กรรมวิธีสำหรับการจัดหมวดหมู่เว็บไซต์ตามมาตราฐานทศนิยมดิวอี้ในรูปแบบสัดส่วน
Method for Website Categorize Using Scale Dewey Decimals Classification Scheme

ภูริวัตร คัมภีรภาพพัฒน์ และอนิราช มิ่งขวัญ***

**ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ*
***ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีและการจัดการอุตสาหกรรม*
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
*Poorivat.K@rmutk.ac.th** *Anirach@ieee.org***

บทคัดย่อ

อินเทอร์เน็ตเป็นแหล่งสารสนเทศขนาดใหญ่ แม้จะมีการพัฒนาเครื่องมือสืบค้นที่มีประสิทธิภาพขึ้น แต่ยังประสบปัญหาในการสืบค้นข้อมูลของผู้ใช้ อันเนื่องมาจากเว็บไซต์มีการเพิ่มขึ้นอย่างรวดเร็ว ส่งผลให้ผู้ใช้ได้รับข้อมูลปริมาณมากที่ไม่ตรงตามความต้องการ ในงานวิจัยนี้ผู้วิจัยมุ่งนำเสนอการใช้กรรมวิธีสำหรับการจัดหมวดหมู่เนื้อหาของเว็บไซต์แบบลำดับชั้น (Hierarchical Classification) และจัดเก็บข้อมูลตามแบบมาตรฐานทศนิยมดิวอี้ (Dewey Decimal Classification) ซึ่งผลการวิจัยในครั้งนี้พบว่าสามารถจัดหมวดหมู่เนื้อหาเว็บไซต์ตามระบบมาตรฐานทศนิยมดิวอี้และแสดงสัดส่วนของหมวดหมู่ได้ทุกระดับ ซึ่งเป็นแนวทางในการพัฒนาเครื่องมือสืบค้นที่มีประสิทธิภาพเพิ่มขึ้น

คำสำคัญ: การค้นคืนสารสนเทศ, การวิเคราะห์หมวดหมู่, ระบบมาตรฐานทศนิยมดิวอี้, การจัดหมวดหมู่อัตโนมัติ, กราฟเรดาร์

Abstract

Internet is the largest information resources, many search engines have been developed their efficiency and accurate of searching, that impact the internet users to face of thousands results that few relevant their interest. . This research proposes technique for hierarchical classification content of the websites and built database in Dewey Decimals Classification scheme. The result found that, this technique can classified websites in four level of DDC scheme, that would increase the efficient of search engine.

KEYWORDS : Search Engines, Web Classification, Dewey Decimal Classification, Automatic Classification, Radar Graph

1. บทนำ

ลักษณะของสังคมสมัยปัจจุบันเป็นสังคมสารสนเทศ (Information Society) ดังนั้นทุกคนจำเป็นต้องใช้ข่าวสารจากแหล่งข้อมูลต่าง ๆ เพื่อพัฒนาชีวิตความเป็นอยู่ ประกอบอาชีพ ตลอดจนร่วมพัฒนาสังคมประเทศชาติ อินเทอร์เน็ตเป็นเครื่องมือที่สำคัญอย่างหนึ่งในกิจการสารสนเทศ เนื่องจากสามารถนำเสนอข้อมูลได้อย่างถูกต้อง รวดเร็ว มีประสิทธิภาพ และสามารถเก็บข่าวสาร ข้อมูลต่าง ๆ ได้ ปัจจุบันมีผู้นิยมใช้อินเทอร์เน็ตเพิ่มขึ้น เนื่องจากเป็นแหล่งสารสนเทศที่มีขนาดใหญ่มีข้อมูลจำนวนมหาศาล แต่ปัญหาที่พบในปัจจุบันคือความยุ่งยากในการค้นหาข้อมูล และข้อมูลที่ได้จากการค้นหายังไม่ตรงความต้องการของผู้ใช้[1][2][3] จึงต้องมีการพัฒนาเครื่องมือสำหรับสืบค้นสารสนเทศ (Search Engine) ที่มีประสิทธิภาพสูง ในด้านการนำข้อมูลจากเว็บไซต์ต่างๆ ที่เติบโตอย่างรวดเร็วมาจัดเก็บในฐานข้อมูลเพื่อให้สามารถสืบค้นได้อย่างถูกต้องและมีความแม่นยำมากขึ้นได้ แม้ว่าจะมีการนำเทคนิคของ Information Retrieval และ Data Mining มา

ใช้ในกระบวนจัดสร้างฐานข้อมูลเพื่อการค้นคืน[4][5][6] แต่เครื่องมือสืบค้นในปัจจุบันได้มีการจัดข้อมูลตามลำดับหรือกลุ่มข้อมูลแตกต่างกันตามเครื่องมือสืบค้น (Search Engine) แม้ว่าจะมีผู้เล็งเห็นประโยชน์ของกรอบองค์ความรู้การจัดหมวดหมู่ของห้องสมุดมาประยุกต์ใช้สำหรับการจัดหมวดหมู่สารสนเทศบนอินเตอร์เน็ตแต่ว่าการวิจัยส่วนใหญ่จะเน้นเฉพาะสาขาวิชาเท่านั้น[7][8] หรือศึกษาเพียงหมวดหมู่หลักเท่านั้น[9]

ในงานวิจัยครั้งนี้ผู้วิจัยมุ่งพัฒนาการจัดเก็บสารสนเทศบนอินเตอร์เน็ตตามกลุ่มเนื้อหาภายใต้กรอบความรู้การจัดหมวดหมู่ในห้องสมุด โดยนำเสนอกรรมวิธีสำหรับสกัดข้อมูลสารสนเทศของเว็บเพจและนำเทคนิคการจัดหมวดหมู่ตามระบบมาตรฐานทศนิยมดิวอี้ (Dewey Decimal Classification) ที่ใช้ในห้องสมุด มาประยุกต์ในการช่วยจัดหมวดหมู่ข้อมูล (Classification) ในเว็บไซต์ต่าง ๆ เนื่องจากเป็นระบบที่นิยมใช้จัดหมวดหมู่สารสนเทศในห้องสมุดเป็นเวลากว่า 100 ปี ใช้ตัวเลขเป็นสัญลักษณ์ง่ายต่อการจดจำและมีการแบ่งลำดับชั้นที่ชัดเจน นอกจากนี้ผู้วิจัยได้นำเทคนิค Information Extraction และ Data Mining มาช่วยในการพัฒนาให้สามารถจัดเก็บคำสำคัญเพื่อสร้างฐานข้อมูลดรรชนีให้ใช้ควบคู่กับหมวดหมู่ระบบมาตรฐานทศนิยมดิวอี้ ซึ่งจะทำให้การจัดเก็บข้อมูลเป็นไปอย่างมีระบบ และเพิ่มประสิทธิผลด้านความถูกต้องของเนื้อหาที่ตรงความต้องการของผู้ใช้

2. เอกสารและงานวิจัยที่เกี่ยวข้อง

2.1 เครื่องมือสืบค้นสารสนเทศ (Search Engine)

อินเทอร์เน็ตเป็นแหล่งข้อมูลอันมหาศาลที่หลายหลายรูปแบบ การเข้าถึงข้อมูลสารสนเทศจึงจำเป็นต้องมีเครื่องมือสืบค้นที่มีประสิทธิภาพ ในปัจจุบันเครื่องมือสืบค้นสารสนเทศที่มีการใช้ใน 4 ประเภทคือ 1) Search Engine เป็นเครื่องมือสืบค้นที่มีระบบการทำงานโดยใช้โปรแกรมที่มีการทำงานลักษณะอัตโนมัติและมีความรวดเร็วที่เรียกว่า Spider หรือ Robot หรือ Crawler เพื่อท่องไปในเว็บไซต์ต่าง ๆ สำหรับอ่านข้อมูลของเว็บไซต์เหล่านั้นและจัดเก็บเนื้อหาของเว็บไซต์ที่พบเข้าสู่ฐานข้อมูล ทำให้ฐานข้อมูลมีขนาดใหญ่ 2) เครื่องมือสืบค้นสารสนเทศประเภท Meta Search Engine เป็นเครื่องมือสืบค้นที่ไม่มีฐานข้อมูลของตนเอง แต่เป็นการค้นหาจากฐานข้อมูลของ Search Engine หลาย ๆ แห่ง 3) เครื่องมือสืบค้นสารสนเทศประเภท Classification Directories Search Engine เครื่องมือสืบค้นประเภทนี้ต้องให้ผู้เชี่ยวชาญเป็นผู้จัดทำหมวดหมู่ของข้อมูลสารสนเทศ ซึ่งต้องใช้ระยะเวลานานในการจัดหมวดหมู่ และต้องใช้ต้นทุนสูง 4) เครื่องมือสืบค้นสารสนเทศประเภท Subject Gateway Search Engine เป็นเครื่องมือสืบค้นที่ให้บริการสืบค้นเฉพาะด้าน มีการแบ่งหมวดและแยกหัวเรื่องโดยบรรณารักษ์

เครื่องมือสืบค้นทั้ง 4 ประเภทไม่ได้ให้ความสำคัญในการจัดเก็บข้อมูลสารสนเทศที่มีการแบ่งหมวดหมู่ที่ชัดเจนเหมือนในระบบห้องสมุดที่ผู้ใช้คุ้นเคย โดยในห้องสมุดได้มีการจัดหมวดหมู่ตามหลักของบรรณารักษ์ศาสตร์ ซึ่งได้มีการยอมรับว่าเป็นการวิเคราะห์แล้วว่าเป็นรูปแบบที่สามารถสืบค้นได้ดี

2.2 ระบบจัดหมวดหมู่ทศนิยมดิวอี้ (Dewey Decimals Classification)

ระบบมาตรฐานทศนิยมดิวอี้ที่ใช้ในห้องสมุด ซึ่งจัดหมวดหมู่องค์ความรู้ของมนุษย์แบบลำดับชั้น(Hierarchical Classification) โดยแบ่งเป็น 10 หมวดหมู่หลัก แต่ละหมวดหมู่หลักจะแบ่งเป็น 10 หมวดหมู่ย่อย และแต่ละหมวดหมู่ย่อยจะแบ่งได้อีก 10 หมวด โดยใช้ตัวเลขเป็นสัญลักษณ์ซึ่งง่ายต่อการจดจำ ซึ่งมีห้องสมุดมากกว่า 200,000 แห่ง ใน 135 ประเทศใช้ระบบนี้ [9] แต่ระบบหอสมุดรัฐสภาอเมริกันได้มีจัดหมวดหมู่องค์ความรู้โดยจำแนกหมวดหมู่ตามตัวอักษรและตัวเลข จึงมีลำดับขั้นของการจัดหมวดหมู่ที่ไม่ชัดเจน ดังนั้นผู้วิจัยขอเสนอการนำระบบจัดหมวดหมู่ทศนิยมดิวอี้มาประยุกต์ในการช่วยจัดหมวดหมู่ข้อมูลสารสนเทศ (Classification) ในเว็บไซต์ต่าง ๆ ให้เป็นกลุ่มของเนื้อหาที่มีลำดับขั้นความสำคัญที่ชัดเจน

2.3 งานวิจัยที่เกี่ยวข้อง

ได้มีผู้ทำการศึกษาด้านกระบวนการสกัดคำด้วยเทคนิคต่าง ๆ เช่น โดยใช้เทคนิค Stochastic Keyword Generation เพื่อพัฒนาสมรรถนะในการทำ text classification โดยวิเคราะห์ใจความสำคัญของ e-mail แล้วแปรค่าใจความสำคัญเป็นตัวเลขเชิง Vector of probability values ในลักษณะ mapped vector แล้วนำไปวิเคราะห์จัดทำเป็น keyword [10] อีกเทคนิคที่นิยมใช้

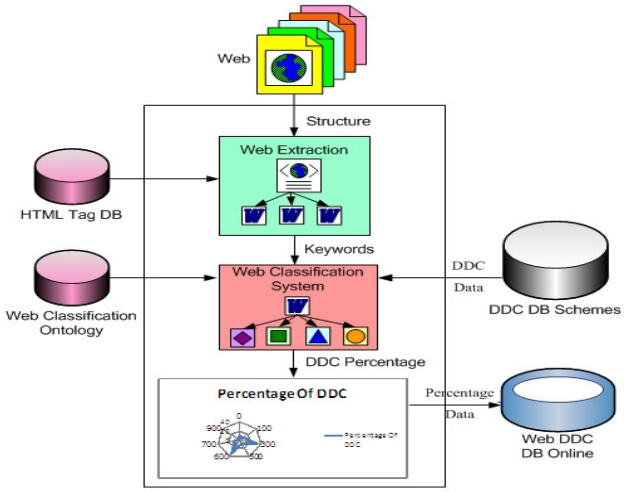
คือเทคนิค SVM (Support Vector Machine) [11] ได้พัฒนา Chinese Meta-Search ที่มีความชาญฉลาด

ส่วนงานวิจัยที่ศึกษาด้านการนำกรอบความรู้มาใช้ในการศึกษาจัดหมวดหมู่อัตโนมัตินั้นพบว่ามีการนำกรอบความรู้มาใช้ในการจัดหมวดหมู่ เริ่มตั้งแต่กรอบความรู้ที่ผู้วิจัยคิดค้นเองเช่น The DocMine Algorithms มีการกำหนดหัวข้อหลักโดยมีการแบ่งเอกสารออกเป็นกลุ่มๆ และมีการใช้วิธีการนำกลับมารวมกันเป็นคลังข้อมูล เพื่อใช้สำหรับค้นหากลุ่มของคำจากข้อมูลแต่ละกลุ่ม ทำให้เพิ่มประสิทธิภาพในการค้นหาข้อมูลที่มีขนาดใหญ่ [12] นอกจากได้มีการนำกรอบความคิด ODP (Open Directory Project) เพื่อนำมาจัดหมวดหมู่เป็นกรอบองค์ความรู้ เพื่อนำมาจัดทำเป็น Meta Search ที่มีประสิทธิภาพในการเชื่อมโยง ฐานข้อมูลต่างๆ [13] นอกจากนี้ได้มีนำกรอบความรู้ด้านห้องสมุดมาใช้ในการจัดหมู่สารสนเทศบนอินเตอร์เน็ตเพื่อศึกษาวิธีการจัดหมวดหมู่สารสนเทศ เพื่อคิดค้นให้ระบบสามารถเกิดการเรียนรู้และทำงานโดยอัตโนมัติ ดังเช่น ศึกษาแนวทางการจัดหมวดหมู่แบบอัตโนมัติ โดยจัดทำกลุ่มคำศัพท์ทางด้านสถาปัตยกรรมด้วยมนุษย์ เพื่อนำไปใช้เชื่อมโยงการจัดหมวดหมู่ ผลพบว่าสามารถจัดหมวดหมู่และลำดับชั้นให้กับคำศัพท์ได้อย่างอัตโนมัติ [14] ในขณะที่มีกลุ่มนักวิจัยได้พัฒนา ontology ที่สามารถจัดหมวดหมู่ตามระบบมาตรฐานทศนิยมดิวอี้ (DDC) และเปรียบเทียบระบบมาตรฐานหอสมุดรัฐสภาอเมริกัน(LLC) ผลการทดสอบพบว่า ontology ที่สร้างขึ้นมีความแม่นยำในการจัดหมวดหมู่ระดับหมวดหลัก[8] แต่ไม่สามารถใช้ ontology จัดหมวดหมู่ในระดับหมวดหมู่ย่อยได้

จากงานวิจัยที่ได้กล่าวมาพบว่างานวิจัยที่ผ่านมายังพบปัญหาในการจัดกลุ่มคำศัพท์ที่มีความหมายคล้ายคลึง และคำศัพท์ที่มีรูปแบบแตกต่างกัน ตลอดจนการจัดหมวดหมู่ตามกรอบองค์ความรู้ตามมาตรฐานระบบดิวอี้นั้นมีความแม่นยำเพียง 10 หมวดหลัก แต่จากงานวิจัย [15] ของผู้วิจัยได้มีการนำเสนอแบบจำลองของการจัดหมวดหมู่เนื้อหาของเว็บไซต์ตามระบบมาตรฐานระบบทศนิยมดิวอี้ และมีงานวิจัย [16] ได้มีการสำรวจหน้าจอภาพสำหรับการแสดงผลลัพธ์ของเครื่องมือสืบค้นโดยการใช้กราฟฟิกเพื่อให้สามารถแสดงผลลัพธ์ได้ชัดเจนนั้น ดังนั้นการวิจัยในครั้งนี้ผู้วิจัยมุ่งที่พัฒนากระบวนการที่สามารถวิเคราะห์เนื้อหา จัดหมวดหมู่คำสำคัญของเว็บไซต์ที่มีความสัมพันธ์ตามการจัดหมวดหมู่ของระบบมาตรฐานทศนิยมดิวอี้ได้ในทุกระดับ เพื่อจัดเก็บเนื้อหา คำสำคัญของเว็บไซต์เป็นฐานข้อมูลไว้สำหรับการสืบค้นบนอินเตอร์เน็ตต่อไป และมีการนำเสนอสัดส่วนความสัมพันธ์ตามการจัดหมวดหมู่ของระบบมาตรฐานทศนิยมดิวอี้ด้วยกราฟเรดาร์

3. กรรมวิธีสำหรับการจัดหมวดหมู่เว็บไซต์ตามมาตรฐานทศนิยมดิวอี้ในรูปแบบสัดส่วน

ในการวิจัยในครั้งนี้ได้นำเสนอผังระบบของกรรมวิธีสำหรับการจัดหมวดหมู่เว็บตามระบบมาตรฐานทศนิยมดิวอี้ในรูปแบบสัดส่วนความสัมพันธ์ของเนื้อหาเว็บไซต์ ดังภาพที่ 1



ภาพที่ 1 : แสดงผังระบบการทำงานของกรรมวิธีสำหรับการจัดหมวดหมู่เว็บตามระบบมาตรฐานทศนิยมดิวอี้ในรูปแบบสัดส่วน

จากภาพที่ 1 เกี่ยวกับผังระบบการทำงานของกรรมวิธีสำหรับการจัดหมวดหมู่เว็บตามระบบมาตรฐานทศนิยมดิวอี้ในรูปแบบสัดส่วนในการวิจัยในครั้งนี้ มีส่วนประกอบดังนี้

3.1 การพัฒนาโปรแกรมเก็บรายละเอียดของเว็บไซต์เพื่อเป็นฐานข้อมูลสำหรับการใช้งานต่อไป

3.2 สร้างฐานข้อมูลกรอบองค์ความรู้ของคำสั่งของภาษา HTML (HTML Tag DB) สำหรับช่วยในการวิเคราะห์และสกัดคำสำคัญของเนื้อหาเว็บไซต์

3.3 พัฒนากรรมวิธีเพื่อนำข้อมูลจากฐานข้อมูลกรอบองค์ความรู้ของคำสั่งของภาษา HTMLมาช่วยในการวิเคราะห์การ

สกัดคำ การแบ่งคำ การหารากศัพท์ ตามเทคนิคกระบวนการ Web Extraction ของ Information Retrieval เพื่อหาคำสำคัญ (Keyword) ของเนื้อหาเว็บไซต์ และหาน้ำหนักของคำ

3.4 นำคำสำคัญจากคู่มือ Dewey decimal classification and relative index [9] ซึ่งมีการจัดหมวดหมู่แบบลำดับชั้น มาสร้างฐานข้อมูลกรอบองค์ความรู้ของระบบมาตรฐานทศนิยมดิวอี้ (DDC DB Schemes) เพื่อนำมาช่วยในการจัดหมวดหมู่ของคำสำคัญตามระบบมาตรฐานทศนิยมดิวอี้

3.5 พัฒนา Ontology กำหนดรูปแบบโครงสร้างข้อมูลขอบเขตความสัมพันธ์ของคำ ใช้สำหรับหาคำที่มีความสัมพันธ์กัน มีความหมายเดียวกัน เพื่อให้คำสำคัญของเนื้อหาเว็บไซต์มีความหมายที่ดีขึ้น (Web Classification Ontology)

3.6 พัฒนากรรมวิธีเพื่อเปรียบเทียบคำสำคัญของแต่ละเว็บไซต์ที่ได้จากการวิเคราะห์มาทำการจัดหมวดหมู่เว็บเว็บไซต์แบบลำดับชั้น (Hierarchical Classification) และหาร้อยละของสัดส่วนความสัมพันธ์ของเนื้อหาตามระบบมาตรฐานทศนิยมดิวอี้ เพื่อจัดเก็บเป็นฐานข้อมูลเว็บไซต์ตามกรอบความรู้ตามระบบมาตรฐานทศนิยมดิวอี้สำหรับการให้บริการสืบค้นบนอินเตอร์เน็ตต่อไป

3.7 ผู้วิจัยได้นำเสนอร้อยละของสัดส่วนของเนื้อหาตามหมวดหมู่ระบบมาตรฐานทศนิยมดิวอี้ของเว็บไซต์ด้วยกราฟเรดาร์ (Radar Graph) ซึ่งสามารถแสดงผลให้เห็นถึงความสัมพันธ์ สัดส่วนร้อยละ และแบ่งแยกชั้นได้ชัดเจน

4. ผลการทดสอบกรรมวิธีสำหรับการจัดหมวดหมู่เว็บตามมาตรฐานทศนิยมดิวอี้ในรูปแบบสัดส่วน

ในการวิจัยในครั้งนี้ผู้วิจัยได้มีการทดสอบกรรมวิธีสำหรับการจัดหมวดหมู่เว็บไซต์แบบลำดับชั้น (Hierarchical Classification) ตามมาตรฐานระบบมาตรฐานทศนิยมดิวอี้ในรูปแบบสัดส่วน ด้วยการทดสอบโดยนำข้อมูลของเว็บไซต์จำนวน 100 เว็บไซต์เพื่อใช้สำหรับการเรียนรู้ตรวจสอบความถูกต้องของโมเดลการจัดหมวดหมู่เว็บไซต์แบบลำดับชั้น (Hierarchical Classification) ตามระบบมาตรฐานทศนิยมดิวอี้ พร้อมสร้างฐานข้อมูลของเว็บไซต์และแสดงสัดส่วนค่าร้อยละของเนื้อหาตามตามกรอบความรู้ระบบมาตรฐานทศนิยมดิวอี้ ซึ่งมีผลการทดสอบดังนี้

4.1 ผลของการพัฒนาโปรแกรมเพื่อเก็บรายละเอียดของเว็บไซต์ ได้ผลดังตารางที่ 1

ตารางที่ 1 : ตารางข้อมูลรายละเอียดของเว็บไซต์

UrlId	Url	DateExtraction
1	www.ram.edu/	05/12/2007
2	www.elearning.nectec.or.th/	05/12/2007
:	:	:

4.2 ผู้วิจัยได้นำกรอบองค์ความรู้ของระบบมาตรฐานทศนิยมดิวอี้มาจัดเก็บเป็นฐานข้อมูลเพื่อใช้งานในการจัดหมวดหมู่ ดังตารางข้อมูลระบบมาตรฐานดิวอี้บางส่วน ในตารางที่ 2

ตารางที่ 2 : ตารางข้อมูลระบบมาตรฐานดิวอี้

DDCNo	DDCTitle
000	generalities
001	knowledge
001.4	research; statistical methods
001.42	research methods
001.422	statistical methods
:	:

4.3 ผลของการวิเคราะห์เทคนิคการสกัดคำตามกระบวนการ Web Extraction ของ Information Retrieval ในหาคำสำคัญ (Keyword) ตามกระบวนการตัดคำ แบ่งคำ การหารากศัพท์ ให้น้ำหนักของคำ ได้ผลลัพธ์บางส่วนดังตารางที่ 3

ตารางที่ 3 : ตารางข้อมูลของคำค้นของเว็บไซต์

WordWeb	WordCount	UrlId	URL
learning	2	1	www.ram.edu/
rights	1	1	www.ram.edu/
learning	7	2	www.elearning.nectec.or.th/
products	2	10	www.learning.com/
:	:	:	:

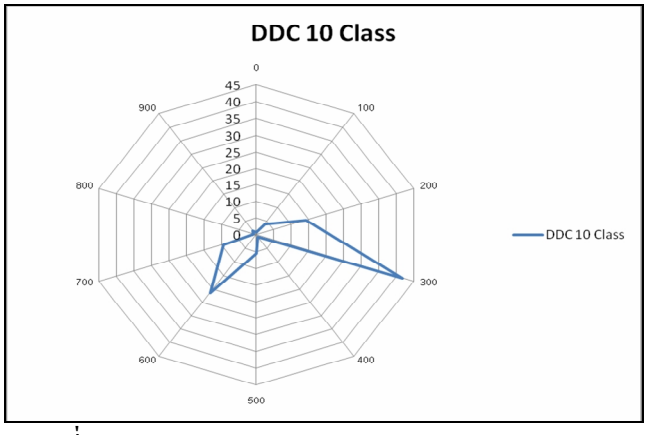
4.4 ผลของการเปรียบเทียบคำสำคัญของแต่ละเว็บไซต์ที่ได้จากการวิเคราะห์ เพื่อนำมาทำการจัดหมวดหมู่เว็บไซต์แบบ

ลำดับชั้น(Hierarchical Classification) และหาร้อยละของสัดส่วนความสัมพันธ์ของเนื้อหาตามระบบมาตรฐานทศนิยมดิวอี้ ได้ผลบางส่วนดังตารางที่ 4

ตารางที่ 4 : ตารางข้อมูลของสัดส่วนร้อยละความสัมพันธ์ของระบบมาตรฐานทศนิยมดิวอี้

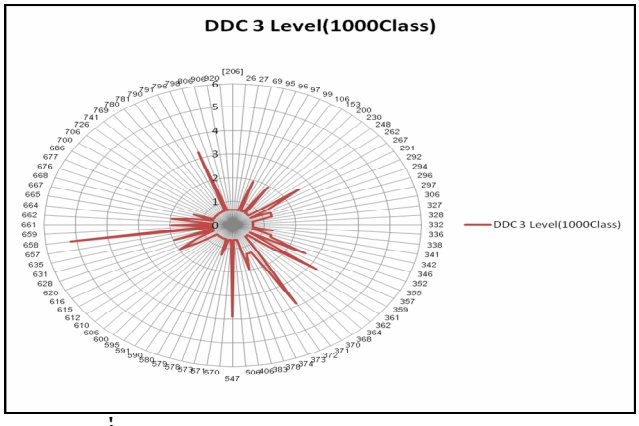
UrlId	Url	DdcNo	DdcPercent
1	www.ram.edu/	000	0.79
1	www.ram.edu/	100	3.94
:	:	:	:

จากตารางที่ 4 ข้างต้นจะพบว่าเว็บไซต์ดังกล่าวมีเนื้อหาที่มีความสัมพันธ์ทุกหมวดหมู่ของระบบมาตรฐานทศนิยมดิวอี้ในสัดส่วนที่แตกต่างกันและมีเนื้อหาส่วนใหญ่สัมพันธ์ด้านการศึกษา (หมวด 300) และด้านวิทยาศาสตร์ประยุกต์ (หมวด 600) และเมื่อนำมาแสดงผลด้วยเรดาร์กราฟจะปรากฏตำแหน่งของสัดส่วนร้อยละในแต่ละหมวดหมู่ทศนิยมดิวอี้เวียนตามเข็มนาฬิกาและแสดงเส้นโยงความสัมพันธ์ของข้อมูลเนื้อหาที่ได้ดังภาพที่ 2



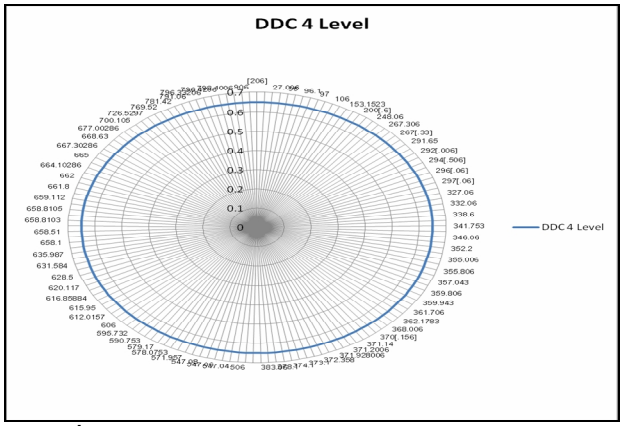
ภาพที่ 2 : แสดงร้อยละสัดส่วนความสัมพันธ์ด้วยกราฟเรดาร์แบบ 10 หมวดหมู่หลัก

ผู้วิจัยได้ทดสอบนำข้อมูลที่ได้จากจัดหมวดหมู่ 3 ระดับ (1000 หมู่) พบว่าการแสดงความสัมพันธ์จะละเอียดมากขึ้น แต่ภาพที่แสดงผลจะไม่สัมพันธ์กับ 10 หมวดหลักทั้งนี้เนื่องจากกราฟเรดาร์จะแสดงผลเฉพาะหมู่ที่ปรากฏเท่านั้นทำให้รูปร่างเป้ไปที่หมวดหมู่ย่อยในหมวด 300 กับหมวด 600 ดังภาพที่ 3



ภาพที่ 3 : แสดงร้อยละสัดส่วนความสัมพันธ์ด้วยกราฟเรดาร์แบบ 3 ระดับ (1000 หมู่)

แต่เมื่อนำข้อมูลที่ได้จากจัดหมวดหมู่ 4 ระดับ พบว่าการแสดงความสัมพันธ์จะละเอียดมากยิ่งขึ้น ส่งผลทำให้ร้อยละของสัดส่วนความสัมพันธ์ที่ได้ใกล้เคียงกันมาก ดังภาพที่ 4



ภาพที่ 4 : แสดงร้อยละสัดส่วนความสัมพันธ์ด้วยกราฟเรดาร์แบบ 4 ระดับของเว็บไซต์ทั้งหมด

จากผลของการวิจัยนี้ผู้วิจัยได้ออกแบบข้อมูลโดยมีการจัดเก็บข้อมูลที่เกี่ยวข้องดังนี้ ลำดับที่ของที่อยู่เว็บไซต์เป็นข้อมูลแบบจำนวนเต็มขนาด 4 ไบต์ ที่อยู่เว็บไซต์เป็นข้อมูลแบบข้อความขนาด 100 ไบต์ เลขหมู่ของระบบมาตรฐานทศนิยมดิวอี้เป็นข้อมูลแบบข้อความขนาด 20 ไบต์ คำสำคัญเป็นข้อมูลแบบข้อความขนาด 50 ไบต์ ความถี่ของคำสำคัญที่ปรากฏในเว็บไซต์เป็นข้อมูลแบบจำนวนเต็มขนาด 4 ไบต์ ความถี่ของคำสำคัญที่จัดหมวดหมู่ 10 คลาสเป็นข้อมูลแบบจำนวนเต็มขนาด 4 ไบต์ ซึ่งมีขนาดของการเก็บข้อมูลต่อ 1 ระเบียนข้อมูลมีค่าประมาณ 182 ไบต์ ถ้ามีการเก็บข้อมูลของเว็บไซต์แบบ 10 คลาสจะมีการเก็บข้อมูลมากที่สุดประมาณ

1820 ไบต์ต่อ ถ้าเก็บข้อมูลแบบ 100 คลาสย่อยจะมีการเก็บข้อมูลมากที่สุดประมาณ 18200 ไบต์ ถ้าเก็บข้อมูลแบบ 1000 หมู่ย่อย จะมีการเก็บข้อมูลมากที่สุดประมาณ 182000 ไบต์ ถ้าเก็บข้อมูลแบบ 4 ระดับย่อย จะมีการเก็บข้อมูลมากที่สุดประมาณ 4680000 ไบต์

5. บทสรุป

จากผลการวิจัยของการสร้างกรรมวิธีสำหรับการจัดหมวดหมู่เว็บไซต์แบบลำดับชั้น(Hierarchical Classification) ตามระบบมาตรฐานทศนิยมดิวอี้ในรูปแบบสัดส่วน ทำให้สามารถจัดเป็นแหล่งสารสนเทศบนอินเตอร์เน็ตได้ตามลำดับขั้นตอนของความสัมพันธ์ของเนื้อหาเว็บไซต์ในแต่ละหมวดหมู่ และหมู่ย่อยของระบบมาตรฐานทศนิยมดิวอี้ ซึ่งจะทำให้เข้าถึงข้อมูลได้ตามความเป็นจริงกว่า การกำหนดให้เว็บไซต์อยู่ในหมวดใดเพียงหมวดเดียว จากผลการทดสอบกรรมวิธีสำหรับการจัดหมวดหมู่เว็บไซต์แบบลำดับชั้น (Hierarchical Classification) ตามระบบมาตรฐานทศนิยมดิวอี้ในรูปแบบสัดส่วนจากเว็บไซต์จำนวน 100 เว็บไซต์ พบว่าสามารถที่จะมาวิเคราะห์หมวดหมู่ตามระบบมาตรฐานทศนิยมดิวอี้ได้หมดทุกลำดับชั้น และมีการใช้พื้นที่สำหรับการจัดเก็บข้อมูลประมาณ 1800 ไบต์ ถึง 4680000 ไบต์ต่อหนึ่งเว็บไซต์ แต่เมื่อนำมาแสดงผลความสัมพันธ์ของสัดส่วนพบว่าสามารถแสดงผลได้ชัดเจนในทุกระดับ โดยเฉพาะแบ่งหมวดหมู่ 10 หมวดหลัก, 100 หมวดย่อยและการแบ่ง 1000 หมู่ ซึ่งจะเป็นการแบ่งระดับชั้นที่มีเกณฑ์ที่แน่นอนและข้อมูลของการแบ่งชั้นมีปริมาณน้อย แต่สำหรับการแบ่งในระดับที่ 4 จะเป็นการแบ่งแบบไม่จำกัด ส่งผลทำให้ร้อยละของสัดส่วนความสัมพันธ์ที่ได้ใกล้เคียงกันมากเนื่องข้อมูลของการแบ่งชั้นมีปริมาณมากขึ้น

แนวทางในการวิจัยในครั้งต่อไป ผู้วิจัยจะพัฒนาเทคนิคกระบวนการการสกัดข้อมูลและการใช้คำที่มีความหมายเหมือนกันและเปรียบเทียบตามความหมายของคำสำคัญของเว็บไซต์

6. การเขียนเอกสารอ้างอิง

[1] Donald T. Hawkins. "Conference circuit : The eighth search enginemeeting." *Information Today.* Vol. 20, no. 6, 2003.

[2] Deborah Fallows. *Search engine users.* Washington, DC: Pew Internet & American Life Project, 2005.

[3] *iProspect search engine user attitudes.* 2004. (online) Available : www.iProspect.com/ Retrieval 2007/07/16.

[4] Susan Dumais and Chen Hao. "Hierarchical classification of web content" *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.* New York : ACM Press, 2000.

[5] Aris Anagnostopoulos, Ansrei Z. Broder and Kunal Punera. "Effective and efficient classification on a search-engine model" *CIKM'06* November5-11 Arlington, Virginia, USA., 2006.

[6] Soumen Chakrabarti. *Mining the web : discovering knowledge from hypertext data.* San Francisco, USA : Elsevier Science, 2003.

[7] K.Golub. "Automated subject classification of textual Web pages, based on a controlled vocaburaly: challenges and recommendations" [online] Available: http://www.it.lth.se/koraljka/Lund/publ/Hypermedia 2006.pdf documents. retrieval 2007/09/21.

[8] R. Prabowo, et al. 2002. Ontology-based automatic classification for the Web. *Proceedings of the 3rd International conference on Web information systems engineering.:* 182-191.

[9] Melvil Dewey. *Dewey decimal classification and relative index.* 21st ed. by John P. Comaromi [and others]. Albany, New York : OCLC Forest Press, 2003.

[10] Cong Li, Wen Ji-Rong and Hang Li.. "Text classification using stochastic keyword generation" *Proceedings of the 20th International Conference on Machine Learning (ICML-2003), 2003.*

[11] Wang Hao-ming and Feng Bo-qin.. "Research of the Chinese meta-search engine model based on intelligent agent" *Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA'05),* 2005.

[12] Barbara Daniel, Domeniconi Carlotta and Ning Kang. Mining relevant text from unlabelled documents. *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)*, 2003.

[13] Venkata Sudhakar and Banshi D. Chaudhary. "A Hierarchical of engines based on ODP concept" *Proceeding of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.* New York : ACM Press, 2006.

[14] Charlotte Jenkins and David Inman. Adaptive automatic classification on the web. *IEEE 0-7695-0680-1/00*, 2000.

[15] Poorivat Kampeerapaappat and Anirach Mingkhwan. "A Propose model for web classification with Dewey decimal classification" *AITETconf2nd (S&T Teaching in Vocational Education based on Sufficient Economy).* Vol. 2, No .1,Jan-Dec, 2007.

[16] Wilaiporn Lertmahakiat and Anirach Mingkhwan. "A propose idea of search engine results page base on DDC classification" AITETconf2nd (S&T Teaching in Vocational Education based on Sufficient Economy). Vol. 2, No .1,Jan-Dec, 2007.