บทที่ 3

การประยุกต์ใช้กับระบบแนะนำทรัพยากรห้องสมุด

เทคนิคอะไพออริอัลกอริทึมมักถูกใช้เพื่อหากฎการรวมกลุ่ม ผู้ศึกษาจึงได้นำมาประยุกต์ใช้ในงานห้องสมุด โดยนำอัลกอริทึมดังกล่าวมาค้นหากฎการรวมกลุ่มของคำสำคัญเพื่อพัฒนาระบบแนะนำสารนิเทศเพื่อแนะนำสารนิเทศอื่นๆ นอกเหนือจากสารนิเทศที่ผู้ใช้บริการต้องการโดยมีรายละเอียดและวิธีการดังต่อไปนี้

3.1 การเตรียมข้อมูล

ข้อมูลที่จำเป็นต้องใช้เพื่อหากฎการรวมกลุ่มคือคำสำคัญของหนังสือแต่ละเล่ม ซึ่งการเลือกคำสำคัญของหนังสือเป็นหน้าที่ของบรรณารักษ์ฝ่ายวิเคราะห์ทรัพยากรสารนิเทศสำนักหอสมุด เมื่อห้องสมุดรับหนังสือใหม่เข้ามาบรรณารักษ์ที่มีหน้าที่จะเป็นผู้ระบุคำสำคัญของหนังสือ และทำการบันทึกไว้ในฐานข้อมูลหนังสือของสำนักหอสมุด ผู้ศึกษาจึงได้ประยุกต์นำข้อมูลดังกล่าวมาค้นหากฎการรวมกลุ่มโดยใช้เทคนิคอะไพออริอัลกอริทึมโดยใช้เลขเรียกหนังสือ (call_no) แทนรายการซื้อสินค้า และใช้คำสำคัญของหนังสือแต่ละเล่มแทนรายการสินค้า ตามขั้นตอนดังต่อไปนี้

3.1.1 การบันทึกคำสำคัญ เป็นขั้นตอนนี้เกิดขึ้นเมื่อห้องสมุดรับหนังสือใหม่เข้ามาบรรณารักษ์จะทำการระบุคำสำคัญของหนังสือแต่ละเล่ม (ตามตัวอย่างตารางที่ 3.1)

ตารางที่ 3.1 ตัวอย่างตารางระบุคำสำคัญ

เลขเรียกหนังสือ (call_no)	คำสำคัญ (keywords)
001	computer JAVA database
002	language JAVA C++
003	computer language JAVA C++
004	language C++

3.1.2 การทำเหมืองข้อมูลโดยใช้หลักการอะไพออริอัลกอริทึม จากตารางที่ 3.1 หากนำมาทำเหมืองข้อมูลโดยใช้เทคนิคอะไพออริอัลกอริทึมโดยตั้งค่าสนับสนุนน้อยสุดเท่ากับสอง (ความถี่ต้องมากกว่าหรือเท่ากับสอง) จะได้ผลลัพธ์ออกมาดังนี้

1) การพิจารณารอบที่หนึ่ง ระบบจะสร้างแคนดิเดตเซต โดยมีจำนวนสมาชิกในเซตเท่ากับหนึ่ง แล้วทำการนับความถี่ของคำสำคัญแต่ละคำ หากแคนดิเดตเซตใดมีความถี่มากกว่าหรือเท่ากับสองจะมีคุณสมบัติเป็นไอเทมเซตขนาดใหญ่ (Large itemset) และจะถูกเก็บไว้ในตารางไอเทมเซตขนาดใหญ่เพื่อนำไปพิจารณาในรอบต่อไป จากผลลัพธ์ในภาพที่ 3.1 จะเห็นได้ว่าเซต {database} มีความถี่เท่ากับหนึ่งซึ่งค่าน้อยกว่าค่าสนับสนุนน้อยสุดไม่มีคุณสมบัติเป็นลาร์จไอเทมเซตและจะไม่ถูกนำไปพิจารณาในรอบต่อไป

เลขเรียกหนังสือ (call_no)	คำสำคัญ (keywords)
001	computer JAVA database
002	Language JAVA C++
003	computer language JAVA C++
004	language C++

ตารางคำสำคัญ

1-itemset	ความถี่
computer	2
language	3
JAVA	3
database	1
C++	3

ตารางแคนดิเดตเซตที่ 1

Large 1-itemset	ความถี่
computer	2
language	3
JAVA	3
C++	3

ตารางไอเทมเซตขนาดใหญ่ที่ 1

ภาพที่ 3.1 ผลลัพธ์การพิจารณารอบที่หนึ่ง

2) การพิจารณารอบที่สอง ระบบจะสร้างแคนดิเดตเซตที่มีค่าเท่ากับสอง (k+1) โดยการนำไอเทมเซตขนาดใหญ่ที่เก็บไว้มาผสมรวมกันเพื่อสร้างเป็นแคนดิเดตเซตที่มีความเป็นไปได้ทั้งหมด

(join) โดยมีเงื่อนไขว่าสมาชิกตัวแรกของเซตต้องเหมือนกันจึงจะสามารถ “join” กันได้ แล้วใช้วิธีพิจารณาเดียวกันกับรอบที่หนึ่งซึ่งได้ผลลัพธ์ดังภาพที่ 3.2

Large 1-itemset	ความถี่
computer	2
language	3
JAVA	3
C++	3

ตารางไอเทมเซตขนาดใหญ่ที่ 1

2-itemset	ความถี่
computer language	1
computer JAVA	2
computer C++	1
language JAVA	2
language C++	3
JAVA C++	2

ตารางแคนดิเดตเซตที่ 2

Large 2-itemset	ความถี่
computer JAVA	2
language JAVA	2
language C++	3
JAVA C++	2

ตารางไอเทมเซตขนาดใหญ่ที่ 2

ภาพที่ 3.2 ผลลัพธ์การพิจารณารอบที่สอง

3) การพิจารณารอบที่สาม ระบบจะใช้วิธีพิจารณาเดียวกันกับรอบที่หนึ่งและรอบที่สอง จากภาพที่ 3.2 เซต {computer,language},{computer,C++} มีความถี่น้อยกว่าสองจึงไม่ถูกนำไปพิจารณาในรอบต่อไป และด้วยเงื่อนไขว่าข้อมูลในเซตแรกจะต้องมีค่าเหมือนกันจึงสามารถทำการ “join” กันได้ ดังนั้นเซตที่เหลือในแคนดิเดตเซตที่สามจึงมีเพียงเซต {language,JAVA,C++} เท่านั้น และเนื่องจากไม่สามารถสร้างแคนดิเดตเซตในระดับที่ “k+1”ได้ระบบจึงหยุดทำงานดังภาพที่ 3.3

Large 2-itemset	ความถี่
computer JAVA	2
language JAVA	2
language C++	3
JAVA C++	2

ตารางไอเทมเซตขนาดใหญ่ที่ 2

3-itemset	ความถี่
language JAVA C++	2

ตารางแคนดิเดตเซตที่ 3

Large 3-itemset	ความถี่
language JAVA C++	2

ตารางไอเทมเซตขนาดใหญ่ที่ 3

ภาพที่ 3.3 ผลลัพธ์การพิจารณารอบที่สาม

3.1.3 การนำผลลัพธ์ไปประยุกต์ใช้งาน จากข้อ 3.1.2 ผู้ศึกษาได้แสดงให้เห็นการนำหลักการอะไพออริอัลกอริทึมมาค้นหากฎการรวมกลุ่มของคำสำคัญ ผลลัพธ์ที่ได้คือกลุ่มของไอเทมเซตขนาดใหญ่ (Large itemset) ทั้งสามตาราง (Large 1-itemset, Large 2-itemset, Large 3-itemset) ซึ่งผู้ศึกษาจะได้นำข้อมูลส่วนนี้ไปแนะนำหนังสือแก่ผู้ใช้บริการต่อไปโดยเรียงตามลำดับตามเซตที่มีขนาดใหญ่ไปหาเซตที่มีขนาดเล็ก ซึ่งเซต {language,JAVA,C++} มีขนาดใหญ่ที่สุด จากตารางระบุคำสำคัญของหนังสือ ผู้ศึกษาจึงสรุปได้ว่าหนังสือที่มีเลขเรียกหนังสือ 002 003 มีความเกี่ยวข้องกันเนื่องจากมีคำสำคัญเหมือนกันถึงสามคำ

3.2 การสร้างโมเดลข้อมูลในโปรแกรมเวก้า

ในการศึกษาครั้งนี้ผู้ศึกษาเลือกใช้โปรแกรมเวก้า (WEKA) เป็นเครื่องมือช่วยในการวิเคราะห์ฐานข้อมูล เพื่อทำเหมืองโดยผู้ศึกษาได้เลือกใช้โมดูลอะไพออริอัลกอริทึมเพื่อช่วยสร้างกฎการรวมกลุ่มของคำสำคัญ โดยมีรายละเอียดดังนี้

3.2.1 โปรแกรมเวก้าเป็นโปรแกรมที่นิยมใช้กันอย่างแพร่หลาย โดยมากมักใช้เป็นเครื่องมือช่วยในการวิเคราะห์ฐานข้อมูลเพื่อทำเหมืองข้อมูล ซึ่งรวบรวมแนวคิดและอัลกอริทึมที่ช่วยในการทำเหมืองข้อมูลไว้มากมาย โดยแยกออกเป็น 4 เมนูหลักคือ

1) การทำเหมืองข้อมูลแบบจัดประเภท (Classification)

2) การทำเหมืองข้อมูลในรูปแบบการเกาะกลุ่ม (Clustering)

3) การทำเหมืองข้อมูลแบบกฎการรวมกลุ่ม (Association)

4) การนำเสนอข้อมูลในรูปแบบของรูปภาพสองมิติ (Visualize)

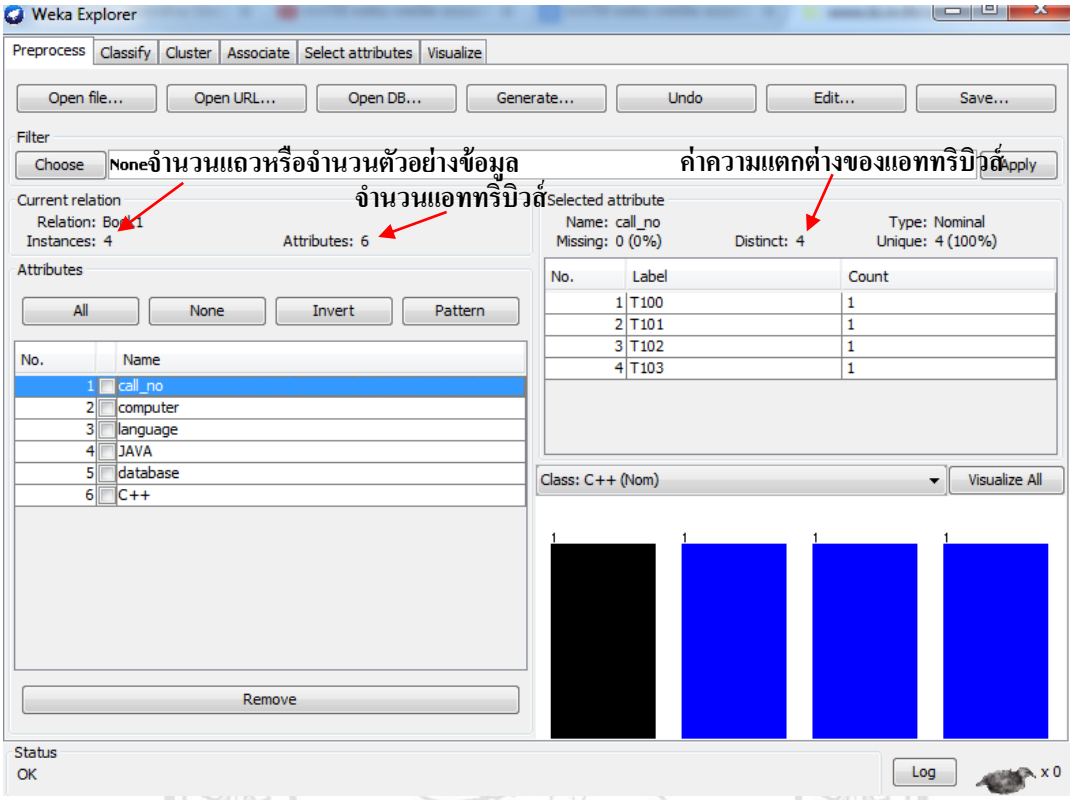
การศึกษาครั้งนี้ ผู้ศึกษาได้เลือกใช้เทคนิคการค้นหากฎการรวมกลุ่ม (Association) โดยใช้อะไพออริอัลกอริทึมเป็นเครื่องมือช่วยวิเคราะห์ข้อมูลเพื่อนำไปพัฒนาระบบแนะนำการสืบค้นทรัพยากรสารนิเทศของห้องสมุดต่อไป

3.2.2 การเตรียมข้อมูล ข้อมูลที่จะสามารถนำเข้าโปรแกรมเวก้าได้จะต้องมีรูปแบบและชนิดไฟล์ตามที่โปรแกรมกำหนด ซึ่งรูปแบบของตารางระบุคำสำคัญในตารางที่ 3.1 นั้นไม่สามารถนำเข้าสู่โปรแกรมเวก้าได้ เนื่องจากโปรแกรมมีลักษณะการทำงานคือจะนับความถี่จากตารางเช็คลิสต์โดยไม่สามารถนับความถี่ของคำสำคัญที่เหมือนกันในตารางระบุคำสำคัญได้ ผู้ศึกษาจึงต้องทำการเปลี่ยนรูปแบบตารางระบุคำสำคัญมาอยู่ในรูปแบบของตารางเช็คลิสต์ก่อน โดยที่ข้อมูลไม่ถูกเปลี่ยนแปลง ตามตัวอย่างภาพที่ 3.4 และเมื่อข้อมูลถูกนำเข้าสู่โปรแกรมแล้ว โปรแกรมจะแสดงรายละเอียดของข้อมูลที่นำเข้าดังภาพที่ 3.5

เลขเรียกหนังสือ (call_no)	คำสำคัญ (keywords)
001	computer JAVA database
002	Language JAVA C++
003	computer language JAVA C++
004	language C++

เลขเรียกหนังสือ (call_no)	**computer**	**language**	**JAVA**	**database**	**C++**
001	x		x	x	
002		x	x		x
003	x	x	x		x
004		x			x

ภาพที่ 3.4 การเปลี่ยนรูปแบบตารางระบุคำสำคัญให้อยู่ในรูปแบบตารางเช็คลิสต์



ภาพที่ 3.5 รูปแสดงการนำข้อมูลเข้าสู่โปรแกรม

3.2.3 การทำเหมืองข้อมูลแบบกฎการรวมกลุ่ม (Association) โดยใช้เทคนิคอะไพออริในเวก้าสามารถทำได้โดยเลือกเมนู “Associat” เลือกอัลกอริทึมเป็น “Aprioori” โดยมีค่าพารามิเตอร์ต่างๆที่จำเป็นในการควบคุมการทำงานของอัลกอริทึมดังนี้

1) lowerBoundMinSupport หมายถึง ปรับค่าสนับสนุนน้อยสุด (minsup) สามารถคำนวณจากสูตร ค่าสนับสนุนน้อยสุด = ตัวอย่างข้อมูล (Instances)*lowerBoundMinSupport

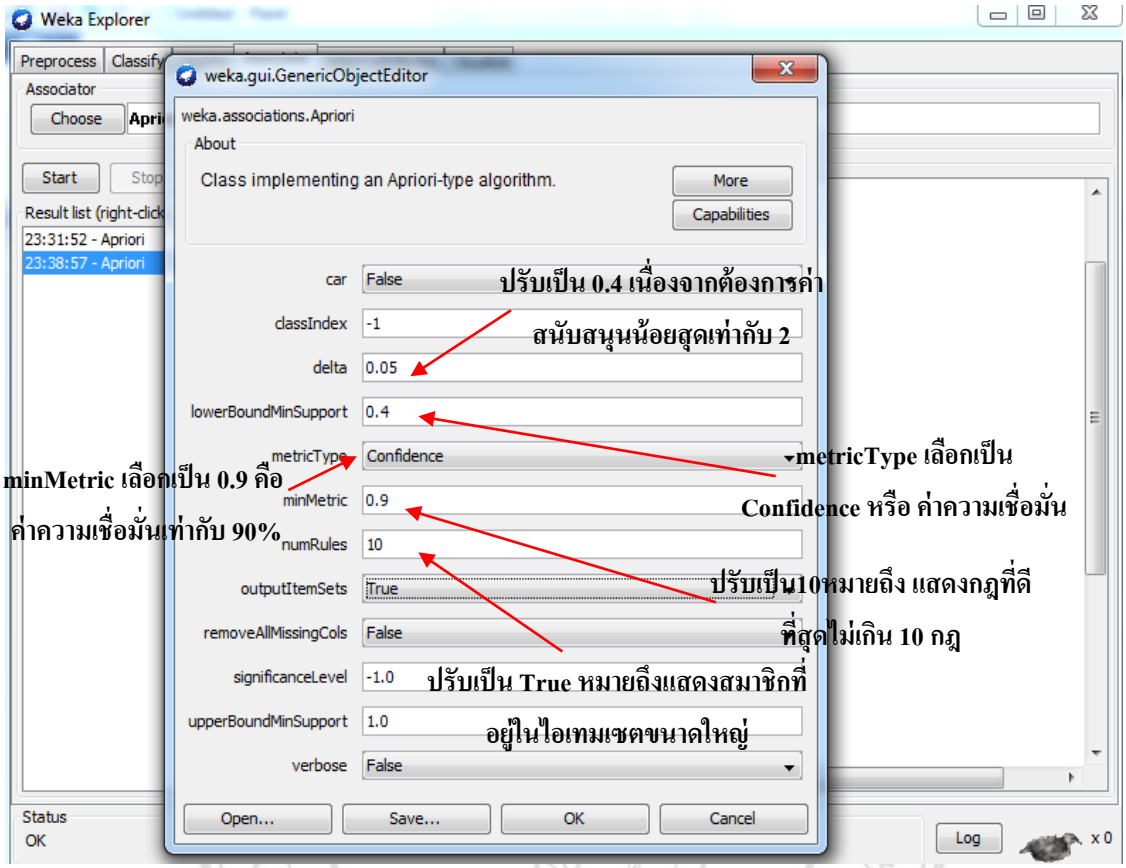
2) metricType หมายถึง ปุ่มเลือกประเภทของค่าความเชื่อมั่น โดยทั่วไปอะไพออริอัลกอริทึมจะใช้ค่า “Confidence”

3) minMetric หมายถึง การปรับค่าความเชื่อมั่น (Confidence) ยกตัวอย่างเช่นระบุตัวเลข 0.9 หมายความว่ามีค่าความเชื่อมั่นเท่ากับร้อยละเก้าสิบ

4) numRules หมายถึง จำนวนกฎที่ดีที่สุดที่ต้องการให้โปรแกรมแสดงผล

5) outputItemSets หมายถึง ต้องการให้ระบบแสดงรายชื่อสมาชิกที่อยู่ในเซตหรือไม่

ตัวอย่างการปรับค่าพารามิเตอร์ของอะไพออริอัลกอริทึมสามารถอธิบายได้ตามภาพที่ 3.6



ภาพที่ 3.6 ตัวอย่างการปรับค่าพารามิเตอร์ของอะไพออริ

3.2.4 ผลลัพธ์ที่ได้จากการวิเคราะห์ของอัลกอริทึมในโปรแกรมเวก้าจะแสดงไอเทมเซตขนาดใหญ่เรียงตามลำดับจำนวนสมาชิกในเซต และชื่อของสมาชิกที่อยู่ในเซตนั้น โดยมีรายละเอียดดังต่อไปนี้

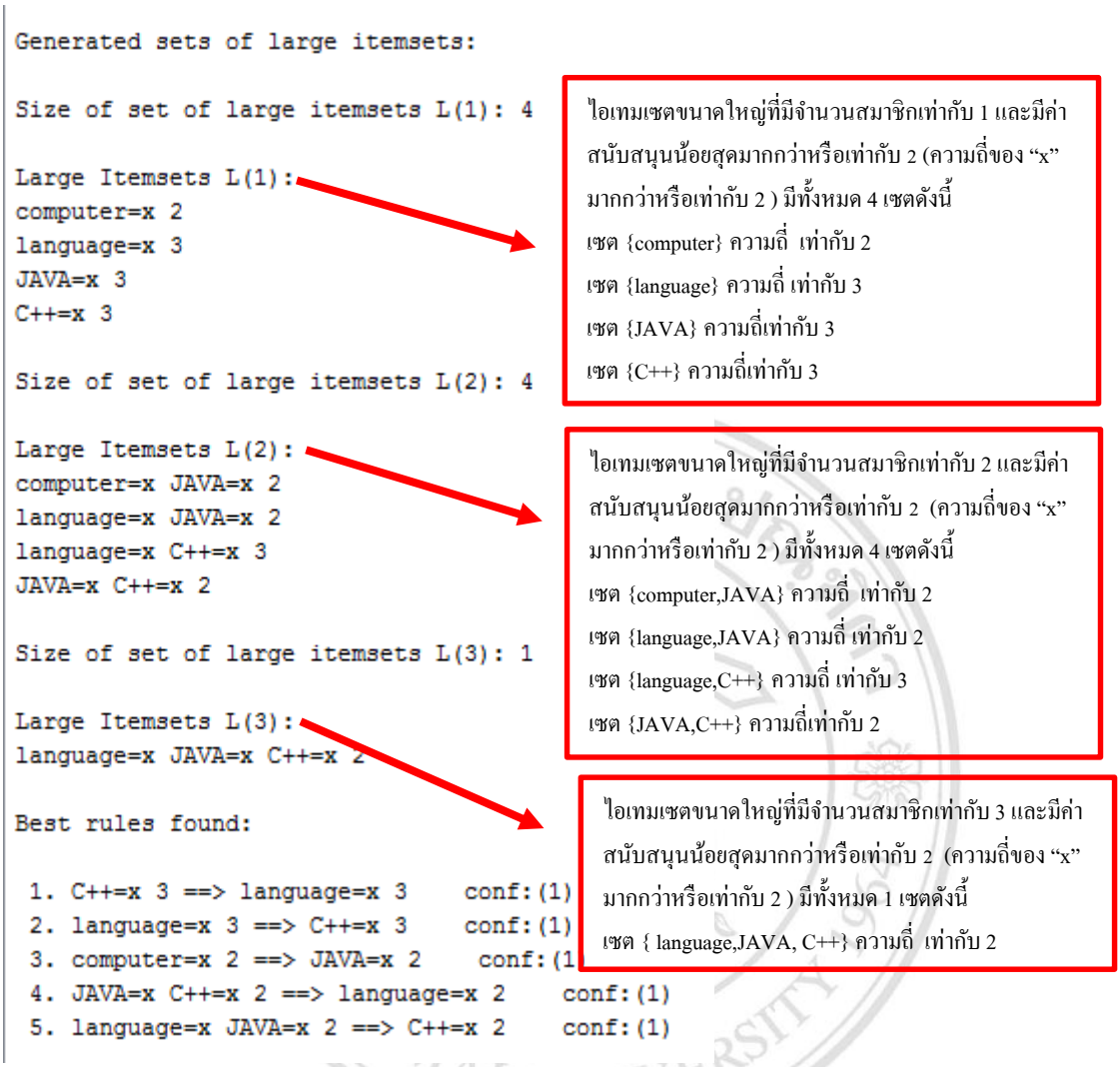
1) จากตัวอย่างข้อมูลหนังสือทั้งหมด 4 เล่ม มีคำสำคัญทั้งหมด 6 คำ โดยปรับค่าสนับสนุนน้อยสุด (min_sup) = 2 ได้ไอเทมเซตขนาดใหญ่ทั้งหมด 9 เซต แบ่งเป็นเซตที่มีสมาชิกเท่ากับหนึ่งจำนวนสี่เซต เซตที่มีจำนวนสมาชิกเท่ากับสองจำนวนสี่เซต และเซตที่มีจำนวนสมาชิกเท่ากับสามจำนวนหนึ่งเซต และที่ค่าความเชื่อมั่น (confidence) = 90% ค้นพบกฎที่ดีที่สุด (Best rules) ทั้งหมด 5 กฎ

2) ความหมายของกฎแต่ละข้อสามารถอธิบายได้ตามตารางที่ 3.2

ตารางที่ 3.2 ตารางแสดงกฎที่ดีที่สุดและความหมาย

Best rules	confidence	ความหมาย
C++=x 3 ==> Language=x 3	1	หนังสือที่มี C++ เป็นคำสำคัญทุกเล่มจะมี Language เป็นคำสำคัญด้วย มีทั้งหมด 3 เล่ม (x=3)
Language=x 3 ==> C++=x 3	1	หนังสือที่มี Language เป็นคำสำคัญทุกเล่ม จะมี C++ เป็นคำสำคัญด้วย มีทั้งหมด 3 เล่ม (x=3)
computer=x 2 ==> JAVA =x 2	1	หนังสือที่มี computer เป็นคำสำคัญทุกเล่มจะมี JAVA เป็นคำสำคัญด้วย มีทั้งหมด 2 เล่ม (x=2)
JAVA=x C++=x 2 ==> Language=x 2	1	หนังสือที่มี JAVA และ C++ เป็นคำสำคัญทุกเล่มจะมี Language เป็นคำสำคัญด้วย มีทั้งหมด 2 เล่ม (x=2)
Language=x JAVA=x 2 ==> C++= x 2	1	หนังสือที่มี Language และ JAVA เป็นคำสำคัญทุกเล่มจะมี C++ เป็นคำสำคัญด้วยมีทั้งหมด 2 เล่ม (x=2)

รูปแบบผลลัพธ์จากการวิเคราะห์โดยโปรแกรมเวก้าแสดงตามตัวอย่างภาพที่ 3.7 ผู้ศึกษาได้นำผลลัพธ์ของเซตขนาดใหญ่ไปบันทึกไว้ในตารางเซตคำสำคัญ (item_set) เพื่อใช้ในการแนะนำหนังสือของห้องสมุดเรียงตามลำดับขนาดของเซต ซึ่งระบบจะแนะนำหนังสือที่มีคำสำคัญอยู่ในเซตที่มีขนาดใหญ่ก่อนเรียงตามลำดับขนาดของเซต ซึ่งจะอธิบายในบทต่อไป



ภาพที่ 3.7 รูปแบบผลลัพธ์และคำอธิบาย