

ระบบหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐาน ทศนิยมดิวอี้

Automated Dewey Decimals Multiple Relation Classification System

ภูริวัตร คัมภีรภาพพัฒน์¹ อนิราข มิ่งขวัญ² และ พยุง มีสัจ³

¹ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

²ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีและการจัดการอุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

³ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

Poorivat.K@rmutk.ac.th¹ Anirach@ieee.org² pym@kmutnb.ac.th³

บทคัดย่อ

ปัจจุบันการจัดหมวดหมู่หนังสือเป็นลักษณะของการจัดหมวดหมู่ที่เจาะจงเพียงหมวดหมู่เดียวเฉพาะหมวดหมู่หลักหรือเฉพาะทางเท่านั้น ทั้งที่เนื้อหาของหนังสือมีความเกี่ยวข้องกับหลายหมวดหมู่ และการจัดหมวดหมู่ในอินเทอร์เน็ตไม่ได้อิงมาตรฐานสากลที่ใช้ในห้องสมุดไม่มีการเชื่อมโยงข้อมูลห้องสมุด

งานวิจัยนี้ผู้วิจัยนำเสนอการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้ มีการสกัดข้อมูล การวิเคราะห์คำหลัก การหาคำน้ำหนัก ทาเปอร์เซ็นต์ของเนื้อหาคำหลัก การสร้างกฎสำหรับการจัดหมวดหมู่ โดยส่วนแรกเป็นการจัดหมวดหมู่อัตโนมัติแบบเจาะจงเพื่อจัดให้ตรงกับหมวดหมู่หนังสือจากเว็บไซต์ OhioLINK มีข้อมูล 2 กลุ่มคือ (1) กลุ่มการเรียนรู้กับการตรวจสอบการทำงานและ (2) กลุ่มทดสอบต้นแบบเพื่อหาประสิทธิภาพค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ส่วนที่สองเป็นการจัดหมวดหมู่แบบสหสัมพันธ์เพื่อให้แสดงหมวดหมู่ที่เกี่ยวข้องกับของหนังสือ และได้ทำการวิเคราะห์ข้อบทความใน Wikipedia เพื่อจะให้เกิดการเชื่อมโยงเนื้อหาหนังสือที่มีอยู่ในห้องสมุด

ซึ่งผลการวิจัยครั้งนี้การจัดหมวดหมู่ทุกหมวดหมู่มี 2 กลุ่ม คือ (1) ใช้ DDC-MR Rule ประสิทธิภาพค่าความถูกต้อง 88 เปอร์เซ็นต์ ค่าความแม่นยำ 90 เปอร์เซ็นต์ ค่าความระลึก 96 เปอร์เซ็นต์ และค่าความถ่วงดุล 93 เปอร์เซ็นต์ และ (2) ใช้โครงข่ายประสาทเทียมประสิทธิภาพค่าความถูกต้อง 73 เปอร์เซ็นต์ ค่าความแม่นยำ 78 เปอร์เซ็นต์ ค่าความระลึก 73 เปอร์เซ็นต์ และค่าความถ่วงดุล 71 เปอร์เซ็นต์ และทำให้สามารถแสดงทุกหมวดหมู่ที่เกี่ยวข้องของหนังสือ และสามารถเชื่อมโยงเนื้อหาบนอินเทอร์เน็ตกับหนังสือในห้องสมุดได้

Abstract

The current book classification system in libraries usually assign only one category to each book, while the content of a book pertains to a variety of categories. The Internet classification also not comply with any international standard, and has no related between the Internet information and the books in library.

This research were presented Automated Dewey Decimals Multiple Relation Classification System consisted of the Information Extraction, the Keyword Analysis, the evaluation of Keyword Weight, the categorization are accordance with the DDC-MR Rules. The first stage is the automatic classification for specific category we compared the classified category with information from OhioLINK. The efficiency of our prototype are measure Accuracy, Precision, Recall and F-Measure. The second innovation is the automatic DDC-MR classification, showing the category pertaining to the book content and was also analyze article title in Wikipedia, which allow to link to the content in the library.

This research of experiment results could separate into 2 groups : (1) Only use DDC-MR Rule the efficiency are Accuracy 88 %, Precision 90%, Recall 96 % and F-Measure 93 % and (2) use Neural networks on efficiency are Accuracy 73 %, Precision 78%, Recall 73 % and F-Measure 71 % . The other results from this work is the analysis of information could be compare values books in the library to classified as the same groups of information.

1. คำนำ

ลักษณะของสังคมในปัจจุบันเป็นสังคมสารสนเทศ (Information Society) [6] ดังนั้นทุกคนจำเป็นต้องใช้ข่าวสารจากแหล่งต่าง ๆ เพื่อพัฒนาชีวิตความเป็นอยู่ การประกอบอาชีพตลอดจนร่วมพัฒนาสังคมประเทศชาติ แต่การให้บริการข่าวสารจากแหล่งต่างๆ นั้นได้มีการจัดหมวดหมู่เจาะจงลงไปที่หมวดหมู่ใดหมวดหมู่หนึ่งเท่านั้น โดยไม่ได้มองถึงภาพรวมของเนื้อหาในข่าวสารและไม่มีเกณฑ์มาตรฐานสากลที่ใช้ในห้องสมุดรองรับ [6] ซึ่งการจัดหมวดหมู่หนังสือในห้องสมุด เป็นการจัดหมวดหมู่เจาะจงลงไปที่หมวดหมู่ใดหมวดหมู่หนึ่งเช่นกัน ทั้งที่เนื้อหาของหนังสือมีความเกี่ยวข้องกับหลายๆ หมวดหมู่ และการจัดหมวดหมู่หนังสือใน

ห้องสมุดแบบอัตโนมัติ [10] เป็นการจัดหมวดหมู่เจาะจงหมวดหมู่ใดหมวดหมู่หนึ่งเฉพาะหนังสือบางกลุ่มเท่านั้น ดังนั้นในงานวิจัยครั้งนี้จึงได้มีการนำเสนอต้นแบบการจัดหมู่แบบอัตโนมัติตามมาตรฐานทศนิยมดิวอี้ ที่ประกอบด้วยการจัดหมวดหมู่แบบอัตโนมัติเพื่อให้ได้หมวดหมู่แบบเฉพาะหมวดหมู่เดียวเช่นเดียวกับบรรณารักษ์พร้อมกับการจัดหมวดหมู่แบบสหสัมพันธ์ [8][9] ที่มีการเชื่อมโยงไปยังทุกหมวดหมู่ที่เกี่ยวข้องกับหนังสือเล่มนั้นๆ และมีการเชื่อมโยงข้อมูลบนอินเทอร์เน็ตกับหนังสือในห้องสมุด เพื่อให้ข้อมูลบนอินเทอร์เน็ตได้มีมาตรฐานที่เป็นสากลรองรับและหนังสือเอกสารในห้องสมุดจะได้ถูกนำไปใช้อ้างอิงมากขึ้นตามปริมาณของข้อมูลบนอินเทอร์เน็ตซึ่งเติบโตเพิ่มขึ้นในทุกๆ วัน

การจัดหมวดหมู่ของบทความหรือหนังสือในปัจจุบันเป็นการจัดหมวดหมู่ที่เจาะจงเพียงหมวดหมู่เดียว ซึ่งในงานวิจัยที่ผ่านมาผู้วิจัยได้นำเสนอการจัดหมวดหมู่ทศนิยมดิวอี้แบบสหสัมพันธ์ (DDC-MR) [17][7] ทำให้มองเห็นเนื้อหาของหมวดหมู่อื่นที่ซ่อนอยู่ในบทความหรือหนังสือ ในงานวิจัยนี้ผู้วิจัยได้ทำการทดลองและการวิเคราะห์เนื้อหาของบทความใน Wikipedia [6] มาใช้ในการจัดหมวดหมู่แบบสหสัมพันธ์ เพื่อทำให้เกิดการเชื่อมโยงเนื้อหาที่มีอยู่ใน Wikipedia และเอกสารในห้องสมุด มีผลให้รูปแบบและความรวดเร็วในการส่งต่อข้อมูลดีขึ้น เกิดการเชื่อมโยงเนื้อหาของบทความใน Wikipedia ภายใต้กรอบความรู้ที่เป็นมาตรฐานสากล แทนการเชื่อมโยงด้วย Keyword เพียงอย่างเดียว และทำให้สามารถค้นหาเอกสารในห้องสมุดที่มีเนื้อหาตรงกัน และให้มีข้อมูลที่มีรายละเอียดที่ชัดเจนขึ้น

ในงานวิจัยครั้งนี้ผู้วิจัยมุ่งการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามกลุ่มเนื้อหาภายใต้กรอบความรู้การจัดหมวดหมู่ในห้องสมุด โดยนำเสนอกรรมวิธีเทคนิคการจัดหมวดหมู่อัตโนมัติตามระบบมาตรฐานทศนิยมดิวอี้ (Dewey Decimal Classification) ที่ใช้ในห้องสมุด เนื่องจากเป็นระบบที่นิยมใช้จัดหมวดหมู่ในห้องสมุดเป็นเวลากว่า 100 ปี ใช้ตัวเลขเป็นสัญลักษณ์ง่ายต่อการจดจำและมีการแบ่งลำดับชั้นที่ชัดเจน นอกจากนี้ผู้วิจัยได้นำเทคนิค

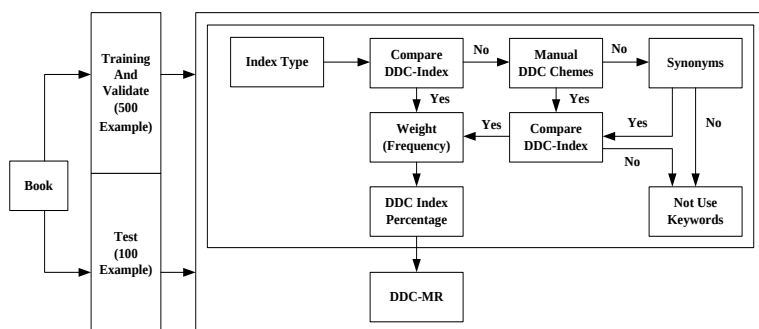
Information Extraction และ Data Mining [15] มาช่วยในการพัฒนาให้สามารถจัดเก็บคำสำคัญเพื่อสร้างฐานข้อมูล ทำให้การจัดเก็บข้อมูลเป็นไปอย่างมีระบบ และมีการสร้างต้นแบบเพื่อให้มีการตรวจสอบเรียนรู้การจัดหมวดหมู่แบบอัตโนมัติทั้งที่เป็นการจัดหมวดหมู่แบบเจาะจงเพื่อเปรียบเทียบกับการจัดหมวดหมู่จากเว็บไซต์ OhioLINK – The Ohio Library and Information Network [4] สำหรับวัดความถูกต้อง ความแม่นยำ ประสิทธิภาพ และได้มีการจัดหมวดหมู่แบบสหสัมพันธ์ ทำให้สามารถทราบปริมาณของหมวดหมู่ที่เกี่ยวข้องกับหนังสือ และสามารถเชื่อมโยงเนื้อหากับข้อมูลบนอินเทอร์เน็ตได้

2. วัตถุประสงค์

ดังนั้นการวิจัยในครั้งนี้นำผู้วิจัยมุ่งที่พัฒนาระบบการที่สามารถวิเคราะห์เนื้อหา จัดหมวดหมู่แบบอัตโนมัติ คำสำคัญ ของชื่อหนังสือที่มีความสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้ได้ในทุกกลุ่มข้อมูลและทุกระดับหมวดหมู่ โดยให้มีความถูกต้องและความแม่นยำ

3. ขอบเขตการวิจัย

ในการวิจัยในครั้งนี้นำเสนอผังระบบของกรรมวิธีสำหรับการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้ ดังผังการทำงานในภาพที่ 1



ภาพที่ 1 : แสดงผังการทำงานการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้

จากภาพที่ 1 ผังการทำงานการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้ เป็นการใช้อัตโนมัติหนังสือ โดยมีการแบ่งข้อมูลออกเป็น 2 ส่วน คือ ส่วนที่ใช้สำหรับเรียนรู้ ตรวจสอบความถูกต้องของการทำงานในระบบใช้อัตโนมัติหนังสือจำนวน 2374 เล่ม และส่วนที่ใช้ในการทดสอบความแม่นยำระบบใช้อัตโนมัติหนังสือจำนวน 500 เล่ม ซึ่งทุกส่วนจะทำงานตามลำดับขั้นตอนดังนี้

1) การวิเคราะห์การสกัดคำ การแบ่งคำ การหารากศัพท์ ตามเทคนิคกระบวนการ Information Extraction (IE) เพื่อหาคำสำคัญ (Keyword) ของหนังสือ

2) นำคำสำคัญเปรียบเทียบดัชนีตามมาตรฐานทศนิยมดิวอี้ ถ้าเปรียบเทียบตรงกันให้หมวดหมู่ของคำสำคัญ และไปทำข้อ 6

3) นำคำสำคัญเปรียบเทียบคำสำคัญของผู้เชี่ยวชาญ ถ้าเปรียบเทียบได้และไปทำข้อ 5

4) นำคำสำคัญเปรียบเทียบคำสำคัญของคำเหมือน ถ้าเปรียบเทียบไม่ได้ให้ตัดคำสำคัญทิ้งไป

5) นำคำสำคัญของผู้เชี่ยวชาญหรือของคำเหมือนเปรียบเทียบดัชนีตามมาตรฐานทศนิยมดิวอี้ ถ้าเปรียบเทียบตรงกันให้หมวดหมู่ของคำสำคัญ

6) หาน้ำหนักของคำสำคัญ

7) หาจำนวนเปอร์เซ็นต์ของคำสำคัญแต่ละตัว ตามสูตรที่ 1

$$P_n = \frac{N_n * 100}{\sum_{n=0}^9 N_n} \quad \text{.....สูตรที่ 1}$$

P_n =จำนวนเปอร์เซ็นต์ของแต่ละคลาส

N_n = จำนวนค่าของแต่ละคลาส

n =คลาสของหมวดหมู่

8) แสดงหมวดหมู่ตามมาตรฐานทศนิยมดิวอี้แบบสหสัมพันธ์

9) คำนวณค่าที่เกี่ยวข้องกับความถูกต้องและความแม่นยำ

- ค่า Precision

$$Pre_i = \frac{T_{ii}}{\sum column} \quad \text{....สูตรที่ 2}$$

Pre_i =ค่า Precision แต่ละคลาส

T_{ii} =จำนวนจัดหมวดหมู่ได้ถูกต้องของคลาสที่ i

$\sum column$ =ผลรวมของจำนวนของหนังสือแนว

คอลัมน์ของคลาสที่ i

i =ลำดับของคลาสมีค่าตั้งแต่ 0 ถึง 9

- ค่า Recall

$$Re_i = \frac{T_{ii}}{\sum row} \quad \text{....สูตรที่ 3}$$

Re_i =ค่า Recall แต่ละคลาส

T_{ii} =จำนวนจัดหมวดหมู่ได้ถูกต้องของคลาสที่ i

$\sum row$ =ผลรวมของจำนวนของหนังสือแนว

ของคลาสที่ i

i =ลำดับของคลาสมีค่าตั้งแต่ 0 ถึง 9

-ค่าความแม่นยำ (Accuracy)

$$Accuracy = \frac{\sum T_{ii}}{N} \quad \text{....สูตรที่ 4}$$

Accuracy =ค่าความแม่นยำ

$\sum T_{ii}$ =ผลรวมของจำนวนจัดหมวดหมู่ได้ถูกต้องในแต่ละคลาส

N =จำนวนทั้งหมดของหนังสือ

i =ลำดับของคลาสมีค่าตั้งแต่ 0 ถึง 9

-ค่า F-Measure

$$F-Measure_i = 2 * \frac{Pre_i * Re_i}{Pre_i + Re_i} \quad \text{....สูตรที่ 5}$$

$F-Measure_i$ =ค่า F-Measure แต่ละคลาส

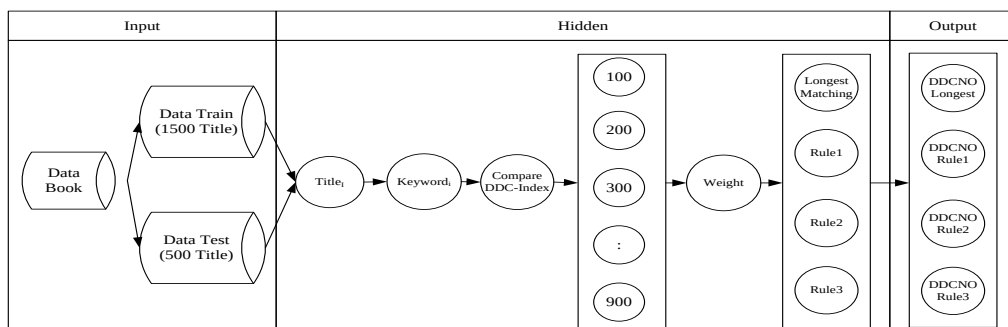
Pre_i =ค่า Precision แต่ละคลาส

Re_i =ค่า Recall แต่ละคลาส

ซึ่งการทำงานทั้ง 9 ขั้นตอนเป็นการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้ ได้มีการแบ่งส่วนการทำงานของระบบออกเป็น 3 ส่วน ดังภาพที่ 2

จากภาพที่ 2 ผังส่วนของระบบการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้ ได้มีการแบ่งส่วนการทำงานของระบบออกเป็น 3 ส่วน คือ

1) ส่วนของการนำข้อมูลเข้า โดยการนำข้อมูลชื่อหนังสือเข้าสู่ระบบการทำงาน ซึ่งได้แบ่งออกเป็น 2 กลุ่ม ได้แก่กลุ่มที่ใช้ทำงานในระบบใช้ข้อมูล



ภาพที่ 2 : แสดงผังส่วนของการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้

4) ส่วนของการนำข้อมูลเข้า โดยการนำข้อมูลชื่อหนังสือเข้าสู่ระบบการทำงาน ซึ่งได้แบ่งออกเป็น 2 กลุ่ม ได้แก่กลุ่มที่ใช้ทำงานในระบบใช้ข้อมูลชื่อหนังสือจำนวน 2374 เล่มและกลุ่มที่ใช้ในการทดสอบระบบใช้ข้อมูลชื่อหนังสือจำนวน 500 เล่ม

จากภาพที่ 3 ได้มีการกำหนดกฎการวิเคราะห์การจัดหมวดหมู่สำหรับบรรณารักษ์แบบเจาะจงหมวดหมู่เดียวเพื่อใช้ทดสอบความแม่นยำของระบบจำนวน 4 กฎ ดังนี้

1) Longest Matching เป็นกฎที่ใช้สำหรับเปรียบเทียบชื่อหนังสือกับดัชนีหมวดหมู่ตาม

ชื่อหนังสือจำนวน 2374 เล่มและกลุ่มที่ใช้ในการทดสอบระบบใช้ข้อมูลชื่อหนังสือจำนวน 500 เล่ม

2) ส่วนของการเรียนรู้ ตรวจสอบความถูกต้องของการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้ เป็นส่วนที่ไม่สามารถมองเห็นได้ จะเป็นการทำงานตั้งขั้นตอนที่ 1 ถึงขั้นตอนที่ 7 ของภาพที่ 1 โดยค่าสำคัญจะต้องมีการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์และมีการตั้งกฎการจัดหมวดหมู่สำหรับบรรณารักษ์แบบเจาะจงหมวดหมู่เพื่อทดสอบความแม่นยำของระบบไว้จำนวน 4 กฎ ดังภาพที่ 3

3) ส่วนแสดงข้อมูลหมวดหมู่สำหรับบรรณารักษ์

มาตรฐานทศนิยมดิวอี้ ถ้าตรงกันให้ใช้หมวดหมู่นั้นเป็นหมวดหมู่สำหรับบรรณารักษ์แบบเจาะจงหมวดหมู่เดียว

2) Rule 1 เป็น กฎ ที่ ต้อง ผ่าน การเปรียบเทียบดังนี้

- เปรียบเทียบจำนวนน้ำหนักของระดับที่ 3 ที่มีจำนวนมากที่สุด และ
- เปรียบเทียบจำนวนน้ำหนักของจุดที่ 1 ของระดับที่ 4 ที่มีจำนวนมากที่สุด และ
- เปรียบเทียบจำนวนน้ำหนักของจุดที่ 2 ของระดับที่ 4 ที่มีจำนวนมากที่สุด

- ให้ใช้หมวดหมู่นั้นเป็นหมวดหมู่สำหรับบรรณารักษ์แบบเจาะจงหมวดหมู่เดียว

3) Rule 2 เป็น กฎ ที่ ต้อง ผ่าน การเปรียบเทียบดังนี้

- เปรียบเทียบจำนวนคำสำคัญของระดับที่ 3 ที่มีจำนวนมากที่สุด

- เปรียบเทียบจำนวนน้ำหนักของระดับที่ 3 ที่มีจำนวนมากที่สุด และ

- เปรียบเทียบจำนวนน้ำหนักของจุดที่ 1 ของระดับที่ 4 ที่มีจำนวนมากที่สุด และ

- เปรียบเทียบจำนวนน้ำหนักของจุดที่ 2 ของระดับที่ 4 ที่มีจำนวนมากที่สุด

- ให้ใช้หมวดหมู่นั้นเป็นหมวดหมู่สำหรับบรรณารักษ์แบบเจาะจงหมวดหมู่เดียว

4) Rule 2 เป็น กฎ ที่ ต้อง ผ่าน การเปรียบเทียบดังนี้

- จำนวนคำสำคัญที่ตรงกันกับดัชนีของระดับที่ 3 ที่มีจำนวนมากที่สุด

- เปรียบเทียบจำนวนคำสำคัญของระดับที่ 3 ที่มีจำนวนมากที่สุด

- เปรียบเทียบจำนวนน้ำหนักของระดับที่ 3 ที่มีจำนวนมากที่สุด และ

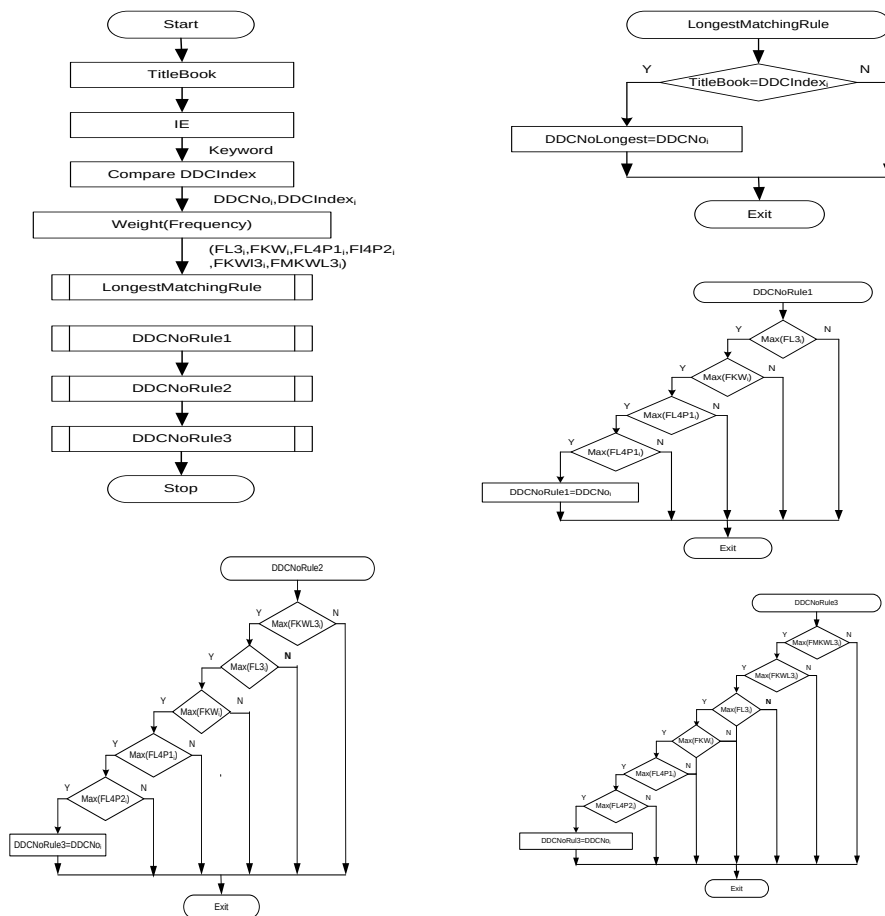
- เปรียบเทียบจำนวนน้ำหนักของจุดที่ 1 ของระดับที่ 4 ที่มีจำนวนมากที่สุด และ

- เปรียบเทียบจำนวนน้ำหนักของจุดที่ 2 ของระดับที่ 4 ที่มีจำนวนมากที่สุด

- ให้ใช้หมวดหมู่นั้นเป็นหมวดหมู่สำหรับบรรณารักษ์แบบเจาะจงหมวดหมู่เดียวจากวิธีการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้ของวิจัยครั้งนี้ทำให้ได้ต้นแบบเครื่องมือการจัดหมวดหมู่แบบอัตโนมัติที่มีความแม่นยำในการทำงานและหมวดหมู่สำหรับบรรณารักษ์

4. ผลการวิจัย

การวิจัยครั้งนี้ผู้วิจัยได้มีการทดสอบวิธีการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้ ด้วยการทดสอบโดยนำข้อมูลหนังสือจำนวน 2000 เล่มเพื่อใช้สำหรับการเรียนรู้ตรวจสอบความถูกต้องของโมเดลการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามระบบมาตรฐานทศนิยมดิวอี้ พร้อมสร้างฐานข้อมูล คำนวณสัดส่วนค่าร้อยละของเนื้อหาตามกรอบความรู้ระบบมาตรฐานทศนิยมดิวอี้ และใช้ข้อมูลหนังสือจำนวน 2500 เล่มสำหรับการทดสอบความแม่นยำ ซึ่งมีข้อมูลหนังสือดังตารางที่ 1



ภาพที่ 3 : แสดงกฎการจัดหมวดหมู่สำหรับบรรณารักษ์แบบเจาะจงหมวดหมู่เดียว

ตารางที่ 1 แสดงข้อมูลชื่อหนังสือ

ลำดับที่หนังสือ	ชื่อหนังสือ	เลขดัชนีของบรรณารักษ์
1	100 years fashion illustration	741.672
2	20th century jewelry complete sourcebook with	739.270 904
3	3-d textile reinforcements composite materials	620.118

ในการวิจัยได้นำข้อมูลข้างต้นเข้าสู่การประมวลผลการทดสอบด้วยการวิเคราะห์การสกัดคำ การแบ่งคำ การหารากศัพท์ แล้วนำคำสำคัญของชื่อหนังสือหรือของผู้เชี่ยวชาญหรือของคำเสมือน ไปเปรียบเทียบกับดัชนีหมวดหมู่ตามมาตรฐานทศนิยมดิวอี้

เพื่อให้หมวดหมู่ตามมาตรฐานทศนิยมดิวอี้หน้าหนักของคำสำคัญ และหาจำนวนเปอร์เซ็นต์ของคำสำคัญแต่ละหมวดหมู่ได้ผลตารางที่ 2

ตารางที่ 2 แสดงข้อมูลหมวดหมู่หนังสือคำสำคัญ

ลำดับหนังสือ	คำสำคัญ	จำนวนคำ	จำนวนความยาวของคำ	หมวดหมู่	เปอร์เซ็นต์
1	fashion	9	1	746.92	25.00
1	illustration	2	1	741.6	8.33
1	years	1	1	529.2	8.33

ได้มีการนำข้อมูลข้างต้นผ่านกฎการจัดหมวดหมู่แบบบรรณารักษ์ ดังตารางที่ 3 และผลการเปรียบเทียบหมวดหมู่ของบรรณารักษ์ ดังตารางที่ 4 เพื่อคำนวณหาความแม่นยำ ดังตารางที่ 3

ตารางที่ 3 แสดงหมวดหมู่แบบบรรณารักษ์ของแต่ละกฎ

ลำดับหนังสือ	เลขดัชนี					
	บรรณารักษ์	โปรแกรม	กฎ Longest	กฎที่ 1	กฎที่ 2	กฎที่ 3
1	741.672	741.672	-	746.92	741.672	391
2	739.270 904	739.27	-	739.27	739.27	391.7
3	620.118	620.118	-	620.118	620.118	620.118

จากตารางที่ 3 เป็นผลของการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามระบบมาตรฐานทศนิยมดิวอี้ และนำผลของหมวดหมู่มาผ่านกฎการวิเคราะห์การจัดหมวดหมู่สำหรับบรรณารักษ์แบบเจาะจงหมวดหมู่เดียว โดยผ่านกฎ Longest ได้หมวดหมู่ในคอลัมน์ เลขดัชนีกฎ Longest, ผ่านกฎ Rule1 ได้หมวดหมู่ในคอลัมน์ เลขดัชนีกฎที่ 1 ผ่านกฎ Rule2 ได้หมวดหมู่ในคอลัมน์ เลขดัชนีกฎที่ 2 และผ่านกฎ Rule3 ได้หมวดหมู่ในคอลัมน์ เลขดัชนีกฎที่ 3 ซึ่งโปรแกรมได้มีการประมวลผลเป็นแบบหมวดหมู่แบบอัตโนมัติให้ในคอลัมน์ เลขดัชนีโปรแกรม

การวิจัยครั้งนี้ผู้วิจัยได้มีการทดสอบวิธีการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้ ได้มีการพัฒนาวิธีการเชื่อมโยงข้อมูลบทความในอินเทอร์เน็ตกับหนังสือในห้องสมุดที่จัดอยู่ในหมวดหมู่เดียวกันได้ โดยการทดสอบนำข้อมูลหนังสือจำนวน 2,000 เล่มเพื่อใช้ในการเรียนรู้ตรวจสอบความถูกต้องของโมเดลการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามระบบมาตรฐานทศนิยมดิวอี้ พร้อมสร้างฐานข้อมูล คำนวณสัดส่วนค่าร้อยละของเนื้อหาตามกรอบความรู้ระบบมาตรฐานทศนิยมดิวอี้ และใช้ข้อมูลหนังสือจำนวน 2,500 เล่ม เพื่อหาประสิทธิภาพค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ซึ่งมีผลการคำนวณดังภาพที่ 4

จากภาพที่ 4 รายชื่อหนังสือที่ใช้สำหรับทดสอบระบบ (Total Data) มีจำนวน 2,500 เล่ม ซึ่งมีจำนวนหนังสือที่มีหมวดหมู่ได้จำนวน 419 เล่ม โดยจัดหมวดหมู่ได้ทุกระดับของบรรณารักษ์ไม่ตรงกับหมวดหมู่ของโปรแกรม (Mismatching Title)

[Confusion Matrix in Section (Level 4)]Data Test					
Conclusion Value					
Total Data	419	Book Matching	352	Tii (Matching Title)	371
MisMatching Title	67	MisMatching Title (MisMatch Section)	48		
Precision	0.9	Recall	0.969	F_Measure	0.93
Accuracy	0.89	Accuracy Percent	88.54		

ภาพที่ 4 ค่าความถูกต้อง ความแม่นยำ และประสิทธิภาพของระดับที่ 4

จำนวน 67 เล่ม แต่เมื่อประมวลผลที่หมวดหมู่ระดับที่ 4 มีจำนวนที่มีหมวดหมู่ระดับที่ 4 ตรงกัน (Tii (Matching Class)) มีจำนวน 371 เล่ม มีจำนวนที่มีหมวดหมู่ระดับที่ 4 ไม่ตรงกัน (Tii (MisMatching Class)) มีจำนวน 48 เล่ม

จากงานวิจัยนี้ได้ใช้ข้อมูลหนังสือจำนวน 2,500 เล่ม สามารถจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามระบบมาตรฐานทศนิยมดิวอี้ ซึ่งมีหนังสือจำนวน 419 เล่มที่สามารถใช้ได้กับกฎเกี่ยวกับการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้ และสามารถจำแนกหมวดหมู่ได้ทุกหมวดหมู่โดยมีประสิทธิภาพค่าความถูกต้อง 88 เปอร์เซ็นต์ ค่าความแม่นยำ 90 เปอร์เซ็นต์ ค่าความระลึก 96 เปอร์เซ็นต์ และค่าความถ่วงดุล 93 เปอร์เซ็นต์ แต่ปรากฏว่ามีหนังสือจำนวน 2,081 เล่มมีความจำเป็นที่จะต้องหมวดหมู่ด้วยโครงข่ายประสาทเทียมแบบ มัลติเลเยอร์เชิงปตรอน (MLP) ซัพพอร์ตเวกเตอร์แมชชีน (SVM) และนาอิวเบย์(NB) ซึ่ง

เป็นการช่วยในการจัดหมวดหมู่การเรียนรู้แบบมีผลเฉลยซึ่งมี การเรียนรู้เพื่อสร้างกลุ่มคำสำคัญของข้อหนังสือต้นแบบและแยกหมวดหมู่ของคำสำคัญของข้อหนังสือที่สนใจด้วยการตรวจสอบหาความคล้ายกับกลุ่มคำสำคัญของข้อหนังสือต้นแบบ โดยข้อมูลนำเข้าอัลกอริธึมเครื่องจักรการเรียนรู้ คือคุณลักษณะของคำสำคัญของข้อหนังสือที่ผ่านการลดคุณลักษณะและให้นำหน้ามาเรียบร้อยแล้ว เพื่อหาประสิทธิภาพค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ของหมวดหมู่หนังสือจำนวน 2,081 เล่ม

1) ผลของการจำแนกหมวดหมู่ด้วยโครงข่ายประสาทเทียม ของหมวดหมู่ระดับที่ 1 ซึ่งได้ผลการวิจัยดังนี้

ตารางที่ 4 เปรียบเทียบความถูกต้องของการจำแนกข้อมูลหนังสือของหมวดหมู่ระดับที่ 1

Type	Correctly Number	Correctly Percent	InCorrectly Number	InCorrectly Percent
SVM	1968	94.56	113	5.43
NB	765	36.76	1316	63.23
MLP	365	17.53	1716	82.46

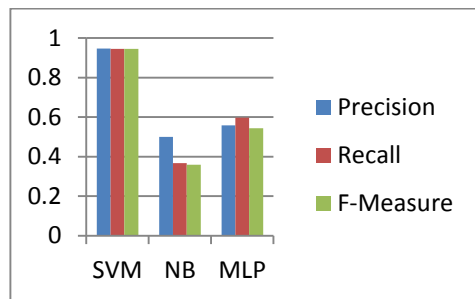
จากตารางที่ 4 คอลัมน์ Type เป็นคอลัมน์ของชนิดโครงข่ายประสาทเทียม คอลัมน์ Correctly เป็นคอลัมน์ของเปอร์เซ็นต์ความถูกต้อง คอลัมน์ InCorrectly และเป็นคอลัมน์ของเปอร์เซ็นต์ไม่ความถูกต้อง ซึ่งมีรายละเอียดของหาประสิทธิภาพค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ดังตารางที่ 4

ตารางที่ 5 เปรียบเทียบค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลของข้อมูลหนังสือหมวดหมู่ระดับที่ 1

Type	Precision	Recall	F-Measure
SVM	0.948	0.946	0.945
NB	0.500	0.368	0.359

MLP	0.559	0.598	0.544
-----	-------	-------	-------

จากตารางที่ 5 คอลัมน์ Type เป็นคอลัมน์ของชนิดโครงข่ายประสาทเทียม คอลัมน์ Precision เป็นคอลัมน์ของความแม่นยำ คอลัมน์ Recall เป็นคอลัมน์ของความระลึก และคอลัมน์ F-Measure เป็นคอลัมน์ของความถ่วงดุล ซึ่งสามารถแสดงเป็นกราฟได้ ดังภาพที่ 5



ภาพที่ 5 เปรียบเทียบค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุลของข้อมูลหนังสือหมวดหมู่ระดับที่ 1

จากภาพที่ 5 แสดงให้เห็นถึงค่าประสิทธิภาพด้วยค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความถ่วงดุล ของโครงข่ายประสาทเทียม

จากผลการวิจัยการจำแนกหมวดหมู่หนังสือจำนวน 2081 เล่ม ซึ่งเป็นหมวดหมู่ระดับที่ 1 สามารถจำแนกหมวดหมู่ด้วยโครงข่ายประสาทเทียมที่ดีที่สุดคือการใช้ SVM โดยมีประสิทธิภาพค่าความถูกต้อง 94 เปอร์เซ็นต์ ค่าความแม่นยำ 94 เปอร์เซ็นต์ ค่าความระลึก 94 เปอร์เซ็นต์ และค่าความถ่วงดุล 71 เปอร์เซ็นต์

ผลของการจำแนกหมวดหมู่ด้วยโครงข่ายประสาทเทียม ของหมวดหมู่ระดับที่ 2 ซึ่งได้ผลการวิจัยดังนี้

ตารางที่ 6 เปรียบเทียบความถูกต้องของการจำแนกข้อมูลหนังสือหมวดหมู่ระดับที่ 2

Type	Correctly Number	Correctly Percent	InCorrectly Number	InCorrectly Percent
SVM	1524	73.23	557	26.76

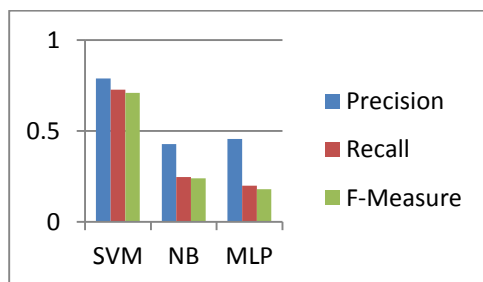
NB	518	24.89	1563	75.10
MLP	365	17.53	1716	82.46

จากตารางที่ 6 คอลัมน์ Type เป็นคอลัมน์ของชนิดโครงข่ายประสาทเทียม คอลัมน์ Correctly เป็นคอลัมน์ของเปอร์เซ็นต์ความถูกต้อง คอลัมน์ InCorrectly และเป็นคอลัมน์ของเปอร์เซ็นต์ไม่ความถูกต้อง ซึ่งมีรายละเอียดของหาประสิทธิภาพค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึกลับ และค่าความถ่วงดุล ดังตารางที่ 7

ตารางที่ 7 เปรียบเทียบค่าความแม่นยำ ค่าความระลึกลับ และค่าความถ่วงดุลของข้อมูลหนังสือหมวดหมู่ระดับที่ 2

Type	Precision	Recall	F-Measure
SVM	0.786	0.732	0.713
NB	0.427	0.249	0.240
MLP	0.456	0.198	0.180

จากตารางที่ 7 คอลัมน์ Type เป็นคอลัมน์ของชนิดโครงข่ายประสาทเทียม คอลัมน์ Precision เป็นคอลัมน์ของความแม่นยำ คอลัมน์ Recall เป็นคอลัมน์ของความระลึกลับ และคอลัมน์ F-Measure เป็นคอลัมน์ของความถ่วงดุล ซึ่งสามารถแสดงเป็นกราฟได้ดังภาพที่ 6



ภาพที่ 6 เปรียบเทียบค่าความแม่นยำ ค่าความระลึกลับ และค่าความถ่วงดุลของข้อมูลหนังสือ

จากภาพที่ 6 แสดงให้เห็นถึงค่าประสิทธิภาพด้วยค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึกลับ

และค่าความถ่วงดุล ของโครงข่ายประสาทเทียมจากผลการวิจัยการจำแนกหมวดหมู่หนังสือจำนวน 2,081 เล่ม ซึ่งเป็นหมวดหมู่ระดับที่ 2 สามารถจำแนกหมวดหมู่ด้วยโครงข่ายประสาทเทียมที่ดีที่สุดคือการใช้ SVM โดยมีประสิทธิภาพค่าความถูกต้อง 73 เปอร์เซ็นต์ ค่าความแม่นยำ 78 เปอร์เซ็นต์ ค่าความระลึกลับ 73 เปอร์เซ็นต์ และค่าความถ่วงดุล 71 เปอร์เซ็นต์

5.สรุป

ปรากฏผลการวิจัยได้มีการแบ่งเป็น 2 ส่วน คือ 1) ส่วนของการสร้างกฎเพื่อจำแนกหมวดหมู่ให้ได้ทุกหมวดหมู่ ข้อมูลจำนวน 419 เล่ม สามารถโดยมีประสิทธิภาพค่าความถูกต้อง 88 เปอร์เซ็นต์ ค่าความแม่นยำ 90 เปอร์เซ็นต์ ค่าความระลึกลับ 96 เปอร์เซ็นต์ และค่าความถ่วงดุล 93 เปอร์เซ็นต์ และ 2) การใช้โครงข่ายประสาทเทียมจำแนกหมวดหมู่ข้อมูลจำนวน 2,081 เล่ม อัลกอริธึมที่ดีที่สุดคือการใช้ SVM โดยมีประสิทธิภาพค่าความถูกต้อง 73 เปอร์เซ็นต์ ค่าความแม่นยำ 78 เปอร์เซ็นต์ ค่าความระลึกลับ 73 เปอร์เซ็นต์ และค่าความถ่วงดุล 71 เปอร์เซ็นต์ จากผลการดำเนินการพัฒนาต้นแบบพบว่างานวิจัยนี้สามารถทำให้

- 1) มีระบบจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้
- 2) มีกฎเกี่ยวกับการจัดหมวดหมู่อัตโนมัติแบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้
- 3) มีต้นแบบการจัดหมวดหมู่แบบสหสัมพันธ์ตามมาตรฐานทศนิยมดิวอี้ แสดงข้อมูลทุกหมวดหมู่ที่เกี่ยวข้อง
- 4) มีพัฒนาวิธีการเชื่อมโยงข้อมูลบทความในอินเทอร์เน็ตกับหนังสือในห้องสมุดที่จัดอยู่ในหมวดหมู่เดียวกันได้

6. เอกสารอ้างอิง

- [1]. Barbara Daniel. Carlotta, D., Kang, N. "Mining relevant text from unlabelled documents." Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03), 2003.
- [2]. ohioLINK – The Ohio Library and Information Network [Online] Available From: <http://www.ohiolink.edu/>
- [3]. Huang Chong and et al. "Towards multi-granularity multi-facet E-book eval." ACM 9 7 8 -1 -5 9 5 9 3 -6 5 4 - 7/07/0005, 2007.
- [4]. Kampeerapaappat Poorivat and Mingkhwan Anirach. "Method for Website Categorize Using Scale Dewey Decimals Classification Scheme.", in Proceedings of 4th National Conference on Computing and Information Technology, NCCIT2 0 0 8 , Bangkok, Thailand, May 23-24, 2008.
- [5]. Kampeerapaappat, P. and Mingkhwan, A., "A Propose Model for Web Classification with Dewey Decimals Classification." Vol.2 No.1 AITETconf2 nd (S&T Teaching in Vocational Education based on Sufficient Economy). Thailand: Rajamangala University of Technology Krungthep.2007.
- [6]. Kampeerapaappat, P. Namvong, N. Mingkhwan, A., "A study of dynamic content DDC-MR Classification : Case Study WIKIPEDIA." The 11th Annual Conference on the Convergence of Telecommunications, Networking & Broadcasting (PGNet 2010), Liverpool, UK, 21-22 June 2010.
- [7]. Lertmahakiat Wilaiporn and Mingkwan Anirach. "A propose idea of search engine results page base on DDC classification." Vol.2 , No.1 , Jan-Dec, 2007, AITET 2 nd conference : S&T Teaching in Vocational Education based on Sufficient Economy. 2007.
- [8]. Lertmahakiat, W. Kampeerapaappat, P. Mingkhwan, A., "A Survey of Search Engine Result Page." The Journal of KMUTNB., Vol.18, No.1, Jan - Apr. 2008a.
- [9]. Lertmahakiat, W. Kampeerapaappat, P. Mingkhwan, A., "A DDC Multiple Relation Content Retrieval Framework. PGNet 2008." the 9thAnnual Postgraduate Symposium on The Convergence of Telecommunications, Networking and Broadcasting, Liverpool John Moores University, UK. 23rd-24th June 2008b.
- [10]. Lertmahakiat, W., Mingkhwan, A., "Estimation Representative Data of Book Contents for Processing with DDC-MR." Proceedings NCCIT 2 0 0 9 . The 5 th National Conference on Computing and Information Technology., King Mongkut's Institute of Technology

- North Bangkok, Thailand, 22 – 23 May 2009.
- [11]. Li, Cong, Ji-Rong Wen and Hang Li. "Text Classification Using Stochastic Keyword Generation." Proceedings of the 20th International Conference on Machine Learning (ICML-2003), 2003.
- [12]. McCallum Andrew, Nigam Kamal. "Text Classification by Bootstrapping with Keywords, EM and Shrinkage." Improving Text Classification by Shrinkage in a Hierarchy of Classes, Andrew McCallum, Roni Rosenfeld, Tom Mitchell and Andrew Y. Ng in Proceedings of the Fifteenth International Conference on Machine Learning, 1998
- [13]. Prabowo, R. and et al. "Ontology-based automatic classification for the Web.", 182-191. Proceedings of the 3rd International conference on Web information systems engineering. 2002.
- [14]. Sankar K. Pal and Pabitra Mitra. "Pattern Recognition Algorithms for Data Mining : Scalability, Knowledge Discovery and Soft Granular Computing. London : A CRC Press Computer." 2004.
- [15]. Sudhakar, Venkata and Banshi D. Chaudhary. 2006. "A Hierarchy of Engines based on ODP Concept.", Proceeding of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. New York : ACM Press.
- [16]. Watthananon, J., Mingkhwan, A., "A connection relationship group of knowledge technique with DDC-Multiple Relations.", Proceedings WUNCA 21st and CIT2010. National Conference on Computer Information Technologies 2010. , Burapa University, Thailand. 2010.