ หรือค่า k นั้นมีการกำหนดค่าที่หลากหลายขึ้นอยู่กับความเหมาะสมสำหรับงานวิจัยนั้น ๆ โดยการเลือกขนาดเมทริกซ์หรือค่า k คำนวณได้จากค่าผิดพลาด (Error) จากเมทริกซ์ A_k ไปยัง เมทริกซ์ A ซึ่งเรียกค่าดังกล่าวนี้ว่าค่าเปลี่ยนแปลงความสัมพันธ์ (Relative Change) (Eckart & Young, 1936) ดังแสดงการคำนวณในสมการ 2.4

งานวิจัยที่ผ่านมาพบว่า การลดขนาดเมทริกซ์ด้วยสูตรของค่าเปลี่ยนแปลงความสัมพันธ์โดยผลลัพธ์ที่ 20% จากการคำนวณของเมทริกซ์ทแยงมุม ให้คำตอบจากการวิเคราะห์ความหมายแอบแฝงได้ดีที่สุด (Tunpaiboon, 2000; อัษฎางค์ แตงไทย, 2548; ศุภวรรณ เปรมกุศลชัย, 2548)

$$\text{Relative Change} = \frac{\|A - A_k\|_F}{\|A\|_F} \qquad (2.4)$$

โดยที่ $\|A - A_k\|_F = \min_{\text{rank(X)} \le \text{k}} \|A - X\|_F = \sqrt{\sigma_{k+1}^2 + \sigma_{k2}^2 \ldots + \sigma_{r_A}^2}$

$\|A\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_{r_A}^2}$

A คือ เมทริกซ์ Σ หรือเมทริกซ์ทแยงมุม

σ คือ ค่าตัวเลขที่เป็นสมาชิกในแนวทแยงมุมของเมทริกซ์ Σ

X, k คือ ตำแหน่งลำดับที่ของค่าตัวเลขที่เป็นสมาชิกในแนวทแยงมุมของเมทริกซ์ Σ

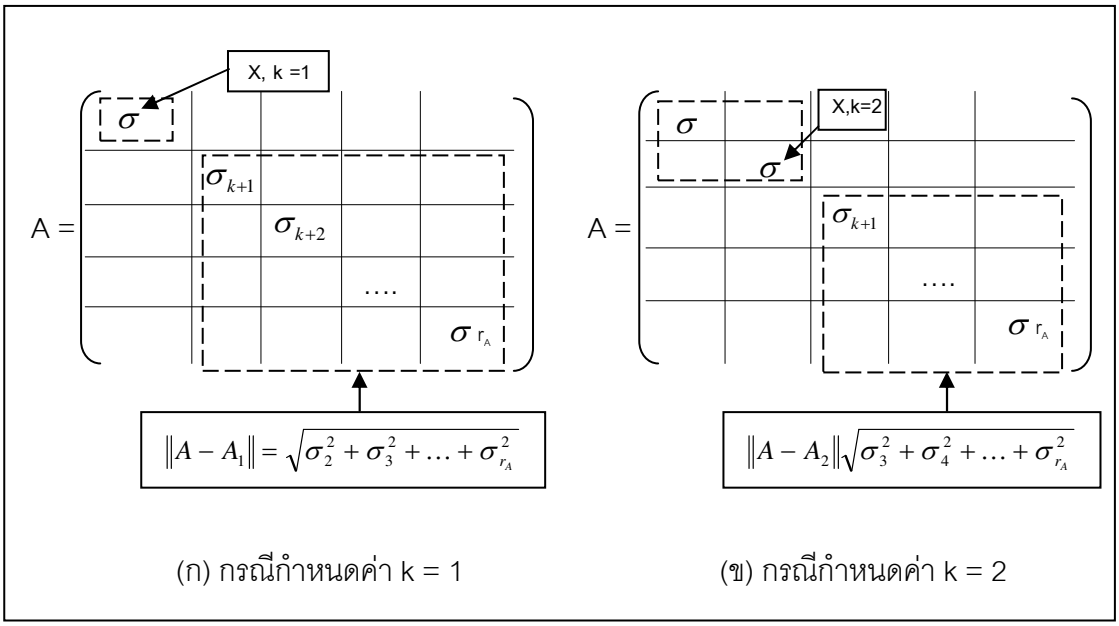
r_A คือ ค่าตัวเลขที่เป็นสมาชิกตัวสุดท้ายในแนวทแยงมุมของเมทริกซ์ Σ

สามารถอธิบายการแทนค่าแต่ละตัวในสมการได้ดังภาพที่ 2.9 โดยรูปภาพ (ก) กรณีกำหนดค่า k = 1 และรูปภาพ (ข) กรณีกำหนดค่า k = 2

ภาพที่ 2.9

ตัวอย่างแสดงการแทนค่าแต่ละตัวในสมการของสูตรหาค่าเปลี่ยนแปลงความสัมพันธ์

(ก) กรณีกำหนดค่า k = 1 และ (ข) กรณีกำหนดค่า k = 2



(ก) กรณีกำหนดค่า k = 1 (ข) กรณีกำหนดค่า k = 2

ตัวอย่างแสดงการวิเคราะห์ของเทคนิคการวิเคราะห์ความหมายแอบแฝง ซึ่งตัวอย่างนี้ได้เลือกใช้ประโยคที่มีการบันทึกข้อความทางด้านเทคนิคทั้งหมด 9 ข้อความโดยเป็นข้อความทางด้านการติดต่อสื่อสารระหว่างมนุษย์กับคอมพิวเตอร์ (Human Computer Interface: HCI) จำนวนห้าข้อความ และข้อความทางด้านทฤษฎีกราฟที่เกี่ยวกับคณิตศาสตร์ (Mathematical Graph Theory) จำนวนสี่ข้อความ ซึ่งจะสังเกตเห็นได้ว่าทั้งสองหัวข้อไม่มีความเกี่ยวข้องกันทางด้านหลักการ ดังแสดงรายละเอียดของข้อความในตารางที่ 2.4

ตารางที่ 2.4
แสดงข้อความบันทึกความรู้ทางด้านเทคนิค

ตัวอย่างข้อมูล ข้อความเกี่ยวกับบันทึกความรู้ทางด้านเทคนิค
c1 : *Human* machine *interface* for ABC *computer* applications
c2 : A *survey* of *user* opinion of *computer system* response time
c3 : The *EPS user* interface management *system*
c4 : *System* and *human system* engineering testing of *EPS*
c5 : Relation of *user* perceived *response time* to error measurement
m1: The generation of random, binary, ordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

ที่มา: “An Introduction to Latent Semantic Analysis,” โดย Landauer, Foltz, and Laham ,1998 , Discourse Processes, p. 265

จากตารางที่ 2.4 นำมาสร้างเมทริกซ์คำ-เอกสารจะมีขนาด 12 x 9 ประกอบด้วยแถวของคำ (Terms) เกิดจากตัวแทนเอกสารหรือคำสำคัญซึ่งนับได้ทั้งหมด 12 คำไม่ซ้ำกันโดยเป็นคำที่ปรากฏขึ้นในข้อความอย่างน้อยสองข้อความ และจำนวนเอกสาร (Documents) มีทั้งหมด 9 เอกสาร โดยข้อมูลสมาชิกแต่ละตัวในเมทริกซ์ a_{ij} คือความถี่ของคำในเอกสารที่นับได้ ดังแสดงในตารางที่ 2.5

ตารางที่ 2.5

แสดงเมทริกซ์ของคำ-เอกสาร

$\{A\}$ =

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1
r(human.user) = -0.38 r(human.minors) = -0.29									

ที่มา: “An Introduction to Latent Semantic Analysis,” โดย Landauer, Foltz, and Laham, 1998 , Discourse Processes, p. 265

จากนั้นระบบก็จะดำเนินการแยกค่าเชิงเส้น (Linear Decomposition) ผ่านกระบวนการแยกค่าแบบเดี่ยว (Singular Value Decomposition) แล้วให้ผลลัพธ์เป็นเมทริกซ์ดังแสดงในตารางที่ 2.6

ตารางที่ 2.6

แสดงเมทริกซ์ผลลัพธ์ที่ได้จากการดำเนินการผ่านกระบวนการแยกค่าแบบเดี่ยว

{U}=								
0.22	-0.11	0.29	-0.41	-0.11	-0.34	0.52	-0.06	-0.41
0.20	-0.07	0.14	-0.55	0.28	0.50	-0.07	-0.01	-0.11
0.24	0.04	-0.16	-0.59	-0.11	-0.25	-0.30	0.06	0.49
0.40	0.06	-0.34	0.10	0.33	0.38	0.00	0.00	0.01
0.64	-0.17	0.36	0.33	-0.16	-0.21	-0.17	0.03	0.27
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.27	0.11	-0.43	0.07	0.08	-0.17	0.28	-0.02	-0.05
0.30	-0.14	0.33	0.19	0.11	0.27	0.03	-0.02	-0.17
0.21	0.27	-0.18	-0.03	-0.54	0.08	-0.47	-0.04	-0.58
0.01	0.49	0.23	0.03	0.59	-0.39	-0.29	0.25	-0.23
0.04	0.62	0.22	0.00	-0.07	0.11	0.16	-0.68	0.23
0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18
{Σ}=								
3.34								
	2.54							
		2.35						
			1.64					
				1.5				
					1.31			
						0.85		
							0.56	
								0.36

ตารางที่ 2.6 (ต่อ)

แสดงเมทริกซ์ผลลัพธ์ที่ได้จากการดำเนินการผ่านกระบวนการแยกค่าแบบเดี่ยว

$\{V\}^T$								
0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53
0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25	0.08
-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01	-0.03
0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15	-0.60
-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00	0.36
0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.15	0.25	0.04
-0.01	0.05	0.01	-0.02	-0.06	0.45	-0.76	0.45	-0.07
-0.06	0.24	0.02	-0.08	-0.26	-0.62	0.02	0.52	-0.45

ที่มา : “An Introduction to Latent Semantic Analysis,” โดย Landauer, Foltz, and Laham, 1998 , Discourse Processes, p. 266

หลังจากการดำเนินการแยกค่าเชิงเส้น (Linear Decomposition) ที่ได้คือเมทริกซ์ U เมทริกซ์ Σ และเมทริกซ์ V^T ทั้งนี้สามารถสร้างเมทริกซ์เริ่มต้น คือเมทริกซ์ A ได้ด้วยสมการ 2.2 แต่เมทริกซ์ A ที่ได้มานั้นยังไม่สามารถหาความสัมพันธ์ของเนื้อหาที่ชัดเจนได้ จึงจำเป็นต้องเลือกขนาดเมทริกซ์หรือลดค่าขนาดของเมทริกซ์ เพื่อประมาณค่า k ที่เหมาะสม จากตัวอย่างข้อมูลในตารางที่ 2.6 ประมาณค่า k ที่เหมาะสมด้วยการคำนวณด้วยสูตรการหาค่าเปลี่ยนแปลงความสัมพันธ์ตามสมการที่ 2.4 ซึ่งแสดงเมทริกซ์ที่ได้หลังจากการคำนวณกำหนดค่า k = 2 ดังภาพที่ 2.10

ภาพที่ 2.10

แสดงเมทริกซ์ที่ได้หลังจากการเลือกขนาดเมทริกซ์หรือค่า k = 2

$\{U_k\}$ =

0.22	-0.11
0.20	-0.07
0.24	0.04
0.40	0.06
0.64	-0.17
0.27	0.11
0.27	0.11
0.30	-0.14
0.21	0.27
0.01	0.49
0.04	0.62
0.03	0.45

$\{\Sigma_k\}$=

3.34	
	2.54

$\{V_k^T\}$ =

0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02	0.08
-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62	0.53

จากนั้นนำเมทริกซ์ U_k เมทริกซ์ Σ_k เมทริกซ์ V_k^T มาสร้างเป็นเมทริกซ์ใหม่ A_k โดยสามารถคำนวณได้จากสมการที่ 2.3 โดยผลจากการคำนวณเพื่อสร้างเมทริกซ์ A_k แสดงได้ดังตารางที่ 2.7

ตารางที่ 2.7

แสดงผลลัพธ์จากการคำนวณเพื่อสร้างเมทริกซ์ A_k หลักจากลดขนาดเมทริกซ์ค่า k = 2

$\{A_k\}$ =	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	0.16	0.40	0.38	0.47	0.18	-0.05	-0.12	-0.16	-0.09
interface	0.14	0.37	0.33	0.40	0.17	-0.03	-0.07	-0.10	-0.04
computer	0.15	0.51	0.36	0.41	0.24	0.02	0.06	0.09	0.12
user	0.26	0.84	0.61	0.70	0.39	0.03	0.08	0.12	0.19
system	0.45	1.23	1.05	1.27	0.56	-0.07	-0.15	-0.21	-0.05
response	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
time	0.16	0.58	0.38	0.42	0.28	0.06	0.13	0.19	0.22
EPS	0.22	0.55	0.51	0.63	0.24	-0.07	-0.14	-0.20	-0.11
survey	0.10	0.53	0.23	0.21	0.27	0.14	0.31	0.44	0.43
trees	-0.06	0.23	-0.14	-0.27	0.14	0.24	0.55	0.77	0.66
graph	-0.06	0.34	-0.15	-0.30	0.20	0.31	0.69	0.98	0.85
minors	-0.04	0.25	-0.10	-0.21	0.15	0.22	0.50	0.71	0.62
r(human.user) = 0.94 r(human.minors)= -0.83									

ที่มา : “An Introduction to Latent Semantic Analysis,” โดย Landauer, Foltz, and Laham, 1998 , Discourse Processes, p. 267

หลังจากทำการเลือกขนาดเมทริกซ์ (ค่า k) และคำนวณหาค่าความคล้ายคลึงระหว่างเอกสารด้วยสูตรการหาค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (r) พบว่า ค่าความคล้ายคลึง ของ human กับ user เป็น 0.94 แสดงให้เห็นว่า human กับ user มีความหมายไปในทิศทางเดียวกัน ถึงแม้ว่าจะไม่ได้อยู่ในหัวข้อเดียวกัน ส่วน human กับ minors ค่าความคล้ายคลึงเป็น -0.83 แสดงให้เห็นว่าไม่มีความเกี่ยวข้องกัน พบข้อน่าสังเกต คือ คำว่า trees กับ survey ในคอลัมน์ m4 พบว่า trees ไม่ปรากฏในหัวข้อ m4 (หัวข้อ m4 มีแต่ graph กับ minors) ก่อนทำการวิเคราะห์ความหมายแอบแฝง ค่า trees มีค่าเท่ากับ 0 แต่เมื่อทำแล้วค่า 0 ถูกแทนที่ด้วย 0.66 ในทางตรงกันข้าม survey ปรากฏในหัวข้อ m4 แต่กลับมีค่า 0.42 ซึ่งแสดงให้เห็นว่า ความถี่ของ

จำนวนคำนั้นที่ปรากฏในหัวข้อ ไม่ได้แสดงถึงความสัมพันธ์คำนั้นกับหัวข้อ การวิเคราะห์ความหมายแอบแฝงอาศัยคำที่อยู่รอบข้างหรือบริบทมาช่วยในการนิยามความหมายของคำ

2.2.1.5 **การหาค่าความคล้ายคลึง** (Similarity Measure)

การคำนวณค่าความคล้ายคลึง (Similarity Measure) ระหว่างเอกสาร สูตรที่นิยมใช้ (Korfhage, 1997) คือ ค่าความคล้ายคลึงแบบโคไซน์ (Cosine) (Salton & McGill,1983) โดยมีสูตรดังนี้

$$\text{Cosine}\theta\ D_i = \text{Sim}(Q,D_i) \qquad (2.5)$$

โดยที่ $$\text{Sim}(Q,D_i) = \frac{Q \cdot D_i}{|Q| * |D_i|} = \frac{\sum_i w_{Q,j} \cdot w_{D_i,j}}{\sqrt{\sum_j w_{Q,j}^{\ 2}} \cdot \sqrt{\sum_i w_{D_i,j}^{\ 2}}}$$

$Q \cdot D_i$ คือ ผลคูณของเวกเตอร์ระหว่างเอกสาร Q กับ D_i

$|Q| * |D_i|$ คือ ผลรวมค่าในเวกเตอร์ของเอกสาร Q คูณกับ ผลรวมค่าในเวกเตอร์ของเอกสาร D_i

w คือ เวกเตอร์ของเอกสาร

i คือ ลำดับที่เอกสาร

j คือ ลำดับที่เวกเตอร์ของเอกสาร

ภาพที่ 2.11

ตัวอย่างแสดงการแทนค่าแต่ละตัวในสมการของสูตรการหาค่าความคล้ายคลึงของเอกสารแบบโคไซน์

	w_1	w_2	w_3	....	w_j
Q	$w_{Q,1}$	$w_{Q,2}$	$w_{Q,3}$	$w_{Q,...}$	$w_{Q,j}$
	w_1	w_2	w_3	....	w_j
D_1	$w_{D_1,1}$	$w_{D_1,2}$	$w_{D_1,3}$	$w_{D_1,...}$	$w_{D_1,j}$
D_2		$w_{D_2,2}$			
D_3			$w_{D_3,3}$		
D..				$w_{D_4,...}$	
D_i	$w_{D_i,1}$	$w_{D_i,2}$	$w_{D_i,3}$	$w_{D_i,4}$	$w_{D_i,j}$

ตัวอย่างการคำนวณค่าความคล้ายคลึงแบบโคไซน์ โดยมีข้อมูลเอกสารหลัก Q ที่ต้องการหาค่าความคล้ายคลึงระหว่างเอกสารอื่น D_i จำนวนสามเอกสาร ซึ่งแต่ละเอกสารมีค่าเวกเตอร์ตัวแทนเอกสารดังแสดงตัวอย่างในตารางที่ 2.8

ตารางที่ 2.8

แสดงตัวอย่างการคำนวณค่าความคล้ายคลึงของเอกสารแบบโคไซน์

	w_1	w_2	w_3
Q	2.29	0.34	2.03
D_1	4.44	0.89	4.69
D_2	-0.86	1.3	1.22
D_3	1.06	1.77	6.08

$$Q.D_1 = \frac{(2.29 \cdot 4.44) + (0.34 \cdot 0.89) + (2.03 \cdot 4.69)}{\sqrt{(2.29^2 + 0.34^2 + 2.03^2)} \cdot \sqrt{(4.44^2 + 0.89^2 + 4.69^2)}} = \frac{19.991}{3.079 \cdot 6.519} = 0.996$$

$$Q.D_2 = \frac{(2.29 \cdot -0.86) + (0.34 \cdot 1.3) + (2.03 \cdot 1.22)}{\sqrt{(2.29^2 + 0.34^2 + 2.03^2)} \cdot \sqrt{(-0.86^2 + 1.3^2 + 1.22^2)}} = \frac{0.949}{3.079 \cdot 1.979} = 0.156$$

$$Q.D_3 = \frac{(2.29 \cdot 1.06) + (0.34 \cdot 1.77) + (2.03 \cdot 6.08)}{\sqrt{(2.29^2 + 0.34^2 + 2.03^2)} \cdot \sqrt{(1.06^2 + 1.77^2 + 6.08^2)}} = \frac{15.372}{3.079 \cdot 6.421} = 0.778$$

ผลที่ได้จากสมการที่ 2.5 จะมีค่าอยู่ระหว่าง 0 ถึง 1 โดยผลลัพธ์ใกล้เคียง 1 แสดงให้เห็นว่าเอกสารสองฉบับนั้นมีความคล้ายคลึงกันมาก แต่ถ้าใกล้เคียง 0 แสดงให้เห็นว่าเอกสารสองฉบับนั้นมีความคล้ายคลึงกันน้อยหรือไม่มีความคล้ายกัน จากตัวอย่างผลการคำนวณค่าความคล้ายคลึงระหว่างเอกสาร Q กับ D_i แสดงให้เห็นว่า เอกสาร Q กับ D_1 มีค่าความคล้ายคลึงกันมากซึ่งผลจากการคำนวณมีค่าเท่ากับ 0.996 ในขณะที่ค่าความคล้ายคลึงของเอกสาร Q กับ D_2 ผลจากการคำนวณมีค่าเท่ากับ 0.156 แสดงให้เห็นว่ามีความคล้ายคลึงกันน้อย แต่อย่างไรก็ตามผลจากการคำนวณค่าความคล้ายคลึงแบบโคไซน์ไม่สามารถบอกได้ว่าความคล้ายคลึงของเอกสารนั้นมีความสัมพันธ์เป็นไปในทิศทางเดียวกันหรือทิศทางตรงกันข้าม หากต้องการหาค่าความคล้ายคลึงที่สามารถบอกทิศทางความสัมพันธ์ได้ควรใช้สูตรการหาค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson's Correlation) ดังสมการที่ 2.6

$$r_{xy} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}} \quad (2.6)$$

โดยที่ r_{xy} คือ ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน

n คือ จำนวนเวกเตอร์ของเอกสาร

x คือ เวกเตอร์ของเอกสารชุดที่ 1

y คือ เวกเตอร์ของเอกสารชุดที่ 2

ผลที่ได้จากสมการที่ 2.6 นั้นมีค่าอยู่ระหว่าง -1 ถึง 1 โดยผลลัพธ์ที่ได้สามารถแบ่งออกเป็นสองลักษณะ คือ ค่าอยู่ระหว่าง 0 ถึง 1 ความสัมพันธ์ที่ได้จะเป็นไปในทิศทางเดียวกัน ส่วนค่าที่อยู่ระหว่าง 0 ถึง -1 ความสัมพันธ์ที่ได้จะเป็นไปในทิศทางตรงข้ามกัน ตัวอย่างการคำนวณค่าความคล้ายคลึงแบบเพียร์สัน โดยมีข้อมูลเอกสารหลัก (x) ที่ต้องการหาค่าความคล้ายคลึงระหว่างเอกสารอื่น (y) จำนวนสามเอกสาร ซึ่งแต่ละเอกสารมีค่าเวกเตอร์ตัวแทนเอกสารจำนวนสามเวกเตอร์ ดังแสดงตัวอย่างในตารางที่ 2.9

ตารางที่ 2.9

แสดงตัวอย่างการคำนวณค่าความคล้ายคลึงของเอกสารแบบเพียร์สัน

	n_1	n_2	n_3	Σx	Σx^2	$(\Sigma x)^2$
x	2.29	0.34	2.03	4.66	$(2.29^2+0.34^2+2.03^2)= 9.48$	21.72
	n_1	n_2	n_3	Σy	Σy^2	$(\Sigma y)^2$
y_1	4.44	0.89	4.69	10.02	$(4.44^2+0.89^2+4.69^2)=42.50$	100.40
y_2	-0.86	1.3	1.22	1.66	$(-0.86^2+1.3^2+1.22^2)=3.92$	2.76
y_3	1.06	1.77	6.08	8.91	$(1.06^2+1.77^2+6.08^2)=41.22$	79.39

$$x, y_1 = \frac{(3*19.99) - (4.66*10.02)}{\sqrt{(3*9.48) - 21.72}\cdot\sqrt{(3*42.50) - 100.40}} = \frac{13.28}{2.59\cdot 5.21} = 0.983$$

$$x, y_2 = \frac{(3*0.95) - (4.66*1.66)}{\sqrt{(3*9.48) - 21.72}\cdot\sqrt{(3*3.92) - 2.76}} = \frac{-4.89}{2.59\cdot 3.00} = -0.209$$

$$x, y_3 = \frac{(3*15.37) - (4.66*8.91)}{\sqrt{(3*9.48) - 21.72}\cdot\sqrt{(3*41.22) - 79.39}} = \frac{4.59}{2.59\cdot 6.65} = 0.266$$

จากตัวอย่างการคำนวณหาความคล้ายคลึงของเอกสารแบบเพียร์สัน ค่าความคล้ายคลึงระหว่างเอกสาร x กับ y_1 แสดงให้เห็นว่ามีความคล้ายคลึงเป็นไปในทิศทางเดียวกันอยู่ในระดับมาก ขณะที่ค่าความคล้ายคลึงระหว่างเอกสาร x กับ y_2 มีค่าความคล้ายคลึงเท่ากับ -2.209 เนื่องจากเป็นค่าติดลบความคล้ายคลึงระหว่างเอกสารจึงเป็นไปในทิศทางตรงกันข้าม

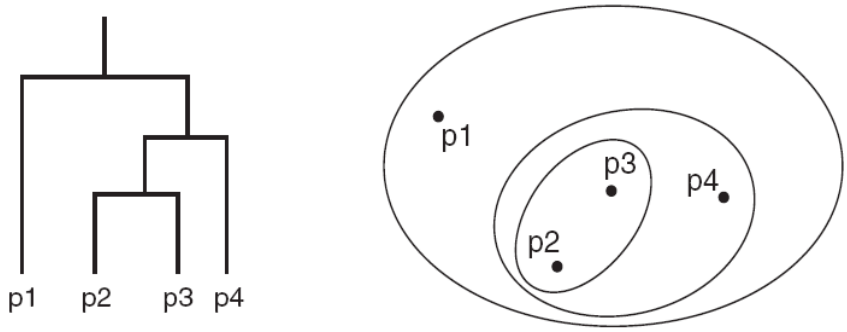
2.2.2 **เทคนิคการจัดกลุ่มข้อความ** (Text Clustering)

การจัดกลุ่มข้อความ คือ การรวมกลุ่มของข้อความที่มีลักษณะเนื้อหาความหมายคล้ายคลึงกันหรือเหมือนกันจะถูกจัดไว้ในกลุ่มเดียวกัน โดยเริ่มจากการหาตัวแทนของกลุ่ม จากนั้นทำการเปรียบเทียบข้อความกับตัวแทนของแต่ละกลุ่ม ถ้าข้อความคล้ายคลึงกับตัวแทนของกลุ่มไหนก็จะถูกจัดให้อยู่กลุ่มนั้น โดยใช้แนวคิดของทฤษฎีการจัดกลุ่ม (Clustering Algorithm) ในการจัดกลุ่มข้อความ ซึ่งการจัดกลุ่มแบ่งออกเป็น 2 ประเภทใหญ่ ๆ (Jain, Murty, & Flynn, 1999) คือ การจัดกลุ่มแบบโครงสร้างลำดับชั้น (Hierarchical Clustering) และ การจัดกลุ่มแบบแบ่งส่วน (Partitional Clustering) โดยมีรายละเอียดดังนี้

2.2.2.1 **การจัดกลุ่มแบบโครงสร้างลำดับชั้น** (Hierarchical Clustering)

การจัดกลุ่มแบบโครงสร้างลำดับชั้นเป็นการจัดแบ่งกลุ่มออกเป็นกลุ่มย่อย ที่มีความสัมพันธ์กันระหว่างกลุ่มใหญ่และกลุ่มย่อย ๆ ลงไป (Tan, Steinbach, & Kumar, 2006) ดังแสดงตัวอย่างในภาพที่ 2.12 โดยรูปภาพ (ก) แสดงภาพของการจัดกลุ่มโครงสร้างลำดับในมุมมองแบบลำดับชั้น (Dendrogram) (ข) แสดงภาพของการจัดกลุ่มโครงสร้างลำดับในมุมมองแบบภาพทับซ้อน (Nested Cluster Diagram)

ภาพที่ 2.12
แสดงภาพของการจัดกลุ่มแบบโครงสร้างลำดับชั้นแบ่งตามลักษณะมุมมอง
(ก) แบบลำดับชั้น และ (ข) แบบภาพทับซ้อน



(ก) แบบลำดับชั้น (Dendrogram) (ข) แบบภาพทับซ้อน (Nested Cluster Diagram)

ที่มา: “Cluster Analysis Basic Concepts and Algorithms,” โดย Tan, Steinbach, and Kumar, ,2006, Pearson-Addison Wesley, pp 516

วิธีการจัดกลุ่มแบบโครงสร้างลำดับชั้นมีอยู่สองวิธีคือ การจัดกลุ่มจากล่างขึ้นบน (Agglomerative, Bottom-Up) และการจัดกลุ่มจากบนลงล่าง (Divisive, Top-Down) โดยวิธีการวัดค่าความเหมือนของทั้งสองวิธีมีหลักการทำงานเหมือนกัน แต่วิธีการจัดกลุ่มแบบบนลงล่างไม่เป็นที่นิยม (ชุลีรัตน์ จรัสกุลชัย และคณะ, 2544) เนื่องจากใช้เวลาในการคำนวณนานกว่า (อรนุช ชัยหมื่น, 2548)

การจัดกลุ่มแบบโครงสร้างมีขั้นตอนการทำงานดังนี้

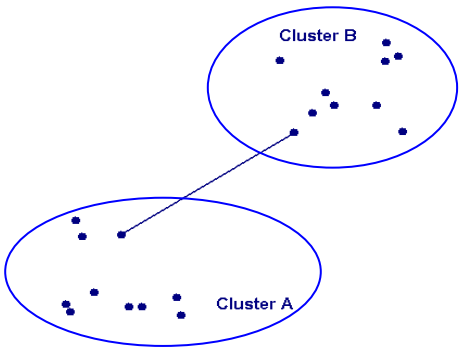
1. กำหนดจำนวนกลุ่ม โดยให้ข้อความหนึ่งตัวต่อหนึ่งกลุ่ม กล่าวคือจำนวนกลุ่มเริ่มต้นมีเท่ากับจำนวนข้อมูล เช่น ถ้ามีข้อมูลอยู่เป็นจำนวน 15 ตัว จำนวนกลุ่มเริ่มต้นจะมีค่าเท่ากับ 15 กลุ่ม

2. ทุกกลุ่มข้อมูลจะทำการคำนวณหาระยะห่างระหว่างตัวข้อความกับกลุ่ม

3. จัดกลุ่มใหม่ โดยพิจารณาจากระยะห่างระหว่างข้อมูลที่ได้จากการคำนวณในขั้นตอน 2 แล้ว จากนั้นทำซ้ำในขั้นตอนที่ 2 และ 3 จนกระทั่งการรวมกลุ่มเหลือเพียงกลุ่มเดียว โดยการคำนวณหาระยะห่างระหว่างข้อความที่นิยมศึกษาวิจัย (Ellen, 1986 อ้างถึงใน ชุลีรัตน์ จรัสกุลชัย และคณะ, 2544; ทินกร คุณาสิทธิ์, 2551) มี 3 วิธีคือ

วิธีที่ 1 การเชื่อมโยงแบบเดี่ยว (Single Link) ในการจัดกลุ่มโดยคำนวณจากค่าระยะห่างของสมาชิกระหว่างกลุ่มตัวที่อยู่ใกล้กันมากที่สุด ดังแสดงในภาพที่ 2.13

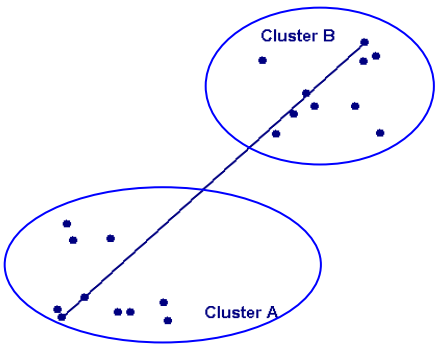
ภาพที่ 2.13
ลักษณะการหาระยะทางการเชื่อมโยงแบบเดี่ยว



ที่มา: “Hierarchical Clustering” โดย Resampling Stats,Inc. (สืบค้น 15 ตุลาคม 2553) จาก http://www.resample.com/xlminer/help/HClst/HClst_intro.htm

วิธีที่ 2 การเชื่อมโยงแบบสมบูรณ์ (Complete Link) วิธีนี้จะตรงกันข้ามกับการเชื่อมโยงแบบเดี่ยว กล่าวคือ การหาระยะทางระหว่างกลุ่มโดยคำนวณจากค่าระยะห่างของสมาชิกระหว่างกลุ่มที่ห่างกันมากที่สุดแสดงดังภาพที่ 2.14

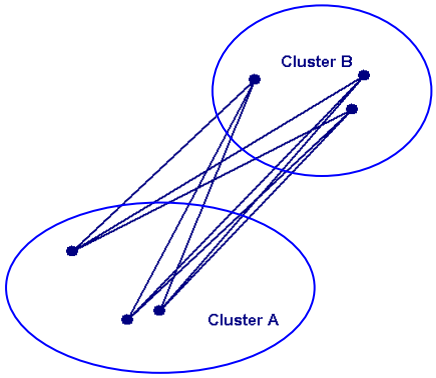
ภาพที่ 2.14
ลักษณะการหาระยะทางการเชื่อมโยงแบบสมบูรณ์



ที่มา: “Hierarchical Clustering” โดย Resampling Stats,Inc. (สืบค้น 15 ตุลาคม 2553) จาก http://www.resample.com/xlminer/help/HClst/HClst_intro.htm

วิธีที่ 3 การเชื่อมโยงแบบค่าเฉลี่ยของกลุ่ม (Group Average Link) เป็นวิธีการผสมการเชื่อมโยงแบบดี่ยวกับการเชื่อมโยงแบบสมบูรณ์ โดยการคำนวณหาระยะห่างระหว่างกลุ่มจากการหาค่าเฉลี่ยระยะห่างระหว่างกลุ่มของสมาชิกทั้งหมด แสดงดังภาพที่ 2.15

ภาพที่ 2.15
ลักษณะการหาระยะทางการเชื่อมโยงแบบค่าเฉลี่ยของกลุ่ม



ที่มา: “Hierarchical Clustering” โดย Resampling Stats,Inc. (สืบค้น 15 ตุลาคม 2553) จาก http://www.resample.com/xlminer/help/HClst/HClst_intro.htm

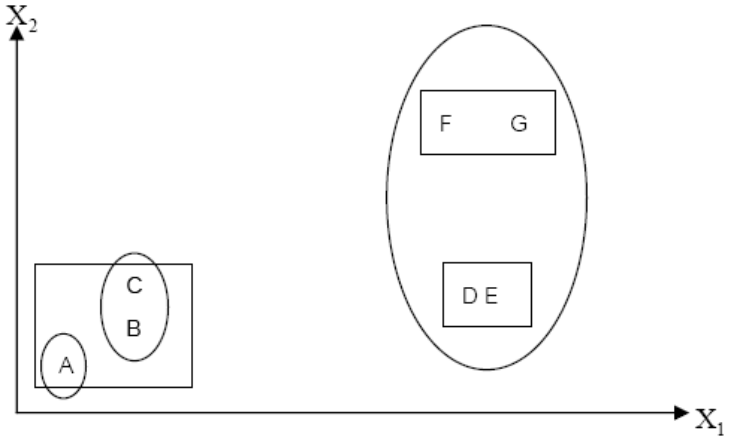
การจัดกลุ่มแบบโครงสร้างลำดับชั้น มีข้อได้เปรียบ คือ สามารถดำเนินการจัดกลุ่มได้แบบอัตโนมัติโดยผู้ใช้ไม่ต้องกำหนดค่าจำนวนกลุ่ม ซึ่งผลการจัดการจัดกลุ่มจะขึ้นอยู่กับลักษณะการเชื่อมโยงที่ใช้ แต่อย่างไรก็ตามขั้นตอนการจัดกลุ่มแบบโครงสร้างลำดับชั้นเปลืองทรัพยากรในเรื่องของการจัดเก็บข้อมูล (Tan, Steinbach, & Kumar, 2006)

2.2.2.2 การจัดกลุ่มแบบแบ่งส่วน (Partitional Clustering)

การจัดกลุ่มแบบแบ่งส่วนมีหลักการทำงานคือแยกหมวดหมู่ข้อความออกเป็นกลุ่มย่อย ๆ ดังแสดงในภาพที่ 2.16 โดยกลุ่มที่จัดได้ในแต่ละกลุ่มจะมีสมาชิกอยู่อย่างน้อยหนึ่งข้อความและหนึ่งข้อความจะถูกจัดให้อยู่ในกลุ่มได้เพียงแค่กลุ่มเดียวเท่านั้น

ภาพที่ 2.16

การจัดกลุ่มแบบแบ่งส่วน



ที่มา: “Data Clustering: A Review,” โดย Jain, Murty, and Flynn, 1999 , ACM Computing Surveys, 31 (3), 279.

การจัดกลุ่มแบบแบ่งส่วนมี K-Means, Fuzzy C-Means และ Bisecting K-Means แต่ละข้อมีรายละเอียดและขั้นตอนที่แตกต่างกันดังต่อไปนี้

1. **เทคนิคการจัดกลุ่มแบบเค-มีน** (K-Means) มีการพัฒนาและนำเสนอโดย Mac Queen ในปี 1967 เป็นอัลกอริธึมที่จัดกลุ่มให้สมาชิกภายในกลุ่ม จะมีระยะใกล้จุดศูนย์กลางหรือตัวแทนของกลุ่ม (Mean) โดยขั้นตอนวิธีการจัดกลุ่มแบบเค-มีนนั้นประกอบด้วย การกำหนดจำนวนกลุ่มเริ่มต้น กำหนดตัวแทนกลุ่ม การจัดข้อความหรือข้อมูลแต่ละตัวเข้ากลุ่ม และสุดท้าย คือ การปรับปรุงตัวแทนกลุ่มในแต่ละกลุ่ม ปัญหาที่พบ คือ การกำหนดจำนวนจุดเริ่มต้น ซึ่งส่งผลต่อประสิทธิภาพของการจัดกลุ่ม หรือการที่กลุ่มบางกลุ่มนั้นมีจำนวนสมาชิกน้อยเกินหรืออาจจะไม่มีสมาชิกในกลุ่มเลย จึงได้มีการกำหนดกฎเกณฑ์ว่ากลุ่มที่จะอยู่ในรอบถัดไปนั้น จะต้องมีสมาชิกอยู่ไม่น้อยกว่าค่าคงที่ค่าหนึ่งที่กำหนดขึ้นมา โดยหลักการของอัลกอริธึมเค-มีน (Tan, 2006) มีขั้นตอนดังนี้

1. กำหนดจำนวนกลุ่มที่ต้องการ โดยกำหนดให้ K เท่ากับจำนวนกลุ่มที่ต้องการแบ่งกลุ่มและเลือกจุดศูนย์กลางของกลุ่มเริ่มต้น

2. คำนวณตัวแทนกลุ่มหรือจุดศูนย์กลางของกลุ่ม (Centroid) ในแต่ละกลุ่ม

3. จัดกลุ่มข้อความใหม่โดยพิจารณาจากค่าความใกล้ชิดหรือระยะห่างของข้อความในกลุ่มกับตัวแทนของกลุ่มต่าง ๆ ว่าข้อความนั้นสมควรจะอยู่กลุ่มใด

4. ทำการปรับปรุงการจัดกลุ่มโดยย้อนกลับไปทำข้อ 2 และหยุดเมื่อข้อความสมาชิกในกลุ่มแต่ละกลุ่มนั้นไม่มีการเปลี่ยนแปลงแล้ว

2. **เทคนิคการจัดกลุ่มแบบฟัซซี่ซี-มีน** (Fuzzy C-Means) เป็นการแบ่งกลุ่มที่พัฒนาเพิ่มเติมจากเค-มีน โดยนำการคำนวณคณิตศาสตร์แบบฟัซซี่มาใช้ในการจัดกลุ่ม ซึ่ง Dunn (1973) ได้นำเสนอวิธีการจัดกลุ่มด้วยฟัซซี่ซี-มีน และ Bezdek (1984) นำมาพัฒนาต่อโดยขั้นตอนการจัดกลุ่มแบบฟัซซี่ซี-มีนของ Bezdek เป็นที่ได้รับความนิยมอย่างมาก (Albayrak & Amasyali, 2003; วีระนุช นิ่มเงิน, 2548) ในการนำมาประยุกต์ใช้จนกระทั้งในปัจจุบันโดยมีการประยุกต์ใช้ขั้นตอนการจัดกลุ่มแบบฟัซซี่ซี-มีนกับงานด้านต่าง ๆ เช่น การรู้จำแบบ (Pattern Recognition) การประมวลผลภาพ (Image Processing) รวมถึงการทำเหมืองข้อมูล (Data Mining)

การจัดกลุ่มแบบฟัซซี่ซี-มีนเป็นการจัดกลุ่มข้อมูลที่ข้อความหนึ่งตัวสามารถเป็นสมาชิกของกลุ่มได้หลายกลุ่มพร้อมกัน โดยหลักการ คือ ศูนย์กลางของกลุ่มข้อมูล (C_i) เป็นค่าตัวแทนเพียงค่าเดียวของกลุ่ม โดยศูนย์กลางถูกกำหนดขึ้นด้วยวิธีการสุ่มค่าและในขั้นตอนการทำงานจะมีการปรับศูนย์กลางใหม่ทุกครั้งที่มีการคำนวณค่าระดับความเป็นสมาชิกของชุดข้อมูล โดยค่าระดับความเป็นสมาชิก (U_{ij}) จะแสดงระดับความใกล้ชิดระหว่างชุดข้อมูลที่จะนำมาจัดกลุ่ม ซึ่งระดับความเป็นสมาชิกจะมีค่าอยู่ระหว่าง 0 ถึง 1 มีขั้นตอนการจัดกลุ่มดังนี้

1.กำหนดจำนวนกลุ่มที่ต้องการจัดเป็น C กลุ่ม เลือกศูนย์กลางของกลุ่มข้อมูลด้วยวิธีการสุ่มค่า

2. คำนวณค่าความเป็นสมาชิกของข้อมูลทุกตัว ด้วยสมการดังต่อไปนี้

$$U_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{(m-1)}}} \qquad (2.7)$$

โดยที่ U_{ij} คือ ค่าความเป็นสมาชิกของข้อมูล

x คือ ข้อมูลที่ยังไม่ได้จัดเข้ากลุ่ม

c คือ ค่าศูนย์กลางของกลุ่ม

m คือ ค่าฟัซซี่พารามิเตอร์

3. คำนวณค่าดัชนีวัดความผิดพลาด cost function ด้วยสมการที่ 2.8 และจัดกลุ่มใหม่อีกครั้ง

$$J = \sum_{i=1}^{c}\sum_{j=1}^{n} U_{ij}{}^{m} \left\| x_i - c_j \right\|^2 \qquad (2.8)$$

โดยที่ J คือ ค่าความผิดพลาด

U_{ij} คือ ค่าความเป็นสมาชิกของข้อมูล

x คือ ข้อมูลที่ยังไม่ได้จัดเข้ากลุ่ม

c คือ ค่าศูนย์กลางของกลุ่ม

4. คำนวณค่าศูนย์กลางของกลุ่มใหม่โดยใช้สูตรที่ 2.9 และปรับค่าความเป็นสมาชิกทุกตัวในกลุ่มอีกครั้งด้วยสมการที่ 2.7

$$C_i = \frac{\sum_{i-1}^{n} U_{ij}{}^{m} x_j}{\sum_{j=1}^{n} U_{ij}{}^{m}} \qquad (2.9)$$

โดยที่ C คือ ค่าศูนย์กลางของกลุ่ม

U_{ij} คือ ค่าความเป็นสมาชิกของข้อมูล

x คือ ข้อมูลที่ยังไม่ได้จัดเข้ากลุ่ม

5. ทำซ้ำตั้งแต่ข้อ 3 จนกระทั้งได้ค่าดัชนีวัดความผิดพลาดไม่เกิดการเปลี่ยนแปลงจากรอบการทำงานก่อน ๆ หรือได้ค่าน้อยที่สุด จบขั้นตอนการทำงาน

3. **เทคนิคการจัดกลุ่มแบบไบเซ็คติ้งเค-มีน** (Bisecting K-Means) เป็นใช้อัลกอริธึมที่มีพื้นฐานมาจากทฤษฎีของเค-มีน และมีลักษณะคล้ายกับการจัดกลุ่มแบบโครงสร้างลำดับชั้น เป็นเทคนิคที่นิยมนำไปใช้ในการจัดกลุ่มเอกสาร (Document Clustering) การทำงานเริ่มจากการจัดให้ข้อมูลอยู่ในกลุ่มเดียวกันทั้งหมด จากนั้นทำการจัดกลุ่มข้อมูลแบ่งออกเป็นสองกลุ่มโดยใช้ขั้นตอนวิธีการจัดกลุ่มแบบเค-มีนในการจัดกลุ่มโดยใช้ K = 2 จากนั้นจะทำการแบ่งกลุ่มต่อด้วยการเลือกกลุ่มที่ได้แบ่งแล้วนำมาแบ่งอีกเป็นสองกลุ่ม ทำต่อไปจนกว่าจะได้กลุ่มที่ต้องการ โดยขั้นตอนวิธีการจัดกลุ่มแบบไบเซ็คติ้งเค-มีน (Tan, Steinbach, & Kumar, 2006) มีรายละเอียดดังนี้

1. กำหนดจำนวนกลุ่มที่ต้องการและจัดข้อมูลให้เป็นกลุ่มเดียว

2. แบ่งกลุ่มข้อมูลโดยใช้เค-มีนอัลกอริธึม กำหนดให้ K = 2

3. เลือกกลุ่มที่จะใช้ในการจัดกลุ่มข้อมูลต่อ ส่วนกลุ่มที่ไม่ได้ถูกเลือกจะถูกจัดเป็นกลุ่มถาวร

4. ทำขั้นตอนที่ 2 และ 3 จนกระทั้งได้จำนวนกลุ่มที่ต้องการ วิธีในการเลือกกลุ่มที่จะใช้แบ่งข้อมูลต่อไป คือ เลือกกลุ่มที่สามารถแบ่งกลุ่มได้อย่างสมบูรณ์ เลือกกลุ่มที่มีจำนวนสมาชิกมากที่สุดในแต่ละรอบและมีค่าความแปรปรวนต่ำสุด

การจัดกลุ่มแบบแบ่งส่วนมีข้อได้เปรียบ คือ ความเร็วในการประมวลผลเมื่อใช้กับข้อมูลจำนวนมาก เพราะไม่ต้องเสียเวลาสร้างโครงสร้างของกลุ่ม แต่อย่างไรก็ตามวิธีการจัดกลุ่มข้อมูลแบบแบ่งส่วนยังมีปัญหาเรื่องการกำหนดจำนวนกลุ่มเริ่มต้น (Jain, Murty, & Flynn, 1999)

2.3 งานวิจัยที่เกี่ยวข้อง

ในหัวข้อนี้จะเป็นส่วนที่รายงานถึงงานวิจัยที่ผ่านมาโดยเริ่มต้นเป็นงานวิจัยที่ทำการศึกษาและประยุกต์ใช้เทคนิคการวิเคราะห์ความหมายแอบแฝง (LSA) กับเอกสารภาษาไทย ส่วนหัวข้อย่อยถัดไปจะเป็นการรายงานถึงงานวิจัยต่าง ๆ ที่เกี่ยวข้องกับการใช้เทคนิคการจัดกลุ่มมาใช้ในการจัดกลุ่มเอกสารภาษาไทย

2.3.1 งานวิจัยที่เกี่ยวข้องกับเทคนิคการวิเคราะห์ความหมายแอบแฝง

Deewester, et al. (1986) ได้ศึกษาการสร้างดัชนีคำสำคัญจากการวิเคราะห์ความหมายแอบแฝง ที่มีการใช้กระบวนการแยกค่าแบบเดี่ยว มาทำการสร้างเมทริกซ์ คำ-เอกสาร ที่มีความสัมพันธ์กันและทำการสร้างมุมมองเนื้อหา (Semantic Space) โดยประกอบไปด้วยค่าแบบเดี่ยวที่ถูกจัดเรียงสะท้อนให้เห็นถึงรูปแบบความสัมพันธ์ของคำกับเอกสาร และโครงสร้างของเนื้อหาที่มีความเกี่ยวข้องกัน ซึ่งจากผลการทดลองพบว่า การวิเคราะห์ความหมายแอบแฝงด้วยกระบวนการแยกค่าแบบเดียวมีความสามารถในการแก้ไขปัญหาเรื่องของคำหลายคำที่เขียนต่างกันแต่มีความหมายเดียวกัน ให้สามารถหาความคล้ายคลึงระหว่างคำที่เขียนต่างกันแต่มีความหมายเดียวกันได้

Landauer, Laham, Rehder, and Schreiner (1987) ได้ศึกษาวิจัยเกี่ยวกับเทคนิคการวิเคราะห์ความหมายแอบแฝงซึ่งพบว่า เทคนิคการวิเคราะห์ความหมายแอบแฝงสามารถทำการจับคู่คำที่มีความหมายเหมือนกันแต่เขียนไม่เหมือนกันได้และยังสามารถตัดเอกสารที่มีคำเหมือนกันแต่ความหมายแตกต่างกันออกไปได้อีกด้วย โดยเทคนิคการวิเคราะห์ความหมายแอบแฝงสามารถดึงข้อมูลที่อยู่ในบทความออกมาได้อย่างอิสระไม่ขึ้นต่อการเรียงลำดับของคำ อีกทั้งยังค้นพบว่า การประยุกต์ใช้เทคนิคการวิเคราะห์ความหมายแอบแฝงนี้สามารถประเมินผลการเขียนบทความที่ไม่มีการเรียงลำดับของคำได้ผลใกล้เคียงกับการประเมินโดยมนุษย์ และการวิเคราะห์ความหมายแอบแฝงในบทความได้ผลเทียบเท่ากับความสามารถในการวิเคราะห์ความหมายแอบแฝงโดยมนุษย์ จากผลดังกล่าวแสดงให้เห็นว่า เทคนิคการวิเคราะห์ความหมายแอบแฝงเป็นเทคนิคที่มีความสามารถในการดึงความหมายที่ซ่อนอยู่ในบทความออกมาได้โดยไม่ต้องอาศัยหลักเกณฑ์ทางไวยากรณ์และสิ่งที่สำคัญก็คือผลลัพธ์ที่ได้จะขึ้นอยู่กับมุมมองความหมาย (Semantic Space) ที่สร้างจากมิติที่เหมาะสม (Optimal Dimension)

Wolfe, Schreiner, Rehder, and Laham (1998) ได้ทำการศึกษาระดับความสามารถของผู้อ่านในการสรุปดึงองค์ความรู้ออกจากบทความนั้นขึ้นอยู่กับระดับความรู้พื้นฐานของผู้อ่านและระดับความยากง่ายของเนื้อหาของบทความที่ทำการศึกษา โดยได้นำเอาเทคนิคการวิเคราะห์ความหมายแอบแฝงมาประยุกต์ใช้เพื่อแสดงให้เห็นถึงข้อสรุปที่ว่าบทความที่ทำให้ผู้อ่านได้รับความรู้มากที่สุดไม่ได้มาจากบทความที่มีระดับความยากหรือง่ายที่สุด แต่เกิดจากการเรียนรู้ที่ได้จากการเชื่อมโยงความรู้ที่ได้เรียนรู้ใหม่กับความรู้เก่าที่มีอยู่ก่อนแล้วเพื่อนำมาใช้ในการตีความเพื่อให้เกิดความเข้าใจในการอ่าน และบทความที่มีเนื้อหาไม่ตรงกับความรู้พื้นฐานของผู้อ่านจะไม่มีความเหมาะสมในการที่จะนำมาเป็นบทความที่ให้ความรู้ และภายในเนื้อหาของงานวิจัยแสดงให้เห็นว่า เทคนิคการวิเคราะห์ความหมายแอบแฝงสามารถนำไปประยุกต์ใช้ในการจับคู่ระหว่างความรู้พื้นฐานของผู้อ่านกับระดับความยากง่ายของบทความได้หลากหลายแนวทาง

Tunpaiboon (2000) ได้ศึกษาวิจัยการใช้เทคนิคการวิเคราะห์ความหมายแอบแฝงกับการค้นคืนข้อมูลภาษาไทย โดยมีวัตถุประสงค์เพื่อปรับปรุงประสิทธิภาพการค้นคืนข้อมูลภาษาไทยเพื่อให้การค้นหา (Query) ข้อมูลสามารถใช้คำที่เป็นคำสำคัญหรือคำที่มีความหมายในเชิงเนื้อหาเดียวกันได้ เช่นคำว่า “บ้าน” และ “เรือนไทย” เมื่อใช้คำว่า “เรือนไทย” ในการค้นคืนเอกสาร ผลของการค้นคืนก็จะได้ทั้งเอกสารที่เกี่ยวกับ “บ้าน” และ “เรือนไทย” คืนกลับมาทำให้ผลการสืบค้นข้อมูลมีประสิทธิภาพเพิ่มขึ้น ข้อมูลที่ใช้ในการทดสอบ คือ คลังข้อมูลของ

สารานุกรมไทยฉบับเยาวชน ประกอบด้วยข้อมูล 185 แฟ้มข้อมูล โดยใช้100 คำมาค้นหาในการทดสอบ ผลการทดลองได้เปรียบเทียบผลของการวิเคราะห์ความหมายแอบแฝงหรือเรียกว่า การจัดทำดัชนีคำไทยแบบแฝงความหมาย (Latent Semantic Indexing) กับแบบ Inversion Technique พบว่าสามารถเพิ่มค่าความถูกต้อง (Precision) ได้ ถึงแม้ว่าผลการทดลองจะแสดงเปอร์เซ็นต์ค่าการเรียกข้อมูลกลับคืน (Recall) ลดลงเล็กน้อย แต่ค่าความถูกต้อง (Precision) เพิ่มขึ้นอย่างมีนัยสำคัญ

อัษฎางค์ แตงไทย (2548) ได้ศึกษาวิจัย การย่อความเอกสารภาษาไทย โดยใช้วิเคราะห์ความหมายแอบแฝง มีวัตถุประสงค์เพื่อให้ผู้อ่านเอกสารภาษาไทยที่มีการเก็บข้อมูลในรูปแบบอิเล็กทรอนิกส์สามารถอ่านเอกสารได้มากขึ้นในเวลาที่น้อยลง โดยนำเสนอแนวทางการย่อความเอกสารสำหรับข้อความภาษาไทยทั้งการย่อความแบบทั่วไปและการย่อความแบบตามความต้องการของผู้ใช้ลักษณะคำถาม-ตอบ ซึ่งทั้งสองวิธีทำการเลือกใจความสำคัญจากการดำเนินการตามขั้นตอนวิธีการวิเคราะห์หาความหมายที่แอบแฝงอัลกอริธึมการแยกค่าแบบเดี่ยว และทำการประเมินผลของการย่อความทั้งสองประเภทโดยเปรียบเทียบผลลัพธ์ที่ได้จากระบบที่พัฒนากับผลลัพธ์ที่ย่อความไว้โดยคนจากสามกลุ่มอาชีพ พบว่าการย่อความเอกสารทั่วไป และการย่อความเอกสารแบบคำถาม-คำตอบ สามารถนำเทคนิคการวิเคราะห์ความหมายแอบแฝงมาประยุกต์ใช้ได้

ศุภวรรณ เปรมกุศลชัย (2548) ได้ศึกษาถึง การประเมินคุณภาพการถอดความภาษาไทยด้วยการวิเคราะห์ความหมายแอบแฝง มีวัตถุประสงค์เพื่อเปรียบเทียบความสอดคล้องของการประเมินคุณภาพระหว่างระบบที่พัฒนาขึ้นกับผลการประเมินของผู้ประเมินที่เป็นผู้เชี่ยวชาญของเนื้อหาบทความ โดยนำเทคนิคการวิเคราะห์ความหมายแอบแฝงมาประยุกต์ใช้ในการประเมินคุณภาพการเขียนถอดความภาษาไทย ผลที่ได้จากระบบประเมินคุณภาพการเขียนถอดความบทคัดย่อวิทยานิพนธ์จำนวน 3 ฉบับ เมื่อนำมาเปรียบเทียบกับผลการประเมินของผู้เชี่ยวชาญ จากผลการทดลองพบว่า ผลการประเมินของผู้ประเมินถ้ามีความสอดคล้องกันไปในทิศทางเดียวกันระบบการประเมินคุณภาพถอดความภาษาไทยจะมีความสอดคล้องในการประเมินไปในทิศทางเดียวกับผู้ประเมิน ดังนั้นเทคนิคการวิเคราะห์ความหมายแอบแฝงมีศักยภาพที่จะนำมาพัฒนาระบบประเมินการเขียนถอดความที่เป็นภาษาไทยได้

ชนัณญา โล่ห์รักษา (2549) ได้ศึกษาวิจัย การให้คะแนนการเขียนเรียงความภาษาไทยแบบอัตโนมัติด้วยเทคนิคการวิเคราะห์ความหมายแอบแฝงและเทคนิคโครงข่ายประสาทเทียม มีวัตถุประสงค์เพื่อทำการประเมินผลการเขียนเรียงความภาษาไทยในรูปแบบคะแนน โดยประยุกต์ใช้เทคนิคการวิเคราะห์ความหมายแอบแฝงและเทคนิคโครงข่ายประสาท

เทียม งานวิจัยนี้ดำเนินการทดลองโดยให้นักเรียนเป็นผู้เขียนเรียงความและให้ผู้ประเมินเป็นผู้ตรวจให้คะแนน จากการศึกษาเปรียบเทียบพบว่า คะแนนที่ได้จากระบบไม่แตกต่างจากคะแนนที่ได้จากผู้ประเมิน และยังพบว่าการให้คะแนนจากการประเมินคุณภาพการเขียนเรียงความภาษาไทยด้วยการประยุกต์ใช้เทคนิคการวิเคราะห์ความหมายแอบแฝงมีค่าความผิดพลาดน้อยกว่าการประยุกต์ใช้ค่าความถี่ของคำที่ปรากฏขึ้น นอกจากนี้จากผลการทดลองยังพบว่าโครงสร้างของโครงข่ายประสาทเทียมที่มีฟังก์ชันกระตุ้นแบบซิกมอยด์มีค่าความผิดพลาดในการให้คะแนนจากการประเมินคุณภาพการเขียนเรียงความภาษาไทยน้อยกว่าโครงสร้างของโครงข่ายประสาทเทียมที่มีฟังก์ชันกระตุ้นแบบเชิงเส้น

จากงานวิจัยที่ผ่านมา กล่าวสรุปได้ว่า เทคนิคการวิเคราะห์ความหมายแอบแฝงมีความสามารถวิเคราะห์เนื้อหาข้อความเพื่อหาความคล้ายคลึงระหว่างข้อความที่มีความหมายในเชิงเนื้อหาหรือการตีความเป็นไปในทิศทางเดียวกันได้โดยไม่ต้องอาศัยหลักเกณฑ์ทางไวยากรณ์

2.3.2 งานวิจัยที่เกี่ยวข้องกับเทคนิคการจัดกลุ่มข้อความ

Steinbach, Karypis and Kumar (2000) ได้ศึกษาเปรียบเทียบเทคนิคการจัดกลุ่มในการจัดกลุ่มเอกสารสามวิธี คือ การจัดกลุ่มแบบโครงสร้างลำดับชั้น (Agglomerative Hierarchical Clustering) ขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน (K-Means) และ ขั้นตอนวิธีการจัดกลุ่มแบบไบเซ็คติ้งเค-มีน (Bisecting K-Means) ซึ่งข้อมูลที่ใช้ในการทดลองเป็นเอกสารจำนวน 8 ชุด ซึ่งผ่านขั้นตอนการวิเคราะห์ด้วยวิธี Vector Space Model และหาความคล้ายคลึงระหว่างเอกสารแบบโคไซน์ก่อนเข้าสู่ขั้นตอนการจัดกลุ่ม โดยใช้ค่า Entropy, F-Measure และ Overall Similarity ในการวัดประสิทธิภาพความถูกต้องของการจัดกลุ่ม ผลการทดลองพบว่า ขั้นตอนวิธีการจัดกลุ่มแบบไบเซ็คติ้งเค-มีน สามารถจัดกลุ่มเอกสารได้มีประสิทธิภาพกว่าการจัดกลุ่มแบบโครงสร้างลำดับชั้นและแบบเค-มีน

ชุลีรัตน์ จรัสกุลชัย เจษฎา กันทะเสนา และสถาพร คิ้วสุวรรณสุข (2544) ได้ศึกษาวิจัยเกี่ยวกับการจัดกลุ่มเอกสารสำหรับข้อความภาษาไทย มีวัตถุประสงค์เพื่อศึกษาขั้นตอนการจัดกลุ่มข้อความข่าวภาษาไทย โดยใช้ขั้นตอนวิธีการจัดกลุ่มทั้งแบบ Complete Link สำหรับการจัดกลุ่มแบบลำดับชั้น และ Single Pass สำหรับการจัดกลุ่มแบบไม่เป็นลำดับชั้น ขั้นตอนวิธีทั้งสองแบบจะเป็นการจัดกลุ่มแบบอัตโนมัติ โดยนำขั้นตอนวิธีดังกล่าวมาประยุกต์ใช้กับข้อความข่าวภาษาไทย นอกจากนี้ งานวิจัยดังกล่าวยังได้ประยุกต์หลักการประมวลผลแบบขนาน เพื่อ

แก้ปัญหาในการคำนวณค่าความเหมือนของเอกสาร โดยผลการวิจัยสรุปว่าขั้นตอนวิธีในการตัดคำที่ศึกษาเปรียบเทียบระหว่างแบบกฎภาษาศาสตร์และแบบเทียบคำยาวที่สุดของ Swath ทั้งสองแบบไม่มีผลต่อการจัดกลุ่มข่าวภาษาไทย และจากการเปรียบเทียบประสิทธิภาพในการจัดกลุ่มด้วยการหาค่าอำนาจจำแนกของการจัดกลุ่มแบบ Complete Link และแบบ Single Pass พบว่าไม่แตกต่างกัน

Albayrak and Amasyali (2003) ได้ศึกษาวิจัยเกี่ยวกับการจัดกลุ่มแบบไม่มีการเรียนรู้มาใช้ในการพัฒนาระบบการวินิจฉัยโรคทางการแพทย์ โดยใช้ขั้นตอนวิธีการจัดกลุ่มแบบฟัซซี่ซี-มีน (Fuzzy C-Means) ในการจัดกลุ่มข้อมูลผู้ป่วยโรคไทรอยด์จำนวน 215 ตัวอย่าง มีการนำผลการจัดกลุ่มไปเปรียบเทียบกับขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน (K-Means) โดยใช้การวัดความถูกต้องของการจัดกลุ่มเป็นตัววัดประสิทธิภาพในการแบ่งกลุ่ม พบว่า ขั้นตอนวิธีการจัดกลุ่มแบบฟัซซี่ซี-มีนสามารถแบ่งกลุ่มได้ถูกต้องถึง 83.71% ส่วนขั้นตอนวิธีการจัดกลุ่มแบบเค-มีนแบ่งกลุ่มได้ถูกต้องเพียง 78.1%

วีรนุช นิ่มเงิน (2548) ได้ศึกษาวิจัย การสร้างฐานขององค์ความรู้จากข้อมูลโดยใช้อัลกอริธึมจัดกลุ่มแบบวนซ้ำ มีวัตถุประสงค์เพื่อหาเทคนิคการจัดกลุ่มที่ให้ผลลัพธ์ที่ดีที่สุดมาใช้ในการเรียนรู้ข้อมูลสำหรับนำไปใช้ในการพัฒนาระบบฐานองค์ความรู้ (Knowledge Base System) โดยเปรียบเทียบประสิทธิภาพการจัดกลุ่มระหว่างเทคนิคการจัดกลุ่มแบบเค-มีนกับแบบฟัซซี่ซี-มีน ทำการทดลองกับกลุ่มข้อมูล 3 ประเภทที่มีเนื้อหาแตกต่างกัน ประกอบด้วย ข้อมูลดอกไม้ไอริสเกี่ยวกับขนาดความกว้างความยาวของกลีบดอก ข้อมูลการจำแนกไวน์ที่เจริญเติบโตในพื้นที่เพาะปลูกที่ต่างกัน และข้อมูลการทำงานของต่อมไทรอยด์ จากผลการทดลองพบว่า ผลการจัดกลุ่มระหว่างเทคนิคการจัดกลุ่มแบบเค-มีน กับ แบบฟัซซี่ซี-มีน กับการจัดกลุ่มข้อมูลทั้งสามประเภทให้ผลในการจัดกลุ่มแตกต่างกันโดยขึ้นอยู่กับเนื้อหาของข้อมูล ในขณะที่ความถูกต้องในการจัดกลุ่มไม่แตกต่างกันแต่การเลือกตัวแทนกลุ่มเริ่มต้นเป็นปัจจัยสำคัญที่ส่งผลต่อประสิทธิภาพของการจัดกลุ่มข้อมูล

อรนุช ชัยหมื่น (2548) ได้ศึกษาเปรียบเทียบการแบ่งกลุ่มข้อมูลลูกค้าสินค้าหัตถกรรมไทย โดยใช้การจัดกลุ่มแบบสองขั้นตอน ระหว่างเทคนิคการจัดกลุ่มแบบ Kohonen's Self Organizing Maps (SOM) ร่วมกับการจัดกลุ่มแบบเค-มีน (K-Means) กับเทคนิคการจัดกลุ่มแบบโครงสร้างลำดับชั้น (Hierarchical Cluster) ร่วมกับการจัดกลุ่มแบบเค-มีน จากการศึกษาพบว่าทั้ง 2 วิธีมีประสิทธิภาพในการแบ่งกลุ่มที่แตกต่างกันออกไป กล่าวคือ กรณีที่มีข้อมูลจำนวนน้อยอัลกอริธึมแบบสองขั้นตอนที่เหมาะสมคือ การจัดกลุ่มแบบลำดับชั้นร่วมกับการจัดกลุ่มแบบ

เค-มีน เนื่องจากเป็นวิธีที่เหมาะสมสำหรับการแบ่งกลุ่มให้ครอบคลุมมีจำนวนกลุ่มไม่มาก และใช้เวลาในการประมวลผลไม่มาก แต่สำหรับกรณีที่มีข้อมูลทดลองจำนวนมาก ประมาณ 15,000 รายการขึ้นไป ควรใช้วิธีการจัดกลุ่มแบบ SOM ร่วมกับการจัดกลุ่มแบบเค-มีน เพราะเป็นวิธีการที่ให้จำนวนการแบ่งกลุ่มที่ละเอียดและใช้เวลาในการประมวลผลน้อยกว่า แต่อย่างไรก็ตาม อัลกอริธึมแบบลำดับชั้นเป็นอัลกอริธึมที่เข้าใจง่าย สามารถนำไปใช้งานได้ง่าย ซึ่งต่างจาก SOM ที่ค่อนข้างมีความซับซ้อน และนำอัลกอริธึมไปพัฒนายากกว่าแบบลำดับชั้น

ศศิธร มงคลศรีพัฒนา (2550) ได้ศึกษาวิจัยเกี่ยวกับการจัดกลุ่มศูนย์บริการและถ่ายทอดเทคโนโลยีทางการเกษตรประจำตำบลในประเทศไทยโดยใช้อัลกอริธึม 2 ขั้นตอนคือ Kohonen's Self Organizing Maps (SOM) ร่วมกับการจัดกลุ่มแบบแบ่งส่วน โดยขั้นตอนแรกนำอัลกอริธึมของ SOM มาใช้ในการจัดกลุ่มเริ่มต้น เพื่อหาจำนวนกลุ่มที่เหมาะสม และขั้นตอนที่ 2 นำจำนวนกลุ่มที่ได้จากขั้นตอนแรกเป็นข้อมูลนำเข้าเพื่อใช้ในการเปรียบเทียบประสิทธิภาพการจัดกลุ่มของขั้นตอนการจัดกลุ่มแบบแบ่งส่วน 3 วิธี คือ แบบเค-มีน แบบไบเซ็คติ้งเค-มีน และ แบบฟัซซี่ซี-มีน จากนั้นทำการคำนวณหาค่าระดับคะแนนรวมประสิทธิภาพการทำงาน (P Score) พบว่า ขั้นตอนวิธีการจัดกลุ่มแบบฟัซซี่ซี-มีน ให้ผลการจัดกลุ่มที่ดีที่สุด

จากงานวิจัยการนำเทคนิคการจัดกลุ่มไปประยุกต์ใช้ในการจัดกลุ่มข้อมูลลักษณะต่าง ๆ ชี้ให้เห็นว่า การจัดกลุ่มแบบแบ่งส่วนและแบบโครงสร้างลำดับชั้นมีประสิทธิภาพในการจัดกลุ่มที่แตกต่างกันออกไปทั้งนี้ขึ้นอยู่กับลักษณะเนื้อหาและจำนวนของข้อมูล โดยการเลือกตัวแทนกลุ่มเริ่มต้นเป็นปัจจัยสำคัญต่อประสิทธิภาพของการจัดกลุ่มทั้งทางด้านจำนวนรอบและเวลาที่ใช้ประมวลผล

2.4 **สรุปทฤษฎีและงานวิจัยที่เกี่ยวข้อง**

จากทฤษฎีและงานวิจัยที่เกี่ยวข้อง เทคนิคเดลฟายเป็นระเบียบวิธีวิจัยที่ให้ข้อมูลในอนาคตที่น่าเชื่อถือสามารถนำไปใช้ประกอบการตัดสินใจได้ดีด้วยเหตุนี้เทคนิคเดลฟายจึงเป็นที่ยอมรับและได้รับความนิยมใช้อย่างแพร่หลายในการวิจัยหลายๆ ด้าน แต่ขั้นตอนการเก็บรวบรวมข้อมูลและสร้างแบบสอบถามมีการดำเนินการหลายรอบ โดยในรอบที่สองเป็นขั้นตอนที่ยากและสำคัญ ผู้ที่ดำเนินการวิจัยด้วยเทคนิคเดลฟายต้องเสียเวลาเป็นอย่างมากกับการวิเคราะห์เปรียบเทียบข้อคิดเห็นของผู้เชี่ยวชาญ อีกทั้งแบบสอบถามรอบที่สองเป็นแบบสอบถามที่ถูกใช้ดำเนินการเก็บรวบรวมข้อมูลในรอบต่อไปจนจบการดำเนินการวิจัย ดังนั้นเพื่อช่วยลดเวลาที่ใช้ในการวิเคราะห์ข้อคิดเห็น ลดความคลาดเคลื่อนหรืออคติที่อาจเกิดขึ้น งานวิจัยนี้จึงมีเป้าหมายเพื่อ

พัฒนาระบบการวิเคราะห์ข้อคิดเห็นของผู้เชี่ยวชาญตามกระบวนการสร้างและพัฒนาแบบสอบถามรอบที่สองของการวิจัยด้วยเทคนิคเดลฟาย ซึ่งขั้นตอนการพัฒนาแบบสอบถามรอบที่สองจะนำข้อคิดเห็นของผู้เชี่ยวชาญที่ได้จากแบบสอบถามรอบแรกของแต่ละท่านมาทำการวิเคราะห์เปรียบเทียบ ตัดข้อความซ้ำซ้อนออก และนำข้อคิดเห็นที่เหมือนกันรวมเข้าเป็นกลุ่มเดียวกันเพื่อตั้งเป็นข้อคำถาม ส่วนข้อคิดเห็นที่ไม่เหมือนกันก็แยกออกนำไปตั้งเป็นข้อคำถามทีละข้อ จากขั้นตอนการวิเคราะห์ข้อคิดเห็นดังกล่าว เพื่อให้คอมพิวเตอร์สามารถวิเคราะห์ข้อคิดเห็นได้ถึงในระดับเนื้อหาและสามารถจัดกลุ่มข้อคิดเห็นได้นั้น ผู้วิจัยจึงแบ่งขั้นตอนการทำงานออกเป็นสองขั้นตอนคือขั้นตอนการวิเคราะห์หาความคล้ายคลึงระหว่างข้อคิดเห็น และขั้นตอนการจัดกลุ่มข้อคิดเห็น

ขั้นตอนการวิเคราะห์หาความคล้ายคลึงระหว่างข้อคิดเห็น เพื่อให้คอมพิวเตอร์สามารถเข้าใจได้ถึงในระดับเนื้อหาและการตีความข้อคิดเห็น จากการทบทวนทฤษฎีเกี่ยวกับการวิเคราะห์ข้อคิดเห็นของผู้เชี่ยวชาญในการวิจัยด้วยเทคนิคเดลฟายพบว่า ข้อคิดเห็นของผู้เชี่ยวชาญแต่ละท่านในแต่ละข้อคำถามมีทั้งที่คล้ายกันและแตกต่างกัน อีกทั้งข้อความภาษาไทยมีคำบางคำที่เขียนไม่เหมือนกันแต่มีความหมายเดียวกัน จากปัญหาดังกล่าวเมื่อศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องพบว่า เทคนิคการวิเคราะห์ความหมายแอบแฝงมีความสามารถวิเคราะห์เนื้อหาข้อความเพื่อหาความคล้ายคลึงระหว่างข้อความที่มีความหมายในเชิงเนื้อหาหรือการตีความเป็นไปในทิศทางเดียวกันได้โดยไม่ต้องอาศัยหลักเกณฑ์ทางไวยากรณ์ ดังนั้นในงานวิจัยนี้จึงเลือกใช้เทคนิคการวิเคราะห์ความหมายแอบแฝงเพื่อหาความคล้ายคลึงระหว่างข้อคิดเห็น หลังจากที่สามารถวิเคราะห์หาความคล้ายคลึงระหว่างข้อคิดเห็นได้ว่าข้อคิดเห็นใดที่มีความหมายคล้ายกันหรือเป็นไปในทิศทางเดียวกันขั้นตอนต่อไปคือจัดข้อคิดเห็นที่มีความคล้ายคลึงกันรวมเข้าเป็นกลุ่มเดียวกัน

ขั้นตอนการจัดกลุ่มข้อคิดเห็น เพื่อหาเทคนิคการจัดกลุ่มข้อความที่มีความเหมาะสมที่สุดในการจัดลุ่มข้อคิดเห็น ผู้วิจัยจึงศึกษาทบทวนทฤษฎีและงานวิจัยที่เกี่ยวข้องด้านเทคนิคการจัดกลุ่มข้อความพบว่า เทคนิคการจัดกลุ่มข้อความใช้หลักการแนวคิดทฤษฎีของการจัดกลุ่มข้อมูลซึ่งแบ่งเป็นสองประเภทหลักคือ การจัดกลุ่มแบบแบ่งส่วนและการจัดกลุ่มแบบโครงสร้างลำดับชั้น ทั้งนี้พบข้อดีและข้อเสียของการจัดกลุ่มทั้งสองประเภท โดยการจัดกลุ่มแบบแบ่งส่วนมีความสามารถในการจัดข้อมูลเข้ากลุ่มได้ดีแต่ผู้ใช้ต้องเป็นผู้กำหนดจำนวนกลุ่มขณะที่การจัดกลุ่มแบบโครงสร้างลำดับชั้นสามารถจัดกลุ่มได้แบบอัตโนมัติโดยที่ผู้ใช้ไม่ต้องกำหนดจำนวนกลุ่มแต่ความสามารถในการจัดข้อมูลเข้ากลุ่มไม่ดี นอกจากนี้การจัดกลุ่มแบบที่ผู้ใช้ไม่ต้องกำหนดจำนวน

กลุ่มอีกประเภทหนึ่งที่นำมาใช้ในการศึกษาวิจัยคือ Kohonen's Self Organizing Maps (SOM) แต่จากงานวิจัยของอรนุช ชัยหมื่น (2548) พบว่า SOM เหมาะสำหรับการจัดกลุ่มข้อมูลที่มีจำนวนมาก ประมาณ 15,000 รายการขึ้นไป ขณะที่การจัดกลุ่มแบบโครงสร้างลำดับชั้นเหมาะสำหรับข้อมูลที่มีจำนวนน้อย ดังนั้นเพื่อให้ระบบสามารถจัดกลุ่มได้โดยอัตโนมัติและเหมาะสมกับปริมาณข้อมูล เนื่องจากการวิจัยด้วยเทคนิคเดลฟายใช้ข้อคิดเห็นจากผู้เชี่ยวชาญประมาณ 17 คน ข้อคิดเห็นจึงมีจำนวนไม่มาก การจัดกลุ่มสองขั้นตอนในงานวิจัยนี้จึงใช้การจัดกลุ่มแบบโครงสร้างลำดับชั้นร่วมกับการจัดกลุ่มแบบแบ่งส่วน ซึ่งการจัดกลุ่มแบบแบ่งส่วนมีด้วยกันสามเทคนิควิธีคือ เค-มีน ฟัซซี่ซี-มีน และไบเซ็คติ้งเค-มีน จากงานวิจัยที่ผ่านมาได้ศึกษาและนำเทคนิคการจัดกลุ่มแต่ละเทคนิควิธีไปประยุกต์ใช้ในการจัดกลุ่มข้อมูลที่แตกต่างกัน เช่น จัดกลุ่มข้อมูลลูกค้า จัดกลุ่มเอกสาร โดยผลการทดลองของงานวิจัยที่เกี่ยวข้องทางด้านการจัดกลุ่มได้ชี้ให้เห็นว่าประสิทธิภาพและผลในการจัดกลุ่มแตกต่างกันออกไปโดยขึ้นอยู่กับลักษณะของเนื้อหาข้อมูล

ดังนั้นเพื่อให้ระบบที่พัฒนาขึ้นในงานวิจัยนี้สามารถจัดกลุ่มได้โดยอัตโนมัติและให้ผลในการจัดกลุ่มที่ได้ผลลัพธ์ถูกต้องมากที่สุด ผู้วิจัยจึงประยุกต์ขั้นตอนการจัดกลุ่มโดยใช้การจัดกลุ่มแบบโครงสร้างลำดับชั้นทำงานร่วมกันกับการจัดกลุ่มแบบแบ่งส่วน และเพื่อหาเทคนิคการจัดกลุ่มแบบแบ่งส่วนที่มีความเหมาะสมและให้ผลในการจัดกลุ่มข้อคิดเห็นที่ดีที่สุดเมื่อทำงานร่วมกับเทคนิคการวิเคราะห์ความหมายแอบแฝงสำหรับวิเคราะห์ข้อคิดเห็นของผู้เชี่ยวชาญในงานวิจัยนี้ จึงศึกษาเปรียบเทียบเทคนิคการจัดกลุ่มข้อความทั้งสามเทคนิควิธีระหว่าง เค-มีน ฟัซซี่ซี-มีน และไบเซ็คติ้งเค-มีน