บทที่ 2

งานวิจัยและทฤษฎีที่เกี่ยวข้อง

สำหรับบทนี้จะเป็นการกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง เพื่อชี้ให้เห็นแนวทางของการประยุกต์การใช้งานทฤษฎีต่าง ๆ กับ เว็บสไปเดอร์แบบจำเพาะกลุ่มเป้าหมายโดย อาศัยหลักทางซัพพอร์ตเวคเตอร์แมชชีน

2.1 งานวิจัยที่เกี่ยวข้องเว็บสไปเดอร์ (Web spider)

ถึงแม้ว่าสารสนเทศบนเวิลด์ไวด์เว็บมีประโยชน์อย่างมาก แต่ไม่มีรูปแบบทางกายภาพที่แน่นอน ไม่มีขอบเขตกำจัด ไม่มีความเป็นถาวร ไม่มีการควบคุมจากหน่วยงานใด เนื่องจากมีความเป็นอิสระในการจัดสร้าง กล่าวคือเอกสารบนเวิลด์ไวด์เว็บเหล่านี้มีโอกาสเปลี่ยนแปลงทั้งในเชิงเนื้อหา และแหล่งสารสนเทศได้ตลอดเวลา รวมถึงปริมาณสารสนเทศที่เพิ่มขึ้นในอัตราที่มหาศาล จากการสำรวจของ Netcraft ในปี 2007 โดยใช้โปรแกรมสำรวจเว็บไซต์ในเดือนพฤษภาคม 2550 พบจำนวนเว็บไซต์กว่า 118 ล้านเว็บไซต์ เพิ่มขึ้นจากเดือนเมษายนปีเดียวกันเกือบ 4.4 ล้านเว็บไซต์ ซึ่งประมาณการกันว่ามีเว็บไซต์ที่ให้บริการบนอินเทอร์เน็ตทั้งหมดกว่า 210 ล้านเว็บไซต์ (เยาวลักษณ์ สุวรรณแข, 2547) และในปี 2543 สำนักบริการคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์ ได้สำรวจพบว่าสารสนเทศที่จัดพิมพ์ในรูปเว็บเพจเพื่อเผยแพร่บนอินเทอร์เน็ตมีมากกว่า 1 พันล้านหน้า

จากปัญหาดังที่กล่าวมาทำให้การสืบค้นสารสนเทศบนเวิลด์ไวด์เว็บจึงไม่ใช่เรื่องง่าย จำเป็นต้องมีเครื่องมือช่วยค้น ที่จะทำให้การสืบค้นและการเข้าถึงสารสนเทศทำได้ง่ายและสะดวก ทำให้ผู้ใช้ได้สารสนเทศที่ตรงกับความต้องการอย่างรวดเร็ว เรียกเครื่องมือช่วยค้นนี้ว่า โปรแกรมค้นคืน (Search engine)

โปรแกรมค้นคืน คือโปรแกรมค้นคืนสารสนเทศจากฐานข้อมูล ที่รวบรวมเว็บเพจต่างๆ ทั่วโลกแล้วนำมาจัดทำเป็นฐานข้อมูล และจัดทำโปรแกรมค้นหาข้อมูลหรือเว็บเพจที่ต้องการให้บริการ โดยจัดทำเป็นเว็บเพจที่อยู่ในรูปของฟอร์มให้ผู้สืบค้นกรอกคำ วลี หรือประโยคที่ต้องการค้น จากนั้นโปรแกรมค้นหาจะทำการค้นจากฐานข้อมูลที่ได้จัดเก็บเว็บเพจไว้ก่อนแล้ว เมื่อพบข้อมูลที่ตรงกับความต้องการก็จะนำรายการที่พบมาแสดงแก่ผู้สืบค้น พร้อมทำรายการเชื่อมโยงไปยังแหล่งที่จัดเก็บเว็บไซต์นั้นๆ ทั้งนี้ผลของการสืบค้นจะพบหรือไม่พบก็ตาม โปรแกรมค้นหาจะส่งข้อมูลกลับมาให้ผู้สืบค้นทราบ (วิเศษศักดิ์ โคตรอาษา และ คณะ, 2542) อังสนา ธงไชย (2550:

100) อธิบายว่า โปรแกรมค้นหาคือโปรแกรมที่ทำหน้าที่ในการสืบค้นเอกสารเว็บในระบบเวิลด์ไวด์เว็บ ซึ่งเป็นส่วนหนึ่งในการจัดการระบบค้นคืนเอกสารเว็บโดยมีโปรแกรมจัดเก็บข้อมูลเอกสารเว็บ (Web Spider) เพื่อนำมาจัดเก็บเป็นฐานข้อมูล และมีโปรแกรมจัดทำดรรชนีคำจากส่วนต่างๆ ของเอกสารเว็บสำหรับสืบค้น

การวิจัยและการพัฒนาเว็บสไปเดอร์นั้นมีการวิจัยน้อยกว่าทศวรรษที่ผ่านมา (Pinkerton, 1994) โชคไม่ดีที่การวิจัยที่เกี่ยวข้องกับการออกแบบของงานส่วนใหญ่ของเว็บสไปเดอร์ ไม่ได้รับเปิดเผยต่อสาธารณชน เนื่องจากลักษณะของการแข่งขันทางด้านธุรกิจ มีเพียงแต่งานวิจัยที่ได้เปิดเผยในงานสิ่งตีพิมพ์ได้แก่ 1.Google crawler 2.อินเทอร์เน็ต และ 3.ระบบ Mercator

Matthew Gray (1993) ได้พัฒนาโปรแกรม Wanderer ขึ้นมาเพื่อใช้สำหรับรวบรวม URL มาเก็บไว้ในฐานข้อมูลที่มีชื่อว่า “Wandex” ข้อมูล URL ที่รวบรวมมาได้จะถูกนำไปวิเคราะห์หาอัตราการเติบโตของเวิลด์ไวด์เว็บ (World Wide Web) โปรแกรม Wanderer นี่เองถือเป็นเว็บโรบอต (Web Robot) ตัวแรกของโลก

Eagan and Bender (1996) ได้ศึกษาเรื่อง Spiders and Worms and Crawlers, Oh my: Searching on the World Wide Web ซึ่งมีวัตถุประสงค์เพื่ออธิบาย และแสดงให้เห็นความแตกต่างระหว่างโปรแกรมค้นหาบนเวิลด์ไวด์เว็บ ผลการศึกษาพบว่า ผู้ใช้บริการมีความสับสนในการสืบค้นข้อมูลบนเวิลด์ไวด์เว็บ เพราะมีโปรแกรมค้นหาอยู่จำนวนมาก และการทำงานของโปรแกรมค้นหามีการทำงานไม่เหมือนกัน ทำให้หาข้อมูลที่ต้องการไม่พบหรือพบเพียงเล็กน้อย รวมถึงข้อมูลบนอินเทอร์เน็ตมีการเปลี่ยนแปลงไปทุกวัน ทำให้ผลการสืบค้นในครั้งก่อนอาจไม่เหมือนกับการสืบค้นในปัจจุบัน

Cho *et al.* (1998) ได้ทำการศึกษาหลักการการรวบรวมข้อมูลของตัวครอว์เลอร์ โดยพวกเขาได้ตั้งเป้าในการรวบรวมข้อมูล 180,000 เพจ จากมหาวิทยาลัย Stanford โดยโดเมนที่ถูกรวบรวมข้อมูลนั้นได้จำลองวิธีการที่แตกต่างกัน ซึ่งการเข้ารวบรวมอาศัยหลักการแบบ Breadth-First, Backlink-Count และ บางส่วนของความสำคัญของเว็บเพจในการคำนวณ อีกอย่างหนึ่งของการวิจัยนี้ที่สรุปได้คือถ้าต้องการ ตัวครอว์เลอร์ในการรวบรวมข้อมูลหน้าเว็บที่มีความสำคัญของเว็บเพจจะสูงในช่วงเริ่มต้นของการบวนการ บางส่วนของวิธีการความสำคัญของเว็บเพจจะดีตามวิธีการค้นหาตาม Breadth-First และ Backlink-Count อย่างไรก็ตามผลลัพธ์เหล่านี้เป็นเพียงโดเมนเดียวเท่านั้น

ต่อมาในปี 1999 เมอร์เคเตอร์ (Mercator) เป็นครอว์เลอร์ที่พัฒนาโดย Heydon และ Najork ซึ่งออกแบบโดยมีลักษณะคล้ายคลึงกับแบบจำลองพื้นฐานนี้มากที่สุด เมอร์เคเตอร์สามารถรองรับการเก็บเว็บเพจจำนวนมากได้โดยในขณะนั้นสามารถเก็บเว็บเพจได้มากกว่าหนึ่งร้อยล้าน

เว็บเพจ จุดเด่นของเมอร์เคเตอร์คือการประสานงานระหว่างการใช้หน่วยความจำและฮาร์ดดิสก์ โดยมีการสลับข้อมูลจากหน่วยความจำไปเก็บไว้ในดิสก์ในเวลาที่ต้องการใช้หน่วยความจำเพิ่มขึ้น และช่วงเวลาในการใช้ดิสก์จะไม่ถ่วงการทำงานของระบบอีกด้วย ขณะที่ Teng *et al.* (1999) เสนอระบบครอว์เลอร์ที่ชื่อว่า CWC โดยเป็นการออกแบบครอว์เลอร์ให้ทำงานร่วมกันได้โดยใช้เครื่องคอมพิวเตอร์มากกว่าหนึ่งเครื่อง เพื่อป้องกันการเกิดการทำงานซ้ำซ้อนกัน CWC ถูกออกแบบให้มีอัลกอริทึมในการแบ่งงานกันเองได้ระหว่างเครื่องโดยปริมาณงานที่แบ่งกันนั้นจะมีความสมดุลกัน นอกจากนี้ Fiedler and Hammer (1998) เสนอวิธีการที่จะทำให้ครอว์เลอร์สามารถเก็บเว็บเพจได้รวดเร็วที่สุด โดยเสนอแนวคิดของ โมบายครอว์เลอร์ (Mobile crawler) หลักการคือโมบายครอว์เลอร์เปรียบเสมือนโมบายเอเจนต์ (Mobile agent) ที่สามารถเคลื่อนย้ายตัวเองไปตามสภาพแวดล้อม โมบายครอว์เลอร์จะเคลื่อนที่ไปฝังตัวอยู่ภายในเครื่องเว็บเซิร์ฟเวอร์ จากนั้นจะคัดเลือกเว็บเพจที่ต้องการจากเว็บเซิร์ฟเวอร์นั้นเพื่อส่งกลับมายังเครื่องหลัก โดยในการส่งเว็บเพจกลับมานั้น โมบายครอว์เลอร์จะบีบอัดเว็บเพจทั้งหมดให้มีขนาดเล็กลงก่อน เพื่อให้การส่งกลับทำได้รวดเร็วขึ้น อย่างไรก็ตามแนวคิดนี้ต้องเผชิญปัญหาในเรื่องความปลอดภัยของเว็บเซิร์ฟเวอร์อย่างหลีกเลี่ยงไม่ได้ อันทำให้ผู้ดูแลเว็บเซิร์ฟเวอร์ไม่อนุญาติให้โมบายครอว์เลอร์เข้ามาฝังตัวได้

Najork and Wiener (2001) ได้ดำเนินการรวมรวมเพจ 328 ล้านเพจ โดยใช้ลำดับการเข้ารวบรวมเป็นแบบ Breadth-First พวกเขาพบว่าการใช้ Breadth-First ในการรวบรวมเพจนั้นความสำคัญของเว็บเพจจะสูงในช่วงต้นของการรวบรวม (แต่พวกเขาไม่ได้เปรียบเทียบกับวิธีการนี้ต่อวิธีการอื่นๆ) การให้คำอธิบายของผู้วิจัยนี้ได้ผลก็คือ “สิ่งสำคัญที่สุดของเพจที่เชื่อมโยงไปยังหลายๆเพจและบรรดาลิงค์ที่ตั้งต้นไม่จำเป็นจะต้องเข้าหน้าเพจแรกของโฮสต์ก็ได้ อาจจะเชื่อมโยงกับเพจที่ไม่ใช่เพจแรกก็ได้”

Abiteboul *et al.* (2003) ได้ออกแบบวิธีการในการเข้ารวบรวมข้อมูลของเพจโดยอาศัยอัลกอริทึมที่เรียกว่า OPIC (On-line Page Importance Computation) ใน OPIC โดยแต่ละเพจเริ่มต้น จะได้รับผลรวมของ “Cash” ซึ่งตัวครอว์เลอร์จะกระจายเข้ารวบรวมอย่างสม่ำเสมอตามเพจที่เชื่อมต่อกัน คล้ายกับการคำนวณความสำคัญของเว็บเพจแต่สามารถทำได้อย่างรวดเร็วและทำได้เพียงทีละขั้น การใช้ OPIC ในการเข้ารวบรวมเพจแรกขึ้นอยู่กับจำนวนของ “Cash” ในการดำเนินการทดสอบใน 100,000 เพจ ซึ่งผลการทดลองนี้ไม่ได้เปรียบเทียบกับการทดสอบอื่นๆ ในปีเดียวกันนี้ นิรันดร์ อังควัฒนวิทย์ ได้เสนองานวิจัยที่เกี่ยวกับการเก็บเว็บแบบเฉพาะเจาะจงหัวเรื่องด้วยเว็บครอว์เลอร์แบบเรียนรู้ได้ ซึ่งการเก็บเว็บเพจแบบเฉพาะเจาะจงหัวเรื่องด้วยเว็บครอว์เลอร์แบบเรียนรู้ได้อาศัยการเรียนรู้จากประสบการณ์ในการเก็บเว็บเพจครั้งก่อนหน้า เพื่อหาหนทาง

ที่จะนำไปสู่ผลลัพธ์ของการเก็บเว็บเพจที่ดีที่สุด โดยประโยชน์ของการใช้เรียนรู้จากประสบการณ์ของเว็บครอว์เลอร์คือ หลีกเลี่ยงการเดินซ้ำเข้าไปในเส้นทางที่ไม่ดี หาเส้นทางที่ดีที่สุดที่ให้ผลลัพธ์ที่ดี และลดปริมาณการใช้ทรัพยากรแบนด์วิดท์ในการเก็บเว็บเพจโดยพยายามเก็บเว็บเพจที่เป็นเป้าหมายให้ได้โดยเร็วเพื่อให้การทำงานเสร็จก่อนระยะเวลาที่คาดการณ์ไว้ ข้อเสียประการหนึ่งของการเรียนรู้คือ ในบางครั้งเว็บครอว์เลอร์อาจต้องใช้ระยะเวลาในการเรียนรู้มากครั้งเกินไปจนเปลืองทรัพยากรแบนด์วิดท์ได้ อย่างไรก็ตามจากผลการทดลอง เราจึงมั่นใจว่าการเก็บเว็บเพจแบบเฉพาะเจาะจงด้วยเว็บครอว์เลอร์แบบเรียนรู้ได้มีประสิทธิภาพที่สูงมาก และฐานความรู้ทั้งสามชนิดที่สร้างขึ้นยังมีบทบาทและเป็นเหตุเป็นผลต่อการเพิ่มประสิทธิภาพให้สูงยิ่งขึ้นไปอีกด้วย

Boldi *et al*. (2004) ได้จำลองสถานการณ์ย่อยของเว็บจาก 40 ล้านหน้า จากโดเมน .IT และ 100 ล้านหน้าจาก WebBase Crawl ใช้การทดสอบแบบ Breadth-First โดยการสุ่มลำดับการการเว็บเพจ และ วิธีการเรียนรู้ด้วยตนเอง วิธีการที่ได้ผลดีที่สุดก็คือแบบ Breadth-First แม้ว่ามีการจะมีการสุ่มลำดับก็ตาม นอกจากนี้งานวิจัยยังได้แสดงให้เห็นถึงวิธีการคำนวณความสำคัญของเว็บเพจที่ไม่ค่อยดีในบางส่วนของเพจ ในระหว่างการเข้ารวบรวมของตัวครอว์เลอร์แต่ยังสามารถคำนวณได้ใกล้เคียงกับวิธีการคำนวณความสำคัญของเว็บเพจจริงได้

Baeza-Yates (2005) ได้จำลองในสองส่วนย่อยของเว็บโดยใช้เพจจำนวน 3 ล้านเพจจากโดเมน .GR และ .CL โดยการทดลองการเข้ารวบรวมเพจหลายวิธีการ พวกเราพบว่าทั้งวิธีการ OPIC และวิธีการที่ใช้ความยาวของการเชื่อมต่อเว็บไซต์ มีประสิทธิภาพดีกว่าวิธีการแบบ Breadth-First งานวิจัยนี้เป็นเป็นทางเลือกอีกทางหนึ่งในงานวิจัยในปัจจุบัน

ปุณณวัฒน์ ธาดาภาคย์ (2009) ได้นำเสนอการวิจัยเกี่ยวกับการสร้างเว็บพร็อกซี่จำลองเพื่อใช้งานฐานข้อมูลเว็บเบสของมหาวิทยาลัยสแตนฟอร์ด (Build a Proxy Simulator for Stanford WebBase Repository) ซึ่งการออกแบบ และทดสอบอัลกอลิทึมเว็บครอว์เลอร์ หรือเว็บสไปเดอร์นั้น นอกจากมีความท้าทายมากมายทางด้านวิศวกรรมคอมพิวเตอร์แล้ว ยังต้องการการเชื่อมต่อกับเครือข่ายอินเตอร์เน็ทอีกด้วย ซึ่งจะทำให้การทดสอบประสิทธิภาพของอัลกอลิทึม หรือสถาปัตยกรรมของ เว็บครอว์เลอร์ที่ออกแบบต้องใช้เวลานานเปลืองแบนด์วิดช์ (Bandwidth) ในการเชื่อมต่อเครือข่าย อีกทั้งยังเป็นการรบกวนเครื่องแม่ข่ายปลายทางที่ถูกทดสอบการเก็บเว็บเพจอีกด้วย ในโครงการวิจัยนี้เสนอการออกแบบเว็บพร็อกซี่จำลอง (Web proxy simulator) โดยใช้ข้อมูลเว็บเพจ จากฐานข้อมูลเว็บเบสของมหาวิทยาลัยแสตนฟอร์ด (Stanford WebBase) เป็นข้อมูลทดสอบ ระบบต้นแบบที่ได้ จะสามารถใช้เป็นเว็บพร็อกซี่ให้ต้นแบบเว็บครอว์เลอร์ที่จะสร้างขึ้นมาภายหลัง สามารถทดสอบประสิทธิภาพการจัดเก็บเว็บเพจได้โดยไม่ต้องเชื่อมต่อกับเครือข่ายอินเตอร์เน็ทอีกต่อไป ลดภาระการรบกวนเครื่องแม่ข่ายทดสอบ อีกทั้งยังทำให้ผู้วิจัยสามารถทำการ

ทดสอบ อัลกอลิทึม หรือเว็บครอว์เลอร์ที่ได้ออกแบบไว้ แบบออฟไลน์ (Offline) เป็นจำนวนกี่ครั้งก็ได้อีกด้วย

2.2 งานวิจัยที่เกี่ยวข้องการจัดกลุ่มเอกสารข้อความ (Text Classification)

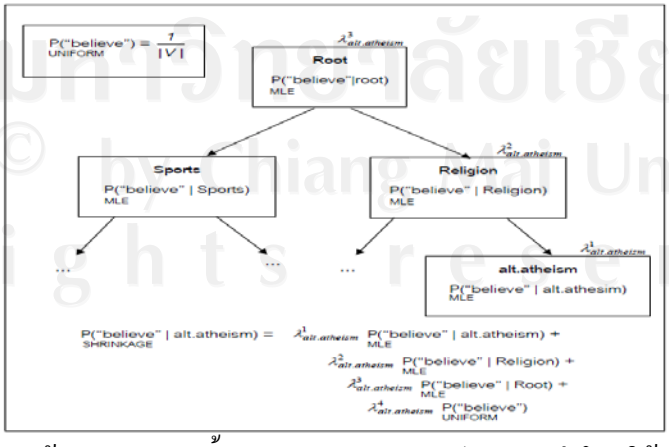
ส่วนงานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มเอกสารในปัจจุบันแบ่งออกเป็น 2 ลักษณะคือ 1) การจัดกลุ่มเอกสารในแนวราบ โดยแยกเอกสารออกเป็นแต่ละประเภทที่ไม่ขึ้นต่อกัน งานวิจัยลักษณะนี้มีอุปสรรคคือ เมื่อจำนวนประเภทของเอกสารเพิ่มมากขึ้น การใช้เวลาในการประมวลผลเพื่อจัดกลุ่มเอกสารก็จะเพิ่มขึ้นเป็นเท่าทวีคูณ จึงทำให้เกิดแนวคิดในการแก้ปัญหาดังกล่าว 2) การจัดกลุ่มเอกสารโดยกำหนดโครงสร้างของประเภทเอกสารเป็นลำดับชั้น เพื่อเป็นการแก้ปัญหาและอุปสรรคที่เกิดขึ้นในงานวิจัยของการจัดกลุ่มที่เป็นแนวราบ

งานวิจัยที่เกี่ยวกับการจัดกลุ่มเอกสารในแนวราบ ส่วนใหญ่จะเป็นการจัดกลุ่มเอกสารโดยใช้หลักของการวัดค่าความถี่ของคำและความน่าจะเป็นของการเกิดของคำในเอกสารประเภทต่าง ๆ หลักการที่นิยมนำมาใช้ในงานวิจัยในลักษณะนี้ได้แก่ หลักการของ TF-IDF (Term Frequency and Inverse Document Frequency) (Joachims, 1997) โดยผสมผสานแนวคิดของความน่าจะเป็นของเบย์สอย่างง่ายเข้าไปด้วย ซึ่งให้ผลการทดลองในการจัดกลุ่มค่อนข้างดีกว่าการจัดกลุ่มด้วยวิธีของเบย์สเพียงอย่างเดียว อีกแนวคิดหนึ่งที่นำมาใช้ในการจัดกลุ่มในแนวราบที่ให้ผลการจัดกลุ่มที่ค่อนข้างน่าเชื่อถือ คือการใช้ขั้นตอนวิธี k-Nearest Neighbor (k-NN Classification) เป็นขั้นตอนวิธีที่นำเอาปัญหาจากขั้นตอนวิธีที่สร้างตัวจัดกลุ่มเอกสาร ให้อยู่ในลักษณะของแผนผัง เช่น ขั้นตอนวิธี Decision Tree C4.5 หรือ RIPPER ซึ่งขั้นตอนวิธีประเภทนี้มักจะไม่มีประสิทธิภาพหากมีจำนวนของคำที่ใช้เป็นลักษณะเด่นมากเกินไป หลักสำคัญของขั้นตอนวิธีนี้คือ ความสามารถในการวัดค่าความใกล้เคียงกันของเอกสารด้วยวิธีของ k-NN ซึ่งสามารถบอกได้ว่า เอกสารใดอยู่ในกลุ่มหรือใกล้เคียงกับเอกสารใด แล้วทำการปรับปรุงน้ำหนักของตัวจัดกลุ่มเพื่อเพิ่มประสิทธิภาพการจัดกลุ่มให้ดียิ่งขึ้น (Weight Adjust k-Nearest Neighbor Classification Algorithm WAKNN) ผลการทดลองจากงานวิจัยนี้จะให้ผลการทำนายที่ดีกว่าวิธีของ C4.5 วิธีของ RIPPER และการจัดกลุ่มด้วยวิธีของ k-NN เพียงอย่างเดียว

ในส่วนของการจัดกลุ่มเอกสารด้วยการกำหนดโครงสร้างของประเภทเอกสารเป็นลำดับชั้นนั้น มีงานวิจัยที่เกี่ยวข้องหลายงานวิจัย ได้แก่ งานวิจัยของ Chuang *et al.* (1999) กล่าวถึงการนำตัวจัดกลุ่มของ TF*IDF ซึ่งเป็นตัวจัดกลุ่มมีขั้นตอนการเรียนรู้ที่ง่าย นำมาจัดโครงสร้างแบบลำดับชั้นจากล่างขึ้นบน โดยโหนดที่อยู่ในลำดับชั้นบนจะเกิดจากการรวมกันของโหนดที่อยู่ในลำดับที่ต่ำกว่า และผู้วิจัยได้มีการกำหนดเวกเตอร์พิเศษขึ้นมาเพื่อช่วยในการปรับจำนวนสมาชิกที่

อยู่ในเวกเตอร์ที่ได้จากการเรียนรู้ในวิธีของ TF*IDF เพื่อให้เกิดความเหมาะสม โดยการเพิ่มเวกเตอร์ที่มีสมาชิกเป็นลักษณะเด่นของเอกสารแต่ละประเภท และทำการตัดลักษณะเด่นที่คิดว่าจะเป็นตัวทำให้เกิดข้อผิดพลาดออก ผลที่ได้จากงานวิจัยนี้ชี้ให้เห็นว่า ลักษณะเด่นที่เกิดการเพิ่มหรือลดด้วยผู้เชี่ยวชาญนั้น จะทำให้ผลการทำนายเอกสารดียิ่งขึ้น แต่ก็ไม่ได้มีหลักเกณฑ์ที่ชัดเจนในการที่จะเพิ่ม หรือลดลักษณะเด่นในตัวจัดกลุ่ม ขึ้นอยู่กับดุลยพินิจของผู้ทำวิจัย และผลการทดลองของงานวิจัยนี้ชี้ให้เห็นว่า การใช้เอกสารที่เป็นเอกสารฝึกมาก ๆ จะทำให้ผลการทำนายมีประสิทธิภาพดีขึ้น

Frommholz (2001) รายงานว่าการจัดโครงสร้างของตัวจัดกลุ่มเอกสารเป็นลำดับชั้น โดยใช้โครงสร้างเป็นแบบกราฟแทนที่โครงสร้างแบบต้นไม้เหมือนกับงานวิจัยอื่น โดยให้เหตุผลว่าการจัดกลุ่มเอกสารแบบลำดับชั้นไม่จำเป็นจะต้องไปสิ้นสุดที่โหนดใบ เพราะเอกสารบางอย่างก็มีโอกาสที่จะเป็นเพียงเอกสารประเภทใดประเภทหนึ่ง ในโหนดที่ไม่ใช่โหนดใบในโครงสร้างของตัวจัดกลุ่ม ดังนั้น การจัดวางโครงสร้างแบบกราฟจะช่วยในการแลกเปลี่ยนข้อมูลกันระหว่างโหนดที่อยู่ในลำดับเดียวกัน และใกล้เคียงกันด้วย ส่วนงานวิจัยที่นำหลักของความน่าจะเป็นมาใช้และให้ผลการทดลองที่ค่อนข้างดีมากงานวิจัยหนึ่ง ก็คืองานวิจัยของ McCallum *et al.* (1998) เป็นการปรับปรุงประสิทธิภาพของการจัดกลุ่มเอกสารที่มีโครงสร้างเป็นลำดับชั้น โดยใช้เทคนิคการรวมกันของตัวประมาณค่าความน่าจะเป็นของคำ (shrinkage-based estimate of the probability) จากตัวจัดกลุ่มที่เป็นโหนดในระดับล่าง ขึ้นไปเป็นตัวจัดกลุ่มในระดับบนจนถึงโหนดที่เป็นรากรูป 2.1 งานวิจัยดังกล่าวมีจุดเด่นคือ สามารถประยุกต์ใช้กับการจัดกลุ่มเอกสารที่มีข้อมูลในการเรียนรู้ของตัวจัดกลุ่มน้อย ๆ ผลการทดลองของงานวิจัยนี้ ให้ค่าความถูกต้องประมาณร้อยละ 84.6 ซึ่งเป็นค่าที่ค่อนข้างสูงและมีข้อดีคือ ไม่จำเป็นต้องใช้คำในการเรียนรู้สำหรับตัวจัดกลุ่มมากนัก



รูป 2.1 การจัดโครงสร้างแบบลำดับชั้นและการคำนวณตัวประมาณค่าโดยใช้เทคนิคการรวมกันของตัวประมาณค่าความน่าจะเป็นของคำ

การนำหลักการของโครงข่ายประสาทเทียมมาประยุกต์ใช้กับการจัดกลุ่มเอกสารแบบมีลำดับชั้น ก็เป็นงานวิจัยที่น่าสนใจ โดยงานวิจัยของ Ruiz and Srinivasan (2002) จะนำโครงข่ายประสาทเทียมมาใช้ในการสร้างตัวจัดกลุ่ม และใช้ข้อมูลทดสอบจากเอกสารทางการแพทย์ที่เกี่ยวกับการวินิจฉัยโรค (MEDLINE) และได้ทำการเปรียบเทียบผลการทดสอบของโครงข่ายประสาทเทียมแบบแนวราบกับแบบมีลำดับชั้น ซึ่งสรุปได้ว่า การจัดกลุ่มด้วยวิธีโครงข่ายประสาทเทียมแบบมีลำดับชั้นให้ผลที่ดีกว่า และเมื่อนำผลการทดลองไปเปรียบเทียบวิธีของ k-NN ก็ให้ผลที่ดีกว่าเช่นกัน ส่วนงานวิจัยของ Koller and Sahami (1997) ได้นำเสนอวิธีการเลือกคำที่เป็นลักษณะเด่นของการจัดกลุ่มเอกสารแบบมีลำดับชั้น โดยใช้วิธีที่เรียกว่า cross-entropy ซึ่งการเลือกลักษณะเด่นนี้จะนำไปใช้ในการสร้างโหนดที่อยู่ในระดับสูงขึ้นไป และจะเลือกเฉพาะคำที่มีค่า cross-entropy สูง ๆ เพียง n ลำดับแรกเท่านั้น เพื่อลดขนาดของคำที่จะใช้ในการเรียนรู้ของตัวจัดกลุ่มเอกสาร ซึ่งการทดลองในงานวิจัยนี้ให้ผลการทดลองที่มีประสิทธิภาพในการทำนายเอกสารค่อนข้างสูง และเป็นงานวิจัยอ้างอิงที่หลายงานวิจัยได้นำไปประยุกต์ใช้

2.3 งานวิจัยที่เกี่ยวข้องซัพพอร์ตเวคเตอร์แมชชีน (Support Vector machines)

ส่วนในงานวิจัยที่อาศัยหลักการซัพพอร์ตเวกเตอร์แมชชีนซึ่งในปี 1998 Thorsten Joachims ได้นำเสนอการจัดกลุ่มเอกสารโดยอาศัยซัพพอร์ตเวกเตอร์แมชชีน (TextCategorization with Support Vector Machines) ในงานวิจัยนี้อาศัยหลักการของซัพพอร์ตเวกเตอร์แมชชีน, Bayes, Rocchio, C4,5, k-NN ในการเรียนรู้การจัดกลุ่มเอกสารจากคลังข้อมูลของ Ohsumed โดยใช้ชุดอบรม 10,000 ชุดอบรม จะทำการวิเคราะห์เฉพาะคุณสมบัติและเอกลักษณ์ของเอกสาร โดยมีคำถามเกี่ยวกับซัพพอร์ตเวกเตอร์แมชชีนว่าเหมาะสมกับงานวิจัยหรือไม่ ผลลัพธ์ของการใช้ Bayes ได้ประสิทธิภาพ $Micro_{avg}$=72% Rocchio ได้ประสิทธิภาพ $Micro_{avg}$=79% C4,5 ได้ประสิทธิภาพ $Micro_{avg}$=79% k-NN ได้ประสิทธิภาพ $Micro_{avg}$=82% และซัพพอร์ตเวกเตอร์แมชชีนได้ประสิทธิภาพ $Micro_{avg}$=86% ซึ่งซัพพอร์ตเวกเตอร์แมชชีนสามารถพัฒนาให้มีประสิทธิภาพดีขึ้นได้อีกและสามารถพัฒนาได้ขึ้นในระดับการเรียนรู้ที่ยากขึ้น นอกจากนี้งานวิจัยนี้ยังทำเป็นแบบอัตโนมัติซึ่งกำจัดการจัดกลุ่มเอกสารที่ทำด้วยมือ ซึ่งในงานของการจัดกลุ่มเอกสารก็ยังมีงานวิจัยของ จันทิมา พลพินิจ (2005) ได้มีการนำเสนอระบบการจัดกลุ่มเอกสารข้อความอัตโนมัติด้วยซัพพอร์ตเวคเตอร์แมชชีน (Automatic document classification by support vector machines) ปัจจุบันเนื่องจากระบบเวิร์ลไวด์เว็บได้รับความนิยมในการใช้งานเป็นอย่างมากเพราะความสามารถในการให้บริการเกี่ยวกับสารสนเทศที่มีความหลากหลาย ทำให้เกิดปัญหาที่ต้องได้รับการแก้ไขเพราะปริมาณของสารสนเทศที่มีจำนวนมหาศาล

ซึ่งปัญหาดังกล่าวก็คือ การค้นหาและการเลือกใช้สารสนเทศจะไม่สามารถกระทำได้ในเวลาอันรวดเร็ว การจัดกลุ่มเอกสารข้อความอัตโนมัติจึงกลายเป็นอีกทางเลือกที่สามารถช่วยให้การเข้าถึงสารสนเทศได้เร็วขึ้น งานวิจัยนี้นำเสนอการจัดกลุ่มเอกสารข้อความภาษาไทยแบบอัตโนมัติบนพื้นฐานของพจนานุกรมและอัลกอริทึมซัพพอร์ตเวคเตอร์แมชชีน ภายหลังจากที่ระบบพัฒนาเสร็จสิ้นได้มีการทดสอบประสิทธิภาพของระบบด้วยการวัดค่าเอฟแล้วพบว่าระบบให้ผลของความถูกต้องในการจัดกลุ่มเอกสารอยู่ระหว่างร้อยละ 77.46% - 84.71% ต่อมาก็ได้มีงานวิจัยที่เกี่ยวกับการจัดกลุ่มเอกสารภาษาไทยของวัลลภ อินทร์ฉ่ำ (2005) ได้มีการนำเสนอระบบการจัดหมวดหมู่เอกสารภาษาไทยอัตโนมัติโดยใช้ซัพพอร์ตเวคเตอร์แมชชีนร่วมกับการประมวลผลภาษา (Automatic Thai document categorization system using SVM with language processing) ผลงานนี้มีวัตถุประสงค์ที่จะใช้เครื่องมือเรียนรู้ซัพพอร์ตเวกเตอร์แมชชีน ซึ่งมีหลายฟังก์ชันให้เลือกใช้งานและมีประสิทธิภาพสูงสำหรับการจัดหมวดหมู่เอกสารในต่างประเทศ เพื่อศึกษาและเปรียบเทียบประสิทธิภาพระบบการจัดหมวดหมู่เอกสารภาษาไทยอัตโนมัติ ร่วมกับการใช้เทคนิคการวิเคราะห์ทางภาษาเข้ามาช่วยในการประมวลผลเพื่อเพิ่มความถูกต้อง และลดขนาดมิติเอกสารลงเพื่อเพิ่มความรวดเร็วในการประมวลผล งานวิจัยนี้ได้ทดสอบการใช้วลีร่วมกับคำเดี่ยว การใช้คำจากชื่อเรื่องเพิ่มค่าความถี่ และทดสอบการใช้งานซัพพอร์ตเวกเตอร์แมชชีน ด้วยฟังก์ชันเคอร์เนลแบบเชิงเส้น แบบ radial basis และแบบ polynomial โดยมีการทดสอบระบบกับเอกสารประเภทข่าวจากหนังสือพิมพ์ผู้จัดการจากเว็บไซต์ 1 ปี จำนวน 5 กลุ่มข่าวคือ ข่าววัฒนธรรมและสิ่งแวดล้อม ข่าวบันเทิง ข่าวการเงิน ข่างการเมือง และข่าวเทคโนโลยี ผลปรากฏว่าการใช้งานของซัพพอร์ตเวกเตอร์แมชชีนกับการจัดหมวดหมู่เอกสารภาษาไทยโดยใช้ เคอร์เนลแบบเชิงเส้นมีประสิทธิผลสูงที่สุดเมื่อเทียบกับเคอร์ radial basis และ polynomial กล่าวคือค่าความแม่นยำเท่ากับ 94.6%,90.1% และ 80.0% ตามลำดับ และเมื่อใช้วลีร่วมกับคำเดี่ยวจะมีผลทำให้ค่าความถูกต้องเพิ่มขึ้นจาก 93.0% เป็น 95.0% และเมื่อใช้ชื่อเรื่องเพิ่มค่านำหนักจะทำให้ประสิทธิผลค่าความถูกต้องเพิ่มจาก 91.0% เป็น 94.0%

Nadeem Ahmed Syed *et al.* (1999) ได้มีการนำเสนอแนวคิดการจัดการเพื่อเพิ่มการเรียนรู้สำหรับซัพพอร์ตเวกเตอร์แมชชีน (Handling Concept Drifts in Incremental Learning with Support Vector Machines) ได้มีการศึกษาการเพิ่มขนาดของฐานข้อมูลในโลกแหล่งความเป็นจริงนั้น ในการเพิ่มขึ้นนี้ทำให้มีอิทธิพล ต่อการเรียนรู้อัลกอริทึม การเพิ่มการเรียนรู้เทคนิคเป็นวิธีหนึ่งที่สามารถที่จะแก้ไขปัญหาได้ ในงานวิจัยนี้ โดยงานวิจัยนี้ได้มีเงื่อนไข 3 เงื่อนไขที่ใช้ทดสอบ คือ จำนวนของข้อมูล (ค่าคงที่) ต้นทุนของการปรับปรุงหน่วยความจำ คุณภาพของการ

เรียนรู้ ในการวิจัยนี้มีการประเมินจุดแข็ง และเพิ่มความน่าเชื่อถือของการเรียนรู้วิธีการศึกษาจุดแข็งของการเพิ่มการเรียนรู้วิธีการของซัพพอร์ตเวกเตอร์แมชชีน

Taku Kudo and Yuji Matsumoto (2000) ได้มีการนำเสนอการวิเคราะห์โครงสร้างของภาษาญี่ปุ่นโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน (Japanese Dependency Structure Analysis Based on Support Vector Machines) งานวิจัยนี้ได้นำเสนอวิธีการของการวิเคราะห์โครงสร้างของภาษาญี่ปุ่นโดยใช้ซัพพอร์ตเวกเตอร์แมช ชีน ตามปกตินิยมการแยกวิเคราะห์เทคนิคโครงสร้างตามการเรียนรู้ของเครื่อง เช่น ต้นไม้ตัดสินใจ และแมกซิมัมเอนโทรปีโมเดล มีความยากที่เลือกใช้ลักษณะเฉพาะตามที่ได้ค้นหามา ในอีกทางหนึ่งเป็นที่รู้จักกันดีนั้นคือซัพพอร์ตเวกเตอร์แมชชีนซึ่งมีประสิทธิภาพสูง แม้กระทั่งกับข้อมูลที่เข้ามานั้นสูงมากก็ตาม นอกจากนี้ขอแนะนำหลักการของเคอร์เนลซึ่งสามารถดำเนินการอบรมข้อมูลหลายมิติโดยใช้ทรัพยากรน้อยในการประมวลผล งานวิจัยนี้ใช้ซัพพอร์ตเวกเตอร์แมชชีนในการแก้ปัญหาวิเคราะห์โครงสร้างภาษาญี่ปุ่น ได้ผลการทดลองของมหาวิทยาลัยเกียวโต แสดงว่าระบบของเราให้ความแม่นยำถึง 89.09% โดยใช้ข้อมูลในการอบรม 7,958 ประโยค

Kengo SATO and Hiroaki SAITO (2002) ได้มีการนำเสนอสกัดลำดับคำโต้ตอบโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน (Extracting Word Sequence Correspondences with Support Vector Machines) ในงานวิจัยนี้ได้นำเสนอการเรียนรู้และการสกัดวิธีการของลำดับคำโต้ตอบจากคลังข้อมูลที่กระจัดกระจายโดยใช้ซัพพอร์ตเวกเตอร์แมชชีนที่มีความสามารถสูงของลักษณะทั่วไป ไม่บ่อยครั้งที่จะอบรมข้อมูลที่พึ่งพาอาศัยกันโดยการใช้ฟังก์ชันของเคอร์เนล วิธีการใช้ลักษณะเฉพาะเพื่อแปลงโมเดลเพื่อแปลพจนานุกรมคือ จำนวนของคำ ชนิดของคำ ส่วนประกอบของคำและคำที่อยู่ใกล้เคียง การทดลองได้ผลลัพธ์การเทียบระหว่างภาษาญี่ปุ่นและภาษาอังกฤษ ได้ค่าความถูกต้อง 81.1% ค่าระลึกได้ 69.0% ของการสกัดลำดับคำ งานวิจัยนี้แสดงให้เห็นว่าวิธีนี้สามารถลดต้นทุนสำหรับการทำแปลพจนานุกรม จากงานวิจัยนี้ได้มีผู้วิจัยทำการวิจัยเกี่ยวการสกัดประโยค Tsutomu HIRAO *et al.* (2002) ได้วิจัยเกี่ยวกับการสกัดประโยคสำคัญโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน (Extracting Important Sentences with Support Vector Machines) การสกัดประโยคที่มีข้อมูลที่สำคัญจากเอกสารที่อยู่ในรูปแบบการสรุปใจความนั้น เทคนิคที่เป็นกุญแจที่จำสกัดประโยคจากข้อความที่แบบสรุปใจความให้เหมือนกับมนุษย์เขียนนั้น เพื่อให้บรรลุดังความตั้งใจ มันเป็นสิ่งที่สำคัญที่จะสามารถผสานรวมความที่ไม่เหมือนของข้อความ แต่มีวิธีหนึ่งโดยใช้พารามิเตอร์ งานวิจัยนี้ได้เสนอวิธีสกัดประโยคโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน เพื่อเป็นยืนยันว่าวิธีการนี้ได้ประสิทธิแค่ไหน ทางผู้วิจัยได้ผลทดสอบที่เปรียบเทียบ 4 วิธีด้วยกัน คือ วิธีการ Lead-based , Decision Tree ,วิธีการ Boosting และ ซัพพอร์ตเวกเตอร์แมชชีน ได้ผลลัพท์โดย

อาศัยข้อมูลจาก TSC (Test Summarization Collection) แสดงให้เห็นว่าการใช้ซัพพอร์ตเวกเตอร์แมชชีนให้ประสิทธิภาพสูงที่สุด ต่อมาได้มีการวิจัยเกี่ยวกับการลดประโยค Minh Le Nguyen *et al.* (2004) ได้นำเสนองานวิจัยเกี่ยวกับความเป็นไปได้ในการลดประโยคโดยอาศัยซัพพอร์ตเวกเตอร์แมชชีน (Probabilistic Sentence Reduction Using Support Vector Machines) งานวิจัยนี้ได้มีการวิวัฒนาการที่ไม่เคยเกิดขึ้นมาก่อนของซัพพอร์ตเวกเตอร์แมชชีนสำหรับการลดประโยค นอกจากนี้ยังได้เสอนความเป็นไปได้ใหม่สำหรับการลดประโยคโดยอาศัยซัพพอร์ตเวกเตอร์แมชชีน โดยใช้อ้างอิงอัลกอริทึมของ (Knighit and Marcu, 2002) ซึ่งใช้คลังข้อมูลของ Benton (http://www.benton.org) มีข้อมูลอบรม 1,035 ประโยค ผลการทดลองแสดงการใช้วิธีการ Proposed มีประสิทธิภาพกว่าวิธีการ earlier ในการทดลองครั้งนี้ ต่อมา ธีรพงศ์ โหมดหิรัญ (2005) นักวิจัยไทยได้วิจัยเกี่ยวกับความกำกวมของคำในภาษาไทย คือ การแก้ไขปัญหาความกำกวมของคำในภาษาไทยโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน (Word sense disambiguation in Thai using support vector machine) ผลงานนี้ได้นำเสนอการแก้ไขปัญหาของคำที่มีความหมายกำกวมในภาษาไทย โดยนำเสนอวิธีการเรียนรู้ที่เรียกว่าซัพพอร์ตเวกเตอร์แมชชีน ซึ่งเป็นอัลกอริทึมในการหาระนาบของการแบ่งข้อมูลที่เหมาะสมที่สุด นำมาใช้ในการเรียนรู้เพื่อแก้ปัญหาความกำกวมของคำ โดยใช้วิธีการเลือกลักษณะ (Feature) 3 แบบ คือคำที่ปรากฏรอบข้างคำกำกวม (Word) คำและหน้าที่ของคำที่ปรากฏรอบข้างคำกำกวม (Word and Part of Speech), คู่ของคำและหน้าที่ของคำที่อยู่ติดกันรอบข้างคำกำกวม (Collocation of Word and Part of Speech) และมีค่าพารามิเตอร์ที่ทำการปรับคือ จำนวนระยะห่างจากคำกำกวมที่แตกต่างกัน (Window size) โดยทดสอบกับคำที่มีความหมายกำกวมในภาษาไทยจำนวน 10 คำ และทำการเปรียบเทียบการเรียนรู้ของเครื่องทั้งหมด 4 แบบคือ SVM (Support Vector Machine), SNOW (Sparse Network of Winnow), Naïve Bayes, Neural Network ผลการทดสอบปรากฏว่าการใช้ซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความถูกต้องมากที่สุด ต่อมา พัฒนชัย เบศรภิญโญวงศ์ (2002) ได้วิจัยเกี่ยวกับการรู้จำของอักษรไทย ซึ่งได้เสนองานวิจัยเกี่ยวกับการรู้จำตัวอักษรไทยโดยใช้ซัพพอร์ตเวกเตอร์แมชชีนและเคอร์เนล (Thai character recognition using Support Vector Machines and Kernels) ได้ปรับปรุงความถูกต้องในการรู้จำของโปรแกรมโอซีอาร์ภาษาไทย โดยได้นำเอาเทคนิคของซัพพอร์ตเวกเตอร์แมชชีน (เอสวีเอ็ม) และเคอร์เนลเข้ามาประยุกต์ใช้ในส่วนของการวิเคราะห์องค์ประกอบสำคัญของข้อมูล ซึ่งเป็นกระบวนการที่สำคัญในการดึงเอาลักษณะสำคัญของข้อมูลรูปภาพตัวอักษร ก่อนที่จะส่งข้อมูลที่ได้ไปยังส่วนรู้จำของโปรแกรม โอซีอาร์ เพื่อแยกแยะว่าเป็นตัวอักษรชนิดใดต่อไป โดยเรียกเทคนิคการวิเคราะห์องค์ประกอบสำคัญของข้อมูลแบบใหม่นี้เรียกว่า การวิเคราะห์องค์ประกอบสำคัญของข้อมูลแบบ

เคอร์เนล ในวิทยานิพนธ์ฉบับนี้ ได้แบ่งรูปภาพที่ใช้ทดสอบออกเป็นสองกลุ่ม คือรูปภาพชุดเรียนรู้ จำนวน 8,544 ตัว และรูปภาพชุดทดสอบจำนวน 1,424 ตัว ประกอบด้วยตัวอักษรแบบ AngsanaUPC, BrowalliaUPC, CordiaUPC, DilleniaUPC, EucrosiaUPC และ FreesiaUPC แต่ละแบบประกอบด้วยตัวอักษรขนาด 14,16,18,20,22,24,24,28 และ 36 พอยต์ ผลของการทดสอบพบว่า ผลของการรู้จำของโปรแกรมโอซีอาร์ภาษาไทย ที่ใช้เทคนิคของการวิเคราะห์องค์ประกอบสำคัญของข้อมูลแบบเคอร์เนล ให้ผลการรู้จำที่ดีขึ้นจากโปรแกรมโอซีอาร์ภาษาไทยตัวเดิม อย่างไรก็ตาม วิธีใหม่นี้กลับใช้หน่วยความจำและเวลาที่เพิ่มขึ้นจากเดิม

Koichi Takeuchi and Nigel (2002) ได้นำเสนอการสกัดข้อมูลชีวการแพทย์โดยใช้ซัพพอร์ตเวกเตอร์แมชชีน (Bio-Medical Entity Extraction using Support Vector Machines) ซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพในงานคัดแยกที่หลากหลาย ในงานวิจัยนี้ได้นำหลักการของซัพพอร์ตเวกเตอร์แมชชีนมานำมาใช้เพื่อการวินิจฉัยและหาความหมาย ตามหลักการทางวิทยาศาสตร์และเทคนิคของระบบคำศัพท์ในโดเมนของชีววิทยาโมเลกุล งานวิจัยนี้แสดงการขยายออกของงานวิจัยแบบเดิมๆ งานวิจัยนี้แสดงถึงซัพพอร์ตเวกเตอร์แมชชีนที่ใช้ตัวอย่าง 100 บทคัดย่อวารสาร จาก {*human, blood cell,transcription factor*} ของโดเมนฐานข้อมูลทางการแพทย์ มีจำนวนคำประมาณ 3,400 คำ และได้ผลของโมเดลที่ใช้ทดสอบคือ F-Score 74 % ในส่วนรายละเอียดของการวิเคราะห์ ซัพพอร์ตเวกเตอร์แมชชีนมีส่วนช่วยให้งานมีประสิทธิภาพ

ด้านการจัดกลุ่มเว็บเพจสามารถนำวิธีการที่ใช้กับการจัดกลุ่มเอกสารมาใช้กับการจัดกลุ่มเว็บเพจได้ ในมุมมองที่เว็บเพจเป็นเอกสารประเภทหนึ่งเช่นกัน แต่ลักษณะของเว็บเพจมีความแตกต่างจากเอกสารทั่วไป คือมีแท็กกำกับคุณสมบัติของวัตถุในเอกสาร และเว็บเพจต่าง ๆ ถูกเชื่อมโยงถึงกันด้วยลิงค์ คุณสมบัติเหล่านี้จึงถูกใช้นำมาใช้กับงานวิจัยที่ใช้ตัวจัดกลุ่มที่ทำงานร่วมกัน เพื่อจัดกลุ่มเว็บเพจโดยอาศัยข้อมูลลิงค์ (Hyperlink) Fürnkranz (1998) ได้เปรียบเทียบลักษณะเด่นจากสองแบบ ลักษณะแรกได้จากคำทั้งหมดที่กำจัดแท็กออกแล้ว (Full-Text) ที่อยู่ในชุดเอกสารเป้าหมาย (Target Page) กับลักษณะที่สองคือ คำบรรยายลิงค์ (Anchor Text) ของทุกหน้าที่มีลิงค์มายังเอกสารเป้าหมาย (Predecessor Page) และใช้เอกสารทดลองจาก WebKB ประกอบด้วย 6 ประเภท และเอกสารเป้าหมาย 1,050 หน้า เอกสารที่มีลิงค์ไปยังเอกสารเป้าหมาย 5,803 หน้า และแบ่งชุดข้อมูล โดยแต่ละตัวจัดกลุ่มทำงานร่วมกันใช้ตัวจัดกลุ่มที่อาศัยการเรียนรู้จากของริปเปอร์ (Ripper Rule Learning Algorithms) จากแต่ละหน้าที่มีลิงค์มายังเอกสารเป้าหมายพบว่าแต่ละตัวจัดกลุ่มให้ความถูกต้องเป็น 51.81% จากนั้นนำทุกตัวจัดกลุ่มที่ได้ผ่านวิธีการรวมเพื่อให้ได้ผลสุดท้ายของการจัดกลุ่ม จากผลการทดลองพบว่าลักษณะเด่นที่ได้จากคำบรรยายลิงค์ ให้ความถูกต้องเป็น 74.67% ซึ่งสูงกว่าลักษณะเด่นที่ได้จากคำทั้งหมด (70.67%) แต่

เมื่อนำคำทั้งหมดผ่านวิธีเอนโทรปี (Entropy) เพื่อเลือกลักษณะเด่น (Feature Subset Selection) ที่ดีที่สุดมา n% จากลักษณะเด่นคำทั้งหมด พบว่าค่า n ที่ดีที่สุดคือ 5-10 ซึ่งให้ความถูกต้องเป็น 74.67% เท่ากับแบบคำบรรยายลิงค์ ในงานวิจัยนี้ใช้ 4 วิธีการสำหรับตัดสินผลการจัดกลุ่ม (Voting, Weighted Sum, Weighted Normalized Sum และ Maximum Confidence) พบว่า Maximum Confidence ให้ค่าความแม่นยำสูงกว่าแบบอื่น นอกจากนั้นยังเปรียบเทียบข้อความในทุกหน้าที่มีลิงค์มายังเอกสารเป้าหมาย 3 ลักษณะเด่นคือ ในส่วนของหัวเรื่อง (Heading) ที่เป็นข้อมูลการเชื่อมโยง คำที่อยู่ในย่อหน้าที่มีข้อมูลการเชื่อมโยง และวลีที่อยู่ในตำแหน่งก่อนหน้าข้อมูลการเชื่อมโยง พบว่าให้ความถูกต้องเป็น 72.95% 66.29% และ 56.57% ตามลำดับ ซึ่งให้ความถูกต้องน้อยกว่าการนำลักษณะเด่นมาใช้ร่วมกัน 3 แบบดังนี้คือ คำบรรยาย ลิงค์ กับหัวเรื่องเป็น 86.57% คำบรรยายลิงค์ กับหัวเรื่อง และย่อหน้า เป็น 86.86% และเมื่อใช้คำบรรยายลิงค์ หัวเรื่อง ย่อหน้า และวลี ร่วมกันเป็น 85.05% ซึ่งแบบนี้ใช้จำนวนคำที่เป็นลักษณะเด่น 8,075 คำ น้อยกว่าจำนวนคำที่ได้จากเอกสารเป้าหมาย (20,322 คำ) ต่อมา Aixin Sun *et al.* (2002) ได้นำเสนอการการจัดกลุ่มเว็บโดยอาศัยซัพพอร์ตเวกเตอร์แมชชีน (Web Classification Using Support Machine) โดยในปัจจุบันเว็บเพจมีจำนวนมากขึ้นและมีเนื้อหาหลายประเภท งานวิจัยนี้จึงมีแนวคิดที่จะจัดกลุ่มเว็บเพจโดยพิจารณาจากเนื้อของเว็บเพจ อย่างเช่น ไฮเปอร์ลิงก์ และ แทกเอชทีเอ็มแอล ในงานวิจัยนี้ได้ใช้ซัพพอร์ตเวกเตอร์แมชชีนในการจัดกลุ่มเว็บเพจโดยใช้เนื้อหาและบริบทของตัวเว็บเพจ การทดลองนี้อาศัยคลังข้อมูลจาก WebKB (http://www2.cs.cmu.edu/~webkb/) ในการทดลอง อาศัยวิธีการ FoIL-PILFS method ของ M. Craven และ S. Slattery ทดสอบบนข้อมูลที่มีอยู่ การทดสอบได้ผลเป็นอย่างดี โดยเฉพาะบริบทของเนื้อหาสามารถจัดกลุ่มได้อย่างมีประสิทธิภาพ โดยได้ค่าประสิทธิภาพคิดเป็นร้อยละเกินกว่า 60 %

ธีรพันธุ์ สุทธิเทพ (2002) ได้มีงานวิจัยที่เกี่ยวกับการจัดกลุ่มของเครื่องมือทางด้านการแพทย์ ซึ่งได้นำเสนอ ระบบจัดกลุ่มเครื่องมือทันตกรรมด้วยวิธีเครื่องเวคเตอร์เกื้อหนุน (Dental Equipment Classification System using Support Vector Machines) ซึ่งงานวิจัยนี้ได้นำเสนอการพัฒนาระบบจัดกลุ่มเครื่องมือที่สามารถบ่งบอกชนิดของเครื่องมือพร้อมทั้งระบุตำแหน่งของเครื่องมือภายในพื้นที่ทำงานได้ โดยใช้การประมวลผลภาพถ่ายจากกล้องวิดีโอ วิธีการจัดกลุ่มเครื่องมือในงานวิจัยนี้เน้นที่วิธีเครื่องเวคเตอร์เกื้อหนุน ซึ่งเป็นวิธีการอย่างหนึ่งที่ใช้ในงานการแยกแยะวัตถุได้อย่างมีประสิทธิภาพและได้รับการยอมรับว่าสามารถทำการแยกแยะวัตถุได้อย่างเหมาะสมที่สุด วิธีเครื่องเวคเตอร์เกื้อหนุนเป็นวิธีที่มีขั้นตอนในการเรียนรู้และจดจำ ดังนั้นระบบที่ได้จึงมีความยืดหยุ่นสูงโดยสามารถนำเอาข้อมูลของวัตถุใหม่ใส่ให้ระบบสามารถเรียนรู้

และจดจำได้ นอกจากนั้นแล้วในงานวิจัยชิ้นนี้ได้นำเอาจีนเนติกอัลกอริทึม ซึ่งเป็นวิธีการค้นหาคำตอบที่ดีที่สุดเชิงปัญญาประดิษฐ์มาทำการปรับพารามิเตอร์ของเครื่องเวคเตอร์เกื้อหนุน เพื่อเป็นการเพิ่มประสิทธิภาพการแยกแยะของระบบขึ้นอีกด้วย ข้อมูลชนิดและข้อมูลตำแหน่งของเครื่องมือที่ได้จากระบบนี้สามารถนำไปใช้เป็นอินพุตให้กับระบบอื่นๆ เพื่อทำการประมวลผลหรือใช้งานกับเครื่องมือนั้นๆ ต่อไปเช่นระบบหุ่นยนต์แขนกลเพื่อให้สามารถหยิบชิ้นเครื่องมือที่ต้องการให้กับผู้ใช้ได้ ระบบต้นแบบที่ได้พัฒนาขึ้นในงานวิจัยนี้เป็นระบบจัดกลุ่มเครื่องมือทันตกรรม ซึ่งในงานทันตกรรมแต่ละครั้งจำเป็นจะต้องใช้ผู้ปฏิบัติหน้าที่อย่างน้อยสองคน ได้แก่ทันตแพทย์และผู้ช่วยทันตแพทย์ โดยหน้าที่อย่างหนึ่งของผู้ช่วยทันตแพทย์คือการหยิบเครื่องมือต่างๆ ตามความต้องการของทันตแพทย์ ดังนั้นระบบต้นแบบนี้สามารถนำไปประยุกต์ใช้เป็นระบบหยิบจับเครื่องมือทันต-กรรมแบบอัตโนมัติ เพื่อที่จะให้ทันตแพทย์สามารถทำงานคนเดียวได้อย่างสะดวกและมีประสิทธิภาพได้

W.M. Campbell et al. (2003) ได้ช่วยกันวิจัยงานด้านการจัดกลุ่มการออกเสียงของผู้พูดโดยอาศัยซัพพอร์ตเวกเตอร์แมชชีน (Phonetic Speaker Recognition with Support Vector Machines) น่าสังเกตการณ์ออกเสียงของผู้พูด ความสัมพันธ์ของการออกเสียง การเปล่งเสียง การสนทนา มีความสำคัญในปัจจุบัน ผู้พูดไม่สามารถจำกัดการฟังแต่ใช้ลักษณะของภาษาในการจัดกลุ่ม คลังข้อมูลขนาดใหญ่มีการใช้ประโยชน์ในการพูดที่มีอยู่ในช่วงไม่กี่ปีที่ผ่านมา ได้มีการทดลองโดยเก็บสถิติ การพูดแบบโทรศัพท์ การพูดแบบถ้อยคำ และรูปแบบอื่นๆ ของแต่ละคน งานวิจัยนี้ได้ใช้ซัพพอร์ตเวกเตอร์แมชชีนและค่าน้ำหนักของคำเพื่อสร้างโมเดลให้กับผู้พูด และยังใช้เทคนิคการจัดกลุ่มเอกสารในการแก้ไขปัญหานี้ด้วย ซึ่งผลการทดลองโดยใช้อาศัยหลักการซัพพอร์ตเวกเตอร์แมชชีนสามารถจัดกลุ่มการออกเสียงของผู้พูดได้เป็นอย่างดีและสามารถลดความผิดพลาดได้ 60 % เมื่อเทียบกับมาตรฐานลอก -ไลคลีฮูด ในงานประเภทนี้ยังมีนักวิจัยได้วิจัยเกี่ยวกับการค้นหาตำแหน่งภาพใบหน้าของมนุษย์ ในปีเดียวกันนี้ พงศักดิ์ ปาละสุทธิกุล ได้เสนอการค้นหาตำแหน่งภาพใบหน้ามนุษย์บนภาพสีด้วยวิธีการซัพพอร์ตเวกเตอร์แมชชีน (Real time face detection on color images by support vector machine) ผลงานนี้ได้นำเสนอเทคนิคการค้นหาตำแหน่งภาพใบหน้ามนุษย์ในเวลาจริง เช่น จากกล้องถ่ายภาพวีดีโอ โดยนำตัวกรองสีผิวที่มีความทนทานต่อแสงเข้ามาช่วยเพื่อเป็นการเพิ่มประสิทธิภาพในการค้าหาภาพให้ดียิ่งขึ้น นอกจากนั้นยังนำวิธีการซัพพอร์ตเวกเตอร์แมชชีนที่ได้ผ่านกระบวนการเทรนข้อมูล มาแยกแยะข้อมูลภาพใบหน้ามนุษย์ที่ได้จากตัวกรองสีผิว จากการทดลองใช้ภาพจำนวน 12,472 ภาพสำหรับเป็นข้อมูลในการเทรนพบว่าผลที่ได้จากการทดลองแสดงให้เห็นถึงประสิทธิภาพและความสามารถของระบบที่ได้นำเสนอและความถูกต้องที่พบว่าสูงกว่าวิธีการอื่น และยังมีงานวิจัยที่

เกี่ยวกับการเคลื่อนไหวของมนุษย์ Hedvig S. *et al.* (2004) ได้นำเสนอการตรวจสอบการเคลื่อนไหวของมนุษย์โดยอาศัยซัพพอร์ตเวกเตอร์แมชชีน (Detecting Human Motion with Support Vector Machines) ในงานวิจัยนี้ได้นำเสนอวิธีการสำหรับการตรวจสอบการเคลื่อนไหวของมนุษย์ในวีดีโอ จุดมุ่งหมายเพื่อตรวจสอบบุคคลที่เดินผ่านอยู่ในท้องถนน ซึ่งเราไม่สามารถจะควบคุมสภาพแวดล้อม ลักษณะเสื้อผ้า สภาพอากาศ และความสว่าง ทิศทางของแสง ที่แตกต่างของมนุษย์ได้ จึงมีแนวคิดเพื่อตรวจสอบรูปแบบของการเคลื่อนไหวของมนุษย์ซึ่งมีขอบเขตและความแตกต่างกัน ดั้งนั้นจึงนำซัพพอร์ตเวกเตอร์แมชชีนมาทำการเรียนรู้อาศัยการเคลื่อนไหวหลายรูปแบบของมนุษย์ในการเรียนรู้ ในการเคลื่อนไหวของมนุษย์นั้นมีการเคลื่อนไหวหลายมุมและกล้องถ่ายภาพยังมีมาตรฐานที่แตกต่างกันในงานวิจัยนี้ได้ขนาดของภาพ 360 x 288 พิกเซลและความถี่อยู่ที่ 25 เฮิรตซ์ จากการเรียนรู้ของซัพพอร์ตเวกเตอร์แมชชีนโดยใช้อัลกอริทึมการตรวจสอบการเคลื่อนไหวสามารถตรวจสอบการเคลื่อนไหวของมนุษย์ได้เป็นอย่างดี และปริญญา สุวรรณศรีคำ (2005) ได้วิจัยเกี่ยวกับการยืนยันของผู้พูด ซึ่งใช้หลักการของซัพพอร์ตเวกเตอร์แมชชีน ได้นำเสนอหัวข้อการยืนยันผู้พูดโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน (Speaker Verification using Support Vector Machine) ในโครงงานนี้ได้เสนอระบบยืนยันผู้พูดโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน ซึ่งระบบนี้เป็นระบบที่ไม่ขึ้นกับข้อความของผู้พูด การยืนยันผู้พูดเป็นปัญหาการแบ่งจัดกลุ่มข้อมูลให้เป็นสองกลุ่ม สัมประสิทธิ์ความถี่ถูกใช้เป็นตัวแทนคุณลักษณะของเสียงพูดและใช้ เป็นอินพุทให้กับส่วนจัดกลุ่ม ซึ่งส่วนจัดกลุ่มในการทดลองนี้ได้ใช้ซัพพอร์ตเวกเตอร์แมชชีนที่มีเคอร์เนลเป็นแบบโพลีโนเมียลอันดับสาม เสียงพูดที่ใช้ในโครงงานนี้มาจากฐานข้อมูลเสียงพูดโลตัสที่พัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

ฐิมาพร เพชรแก้ว (2004) ได้มีการนำเสนอซัพพอร์ตเวกเตอร์แมชชีนแบบหลายกลุ่มโดยใช้กราฟไม่มีวงมีทิศทางที่ปรับได้แบบจัดเรียงใหม่ (Multiclass support vector machines using reordering adaptive directed acyclic graphs) ปัญหาการพัฒนาซัพพอร์ตเวกเตอร์แมชชีนให้สามารถจัดกลุ่มข้อมูลได้หลายกลุ่มยังคงอยู่ในขั้นตอนการวิจัย วิธีดีดีเอจี (Decision Directed Acyclic Graph-DDAG) ให้ความถูกต้องเทียบได้กับวิธีแมกซ์วิน (Max Wins) ที่เป็นอัลกอริทึมที่ให้ค่าความถูกต้องสูงที่สุดในปัจจุบัน แต่ใช้เวลาในการสอนและประมวลผลต่ำกว่า วิธีเอดีเอจี (Adaptive Directed Acyclic Graph-ADAG) สามารถลดปัญหาที่เกิดจากโครงสร้างของดีดีเอจีได้ อย่างไรก็ตามลำดับของโหนดที่แตกต่างกันในวิธีเอดีเอจีอาจให้ความถูกต้องที่แตกต่างกัน งานวิจัยนี้ได้เสนอวิธีการใหม่สำหรับการจัดกลุ่มข้อมูลแบบหลายกลุ่ม เรียกว่าอาร์เอดีเอจี (Reordering Adaptive Directed Acyclic Graph-RADAG) ซึ่งเป็นการปรับปรุงวิธีเอดีเอจีเดิมและได้เสนออัลกอริทึมสำหรับการเลือกลำดับที่เหมาะสมของโหนดในวิธีเอดีเอจีเพื่อนำมาใช้

ในการจัดกลุ่มข้อมูล โดยพิจารณาจากค่าขอบเขตของความผิดพลาดของตัวจัดกลุ่มข้อมูลทั้งหมด และได้นำอัลกอริทึมการจับคู่สมบูรณ์แบบน้ำหนักน้อยสุด (Minimum-weight perfect matching) มาประยุกต์ใช้กับอัลกอริทึมที่ทำการจัดเรียงลำดับเพื่อเลือกตัวจัดกลุ่มที่มีค่าขอบเขตของความผิดพลาดต่ำมาใช้ในการจัดกลุ่มข้อมูล และเพื่อเลือกลำดับของโหนดที่เหมาะสมให้ได้ภายในเวลาพหุนาม (Polynomial time) งานวิจัยนี้ได้เปรียบเทียบประสิทธิภาพของวิธีการใหม่กับวิธีดีดีเอจี เอดีเอจี และแมกซ์วิน ผลการทดลองที่ได้แสดงให้เห็นว่าวิธีการใหม่ให้ความถูกต้องที่สูงกว่า และมีการประมวลผลเร็วกว่าวิธีแมกซ์วิน โดยเฉพาะอย่างยิ่งเมื่อมีจำนวนกลุ่ม (Class) และจำนวนมิติของข้อมูล (Dimension) สูง งานวิจัยนี้ได้เสนอแนวทางในการเพิ่มประสิทธิภาพของวิธีอาร์เอดีเอจี และวิธีดีดีเอจีด้วย

จิรา แก้วสุวรรณ์ (2006) ได้มีการนำเสนอการตรวจจับและการแก้ไขการวางตัวของภาพโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน (Image orientation detection and correction using support vector machine)ในปัจจุบันการปรับทิศทางภาพเพื่อให้ได้ภาพในทิศทางที่ถูกต้อง มักใช้การปรับทิศทางภาพด้วยมือ แม้ว่าการปรับทิศทางภาพจะสามารถทำได้ง่ายและใช้กระบวนการทำเพียงไม่กี่ขั้นตอน แต่ในกรณีที่มีภาพจำนวนมาก เวลาที่เสียไปและกำลังแรงงานที่ใช้ไปจะเพิ่มมากขึ้น ดังนั้นในงานวิจัยนี้ จึงได้นำเสนอเทคนิคการตรวจจับและแก้ไขทิศทางภาพให้ถูกต้อง โดยใช้วิธีการหาลักษณะเด่นของภาพจากค่าความ สำคัญของสี และการหาค่าฮิสโตแกรมของขอบวัตถุภาพจากภาพย่อย จากนั้นทำการวิเคราะห์และรวมลักษณะเด่นของภาพจากข้อมูลภาพในทิศทางที่แตกต่างกันในแต่ละภาพ เพื่อแบ่งประเภทของภาพจากลักษณะเด่นที่ได้ ทำให้รู้ถึงทิศทางการวางตัวของภาพที่ป้อนเข้ามา สามารถตรวจจับและแก้ไขทิศทาง การวางตัวของภาพได้ถูกต้องตามทิศทางที่เป็นจริงโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน ที่ได้ผ่านกระบวน การสอนให้ระบบเรียนรู้ เพื่อแยกแยะข้อมูลความแตกต่างของภาพ จากข้อมูลภาพที่ใช้ในการสอนให้ระบบเรียนรู้ จำนวน 350 ภาพ และจำนวนภาพที่ใช้ในการทดสอบจำนวน 350 ภาพ สามารถให้ค่าความถูกต้องมากกว่าร้อยละ 88.00 ในปีเดียวกันนี้ วีรพล จิรจริต (2006) ได้มีงานวิจัยการคัดแยกประเภทภาพก้อนหินปูนโดยใช้ลักษณะเด่นที่สัมพันธ์กันจากสองมุมมอง หลังจากการคัดกรองภาพเอกซเรย์เต้านมปกติ ด้วยการแปลงผลต่างความน่าจะเป็นเฉพาะ และการเรียนรู้โดยใช้เวกเตอร์สนับสนุน (Microcalcification Classification Using Two-view Corresponding Features after Normal Mammogram Screening by LPD Transform and SVM) ซึ่งงานวิจัยนี้นำเสนอวิธีการใหม่ในการคัดแยกประเภทภาพกลุ่มก้อนหินปูนที่ดีและร้าย ซึ่งเป็นสัญลักษณ์สำคัญของมะเร็งเต้านมในระยะแรก ที่ปรากฏอยู่บนภาพเอกซเรย์เต้านมแบบดิจิตอล วิธีการนี้ถูกแบ่งออกเป็นสองส่วน คือ การตรวจหาภาพเอกซเรย์เต้านมปกติ ซึ่งเป็นการคัดกรองขั้นแรก เพื่อเพิ่ม

ความจำเพาะในการวินิจฉัย และการคัดแยกประเภทภาพก้อนหินปูน ซึ่งเป็นความเห็นที่สอง เพื่อเพิ่มความไวต่อความผิดปกติ ในส่วนแรกนั้น การตรวจหาภาพเอกซเรย์เต้านมปกติจะมีปัญหาหลักสองประการ ประการแรกมาจากการเหลื่อมกันของค่าคุณลักษณะที่คำนวณจากภาพเอกซเรย์เต้านมที่ปกติและที่ผิดปกติ ขณะที่ประการที่สองมาจากการกระจายแบบไขว้กันของค่าคุณลักษณะแต่ละคู่ ซึ่งทำให้ยากต่อการคัดแยก ดังนั้นวิทยานิพนธ์นี้จึงเสนอวิธีซึ่งเป็นการเรียนรู้โดยใช้เวกเตอร์สนับสนุนร่วมกับการแปลงผลต่างความน่าจะเป็นเฉพาะ เพื่อแก้ปัญหาสองประการดังกล่าว โดยในวิธีการที่เสนอนี้ ค่าคุณลักษณะที่มีเหลื่อมกันจะถูกประมวลผลขั้นต้น ด้วยการสุ่มค่าฟังก์ชันความหนาแน่นความน่าจะเป็นของค่าคุณลักษณะของภาพเอกซเรย์เต้านมที่ปกติและที่ผิดปกติ แล้วทำการหาฟังก์ชันผลต่างความน่าจะเป็นเฉพาะ ค่าคุณลักษณะที่มีการกระจายแบบไขว้จะถูกแปลงไปเป็นค่าคุณลักษณะชุดใหม่ที่สามารถแยกการไขว้กันได้ด้วยระนาบศูนย์ของค่าใหม่ จากนั้นจึงทำการหาขอบเขตการคัดแยกที่ดีที่สุดเพื่อตรวจหาภาพเอซเรย์เต้านมที่ปกติ หลังจากที่คัดกรองภาพเอกซเรย์เต้านมที่ปกติออกไปแล้ว ภาพเอกซเรย์เต้านมที่ผิดปกติที่เหลือก็จะถูกส่งเข้าไปในส่วนที่สอง ซึ่งเป็นการคัดแยกประเภทภาพก้อนหินปูน โดยในส่วนที่สองจะมีปัญหาหลักประการเดียวคือการตัดสินใจผิดพลาดอันเนื่องมาจากความผิดเพี้ยนของรูปร่าง ขนาด จำนวน และการกระจายตัว ของภาพกลุ่มก้อนหินปูน ซึ่งเป็นผลมาจากการฉายภาพทางรังสี ของภาพจากมุมมองบนลงล่าง กับภาพจากมุมมองทแยงมุมดังนั้นวิทยานิพนธ์นี้จึงเสนอวิธีในการปรับแต่งค่าคุณลักษณะแบบเดิม เพื่อให้ได้ค่าคุณลักษณะใหม่ที่มีความสัมพันธ์กันระหว่างสองมุมมอง โดยจะทำการหาคุณลักษณะในสามมิติของภาพกลุ่มก้อนหินปูน ซึ่งอาศัยเงื่อนไขบังคับของการเข้าคู่แบบสเตริโอ จากการจัดวางท่าในการถ่ายภาพ แล้วทำการหาค่าคุณลักษณะใหม่ซึ่งเป็นค่าคุณลักษณะที่แทนความผิดปกติ จากคุณลักษณะในสามมิติบางส่วนในการทดลองจะใช้โครงข่ายประสาทเทียมที่มีการเรียนรู้แบบแพร่กระจายย้อนกลับ ป้อนเข้าด้านหน้าสามชั้น เป็นตัวคัดแยกประเภท ผลการทดลองในด้านความจำเพาะในการวินิจฉัยและความไวต่อความผิดปกติ มีประสิทธิภาพที่เพิ่มขึ้น

สุชาวดี ตันติศักดิ์ (2007) ได้มีการนำเสนอการตรวจจับฮาร์โมนิกในระบบจำหน่ายโดยใช้ซัพพอร์ตเวกเตอร์แมชีน (Harmonic detection in distribution systems using support vector machine) เมื่อการใช้อุปกรณ์อิเล็กทรอนิกส์ในระบบจำหน่ายมีความแพร่หลายเพิ่มมากขึ้น ปัญหาคุณภาพไฟฟ้าเกี่ยวกับการเกิดความผิดเพี้ยนของสัญญาณเนื่องจากฮาร์โมนิกจึงเป็นหนึ่งในสาเหตุสำคัญที่ทำให้คุณภาพไฟฟ้าลดลง วิทยานิพนธ์นี้จึงได้นำเสนอวิธีการสำหรับการตรวจจับฮาร์โมนิกในสัญญาณของระบบจำหน่ายด้วยซัพพอร์ตเวกเตอร์แมชีน โดยในกระบวนการทำงานจะใช้การแปลงสัญญาณเวฟเล็ตเป็นตัวสกัดจุดเด่นในสัญญาณของฮาร์มอนิกแต่ละลำดับ โดยใช้ค่าสัมประสิทธิ์เวฟเล็ตในแต่ละระดับของการจัดกลุ่มองค์ประกอบหลายระดับความละเอียด เมื่อได้

ข้อมูลของสัญญาณดังกล่าวแล้ว จะนำข้อมูลดังกล่าวไปเป็นข้อมูลในการสอนซัพพอร์ตเวกเตอร์แมชชีนเพื่อใช้ในการตรวจจับฮาร์โมนิกที่เกิด ขึ้นในระบบจำหน่าย ในการทดลองจะใช้สัญญาณที่จำลองขึ้นจากโปรแกรมและสัญญาณที่วัดได้จริงมาทดสอบกับอัลกอริทึมที่ทำการออกแบบไว้ โดยจะพบว่าวิธีการที่นำเสนอมานั้นสามารถใช้ในการตรวจจับฮาร์โมนิกในระบบจำหน่ายได้ดีโดยมีความถูกต้องในการตรวจจับฮาร์โมนิกในแบบจำลองที่ไม่มีสัญญาณรบกวน มีค่าร้อยละ 96.22 และแบบมีสัญญาณรบกวน ร้อยละ 95.15 นอกจากนี้ยังมีความถูกต้องสูงในการตรวจจับสัญญาณฮาร์โมนิกในสัญญาณจริงอีกด้วย ทำให้วิธีการที่นำเสนอมานั้นสามารถนำไปพัฒนาศักยภาพของงานด้านการวิเคราะห์คุณภาพต่อไปได้ในอนาคต

อริยะ นามวงศ์ (2008) ได้นำเสนอระบบการรู้จำใบหน้า 3 มิติด้วยวิธีเซลลูลาร์ออโตมาตาและซัพพอร์ตเวกเตอร์แมชชีน (3D face recognition system using cellular automata and support vector machines) ผลงานนี้ได้นำเสนอวิธีการรู้จำใบหน้า 3 มิติด้วยตัวแบบเซลลูลาร์ออโตมาตาและซัพพอร์ตเวกเตอร์แมชชีน โดยใช้ภาพแสดงระยะ 3 มิติซึ่งสร้างจากฐานข้อมูลภาพใบหน้า 3 มิติตัวแบบเซลลูลาร์ออโตมาตาใช้สำหรับหาตำแหน่งสำคัญบนใบหน้า 9 ตำแหน่ง เพื่อเป็นข้อมูลคุณลักษณะของแต่ละบุคคลซึ่งจะถูกนำเข้าสู่ขบวนการรู้จำโดยใช้ตัวแบบซัพพอร์ตเวกเตอร์แมชชีนต่อไป ภาพใบหน้าที่นำมาฝึกสอนประกอบด้วยภาพใบหน้าของ 100 คน ในท่าทางแตกต่างกัน 9 ท่าทาง ได้แก่ ภาพท่าทางหน้าไปทางซ้ายและทางขวา 5 องศา 10 องศา 15 องศา และ 20 องศา ตามลำดับ ในการวัดประสิทธิภาพใช้ข้อมูลทดสอบ 5 ชุด แต่ละชุดประกอบด้วยภาพใบหน้า 100 ภาพของ 100 คน ในท่าทางแตกต่างกัน จากการทดลองพบว่าขั้นตอนวิธีที่นำเสนอให้ผลลัพธ์ในการระบุตัวบุคคลได้ถูกต้องเฉลี่ยสูงถึง 97.00 เปอร์เซ็นต์

2.4 ทฤษฎีที่เกี่ยวข้องเว็บสไปเดอร์ (Web spider)

เว็บสไปเดอร์ (Web spider) หรือเรียกว่า ครอว์เลอร์ (Crawler), โรบอต (Robot) เป็นซอฟต์แวร์ชนิดหนึ่งที่ใช้สำหรับวิ่งไปบนอินเตอร์เน็ต โดย ใช้ซอฟต์แวร์อัตโนมัติที่เรียกว่า " Bots" หรือ " Spiders" สำหรับวิ่งไปยังเว็บไซต์ต่างๆ ซึ่งทำหน้าที่ในการเก็บรวบรวมเอกสารอิเล็กทรอนิกส์บนเว็บทั้งหมด เช่น แฟ้มเอกสารประเภท HTML, PHP, PDF, DOC และอื่นๆ บนเว็บ ซึ่งจะเรียกสั้นๆ ว่า เอกสารบนเว็บ หรือเอกสาร (Web Documents) เพื่อเก็บรวมรวมข้อมูลและส่งข้อมูลที่รวมรวมได้กลับมายังฐานข้อมูล ซึ่งมีลักษณะในการทำงานดัง รูป 2.2 เพื่อทำการประมวลผลตามโปรแกรมการจัดทำดัชนีของแต่ละกลไกการสืบค้น กลไกการสืบค้นข้อมูลหนึ่งๆ มักมีตัว ครอว์เลอร์ หลายตัวเพื่อความรวดเร็วในการสำรวจและเก็บข้อมูล ซึ่งมีผลต่อความทันสมัยของข้อมูล เนื่องจากข้อมูลบนอินเทอร์เน็ตมีลักษณะเฉพาะตัวที่สำคัญอย่างนึ่งคือ มีการ

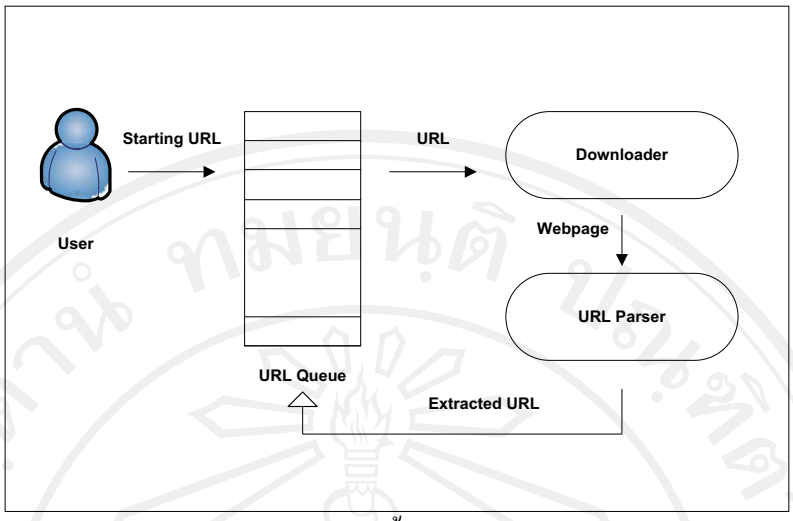
เปลี่ยนแปลงสูงและเกือบตลอดเวลา นอกจากนี้ยังมีข้อมูลใหม่เพิ่มขึ้นอย่างรวดเร็ว วิธีการและระยะเวลาในการออกสำรวจข้อมูลของเว็บ สไปเดอร์แต่ละกลไกจะแตกต่างกันขึ้นอยู่กับนโยบายหรือความสามารถของตัว ครอว์เลอร์ของแต่ละกลไก การค้นหาข้อมูลที่ เสิร์ชเอนจิน (Search Engine) แต่ละตัวใช้ในการค้นหาข้อมูลจากเว็บไซต์ต่างๆ เพื่อมาทำเป็นระบบฐานข้อมูลมี 3 วิธี คือ

1. การค้นแบบคำสำคัญ (Keyword Searching) เป็นการค้นหาข้อมูลแบบที่ง่านที่สุดจากเว็บไซต์ในระบบอินเทอร์เน็ต ซึ่ง เสิร์ชเอนจิน ส่วนมากจะใช้วิธีนี้ในการเข้าไปตรวจสอบเว็บไซต์หรือคำต่างๆ ที่อยู่ภายในโฮมเพจ หรือเว็บไซต์ที่อยู่ในตอนต้นๆ ของแฟ้มข้อมูล หรือคำใดๆ ก็ตามที่ปรากฏอยู่และมีซ้ำกันมากก็จะถือว่าคำๆ นั้นเป็นคำสำคัญสำหรับโฮมเพจนั้น

2. การค้นแบบแนวความคิด (Concept-Based Searching) ในการค้นหาแบบแนวความคิดนี้จะแตกต่างจากการค้นหาแบบคำสำคัญ เพราะไม่ใช่การค้นหาคำที่ผู้ใช้ต้องการ แต่จะพยายามค้นหาว่าสิ่งที่ผู้ใช้ต้องการคืออะไร โดยจะค้นหาข้อมูลที่เกี่ยวกับข้อความหรือหัวเรื่องที่ผู้ใช้ต้องการค้นหา ถึงแม้ว่าคำที่อยู่ในโฮมเพจอาจจะไม่ตรงหรือไม่เหมือนกับคำที่ต้องการค้นหาแต่ละค้นหาในความหมายที่คล้ายๆ กัน วิธีการค้นหาข้อมูลแบบนี้เป็นวิธีที่ยังใช้ไม่ค่อยได้ผลเท่าที่ควร เนื่องจากในความเป็นจริงแล้วเครื่องคอมพิวเตอร์ยังต้องอาศัยคนในการทำงาน ถ้าคนไม่สั่งคอมพิวเตอร์ก็จะไม่ทำงาน และที่สำคัญคอมพิวเตอร์ไม่สามารถที่จะเข้าใจความหมายของข้อความได้ดีเหมือนกับมนุษย์ ทำให้ความสมบูรณ์ในการค้นหาแบบนี้มีแต่ในทฤษฎีเท่านั้น

3. การค้นแบบพิเศษ (Refining your Search) หมายถึง การเลือกขั้นตอนการค้นหาที่เหมาะสมในเว็บไซต์ของ Search Engine มักจะมีวิธีการค้นหาข้อมูลที่ต้องการ 2 วิธีคือ วิธี Basic และ Refined ในการค้นหาข้อมูลแบบ Basic นั้น ผู้ใช้เพียงแต่กำหนดหรือเลือกวิธีใช้คำสำคัญธรรมดา โดยไม่มีการใช้วิธีค้นหาข้อมูลแบบพิเศษ ในส่วนของการค้นหาแบบ Refined นั้น ผู้ใช้งานสามารถเลือกใช้ความสามารถพิเศษของเสิร์ชเอนจินนั้นช่วยในการค้นหาข้อมูลได้

ครอว์เลอร์เป็นโปรแกรมที่พัฒนาขึ้นเพื่อใช้ประโยชน์ในการรวบรวมเว็บเพจจากอินเทอร์เน็ต กลุ่มวิจัยมากมายหลากหลายกลุ่มต่างพยายามออกแบบและสร้างครอว์เลอร์ให้สามารถทำงานได้อย่างมีประสิทธิภาพสูงสุด อย่างไรก็ตามการออกแบบและลักษณะการทำงานยังคงยึดอยู่บนพื้นฐานเดียวกัน ปรับปรุงเปลี่ยนแปลงมาจากแบบจำลองพื้นฐานแบบเดียวกัน ในหัวข้อนี้กล่าวถึงแบบจำลองพื้นฐานในการออกแบบและพัฒนาครอว์เลอร์ และชี้แนะในส่วนของการเพิ่มประสิทธิภาพเบื้องต้น

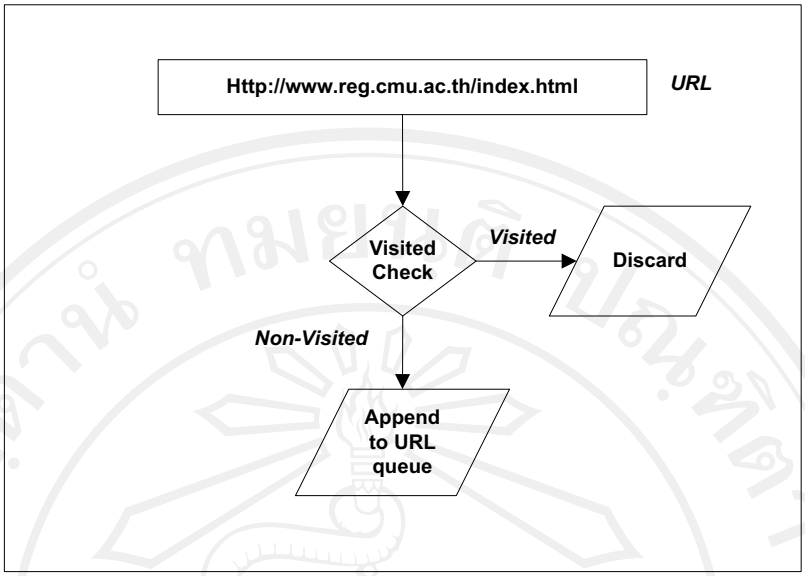


รูป 2.2 แบบจำลองพื้นฐานของครอว์เลอร์

ในรูป 2.2 ครอว์เลอร์เป็นโปรแกรมที่สามารถรวบรวมเว็บเพจได้โดยอัตโนมัติ โดยเริ่มเก็บเว็บเพจจากยูอาร์แอลเริ่มต้น จากนั้นยูอาร์แอลใหม่ที่อยู่ในเว็บเพจนั้นจะถูกแยกออกมาและนำไปตรวจสอบว่าก่อนหน้านี้ได้เก็บเว็บเพจจากยูอาร์แอลนี้มาก่อนหรือไม่ หากตรวจสอบพบว่าเคยเก็บมาก่อนจะไม่เก็บซ้ำซ้อนอีก แต่หากพบว่าไม่เคยเก็บมาก่อนครอว์เลอร์จะต้องเก็บเว็บเพจจากยูอาร์แอลนั้น ขั้นตอนเหล่านี้จะถูกทำงานวนรอบเสมอจนกว่าจะถึงจุดสิ้นสุดการทำงานซึ่งมี 2 กรณี คือ เก็บเว็บเพจได้ครบตามจำนวนที่กำหนด ไม่พบยูอาร์แอลที่สามารถเก็บต่อไปได้

แบบจำลองพื้นฐานของครอว์เลอร์ประกอบด้วย 3 ส่วนหลัก ได้แก่ ยูอาร์แอลคิว (URL queue) ตัวดาวน์โหลด (Downloader) และตัวพาร์สเซอร์ยูอาร์แอล (URL Parser) โดยแต่ละส่วนประกอบจะมีลักษณะดังนี้

1. ยูอาร์แอลคิว ทำหน้าที่เก็บยูอาร์แอลที่ครอว์เลอร์พบในเว็บเพจ แต่ยังไม่ได้ไปเก็บรวบรวมมามีลักษณะเป็นยูอาร์แอลคิวแบบเข้าก่อนออกก่อน ยูอาร์แอลคิวนี้จะมีกลไกในการตรวจสอบว่า ยูอาร์แอลที่เข้ามาเป็นยูอาร์แอลที่ครอว์เลอร์เคยเก็บมาก่อนหรือไม่ หากเคยเก็บมาแล้วยูอาร์แอลนั้นจะไม่ถูกนำมาใส่ในยูอาร์แอลคิว ทั้งนี้เพื่อป้องกันความซ้ำซ้อนในการรวบรวมเว็บเพจ โดยกลไกของยูอาร์แอลคิวนี้แสดงดังรูป 2.3



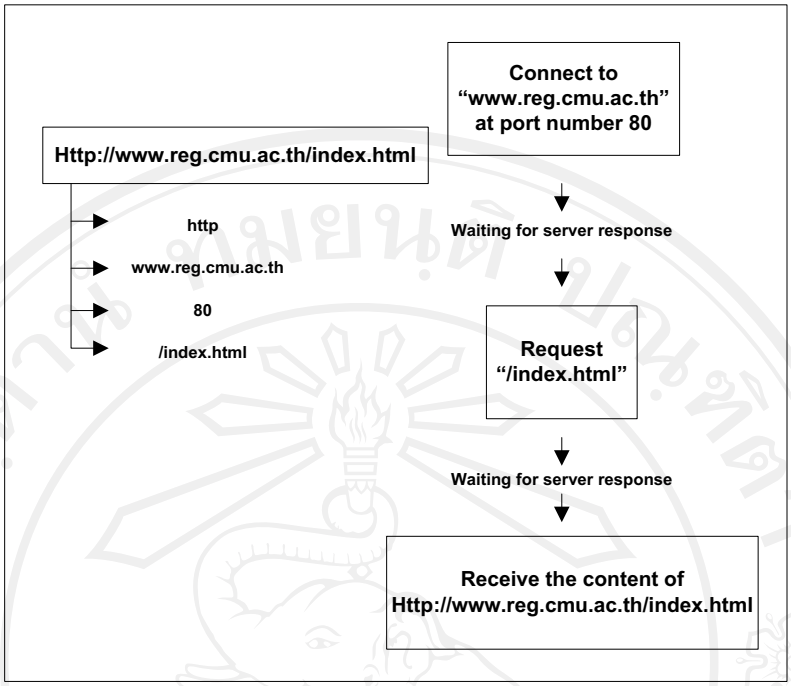
รูป 2.3 กลไกการตรวจสอบยูอาร์แอลก่อนนำไปใส่ในยูอาร์แอลคิว

2. ตัวดาวน์โหลด ทำหน้าที่เก็บเว็บเพจจากอินเทอร์เน็ต โดยตัวดาวน์โหลดจะดึงยูอาร์แอลออกจากยูอาร์แอลคิว ยูอาร์แอลจะถูกตัวดาวน์โหลดแยกออกเป็น 4 ส่วนประกอบ ได้แก่ โปรโตคอล (Protocol) เซิร์ฟเวอร์ (Server) พอร์ต (Port) และพาธ (Path) ตัวอย่างในการแยกส่วนของยูอาร์แอลดังตารางที่ 2.1

การแยกส่วนประกอบของยูอาร์แอลจากตารางที่ 2.1 มีประโยชน์ต่อตัวดาวน์โหลดมาก โดยโปรโตคอลช่วยให้แยกและเลือกประเภทของข้อมูลได้ เช่น โปรโตคอล http จะเป็นข้อมูลเว็บเพจ เป็นต้น ส่วนของเซิร์ฟเวอร์และพอร์ตทำให้ครอว์เลอร์ทราบว่าเว็บเพจนี้ต้องร้องขอจากเว็บเซิร์ฟเวอร์ใดและที่พอร์ตบริการหมายเลขใด และส่วนของพาธจะเป็นการอ้างอิงถึงเว็บเพจที่ต้องการ ตัวอย่างกลไกการร้องขอเว็บเพจของตัวดาวน์โหลดแสดงดังรูป 2.4

ตารางที่ 2.1 ตัวอย่างการแยกส่วนประกอบของยูอาร์แอล

ยูอาร์แอล	โปรโตคอล	เซิร์ฟเวอร์	พอร์ต	พาธ
http://www.reg.cmu.ac.th/index.html	http	www.reg.cmu.ac.th	80	/index.html
http://www.reg.cmu.ac.th:8080/	http	www.reg.cmu.ac.th	8080	/
ftp://ftp.sei.cmu.edu/pub/wwwadm/ftp.tar	ftp	ftp.sei.cmu.edu	21	/pub/wwwadm/ftp.tar



รูป 2.4 กลไกการร้องขอเว็บเพจจากเซิร์ฟเวอร์

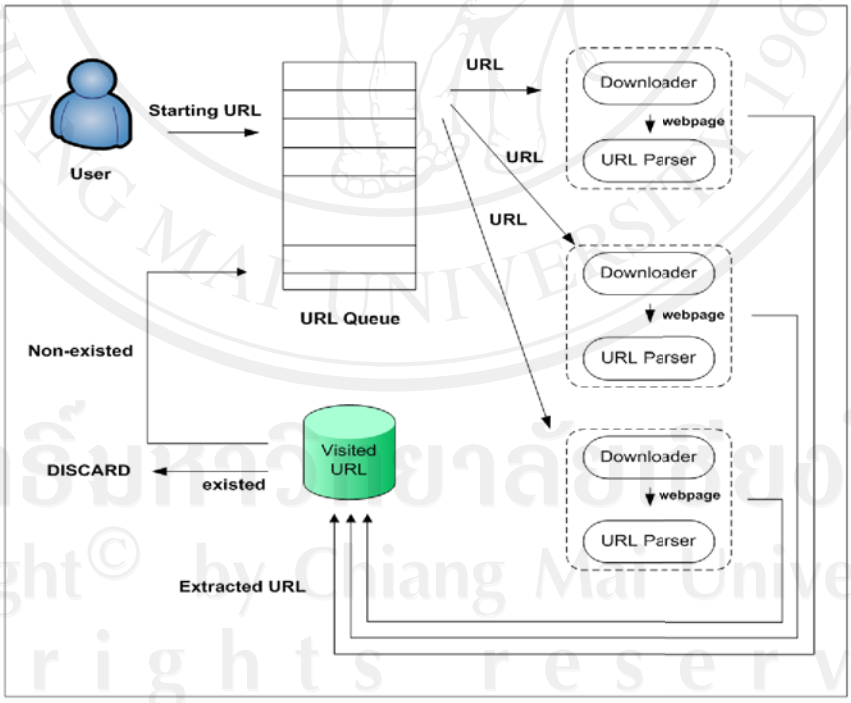
3. ตัวพาร์สเซอร์ยูอาร์แอล ทำหน้าที่แยกส่วนของยูอาร์แอลที่พบในเว็บเพจออกมาโดย ยูอาร์แอลที่ปรากฏในเว็บเพจจะถูกกำกับด้วยแท็กของภาษาเอชทีเอ็มแอล (HTML) ดังตารางที่ 2.2 ตัวพาร์สเซอร์ยูอาร์แอลจะอ่านเว็บเพจและแยกส่วนของยูอาร์แอลที่ถูกแท็กในตารางที่ 2.2 กำกับอยู่ ออกมาเพื่อนำไปใส่ในยูอาร์แอลคิวต่อไป

ตารางที่ 2.2 แท็กในภาษาเอชทีเอ็มแอลที่กำกับยูอาร์แอล

แท็ก	ตัวอย่างการใช้แท็ก
A	<A href="http://www.reg.cmu.ac.th:>นักศึกษาลงทะเบียน</A>
IMG	<IMG src="images/president_40.jpg">
AREA	<AREA shape="rect" coords="8,28,234,102" href="http://autonomous.cmu.ac.th/">
FRAME	<FRAME SRC="http://www.reg.cmu.ac.th/index.html">
META	<META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=tis-620">

จากรูป 2.2 ผู้ใช้กำหนดยูอาร์แอลเริ่มต้นให้กับครอว์เลอร์ โดยจะนำยูอาร์แอลเริ่มต้นไปใส่ไว้ในยูอาร์แอลคิวซึ่งมีลักษณะเป็นคิวแบบเข้าก่อน -ออกก่อน (FIFO: First-in First-out) จากนั้นโปรแกรมดาวน์โหลดจะดึงยูอาร์แอลออกจายูอาร์แอลคิวเพื่อเก็บเว็บเพจจากอินเทอร์เน็ต หลังจากได้เก็บเสร็จเรียบร้อยแล้ว เว็บเพจจะถูกส่งต่อไปยังตัวพาร์สเซอร์ยูอาร์แอลเพื่อตัดแยกนำส่วนของยูอาร์แอลใหม่ที่อยู่ในเว็บเพจนั้นออกมา ยูอาร์แอลใหม่ที่ได้มาจะส่งต่อให้กับยูอาร์แอลคิว ซึ่งในยูอาร์แอลคิวจะมีกลไกในการตรวจสอบยูอาร์แอลที่เคยเก็บมาแล้วอยู่ภายในเรียบร้อยแล้ว

อย่างไรก็ตามแบบจำลองพื้นฐานใน รูป 2.2 เป็นเพียงการนำเสนอการทำงานของครอว์เลอร์เบื้องต้นเรียกว่า การทำงานแบบเดี่ยว (Sequential processing) อย่างไรก็ตามในระยะเวลา 4-5 ปีที่ผ่านมา งานวิจัยเกี่ยวกับการพัฒนาครอว์เลอร์ให้มีประสิทธิภาพสูงนั้นต่างพากันนำเสนอการออกแบบ ครอว์เลอร์ให้ทำงานแบบขนาน (Parallel processing) เนื่องจากมีเว็บเพจจำนวนมหาศาลอยู่ในอินเทอร์เน็ต จากรายงานของเสิร์จเอ็นจินชื่อ Inktomi (Inktomi, 2000) พบว่า ในปี 2000 จำนวนเว็บเพจในอินเทอร์เน็ตได้เกินหลักพันล้านไปเรียบร้อยแล้วและยังมีแนวโน้มเพิ่มสูงขึ้นเป็นจำนวนพันล้านต่อปี ด้วยเหตุนี้การใช้ครอว์เลอร์แบบขนานเพื่อให้สามารถการเก็บรวบรวมเว็บเพจ จากอินเทอร์เน็ตใช้ระยะเวลาสั้นลง



รูป 2.5 แบบจำลองครอว์เลอร์แบบขนาน

การทำงานแบบขนานของครอว์เลอร์ช่วยให้ประสิทธิภาพในการเก็บเว็บเพจสูงขึ้น เพื่อให้ง่ายต่อการเข้าใจในการทำงานแบบขนาน รูป 2.5 ได้ปรับปรุงมาจาก รูป 2.2 เพื่อแสดงให้เห็นถึง

ความแตกต่างของการทำงานแบบเดี่ยวและการทำงานแบบขนาน จากแบบจำลองใน รูป 2.5 ส่วนการทำงานของตัวดาวน์โหลดและตัวพาร์สเซอร์จะถูกกำหนดให้มีมากกว่าหนึ่งหน่วย ดังนั้นตัวดาวน์โหลดและตัวพาร์สเซอร์จะทำงานขนานกันไปอย่างเป็นอิสระต่อกัน แต่ยังคงติดต่อกับยูอาร์แอลคิวหน่วยเดียวกันอยู่ เพื่อลดปัญหาความซ้ำซ้อนของการทำงาน ยูอาร์แอลคิวตัวนี้จึงเปรียบเสมือนตัวแจกจ่ายงานและควบคุมทิศทางของงาน

2.5 การเปลี่ยนทิศทางใหม่ของยูอาร์แอล (URL Redirect)

ปกติแล้วในผู้ให้บริการเว็บ (Web Server) จะมีไฟล์ที่เป็นไฟล์แรกสำหรับการแสดงเพจหน้าแรกซึ่งจะวางอยู่ในตำแหน่ง (Document Root) แต่ในบางครั้งเมื่อผู้ใช้มีการเรียกเข้ามาที่ยูอาร์แอลหรือไฟล์ดังกล่าวบนผู้ให้บริการเว็บอาจจะมีความจำเป็นที่ยูอาร์แอลดังกล่าวยังไม่พร้อมที่จะให้บริการ เราก็สามารถจะเขียนโค้ด (Code) ในไฟล์ดังกล่าวให้มีการเปลี่ยนเส้นทางไปเรียกไฟล์อื่นซึ่งอาจจะอยู่ในอีกไดเร็คทอรี่ (Directory) หรืออีกโฟลเดอร์บนผู้ให้บริการเว็บตัวเดียวกัน หรืออาจจะเปลี่ยนเส้นทางไปเป็นผู้ให้บริการเว็บอื่น การเขียนโค้ดของ URL Redirect ดังรูป 2.6 การเปลี่ยนทิศทางใหม่ของยูอาร์แอลแบบง่ายสามารถทำได้ด้วยการใช้ meta tag ของ html

```
<html>
<head>
<META HTTP-EQUIV="Refresh"
CONTENT="0;URL=http://www.yourname.com">
</head>
</html>
```

รูป 2.6 การเปลี่ยนทิศทางใหม่ของยูอาร์แอลโดยใช้ meta tag ของ HTML

จากรูป 2.6 ซึ่งถ้านำโค้ดไปเป็นไฟล์หลักในตำแหน่งไฟล์แรกสำหรับการแสดงเพจหน้าแรกบนผู้ให้บริการเว็บก็จะทำให้เว็บเพจถูกเปลี่ยนทิศทางใหม่ไปยัง http://www.yourname.com โดยทันที เพราะค่าของ CONTENT=0 แต่ถ้าต้องการหน่วงเวลาให้ผู้ใช้ได้อ่านข้อความบางอย่างก่อนการ เปลี่ยนทิศทางใหม่ก็สามารถทำได้ด้วยการกำหนดค่า CONTENT ไม่เป็น 0 ดังรูป 2.7

```
<html>
<head>
<META HTTP-EQUIV="Refresh"
CONTENT="5;URL=http://www.yourname.com">
</head>
<body>
<center>
เว็บไซต์ของเรามีการเปลี่ยนเป็นชื่อใหม่แล้ว การเข้ามาเว็บไซต์ของเราครั้งต่อไป กรุณาใช้ชื่อ
<a href="http://www.yourname.com">www.yourname.com</a>
</center>
</body>
</html>
```

รูป 2.7 การเปลี่ยนทิศทางใหม่ของยูอาร์แอลโดยการหน่วงเวลา

2.6 คำนวณน้ำหนักของคำหลักของเอกสารบนเว็บ

การกำหนดค่าน้ำหนักของแต่ละคำหลักในเอกสารหนึ่งๆ เป็นการกำหนดน้ำหนักเบื้อง ต้นให้แต่ละคำในแต่ละเอกสาร โดยพิจารณาจากแหล่ง (ตำแหน่ง) ของคำในเอกสาร ซึ่งเอกสารบนเว็บจะมีระดับความสำคัญของคำในแต่ละแหล่งแตกต่างกันดังตารางที่ 3.1 เพื่อประกอบการคำนวณดังสมการที่ (2.1) (สมจิตร อาจอินทร์, 2006)

$$pw_{i,j} = s_{i,k} * kf_{i,k} \qquad (2.1)$$

เมื่อ

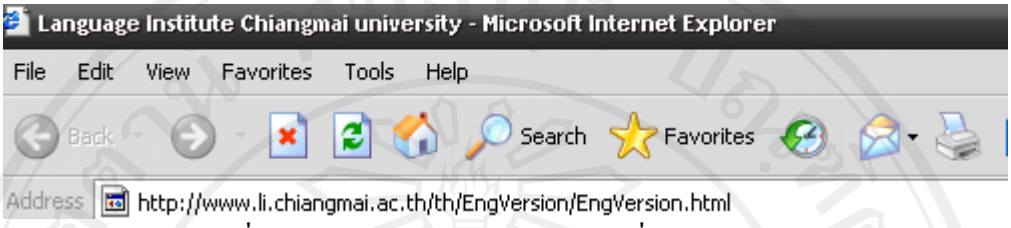
$pw_{i,j}$	คือ ค่าน้ำหนักคำหลักที่ i ในเอกสาร j
$s_{i,k}$	คือ ระดับความสำคัญ k ของคำหลักที่ i
$kf_{i,k}$	คือ ความถี่ของคำหลักที่ i ที่มีระดับความสำคัญ k

ในแต่ละเอกสาร ประกอบด้วยข้อมูลที่ถูกจัดไว้ในตำแหน่งและรูปแบบที่หลากหลาย โดยแบ่งออกเป็นกลุ่มและกำหนดความสำคัญของคำที่อยู่ในกลุ่มดังตารางที่ 2.3

ตารางที่ 2.3 การจัดระดับความสำคัญของข้อมูลในเอกสารบนเว็บ

ตำแหน่งของคำหลัก	ระดับความสำคัญ
- Title	100
- META	50
- Anchor, **Bold**, *Italic*, Underline, H1,H2,H3	20
- Body	10

รูป 2.8 เป็นตัวอย่างหัวข้อของเอกสาร (Web Document Title) เป็นตัวอย่างเว็บ ไซต์ของ http://www.li.chiangmai.ac.th/th/EngVersion/EngVersion.html มี Tag Title คือคำว่า “Language Institute Chiangmai university” ซึ่งค่าระดับความสำคัญของคำหลักที่มาจาก tag นี้เท่ากับ 100 เป็นค่าที่ใช้ในการถ่วงน้ำหนักกับความถี่ของคำหลัก (ตามสมการ 2.1)



รูป 2.8 ตัวอย่างชื่อหัวข้อของเอกสาร (Tag Title) ที่อ่านมาจากเอกสาร HTML

รูป 2.9 คือ ส่วนของโปรแกรมในส่วนของ Tag Meta ที่อยู่ภายในเอกสาร HTML ซึ่งค่าน้ำหนักของคำหลักคือคำว่า “การรับตรง มช” จะให้ระดับความสำคัญที่ระดับ 50

```
<META HTTP-EQUIV="Pragma" CONTENT="no-cache">
<META HTTP-EQUIV="Expires" CONTENT="-1">
```

รูป 2.9 แสดง Tag Meta ที่คัดลอกมาจากเอกสาร HTML

รูป 2.10 และรูป 2.11 คือ ส่วนของโปรแกรมในส่วนของ Tag Anchor และ Tag Body ที่อยู่ภายในเอกสาร HTML จะกำหนดระดับความสำคัญไว้ที่ 20 และ 10 ตามลำดับ

```
<a class="toolLnk" href="AboutUs.html"><font face="MS Sans Serif" size="2"
color="#ffffff">
```

รูป 2.10 แสดง Tag Anchor ที่คัดลอกมาจากเอกสาร HTML

```
<BODY>
<TR><TD align=middle width="100%" bgColor=#ECF3FF height="55">
<p><font size="2" face="MS Sans Serif, Microsoft Sans Serif" color="#003300">
                On September 3, 2004, Chiang Mai University initiated the
                regulations pertaining to the operation and objectives of
                the Language Institute, CMU.<br>
                </font> </p>
</TD></TR>
</BODY>
```

รูป 2.11 แสดง Tag Body ที่คัดลอกมาจากเอกสาร HTML

2.7 ทฤษฎีที่เกี่ยวข้องการจัดกลุ่มเอกสารข้อความ (Text Classification)

การจัดกลุ่มเอกสารข้อความแบบอัตโนมัติ (Automatic Text Classification)(Polpinij and others, 2005) กลายเป็นสิ่งที่มีความสำคัญและจำเป็น เมื่อจำนวนเอกสารได้มีการเพิ่มขึ้นอย่างรวดเร็ว จนยากเกินกว่าที่จะทำการจัดกลุ่มเอกสารด้วยคน โดยเฉพาะอย่างยิ่งเมื่อเอกสารที่อยู่บนระบบอินเทอร์เน็ต การประยุกต์เทคนิคต่างๆ เพื่อสร้างระบบการจัดกลุ่มเอกสารส่วนใหญ่มักจะมีพื้นฐานอยู่บนอัลกอริทึมกลไกเชิงเรียนรู้ (Machine Learning Algorithms) เช่น นาอีฟเบส์ (Naïve Bayes) ซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machines) โครงข่ายประสาทเทียม (Artificial Neural Network) ต้นไม้ตัดสินใจ (Decision Tree) เป็นต้น

แนวคิดพื้นฐานในการจัดกลุ่มเอกสารคือ การสร้างโมเดลที่เป็นระเบียบแบบแผนในการประมาณผลลัพธ์ที่ไม่อาจจะทราบได้ โดยมีฟังก์ชันของการประมาณคือ $\Phi : D \times C \rightarrow \{T, F\}$ ซึ่งก็คือ ฟังก์ชันที่ใช้ในการประมาณค่าของการจัดกลุ่มเอกสารข้อความแบบอัตโนมัติ โดย C คือกลุ่มของเอกสารที่จะเป็นไปได้ นั่นคือ $C = \{c_1, c_2, ..., c_{|c|}\}$ และ D คือเซตของเอกสารที่จะนำมาสร้างโมเดลของการจัดกลุ่มเอกสาร โดยผลลัพธ์เบื้องต้นของการจัดกลุ่มเอกสารแบบอัตโนมัติสามารถกำหนดได้ 2 กลุ่มคือ กลุ่มที่ตรงตามความต้องการหรือกลุ่มที่สนใจเรียกว่า *Positive* ซึ่งแทนด้วย *True*(T) นั่นคือ $\Phi : (d_i, c_j) = T$ และ กลุ่มที่ไม่ตรงตามความต้องการหรือกลุ่มที่ไม่สนใจจะเรียกว่า *Negative* ซึ่งแทนด้วย คือ $\Phi : (d_i, c_j) = F$

2.8 ทฤษฎีที่เกี่ยวข้องการตัดคำและการสกัดคำหลัก (Word Segmentation)

เป็นขั้นตอนการตัดข้อความ (Text) ออกเป็นกลุ่มของคำ ซึ่งเป็นคำสำคัญ-เอกสาร (Keyword) ที่มีความหมาย ในที่นี้คือ คำนาม (Noun) หรือวลี (Phrase) เป้าหมายของการตัดคำคือ การได้มาซึ่งคำหรือวลีเพื่อใช้เป็นตัวแทนของเอกสาร นั่นคือ ได้กลุ่มคำหลักหรือวลีที่สามารถนำไปใช้สืบค้นเชิงความหมายได้

Text Processing คือการประมวลผลข้อความบนพื้นฐานของระบบการค้นคืนสารสนเทศ (Information Retrieval) ซึ่งเป็นขั้นตอนการเตรียมข้อมูลเบื้องต้นสำหรับในกระบวนการนำเอกสารที่รวบรวมได้เข้าสู่กระบวนการเตรียมข้อความก่อนจัดทำดัชนี เช่น การแยกคำ การตัดคำ การเลือกกลุ่มคำที่เป็น คำสำคัญ-เอกสาร การตัดคำที่เป็นภาษาอังกฤษจะถูกสกัดออกเป็นข้อความ (Text) และสิ่งที่เป็นพื้นฐานที่จำเป็นอย่างยิ่งคือ “หน่วยคำ” ดังนั้นการหาขอบเขตของแต่ละคำจึงเป็นสิ่งแรกที่ต้องคำนึงถึง เพราะหากเลือกการหาขอบเขตคำไม่เหมาะสม อาจนำมาสู่ ระบบการประมวลผลข้อความที่ไม่ถูกต้อง มีกระบวนการแบ่งส่วนการประมวลผลข้อความนี้ เป็น 3 ส่วน ดังนี้

1) การตัดคำที่ไม่จำเป็นออก

การกำจัดคำที่อาจทำให้เกิดข้อผิดพลาดในการจัดกลุ่มเอกสาร หรือทำให้ประสิทธิภาพโดยรวมของการจัดกลุ่มลดลง เพื่อหลีกเลี่ยงปัญหาเหล่านี้ จึงจำเป็นต้องมีขั้นตอนในการตัดหรือคัดเลือกคำที่ไม่จำเป็นออกไปจากระบบ ซึ่งประเภทของคำดังกล่าวคือคำที่ไม่ใช่คำนาม เช่น สันธาน คำสรรพนาม คำบุพบท คำคุณศัพท์ เป็นต้น รวมถึงคำที่เป็นชื่อเฉพาะ เช่น คำแสดงจำนวน one two three ชื่อเดือนและชื่อวัน เช่น June, July, Monday, Tuesday เป็นต้น รวมทั้งอักขระพิเศษต่าง ๆ เช่น ตัวเลข เครื่องหมายและสัญลักษณ์ (1, 2, 3,…, +, -, *, /, ….)

2) การแปลงคำศัพท์ให้อยู่ในรูปของรากศัพท์ (Stemming words)

การแปลงคำต่าง ๆ ให้กลับไปเป็นรากศัพท์ของคำ ๆ นั้น เพื่อลดจำนวนของคำที่ใช้ในการสร้างตัวจำแนกให้น้อยลง ซึ่งจะทำให้ประสิทธิภาพในการจัดกลุ่มเอกสารดีขึ้น เช่น คำว่า “measure”, “measured”, “measurer” และคำว่า “measurement” มีความหมายเกี่ยวกับ “การวัด” ทั้งสิ้น คำทั้งหมดนี้ในงานวิจัยเกี่ยวกับการจัดกลุ่มเอกสารบางงานวิจัยถือว่าเป็นคำเดียวกันโดยใช้เป็นคำว่า “measure” แทนคำทั้งหมด โดยขั้นตอนวิธีที่ใช้ในการแปลงคำศัพท์ให้อยู่ในรูปของรากศัพท์ที่นิยมใช้ในงานวิจัยส่วนใหญ่จะใช้ขั้นตอนวิธีการแปลงรากศัพท์ของพอร์เตอร์ (Porter’s Stemming Algorithm) (Porter, 1980)

3) การทำดัชนีคำ (Document Indexing)

การจัดคู่ลำดับของคำศัพท์ที่ใช้เป็นลักษณะเด่น (Features) และน้ำหนักของลักษณะเด่นนั้น ในการสร้างเวกเตอร์ของตัวจัดกลุ่มเอกสารแต่ละประเภท โดยการจัดคู่ลำดับจะเรียงตามตัวอักษรของคำที่เป็นลักษณะเด่น เพื่อความสะดวกรวดเร็วในการค้นหาและนำไปใช้ในการจัดกลุ่มเอกสาร น้ำหนักของคำที่เป็นลักษณะเด่นนั้นจะแตกต่างกันไปขึ้นกับจำนวนของคำในเอกสารแต่ละประเภทและวิธีการกำหนดน้ำหนักของงานวิจัยต่าง ๆ ด้วย

2.9 ทฤษฎีที่เกี่ยวข้องการสร้างตัวแทนเวกเตอร์หรือการหาน้ำหนักของคำ (Term word weighting)

สิ่งสำคัญในการค้นหาสารสนเทศส่วนใหญ่ที่ใช้ในระบบการค้นคืนสารสนเทศนั้นจะใช้วิธีสร้างตัวแทนเวกเตอร์ เพื่อเข้าไปค้นหาข้อมูลในเอกสารนั้นก็คือเมื่อมีการป้อนเอกสารเข้าไปในระบบ ระบบก็จะทำการวิเคราะห์เอกสารและหาคำ (Unique terms) ทั้งหมดที่อยู่ในเอกสาร จากนั้นจะทำการหาความถี่ของคำ (Term Frequency : TF) และส่วนกลับของเอกสาร (Inverse Document Frequency : IDF) ซึ่งส่วนกลับของเอกสารได้จาก $IDF = \log(\left|\frac{N}{DF}\right|)$ โดยที่ N คือจำนวนของเอกสารทั้งหมดในกลุ่มและ DF คือจำนวนเอกสารที่มีคำนั้นปรากฏอยู่ เพื่อนำไปสร้างตัวเวกเตอร์เพื่อที่จะนำไปค้นคืนเอกสาร ด้วยการให้น้ำหนักคำภายใต้สมการ

$$TF\text{-}IDF = TF \times IDF \tag{2.2}$$

อย่างไรก็ตามหากค่า N และค่า DF มีค่าเท่ากัน ค่า log 1 จะมีค่าเป็น 0 และเมื่อนำค่านี้ไปคูณกับค่า TF ผลลัพธ์ที่ได้ก็จะมีค่าเป็น 0 ด้วย ดังนั้นจึงได้มีการปรับค่า IDF ใหม่ โดยสามารถใช้เป็น $IDF = 1 + \log(\left|\frac{N}{DF}\right|)$ หรือ $IDF = 1 - \log(\left|\frac{N}{DF}\right|)$ (2.3)

จากสมการดังกล่าวเป็นวิธีการหาตัวแทนเวกเตอร์เพื่อนำไปค้นคืนสารสนเทศที่เป็นกลุ่มของเอกสาร โดยเมื่อมีการป้อนเอกสารเข้าไปในระบบ ระบบก็จะทำการวิเคราะห์เอกสารและหาคำทั้งหมดที่มีอยู่ในเอกสาร

2.10 ทฤษฎีที่เกี่ยวข้องลกอริทึมซัพพอร์ตเวคเตอร์แมชชีน (Support Vector machines)

ซัพพอร์ตเวกเตอร์แมชชีน หรือ SVMs ถูกนำเสนอขึ้นโดย Vapnik ในปี 1960 (Joachim, 1997) จุดมุ่งหมายที่สำคัญของแนวคิด ซัพพอร์ตเวกเตอร์แมชชีน คือการหาเส้นแบ่ง Hyper-planes ซึ่งใช้แบ่งข้อมูลสองคลาสเพื่อให้ได้ผลลัพธ์ที่ดีโดยพิจารณาจากสมการเส้นตรง Hyper planes และ ซัพพอร์ตเวกเตอร์แมชชีนจะทำการค้นหาจุดของข้อมูลที่อยู่ใกล้เส้นแบ่ง Hyper planes ซึ่งจุดนั้นเรียกว่า “Support Vector” และได้ถูกประยุกต์มาสู่การจัดกลุ่ม ข้อมูลเอกสาร (Text Classification) โดย Joachim (1999) ซึ่ง ซัพพอร์ตเวกเตอร์แมชชีน จะมีหลักการคือ

1. นำเอกสารที่อินพุตเข้าคำนวณหาค่า y ซึ่งค่าของ $y \in \{-1,1\}$ ได้จากสมการ

$$y = w^T x + b \tag{2.4}$$

ถ้าค่าของ $w^T x + b > 0$ จะกำหนดให้ค่า $y = 1$ ซึ่งจะจัดอยู่ในคลาสที่ 1 ถ้าค่าของ $w^T x + b < 0$ จะกำหนดให้ค่า $y = -1$ ซึ่งจะจัดอยู่ในคลาสที่ 2

2. คำนวณหาเส้นตรงที่แบ่งเอกสารซึ่งเรียกว่า เส้น *Optimal Hyperplane* จากสมการ

$$w^T x + b = 0 \tag{2.5}$$

3. นำค่าที่ได้จากข้อที่ 1 และ 2 ไปเขียนบนเส้นตรงตามแนวแกนตั้งและแกนนอนจะได้ดังรูป 2.12

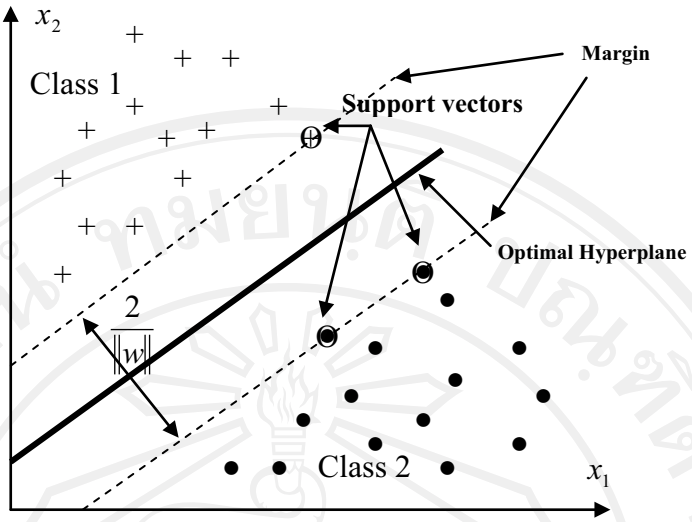
โดยระยะทาง (d) หรือ *maximum margin* จากเส้นขอบ ณ จุด x_i ไปยัง *hyperplane* สามารถแสดงได้ดังสมการ

$$d = \frac{\left|w^T x_i + b\right|}{\|w\|} \tag{2.6}$$

โดยกำหนดให้ w คือ เวกเตอร์น้ำหนัก (Weight Vector)

x_i คือ Input Vector ของเอกสาร

b คือ ค่าคงที่ที่กำหนดขึ้นเพื่อให้เหมาะสมกับการจัดกลุ่มเอกสาร



รูป 2.12 การแบ่งข้อมูลโดยซัพพอร์ตเวกเตอร์แมชชีน (Bennett and Campbell, 2000)

4. เลือกจุดที่อยู่ใกล้เส้นตรง *Optimal Hyperplane* ทั้งเหนือเส้นซึ่ง เรียกว่า "ขอบล่าง" ซึ่งเป็นขอบล่างสุดของ class เอกสารที่อยู่เหนือเส้นตรง *Optimal Hyperplane* และใต้เส้น เรียกว่า "ขอบบน" ซึ่งเป็นขอบบนสุดของ class เอกสารที่อยู่ใต้เส้นตรง *Optimal Hyperplane* เพื่อที่จะหาระยะทางระหว่างเส้นขอบทั้งสองโดยจะเลือกเอาค่าระยะทางที่ห่างจากเส้นตรง *Optimal Hyperplane* ที่น้อยที่สุดเป็นตัวเลือกในการจัดกลุ่มเอกสาร

อย่างไรก็ตามโดยพื้นฐานของ ซัพพอร์ตเวกเตอร์แมชชีนนั้น จะสามารถแบ่งกลุ่มข้อมูลได้เพียง 2 กลุ่ม ดังนั้นการปรับเทคนิคของการเรียนรู้ด้วย ซัพพอร์ตเวกเตอร์แมชชีน เพื่อให้ได้เป็นการจัดแบบหลายกลุ่มจึงเป็นสิ่งจำเป็น สำหรับงานวิจัยฉบับนี้จะปรับปรุงขั้นตอนการเรียนรู้ โดยเป็นการสร้างโมเดลการจัดกลุ่มด้วย One Class SVMs นั่นคือ ให้แต่ละกลุ่มข้อมูลที่กำลังสนใจนั้นเป็น $w^T x + b > 0$ โดยค่า $y = 1$ เมื่อผ่านข้อมูลแต่ละชุด (ที่ผ่านการจัดกลุ่มด้วยมือไว้ก่อนหน้า) เข้าสู่กระบวนการเรียนรู้ ก็จะสร้างโมเดลของการจัดกลุ่มเอกสารแต่ละกลุ่มด้วย ซัพพอร์ตเวกเตอร์แมชชีน นอกจากนี้การจัดกลุ่มเอกสารด้วย ซัพพอร์ตเวกเตอร์แมชชีน หากต้องการความถูกต้องอย่างมาก จะต้องคำนึงถึงลักษณะของข้อมูลที่มี เพื่อให้สามารถเลือก Kernel Function ของการทำงานได้อย่างเหมาะสม เพราะ Kernel Function จะเป็นปัจจัยในการทำงานที่สำคัญของซัพพอร์ตเวกเตอร์แมชชีน

โดยทั่วไป Kernel Function ที่ใช้งานกับซัพพอร์ตเวกเตอร์แมชชีน มี 4 ประเภทคือ

1. Linear

$$k(x,z) = (x^T z)^d \tag{2.7}$$

2. Polynomial

$$k(x,z) = ((x^T z) + \theta)^d \tag{2.8}$$

3. RBF (Radial Basis Function)

$$k(x,z) = \exp(\frac{-\|x-z\|^2}{c}) \quad (2.9)$$

Kernel RBF $= \exp(-gamma * \|x-y\|^2)$

$= \exp(-gamma * (x^2 + y^2 - 2xy))$

Gamma $= 1/k$

k คือจำนวนคำทั้งหมดในเอกสาร

x คือผลรวมของน้ำหนักคำทั้งหมด

y คือผลรวมของน้ำหนักคำทั้งหมดของเอกสารตัวที่ 2 หรือตัวถัดไป

4. Sigmoid

$$k(x,z) = \tanh(k(x^T z) + 0) \quad (2.10)$$

ด้วยแนวคิดของการจัดกลุ่มด้วย ซัพพอร์ตเวกเตอร์แมชชีน จะเป็นการสร้าง Hyperplane เพื่อแยกกลุ่มเป็น 2 กลุ่ม แบบที่ต้องดูค่า Maximum Margin ที่เหมาะสมที่สุดในการจัดกลุ่ม การใช้ Maximum Margin ตามทฤษฎีของ Vapnik Chervonenkis ด้วยการดูค่าความผิดพลาดที่น้อยที่สุดเมื่อได้ค่า Margin ที่มากที่สุด แต่การทำลักษณะเช่นนี้แม้จะมีประโยชน์ แต่อาจจะก่อให้เกิดเส้นแบ่งเขตแดนที่ไม่เหมาะสมเพราะเกิดค่าความ ผิดพลาดสูง ดังนั้นจึงต้องกำหนดพารามิเตอร์ของค่า Maximum Margin hyperplane ดังนั้นจึงต้องมีการนำการทำงานอื่นๆ เข้ามาช่วย ซึ่งก็คือการนำเอาหลักการเรื่อง Kernel เข้ามาใช้เพื่อช่วยในการหาค่า Maximum Margin hyperplane นั่นเอง

ถึงแม้ว่าซัพพอร์ตเวกเตอร์แมชชีนจะสามารถแบ่งกลุ่มข้อมูลได้เพียง 2 กลุ่ม แต่ซัพพอร์ตเวกเตอร์แมชชีนสามารถจัดกลุ่มข้อมูลแบบหลายกลุ่มด้วยวิธีหนึ่งต่อหนึ่งซัพพอร ์ตเวกเตอร์แมชชีน (One-Against-One Support Vector Machines : OAO SVM) (Pawan และ Cory, 2007) เป็นวิธีจัดกลุ่มประเภทข้อมูลแบบเส้นตรงที่ทำงานโดยใช้เคอร์เนลฟังก์ชันโดยใช้หลักการการหาขอบที่กว้างที่สุด ในการทำงานสำหรับจัดกลุ่มหลายกลุ่มข้อมูลจะใช้เทคนิคหนึ่งต่อหนึ่ง (One-Against-One : OAO) เข้ามาช่วย ซึ่งเป็นวิธีที่ช่วยลดจำนวนข้อมูลที่ไม่สามารถจัดกลุ่มกลุ่มได้และมีประสิทธิภาพดีกว่าวิธีหนึ่งต่อทั้งหมด (One-Against-All : OAA) โดยใช้การตัดสินใจทำงานทั้งหมดเท่ากับ $\frac{n(n-1)}{2}$ โดยที่ n คือจำนวนกลุ่มข้อมูลทั้งหมด เมื่อใช้ฟังก์ชันการตัดสินใจเพื่อจัดกลุ่มกลุ่มข้อมูลที่ i กับกลุ่มข้อมูลที่ j จะได้ค่าระยะห่างเส้นแบ่งที่มากที่สุดระหว่างข้อมูลที่เหมาะสมที่สุด (D_{ij}) ตามสมการนี้

$$D_{ij}(x) = w_{ij}^{T}\Phi(x) + b_{ij} \tag{2.11}$$

โดยที่ $w_{ij}^{T}\Phi$ คือเวกเตอร์ m มิติ $\Phi(x)$ คือฟังก์ชันที่นำค่าข้อมูล x ไปสู่มิติที่ m และ b_{ij} เป็นค่าคงที่และ $D_{ij}(x) = -D_{ij}(x)$

กำหนดให้บริเวณ (Region) R_i ที่ไม่มีการทับซ้อนกัน ดังสมการ 2.11

$$R_i = \{x \mid D_{ij}(x) > 0, j = 1,...,n; j \neq i\} \tag{2.12}$$

โดยถ้าอินพุต x อยู่ R_i แล้วจะระบุว่า x อยู่ในกลุ่มที่ i แต่ถ้า x ไม่อยู่ใน $R_i(i = 1,....,n)$ แล้วจะระบุกลุ่มของ x โดยหาค่า $D_i(x)$ จากสมการ 2.13

$$D_i(x) = \sum_{j \neq i, j=1}^{n} sign\left(D_{ij}(x)\right) \tag{2.13}$$

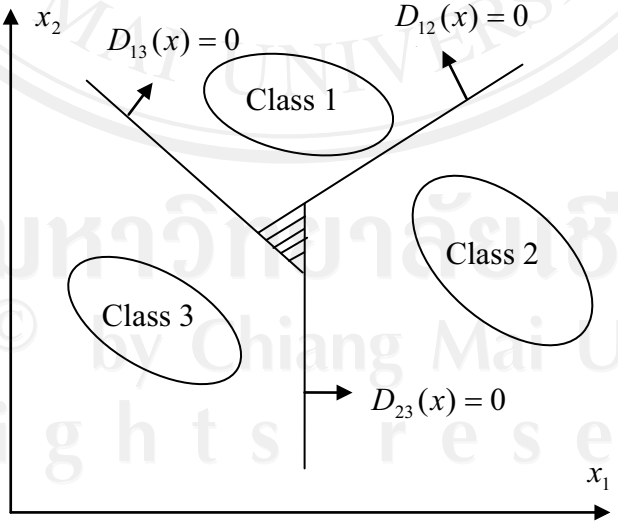
โดยที่

$$sign(x) = \begin{cases} 1 \ \ for\, x \geq 0 \\ -1\, for\, x < 0 \end{cases}$$

และสามารถระบุกลุ่มของ x จากสมการ2.14

$$D(x) = \arg\max_{i=1,...n} D_i(x) \tag{2.14}$$

จากรูป 2.13 แสดงพื้นที่ที่มีค่า $D_i(x) = 0, (i = 1,2,3)$ ตัวอย่างเช่น $sign\left(D_{13}(x)\right) = 1$ และ $sign\left(D_{12}(x)\right) = -1$ จะได้ว่า $D_1(x) = 0$ ในทำนองเดียวกัน $D_2(x) = 0$ และ $D_3(x) = 0$ ซึ่งทำให้พื้นที่บริเวณนี้เป็นพื้นที่ของเวกเตอร์ที่ระบุกลุ่มไม่ได้ โดยระบุเวกเตอร์ x อยู่กลุ่ม i ก็ต่อเมื่อ $D_{ij}(x)$ มีค่ามากที่สุดโดยที่ $(i, j = 1,2,3)$ และ $(i \neq j)$



รูป 2.13 พื้นที่ข้อมูลที่แบ่งแยกไม่ได้ด้วยวิธี OAO SVM

2.11 การวัดประสิทธิภาพของการจัดกลุ่มเอกสาร (Measurement)

การวัดประสิทธิภาพของการจัดกลุ่มเอกสารมีอยู่หลายวิธี แต่สองวิธีที่นิยมตามมาตรฐานของระบบค้นคืนสารสนเทศ (Frankes and Baeza, 1992) ก็คือการใช้การวัดค่าความแม่นยำ (Precision) และค่าความระลึก (Recall)

ค่าความแม่นยำ (Precision: P) เป็นอัตราส่วนของการค้นพบเอกสารที่ถูกต้องจากจำนวนเอกสารทั้งหมดที่ทำการค้นหามาได้

$$\text{ค่าความแม่นยำ} = \frac{\text{จำนวนเอกสารที่ถูกนำมาจัดกลุ่มและถูกต้อง}}{\text{จำนวนเอกสารทั้งหมดที่จัดอยู่ในกลุ่ม}} \quad (2.15)$$

ค่าความระลึก (Recall: R) เป็นอัตราส่วนของการค้นพบเอกสารที่ถูกต้องจากจำนวนเอกสารที่ถูกต้องทั้งหมด

$$\text{ค่าความระลึก} = \frac{\text{จำนวนเอกสารที่จัดอยู่ในกลุ่ม}}{\text{จำนวนเอกสารทั้งหมดในฐานข้อมูล}} \quad (2.16)$$

โดยทั่วไปแล้วสำหรับฐานข้อมูลสารสนเทศที่มีขนาดใหญ่มาก ๆ มักจะไม่ทราบว่าเอกสารที่ถูกต้องทั้งหมดมีอยู่เท่าใด ทำให้ต้องทำการประมาณโดยใช้การสุ่มตัวอย่าง (Sampling) ตามหลักทางสถิติหรือด้วยวิธีอื่น ๆ ด้วย โดยทั่วไปจะเป็นการหาค่า F-measure ซึ่งเป็นการวัดค่าความสัมพันธ์ระหว่างค่าความระลึกและค่าความแม่นยำในเชิงฮาร์โมนิค (Harmonic) โดยที่ค่า F-measure จะมีค่าระหว่าง 0 ถึง 1 ซึ่งถ้า F-measure มีค่าที่ได้ใกล้เคียง 1 มากเท่าไรก็จะแสดงว่าการให้ผลในการจัดกลุ่มเอกสารมีประสิทธิภาพมากขึ้นเท่านั้น แสดงถึงค่าความแม่นยำ การค้นพบเอกสารที่ถูกต้องจากจำนวนเอกสารทั้งหมดที่ทำการค้นหามาได้ (ค่า P)และค่าความระลึก การค้นพบเอกสารที่ถูกต้องจากจำนวนเอกสารที่ถูกต้องทั้งหมด (ค่า R) ทั้งสองค่ามีค่ามากเท่าไรจะทำให้ค่าของการวัดประสิทธิภาพของการจัดกลุ่มเอกสารมากขึ้น ซึ่งแสดงได้ดังสมการ 2.17

$$F = \frac{2xPxR}{P+R} \quad (2.17)$$

ขั้นตอนการวัดประสิทธิภาพเป็นขั้นตอนของการนำเอากลุ่มของเอกสารที่จัดได้มาทำการประเมินประสิทธิภาพ โดยจะตรวจสอบดูว่ากลุ่มของเอกสารที่จัดได้มีค่าเป็นอย่างไร เมื่อเทียบกับกลุ่มของเอกสารที่ถูกต้องซึ่งวัดจากค่าความระลึก (Recall) และค่าความแม่นยำ (Precision) ค่าความแม่นยำจะเป็นค่าที่แสดงว่า การค้นพบเอกสารได้ตรงกับความต้องการเพียงใด ส่วนค่าความ

ระลึกจะเป็นค่าที่แสดงถึงความครอบคลุมในการจัดกลุ่มเอกสาร ในงานค้นคว้าแบบอิสระนี้ได้จัดประสิทธิภาพการจัดกลุ่มเอกสารออกเป็น 3 กลุ่มคือ 1.กลุ่มวัดประสิทธิภาพการจัดกลุ่มเอกสารได้ดีที่สุด คือกลุ่มที่มีค่าความแม่นยำและค่าความระลึกสูง แสดงว่าการจัดกลุ่มเอกสารได้ตรงกับกลุ่มเอกสารและถูกต้องมากที่สุด 2.กลุ่มวัดประสิทธิภาพการจัดกลุ่มเอกสารได้ปานกลาง คือกลุ่มที่ค่าความแม่นยำสูงแต่ค่าความระลึกต่ำ แสดงว่าการจัดกลุ่มเอกสารได้ตรงกับกลุ่มเอกสารแต่มีเอกสารบางส่วนมีความคล้ายคลึงกับกลุ่มเอกสารอื่น 3.กลุ่มวัดประสิทธิภาพการจัดกลุ่มเอกสารได้ต่ำ คือกลุ่มที่ค่าความแม่นยำต่ำแต่ค่าความระลึกสูง แสดงว่าการจัดกลุ่มเอกสารได้ไม่ตรงกับกลุ่มเอกสารและมีเอกสารที่ความคล้ายคลึงกับกลุ่มเอกสารอื่น เนื่องจากเอกสารมีการใช้คำสำคัญ-เอกสารที่ให้ความหมายที่ต่างกัน

สมมติตัวอย่าง ถ้ามีเอกสาร 100 เอกสาร และมีเอกสารที่จัดอยู่ในกลุ่มค้นออกมาได้ 60 เอกสาร ซึ่งเป็นเอกสารที่เกี่ยวข้องและถูกต้อง 30 เอกสาร แต่เอกสารที่จัดอยู่ในกลุ่มค้นออกมาได้และเป็นเอกสารที่ถูกต้องมี 20 เอกสารสามารถคำนวณค่าความแม่นยำและค่าความระลึกได้ดังนี้

$$\text{ค่าความแม่นยำ} = \frac{\text{จำนวนเอกสารที่ถูกนำมาจัดกลุ่มและถูกต้อง}}{\text{จำนวนเอกสารทั้งหมดที่จัดอยู่ในกลุ่ม}} = \frac{20}{60} = 0.34$$

$$\text{ค่าความระลึก} = \frac{\text{จำนวนเอกสารที่จัดอยู่ในกลุ่ม}}{\text{จำนวนเอกสารทั้งหมดในฐานข้อมูล}} = \frac{60}{100} = 0.60$$

$$\text{และจากสมการ F} = \frac{2xPxR}{P+R} = \frac{2x0.34x0.6}{0.34+0.6} = 0.4286$$

นั่นหมายความว่าระบบให้ประสิทธิภาพของการจัดกลุ่มเอกสารคิดเป็นร้อยละ 42.86 แสดงให้ว่าการวัดประสิทธิภาพจัดอยู่ในกลุ่มวัดประสิทธิภาพการจัดกลุ่มเอกสารได้ต่ำ