การจัดเก็บและค้นคืนสารสนเทศ

ผศ.นิพนธ์ เจริญกิจการ*

มนุษย์พึ่งพาสารสนเทศในกิจกรรมของแต่ละวัน ในทุกภาระกิจตั้งแต่การข้ามถนนไปจนถึงการวิคราะห์โรคภัยไข้เจ็บอันสลับซับซ้อนจำเป็นจะต้องมีสารสนเทศ สารสนเทศบางชิ้นเช่น ไฟเขียวไฟแดง เราสามารถทราบได้ทันทีด้วยการมอง แต่ข้อมูลบางอย่างอาจจะต้องสืบคืนหรือต้องผ่านการวิเคราะห์รวบรวมจากหลายๆ แหล่ง มนุษย์เรียนรู้ตั้งแต่เยาว์วัยที่จะเสาะหาข้อมูลเพื่อตอบสนองต่อความต้องการสารสนเทศเบื้องต้น เช่น ด้วยการอ่านหนังสือแบบเรียนหรือการสอบถามจากบิดามารดา ครูบาอาจารย์ สำหรับการสืบค้นสารสนเทศเพื่อตอบสนองต่อความต้องการที่สลับซับซ้อน เช่น การค้นหาหัวข้อวิทยานิพนธ์จำเป็นที่จะต้องใช้ทั้งการศึกษาและประสบการณ์

การค้นคืนสารสนเทศ (Information Retrieval : IR) จะมีการเข้าถึง (Access) สารสนเทศที่ทำการจัดเก็บ (Store) ไว้ล่วงหน้าก่อนแล้ว โดยผ่านขบวนการคัดเลือก (Selectivity) ข้อมูล การค้นคืนสารสนเทศอาจแบ่งเป็นสองประเภทคือ การค้นคืนสารสนเทศโดยที่มีและไม่มีเครื่องคอมพิวเตอร์เข้ามาเกี่ยวข้อง การค้นคืนแบบไม่มีคอมพิวเตอร์อาจจะใช้สมองของมนุษย์ โดยอาศัยการเรียนรู้และการจดจำของสมอง แต่ก็มีข้อเสียตรงที่ว่าสมองมีความจุที่จำกัด อีกทั้งข้อมูลที่ถูกบรรจุอยู่มักจะลบเลือนตามกาลเวลา มนุษย์จึงได้คิดค้นอุปกรณ์ช่วยเข้ามาช่วยจัดการ เช่น การบันทึกภาพเขียนและตัวอักษรลงตามผนังถ้ำหลักศิลาจารึก หนังสัตว์ กระดาษ เป็นความพยายามที่จะเก็บสารสนเทศที่เชื่อว่าสำคัญไว้ ซึ่งอุปกรณ์เหล่านี้เป็นอุปกรณ์ช่วยเสริมสำหรับการเก็บบันทึกข้อมูล วิธีนี้ถึงแม้ว่าจะสามารถจดบันทึกข้อมูลต่างๆ ได้ครบถ้วนเพียงใด แต่เนื่องด้วยสภาพทาง กายภาพของสิ่งของเหล่านี้ เช่น การโยกย้ายและพกพายากของก้อนหิน การเน่าเปื่อยของหนังสัตว์ การฉีกขาดได้ง่ายของกระดาษส่งผลให้ข้อมูลที่เก็บบันทึกด้วยอุปกรณ์ช่วยเหล่านี้สามารถสะดวกในการบันทึกและการค้นคืนอีกทั้งอาจสูญหายไปได้พร้อมกับกาลเวลาเช่นกัน

การค้นคืนสารสนเทศโดยใช้เครื่องคอมพิวเตอร์เข้ามาช่วยได้ทวีความสำคัญมากขึ้นเรื่อยๆ เนื่องจากการจัดเก็บด้วยคอมพิวเตอร์สามารถจัดเก็บข้อมูลได้คงทนถาวรครบถ้วนในปริมาณมากๆ อีกทั้งสารสนเทศจำนวนมากในปัจจุบันอยู่ในรูปแบบสื่ออิเลกทรอนิคส์ที่เครื่องสามารถอ่านได้ (Electronics Media) รวมถึงการค้นคืนโดยคอมพิวเตอร์ปัจจุบันทำได้อย่างสะดวก รวดเร็ว การออกแบบระบบคอมพิวเตอร์เพื่อเป็นเครื่องมือช่วยในการค้นคืนสารสนเทศนั้นเป็นงานที่มีความพิเศษ จำเป็นจะต้องใช้ความรู้ว่าสารสนเทศถูกจัดอยู่ในรูปแบบใด และมนุษย์ทำการค้นคืนสารสนเทศอย่างไรระบบสารสนเทศปัจจุบันมาไกลกว่าภาพผนังถ้ำโบราณหรือหลักศิลาจารึกมาก ปัจจุบันระบบต้องเกี่ยวข้องกับทั้งคอมพิวเตอร์และ

* คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

อุปกรณ์สื่อสารที่มีทั้งราคาแพงและสลับซับซ้อน แม้ว่าคนจำนวนมากจะสามารถเรียนรู้ ที่จะใช้คอมพิวเตอร์หรือเครื่องมือสื่อสาร สำหรับ งานพื้นๆได้ แต่การออกแบบและการใช้ระบบ เพื่องานสารสนเทศที่ซับซ้อนก็ยังเป็นงานเฉพาะของมืออาชีพ

เราจะสามารถค้นหาข้อมูลได้ก็ต่อเมื่อข้อมูลนั้นได้ถูกทำการจัดเก็บไว้ก่อนในรูปแบบใดรูปแบบหนึ่งรูปแบบโดยทั่วไปซึ่งสารสนเทศถูกจัดอยู่คือข้อความและรูปภาพ รูปแบบเหล่านี้ทำให้การตอบสนองต่อคำขอของผู้ใช้งานเป็นไปอย่างลำบาก (หรือแม้แต่เป็นไปไม่ได้ในหลายๆ กรณี) ในหลายๆ กรณีการค้นชุด ของเอกสารอาจจะง่ายหรือยากขึ้นกับว่าชุดของข้อมูลนั้นถูกจัดอยู่ในรูปแบบใด

ในหัวข้อถัดไปจะแสดงความแตกต่างระหว่างคำว่า ข้อมูล (Data) สารสนเทศ (Information) และความเป็นจริง (Reality) องค์กรที่เกี่ยวข้องกับระบบ รวมถึงปัญหาของการวัดคุณภาพของระบบสารสนเทศ

1. นามธรรม (Abstraction)

ความเป็นจริง (Reality) เป็นสิ่งที่เป็นอิสระจากมนุษย์แต่ละคน นั่นคือความเป็นจริงก็คือสิ่งที่เป็นจริงเป็นอย่างนั้นกับทุกๆ คน ส่วนหนึ่ง ของความเป็นจริงเป็นสิ่งที่จับต้องได้ที่เราจับต้องและโต้ตอบอยู่ทุกๆ วัน อีกส่วนหนึ่งของความเป็นจริงเป็นส่วนที่เป็นนามธรรมมากขึ้น ประกอบด้วยแนวความคิดซึ่งประกอบเป็นคณิตศาสตร์ ดนตรี ศิลปะ และสาขาวิชาความรู้ต่างๆ จุดสำคัญที่อยากจะชี้ให้เห็นคือเราไม่มีทางที่จะรู้ความเป็นจริงทั้งหมดได้ แต่มนุษย์มีความเข้าใจในสิ่งพื้นฐานบางส่วนร่วมกันทำให้เราสามารถสื่อสารกันได้อย่างมีประสิทธิภาพ

ระบบสารสนเทศใดๆ ก็ตาม ในส่วนที่เป็นหัวใจของมันก็คือมันเป็นชุด (Collection) ของข้อมูลเกี่ยวกับความเป็นจริง ชุดของข้อมูลไม่มีทางที่จะสมบูรณ์ได้ เมื่อมีข้อมูลใหม่ๆ เข้ามาข้อมูลใหม่ถูกรวมเข้ากับข้อมูลเดิม ปรับปรุงให้มันใกล้เคียงกับความเป็นจริงมากขึ้นๆ ระบบสารสนเทศจะนำชุดของข้อมูลเหล่านี้มาใช้ สรุปว่าเราได้นามธรรมอันแรกคือ

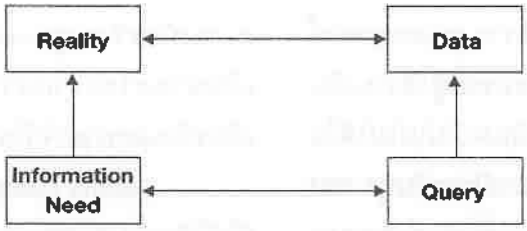
"ในระบบสารสนเทศใดๆ โลกแห่งความเป็นจริงจะถูกแสดงด้วยชุดของข้อมูลอันเป็นนามธรรมจากการสังเกตโลกแห่งความเป็นจริงและนำมาไว้ในระบบ"

ลองพิจารณาผู้ใช้ (User) ของระบบ ผู้ใช้ใช้ระบบสารสนเทศในสองลักษณะหลักๆ คือ ในการจัดเก็บสารสนเทศ (หรือข้อมูล) โดยคาดหวังความต้องการใช้ในอนาคต และในการค้นหาสารสนเทศ (ข้อมูล) เพื่อตอบสนองต่อความต้องการในปัจจุบัน ไม่ว่าจะเป็นไปลักษณะใดผู้ใช้จะมีความต้องการสารสนเทศเป็นตัวผลักดันให้มีการใช้ระบบสารสนเทศ ความต้องการนี้จะต้องแสดงให้แก่ระบบ (Debons, Home and Croenweth 1988) ถ้าผู้ใช้ต้องการจัดเก็บสารสนเทศ ก็จะต้องอยู่ในรูปแบบที่สะท้อนการคาดหวังความต้องการที่จะเกิดขึ้นในอนาคต ผู้จัดเก็บจะพยายามจัดเก็บสารสนเทศในรูปแบบที่คาดว่าจะเป็นประโยชน์หรือง่ายต่อการค้นคืนในภายหลัง นั่นคือระบบสารสนเทศจะมีอิทธิพลต่อรูปแบบที่สารสนเทศจะถูกจัดเก็บ ยกตัวอย่างเช่น สารสนเทศจะต้องถูกนำไปไว้ในรูปแบบสื่ออิเล็กทรอนิคส์เพื่อที่จะใช้กับคอมพิวเตอร์ได้ ถ้าผู้ใช้งานต้องการที่จะค้นคืนสารสนเทศ ก็จะต้องแสดงความต้องการในรูปแบบของคำขอ (Query) ซึ่งระบบสามารถเข้าใจได้ ระบบมีอิทธิพลอย่างมากต่อรูปแบบซึ่งคำขอสารสนเทศ

จะสามารถเป็นได้ ทำให้ได้นามธรรมที่สองคือ

"ความต้องการสารสนเทศของผู้ใช้ ไม่ว่าจะเพื่อการสร้าง จัดเก็บ หรือค้นคืนสารสนเทศ ถูกทำเป็นนามธรรมในรูปแบบซึ่งเหมาะสม กับระบบสารสนเทศที่จะนำมันไปใช้"

หลักการนามธรรมสองอันนี้แสดงให้เห็นถึงปัญหาพื้นฐานของผู้พัฒนาระบบ ผู้ใช้มีความต้องการสารสนเทศ ซึ่งควรที่จะได้มาจากโลกแห่งความเป็นจริง แต่ระบบสารสนเทศสามารถทำงานได้ในระดับนามธรรม (เปรียบเทียบคำขอของผู้ใช้กับข้อมูลที่จัดเก็บ) ดังแสดงในรูปที่ 1 ในทำนองเดียวกันเมื่อผู้ใช้พยายามที่จะจัดเก็บสารสนเทศปัญหาของระบบคือการที่จะต้องพยายามหารูปแบบที่ทำให้การบิดเบือนผิดเพี้ยนเนื่องจากกระบวนการนามธรรมเกิดขึ้นน้อยที่สุด



รูปที่ 1 นามธรรมและการคะเน (Mapping) ของระบบสารสนเทศ (Korthage 1997)

2. ระบบสารสนเทศ

ระบบสารสนเทศอาจจะแบ่งได้เป็นสองส่วนที่สำคัญคือ ส่วนที่ควบคุมไม่ได้ของระบบ (Ectosystem) และส่วนที่ควบคุมได้ของระบบ *(Endosystem, Korfhage and Delutis 1969)* จากมุมมองของนักออกแบบระบบ ส่วนที่ควบคุมไม่ได้ของระบบจะประกอบด้วยปัจจัยซึ่งไม่อยู่ภายใต้การควบคุมของนักออกแบบ อันรวมถึงผู้คนที่เกี่ยวข้องกับระบบ รูปแบบของสารสนเทศ อุปกรณ์ และเทคโนโลยีที่มี ณ ขณะนั้น ส่วนที่ควบคุมได้ของระบบประกอบด้วยปัจจัยซึ่งนักออกแบบสามารถระบุและควบคุมเช่น อุปกรณ์ อัลกอริธึม และโปรซีเยอร์ที่เลือกใช้

ส่วนที่ควบคุมไม่ได้ของระบบประกอบด้วยส่วนที่เป็นมนุษย์สามกลุ่ม คือผู้ใช้ (User) นายทุน (Funder)และเจ้าหน้าที่บริการ *(Server, Nance 1967; Baker 1968)* ผู้ใช้คือบุคคลซึ่งปรารถนาที่จะจัดเก็บสารสนเทศลงในระบบหรือค้นคืนสารสนเทศจากระบบ บ่อยครั้งที่ผู้ใช้มีความรู้ทางด้านโครงสร้างของระบบน้อยหรืออาจจะไม่รู้ระบบทำงานได้อย่างไร ผู้ใช้อาจจะตัดสินประสิทธิภาพของระบบโดยอยู่บนพื้นฐานความรู้เท่าที่มีการตัดสินของผู้ใช้มีความสัมพันธ์โดยตรงกับความคาดหวังของผู้ใช้ ระบบสารสนเทศส่วนมากจะมีผู้ใช้มากกว่าหนึ่งคน ลักษณะของกลุ่มผู้ใช้ว่ามีความเหมือนกันหรือแตกต่างกันของสมาชิกในกลุ่ม ความสนใจของสารสนเทศของพวกเขา และความคุ้นเคยของพวกเขาในระบบ จะต้องถูกนำมาพิจารณาในการพัฒนาและใช้งานส่วนที่ควบคุมได้ของระบบ

นายทุนเป็นบุคคลหรือองค์กรซึ่งจ่ายค่าดำเนินการของระบบสารสนเทศ ไม่มีระบบใดที่ไม่มีค่าใช้จ่าย และผู้สนับสนุนก็ปรารถนาที่จะให้ได้ผลตอบแทนจากระบบสูงกว่ารายจ่าย ค่าใช้จ่ายของระบบอาจจะถูกสนับสนุนได้ใน

หลายรูปแบบ อาจจะไม่แปลกที่จะเห็นองค์กรแบกรับค่าใช้จ่ายนี้ในฐานะเป็นส่วนหนึ่งของค่าโสหุ้ย (Overhead) ขององค์กร ทำให้ผู้ใช้อาจจะใช้ระบบได้ฟรี เช่น การที่เราใช้ระบบค้นคืนหนังสือในห้องสมุด องค์กรบางองค์กรอาจจะคิดค่าบริการกับผู้ใช้งานเพื่อครอบคลุมค่าใช้จ่ายบางส่วนหรือทั้งหมด ถ้าผู้ใช้เสียค่าบริการ เขาต้องเห็นว่าประโยชน์ที่ได้รับจากการใช้ระบบคุ้มค่ากับค่าใช้จ่ายที่เสียไป ในลักษณะเดียวกันนายทุนก็ต้องแน่ใจว่าประโยชน์ของระบบสูงกว่าค่าใช้จ่ายสุทธิของการดำเนินระบบ ขณะที่ค่าใช้จ่ายอาจจะวัดได้ค่อนข้างชัดเจน ผลประโยชน์จากระบบนอกเหนือจากการตอบแทนทางด้านการเงินอาจจะวัดได้ยากกว่า ซึ่งอาจจะรวมถึงการที่เราสามารถรับสารสนเทศที่ต้องการได้โดยสะดวก เวลาที่ใช้ในการค้นสารสนเทศน้อยลง หรือแม้แต่การเป็นภาพพจน์หน้าตาขององค์กร

เจ้าหน้าที่บริการเป็นผู้ที่ประกอบอาชีพทางด้านสารสนเทศ คนพวกนี้ทำให้ระบบดำรงอยู่และจัดเสนอบริการให้แก่ผู้ใช้ องค์กรบางองค์กรเช่น ระบบห้องสมุดขนาดใหญ่หรือองค์กรที่ให้บริการทางด้านสารสนเทศมีเจ้าหน้าที่บริการเป็นจำนวนมากทำงานอยู่ (เช่นห้องสมุดของ University of Toronto ซึ่งมีการบริหารงานห้องสมุดแบ่งเป็นห้องสมุดย่อยๆ มากกว่า 40 ห้องสมุด มีเจ้าหน้าที่บริการหลายร้อยคน) โดยมีลักษณะงานที่หลากหลาย เจ้าหน้าที่บริการบางรายจะทำงานติดต่อกับผู้ใช้งานโดยตรง ขณะที่บางส่วนจะทำงานอยู่หลังฉาก เพื่อสนับสนุนให้การดำเนินงานเป็นไปอย่างราบรื่น อันรวมถึงเจ้าหน้าที่ทางด้านเทคนิค นักวิเคราะห์ระบบ นักออกแบบระบบ และเจ้าหน้าที่ฝ่ายดำเนินการและฝ่ายบำรุงรักษา

นักออกแบบระบบไม่สามารถควบคุมคนสามกลุ่มนี้ แต่อย่างไรในการวิเคราะห์และออกแบบระบบเราจะต้องนำพวกเขาเหล่านั้นเข้ามาเป็นปัจจัยในการพิจารณาด้วย คนทั้งสามกลุ่มจะมีผลต่อความสำเร็จของระบบได้ บางครั้งคนคนเดียวอาจจะทำหน้าที่ทั้งสามบทบาทนี้ เช่นในระบบสารสนเทศที่เล็กมากๆ เช่น ระบบที่เก็บสูตรอาหารสำหรับครอบครัว หรือระบบงบประมาณค่าใช้จ่ายในบ้าน ผู้ใช้ นายทุน และผู้ให้บริการอาจจะเป็นบุคคลคนเดียวกันได้ ในระบบที่มีขนาดใหญ่ขึ้นบทบาทซอบเขตอาจจะคาบเกี่ยวกันได้ เช่น ถ้าผู้ใช้จ่ายค่าบริการ ดังนั้น ผู้ใช้ในบทบาทหนึ่งก็เป็นนายทุน ระบบสารสนเทศส่วนบุคคลหลายอันผู้ใช้เป็นเจ้าหน้าที่บริการด้วยทำการค้นคืนสารสนเทศและการประมวลผลสารสนเทศโดยตรง

ในขณะที่นักออกแบบระบบไม่สามารถควบคุมส่วนที่ควบคุมไม่ได้ของระบบ เขาสามารถที่จะควบคุมส่วนที่ควบคุมได้ของระบบได้อย่างบริบูรณ์ ส่วนที่ควบคุมได้ของระบบมีอยู่สี่องค์ประกอบ สื่อที่จะใช้จัดเก็บสารสนเทศอุปกรณ์ที่จะใช้ในการประมวลสารสนเทศ อัลกอริธึมซึ่งทำงานอยู่บนอุปกรณ์ และโครงสร้างข้อมูลซึ่งใช้จัดรูปแบบของสารสนเทศ

สื่อครอบคลุมถึงสิ่งพิมพ์ต่างๆ ข้อความ รูปภาพ แผนที่บนกระดาษ หรือสื่ออิเล็กทรอนิกส์ที่เครื่องสามารถอ่านได้ เช่น ไมโครฟอร์ม เทปแม่เหล็ก แผ่นดิสต์ ฮาร์ดดิสต์ และซีดีรอม ชนิดของการประมวลผลซึ่งสามารถกระทำได้บนข้อมูลเหล่านี้ ส่วนหนึ่งขึ้นอยู่กับชนิดของสื่อที่เลือกใช้ด้วย

อุปกรณ์ที่สามารถนำมาใช้ได้ในระบบสารสนเทศมีอยู่มากมาย ตั้งแต่ตู้เอกสารหรือชั้นวางหนังสือไปจนถึงเครื่องคอมพิวเตอร์ แสกนเนอร์และอุปกรณ์ประมวลผลซึ่งใช้เทคโน

โลยีเลเซอร์ นักออกแบบระบบจะต้องพยายามที่จะจับคู่อุปกรณ์และสื่อให้เหมาะสมเพื่อให้เกิดประโยชน์สูงสุดต่อระบบ การเลือกคู่ที่ไม่ดีอาจจะทำให้ทั้งระบบทำงานได้อย่างไม่มีประสิทธิภาพและมีค่าใช้จ่ายที่สูงเกินจำเป็น

การเลือกอัลกอริธึมที่ดีเพื่อประมวลผลสารสนเทศที่ถูกจัดเก็บไว้ ประมวลคำขอและประมวลข้อมูลอื่นๆ สามารถบันดาลให้ระบบสารสนเทศมีทั้งประสิทธิภาพและประสิทธิผล หรือการจะทำให้ระบบมีค่าใช้จ่ายที่สูงเกินไป เกิดความล้มเหลวในแง่การให้บริการ ผู้ใช้โดยทั่วไปสนใจในรูปแบบของบริการ (ฟังก์ชัน) ซึ่งระบบสามารถนำเสนอได้ ความสามารถต่างๆ ที่ถูกนำเสนอให้กับผู้ใช้ส่วนใหญ่เป็นผลมาจากอัลกอริธึมที่เลือกใช้มากกว่าที่จะมาจากอุปกรณ์หรือสื่อที่เลือก

นักออกแบบระบบสามารถเลือกโครงสร้างข้อมูลได้มากมายที่เหมาะสมกับอัลกอริธึมและโปรแกรมที่เขียน ในหลายๆ ครั้งสื่อหรืออุปกรณ์ที่เลือกใช้กลายเป็นข้อจำกัดในการเลือกโครงสร้างข้อมูลเหล่านี้ ยกตัวอย่างเช่น ถ้าสารสนเทศถูกจัดเก็บอยู่บนเทปแม่เหล็ก การเข้าหาข้อมูลเหล่านี้ก็จะเป็นแบบเรียงทีละลำดับ (Sequential) การจัดองค์กรของข้อมูลในรูปแบบใดๆ ที่ไม่ใช่การจัดแบบเรียงลำดับก็เป็นการสูญเปล่า ปัจจุบันการค้นคืนสารสนเทศโดยทั่วไปจะเป็นแบบการเข้าหาข้อมูลโดยสุ่ม (Random) เทปแม่เหล็กมักจะใช้สำหรับการทำสำรอง (Backup) เท่านั้น ขณะที่เราใช้ดิสก์แม่เหล็กและซีดีรอมเป็นหลักในการจัดเก็บข้อมูลจำนวนมาก นักออกแบบระบบที่ดีจะต้องพิจารณาเลือก อุปกรณ์ สื่อ อัลกอริธึม และโครงสร้างข้อมูลที่เหมาะสม สอดคล้องกับความต้องการหรือความคาดหวังของคนกลุ่มต่างๆที่มีต่อระบบ

3. การวัดผล (Measures)

นักออกแบบระบบสามารถกำหนดสมรรถภาพของระบบได้โดยการเลือกใช้ส่วนที่ควบคุมได้ของระบบซึ่งครอบคลุมสื่อ อุปกรณ์ อัลกอริธึม และโครงสร้างข้อมูล อย่างไรก็ตามก็ต้องมีการประเมินสมรรถภาพจากทางด้านส่วนที่ควบคุมไม่ได้ของระบบด้วย ผู้ใช้อาจจะพึงพอใจหรือไม่พึงพอใจต่อผลที่ได้จากการค้นคืน นายทุนอาจจะรู้สึกว่าระบบไม่คุ้มทุน หรือเจ้าหน้าที่บริการอาจจะเชื่อว่า บางส่วนของการดำเนินระบบสามารถทำให้มีประสิทธิภาพมากขึ้น ถ้าความต้องการหรือความคาดหวังของใครก็ตามในคนสามกลุ่มนี้ไม่ได้ถูกตอบสนองโดยระบบ โอกาสที่ระบบจะล้มเหลวก็เป็นไปได้ *(Meadow 1973; Kraft and Bookstein 1978; Blaie and Maron 1985; Losee 1991; Turtle and Croft 1991; Dumais 1994)*

สมาชิกแต่ละกลุ่มตัดสินสมรรถนะของระบบจากมุมมองที่แตกต่างกัน ผู้ใช้อาจจะสนใจในแง่ประสิทธิภาพ (Effectiveness) ของระบบว่ามันตอบสนองต่อความต้องการสารสนเทศของเขามากน้อยเพียงใด มีปัจจัยที่ส่งผลประสิทธิภาพของระบบ อันรวมถึงและความแม่นยำและความสมบูรณ์ที่ระบบตอบสนองต่อคำขอและจำนวนของเอกสารที่ไม่เกี่ยวข้อง (Irrelevance) ซึ่งถูกดึงออกมาด้วย ผู้ใช้ซึ่งรู้สึกว่าระบบไม่ตอบสนองต่อความต้องการของเขาอาจจะเลิกใช้ระบบนั้น เราอาจจะเรียกได้ว่าระบบนั้นล้มเหลว (Bottle 1965) บางครั้งเราอาจจะถามผู้ใช้โดยตรงเลยก็ได้ เช่น การถามผู้ใช้งานว่ามีความพึงพอใจในระบบที่ใช้สืบค้น ความพึงพอใจต่อผลที่ได้ค้น และความมั่นใจต่อผลที่ได้ค้นหา เป็นต้น *(Bottle 1965; Hamilton and Chervany 1971; Heaps 1971; Eisenberg and Hu 1987; Frei and*

Schauble 1991; Frei and Wyle 1991; Charoenkitkarn 1996)

เจ้าหน้าที่บริการมีแนวโน้มที่จะสนใจในแง่ของประสิทธิผล (Efficiency) ของระบบมากกว่า พวกเขาอาจจะถามว่า ระบบได้ถูกออกแบบมาให้ผู้ใช้สามารถแสดงหรือแจกแจงความต้องการของเขาได้อย่างมีประสิทธิผลหรือเปล่าหรือระบบได้ตอบสนองต่อคำขออย่างมีประสิทธิผลหรือเปล่า หรือมีการสูญเปล่ามากน้อยเพียงใดในการตอบสนองต่อผู้ใช้ หรือระบบสามารถหาสารสนเทศที่ต้องการได้อย่างรวดเร็วหรือไม่ หรือโครงสร้างข้อมูลและอัลกอริธึมได้ถูกเลือกมาดีพอหรือไม่ ความพึงพอใจของเจ้าหน้าที่บริการจะลดลงถ้าพวกเขาต้องรอการทำงานระบบนานขึ้น *(Lancaster and Climenson 1968)* ประสิทธิผลทางด้านความเร็วในการทำงานเป็นการวัดด้วยเวลาซึ่งระบบจะต้อง ใช้ในการทำงาน (สำหรับผู้ที่มีความคุ้นเคยกับระบบ UNIX อาจจะเลือกใช้โปรแกรม prof เป็นต้น) การวัดทางด้านความเร็วเป็นสิ่งที่สำคัญมาก โดยเฉพาะอย่างยิ่งระบบค้นคืนสารสนเทศส่วนใหญ่จะเป็นระบบที่มีการโต้ตอบแบบทันที (Interactive) การค้นคืนที่ใช้เวลานานอาจจะทำให้ระบบไม่น่าใช้และไร้ประโยชน์ ประสิทธิผลทางด้านเนื้อที่ที่ใช้จะวัดด้วยจำนวนของไบต์ซึ่งจะต้องใช้ในการจัดเก็บข้อมูลเนื้อที่ส่วนเกิน (Space Overhead) โดยวัดเป็นอัตราส่วนของขนาดของดรรชนีรวมกับขนาดของเอกสารต่อขนาดของเอกสาร อัตราส่วนนี้มักจะอยู่ที่ 1.5 ถึง 3 สำหรับระบบค้นคืนชนิดใช้ Inverted File

นายทุนจะสนใจทางด้านเศรษฐศาสตร์ของระบบ ถึงแม้ระบบจะมีประสิทธิภาพและประสิทธิผลที่ดีแต่ถ้ามันไม่คุ้มทางด้านเศรษฐศาสตร์ระบบอาจจะถูกมองว่าล้มเหลวได้ เนื่องจากประสบการขาดทุน *(Bourne and Ford 1964; Korfhage and DeLutis 1969, Lancaster 1971; Cooper 1972; Taylor 1986)*

คนแต่ละกลุ่มที่เกี่ยวข้องกับระบบตัดสินมันในแง่มุมที่ต่างกัน การตัดสินเหล่านี้มีความสัมพันธ์ระหว่างกัน ส่วนที่ควบคุมได้ของระบบที่ไม่มีประสิทธิผลอาจจะทำให้ผู้ใช้ไม่พึงพอใจทั้งๆที่มันอาจจะให้ผลการค้นคืนที่ตอบสนองความต้องการสารสนเทศได้เป็นอย่างดี (มีประสิทธิภาพดี) การที่ผู้ใช้ลดลงก็จะส่งผลให้เกิดรายได้ที่ลดน้อยลงตามด้วยทำให้ระบบไม่คุ้มทุน ในทางตรงกันข้ามระบบที่มีประสิทธิผลที่แม้จะให้ผลการค้นคืนที่ไม่ตอบสนองต่อความต้องการสารสนเทศของผู้ใช้อย่างดีนัก แต่ผู้ใช้อาจจะเห็นว่าระบบนั้นมีประสิทธิภาพที่ดีทำให้เกิดการใช้งานระบบมากขึ้น เกิดรายได้มากขึ้น ซึ่งนายทุนอาจจะสนับสนุนให้ทำการขยายระบบหรือเพิ่มเติมบริการในรูปแบบต่างๆมากขึ้น

เราสรุปได้ว่ามีการวัดผลในสามแง่มุม ทั้งสามมีความสำคัญ แต่ในหนังสือเล่มนี้เราจะสนใจในเรื่องของประสิทธิภาพเป็นหลัก การวัดด้วยวิธีนี้จะมีหลักที่สำคัญคือจะต้องมีการตัดสิน Relevance Judgement) ว่ามีเอกสารใดบ้างที่ตรงหรือเกี่ยวข้อง (Relevance) กับความต้องการ ซึ่งวิธีการนี้อาจจะเป็นปัญหาได้เนื่องจากการพิจารณาว่า เอกสารสารตรงกับความต้องการหรือไม่ ไม่มีหลักตายตัว เชื่อถือไม่ได้ ผู้ที่ทำการพิจารณาคนละคนกัน (หรืออาจจะมีระดับชั้นความเห็นที่แตกต่างกันในกรณีที่มีความเห็นไม่เป็นแบบไบนารีไม่ใช่เพียงใช่ หรือ ไม่ใช่ เท่านั้น) นักวิจัยทางด้านการค้นคืนสารสนเทศมีความเห็นขัดแย้งกันมาโดยตลอดว่าควรหรือไม่ควรที่จะใช้การวัดผลวิธีนี้ โดยนักวิจัยจำนวนมากเห็นว่าแม้วิธีนี้อาจจะมีปัญหาเรื่องการตัดสินที่ไม่ตรงกัน แต่วิธีนี้ก็ยัง

เป็นการวัดผลที่ใช้ได้อยู่ (Valid) ผู้สนใจในรายละเอียดการอภิปรายเรื่องนี้อาจจะหาเพิ่มได้จาก Silton and McGitt (1983) และ Sparck-Jones (1981) ในการค้นคืนเมื่อมีการค้นคืนเอกสารออกมาได้ก็จะมาทำการเปรียบเทียบว่า มีความถูกต้องตรงกับความต้องการมากน้อยเพียงใด การวัดประสิทธิภาพของการค้นคืนมีอยู่หลายวิธี แต่สองวิธีที่มักนิยมใช้กันคือ ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall, Salton 1988)

ค่าความระลึก(Recall) เป็นอัตราส่วนของการค้นพบเอกสารที่ถูกต้องจากจำนวนเอกสารที่ถูกต้องทั้งหมด

$$\text{ค่าความระลึก} = \frac{\text{จำนวนเอกสารที่ถูกต้องที่ค้นคืนได้}}{\text{จำนวนเอกสารที่ถูกต้องทั้งหมดในฐานข้อมูล}}$$

โดยทั่วไปแล้วสำหรับฐานสารสนเทศ ที่มีขนาดใหญ่พอ เรามักจะไม่ทราบว่าเอกสารที่ถูกต้องทั้งหมดมีอยู่เท่าใด ทำให้เราต้องทำการประมาณโดยใช้การสุ่มตัวอย่าง (Sampling) ตามหลักทางสถิติหรือด้วยวิธีอื่นๆ

ค่าความแม่นยำ (Precision) เป็นอัตราส่วนของการค้นพบเอกสารที่ถูกต้องจากจำนวนเอกสารทั้งหมดที่ทำการค้นหามาได้

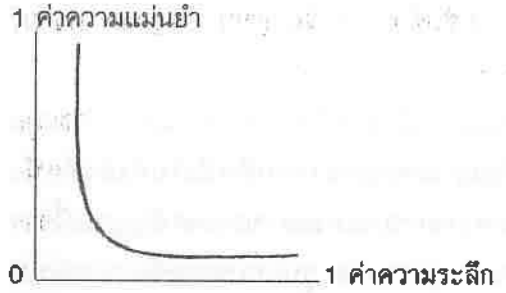
$$\text{ค่าความแม่นยำ} = \frac{\text{จำนวนเอกสารที่ถูกต้องที่ค้นคืนได้}}{\text{จำนวนเอกสารทั้งหมดที่ค้นคืนออกมาได้}}$$

ค่าความแม่นยำจะเป็นค่าที่แสดงว่าการค้นคืนข้อมูลได้ตรงกับความต้องการเพียงใด เช่นถ้าค้นคืนเอกสารออกมาได้ N เอกสาร และมีเอกสาร R เอกสารที่ถูกต้อง ดังนั้น ค่าความแม่นยำจะเป็น R/N หรือเป็นโอกาสของเอกสารที่ค้นคืนออกมาได้ตรงกับความต้องการ ส่วนค่าความระลึกจะเป็นค่าที่แสดงความครอบคลุม เช่น ถ้าเอกสารที่ตรงกับความต้องการมีทั้งสิ้น T เอกสาร และการค้นคืนสามารถดึงเอกสารที่ตรงกับความต้องการได้ R เอกสาร ค่าความระลึกจะเป็น R/T ทั้งค่าความแม่นยำและค่าความระลึกจะมีค่าอยู่ระหว่าง 0 ถึง 1

ในการค้นคืนสารสนเทศถ้าได้เฉพาะ เอกสารที่ตรงกับความต้องการออกมาทั้งหมดและไม่มีเอกสารที่ไม่เกี่ยวข้องออกมาด้วย ค่าของความแม่นยำและความระลึกจะมีค่าเป็น 1.00 ค่าความแม่นยำและความระลึกมักมีความ สัมพันธ์เป็นปฏิภาคผกผัน อาจกล่าวได้ว่าโดยทั่วไปเมื่อมีค่าความแม่นยำและความระลึกที่ได้จากการค้นคืนด้วยคำขอ (Query) อันหนึ่ง หากต้องการให้ค่าความแม่นยำสูงขึ้น ค่าความระลึกก็มักจะลดลง (ไม่จำเป็นเสมอไป)และทางตรงกันข้ามหากต้องการให้ค่าความระลึกสูงขึ้นค่าความแม่นยำก็มักจะลดลง(ไม่จำเป็นเสมอไป)

โดยทั่วไป เรามักจะต้องการวัดผลการ ค้นคืนทั้งค่าความแม่นยำและความระลึกพร้อมกันทั้งสองตัว จึงมีการนำค่าของทั้งคู่ (สองตัวแปรมาบันทึกลงกราฟสองระบบเพื่อแสดงผล ดังตัวอย่างในรูปที่ 2 ซึ่งแสดงความสัมพันธ์โดยทั่วไปของตัวแปรทั้งสองอยู่ในรูปปฏิภาคผกผัน นั่นคือ เมื่อค่าความแม่นยำเพิ่มขึ้นค่าความระลึกจะลดลง เมื่อค่าความแม่นยำลดลงค่าความระลึกเพิ่มขึ้น

การลากกราฟนี้อาจจะสร้างขึ้นจากการค้นคืนหลายๆ ครั้ง หรืออาจจะสร้างจากการหาค่าเฉลี่ยการค้นคืนจำนวนหนึ่งดังที่อธิบายใน Salton and McGill (1983) และ van Rijsbergen (1979)



รูปที่ 2 ความสัมพันธ์โดยทั่วไประหว่างค่าความแม่นยำและความระลึก

มีผู้เสนอวิธีการรวมค่าของตัวแปรทั้งสองออกมาเป็นค่าๆ เดียว ดังตัวอย่างที่เสนอไว้โดย van Rijsbergen (1979) เป็นค่า E

$$E = 1 - \frac{(1+b^2)PR}{(b^2 P+R)}$$

โดยที่ P คือค่าของความแม่นยำ และ R คือค่าของความระลึก และ b เป็นการวัดความสำคัญเชิงสัมพันธ์ของค่าความระลึกและความแม่นยำต่อผู้ใช้ เช่นถ้าเราให้ความสำคัญของ b = 0.5 หมายความว่าผู้ใช้สนใจในค่าของความแม่นยำเป็นสองเท่าของค่าความระลึก และถ้า b = 2 แสดงว่าผู้ใช้สนใจค่าความระลึกเป็นสองเท่าของค่าความแม่นยำ ผลของการค้นคืนที่เป็นตัวเลข (ตัวแปร) เดียวนี้ทำให้เราสามารถวัดการค้นคืนสองครั้งจากระบบเดียวกัน (หรือการสืบค้นคำขอๆ เดียวจากต่างระบบกันก็ได้) แล้วบอกได้ว่าการค้นคืนอันไหนดีกว่ากัน

การทดลองทางด้านการค้นคืนสารสนเทศมักจะใช้ชุดทดสอบซึ่งจะประกอบด้วยฐานเอกสารข้อความ (Textbase) และชุดของการระบุความต้องการสารสนเทศ (บางครั้งเรียกคำถาม) และเอกสารที่สัมพันธ์กับความต้องการนั้น (คำตอบ) จำนวนของเอกสารในชุดทดสอบมักมีจำนวนน้อย โดยทั่วไปมักจะมีจำนวนไม่กี่ร้อยจนถึงไม่กี่พันเอกสาร ในระยะหลังข้อมูลเหล่านี้ได้ถูกจัดเก็บไว้บนแผ่นซีดีรอมแล้ว การทดลองที่ใช้จำนวนเอกสารน้อยๆ นี้ ได้ถูกวิพากษ์วิจารณ์ว่าไม่มองความเป็นจริงเพราะว่าฐานข้อมูลของระบบที่ใช้งานจริง (ในเชิงพาณิชย์) จะมีขนาดใหญ่มากอย่างเช่น ฐานข้อมูลของระบบสืบค้นห้องสมุดที่ University of Toronto มีขนาด (เมื่อปี 2538) กว่า 1 เทราไบต์ (Terrabyte, ล้านล้านตัวอักษร) หรือ ฐานข้อมูลของบริษัท Lexus Nexus ก็มีขนาดใหญ่เช่นกัน การทดลองบนฐานข้อมูลขนาดเล็กอาจจะไม่มีประโยชน์หรือไม่สามารถนำมาประยุกต์ใช้กับระบบขนาดใหญ่ได้

แต่ในระยะหลังได้มีการทดลองกับชุดทดสอบขนาดใหญ่ขึ้น เช่น การทดลองซึ่งสนับสนุนโดยสถาบันมาตรฐานและเทคโนโลยีแห่งชาติของสหรัฐอเมริกา (National Institute of Standards and Technology: NIST) ที่รู้จักกันอย่างแพร่หลายในหมู่นักวิจัยสาขาการค้นคืนสารสนเทศในนาม TREC *(Text REtrieval Conference, Harman 1993-1997)* การทดลองนี้ทำกันปีละครั้งโดยเริ่มมาตั้งแต่ปี 1992 จนถึงปัจจุบัน จำนวนเอกสารที่ใช้ในการทดลอง นี้มีจำนวนมากกว่า 1 ล้านเอกสาร (ประมาณ 2 กิกะไบต์)

สำหรับในประเทศไทยได้มีการทดลองทางด้านการค้นคืนสารสนเทศในลักษณะนี้บ้าง เช่น กนกวรรณ โสตถีวรกุล (2540) และสุวัฒน์

สถาพรพิริยะเดช (2540) โดยสุวัฒน์ได้ทำการสร้างชุดทดสอบขนาดเล็กขึ้น โดยเป็นข่าวภาษาไทยที่ได้จากสำนักข่าวไอเอ็นเอ็น เป็นข่าวที่มีความหลากหลายเช่นทางด้านธุรกิจ การเมือง และการกีฬา เป็นต้น จำนวนข่าวที่รวบรวมไว้ประมาณ 2,700 เอกสารมีความสมมุติความต้องการสารสนเทศไว้สี่ข้อ โดยแต่ละข้อจะมีคำตอบที่ได้ทำการตัดสินใจไว้แล้ว สำหรับกนกวรรณได้ใช้พระคริสต์ธรรมคัมภีร์ฉบับภาษาอังกฤษเสมือนฐานสารสนเทศ โดยข้อแต่ละข้อในพระคัมภีร์จะเป็นเสมือนเอกสารหนึ่งเอกสาร กนกวรรณได้สร้างคำถามสี่ข้อให้ผู้เข้าทำการทดลองได้ทำการค้นเอกสารเพื่อหาคำตอบที่ถูกต้อง คำถามแต่ละข้อจะมีคำตอบที่ได้ถูกพิจารณาไว้แล้ว เช่นกัน

4. จากข้อมูลถึงความฉลาด

ในหน้าก่อนๆ คำว่า ข้อมูลและสารสนเทศ ได้ถูกใช้อย่างไม่เป็นทางการ ตอนนี้ถึงเวลาที่เราจะแสดงให้เห็นความแตกต่างระหว่างคำสองคำนี้ คำอธิบายในที่นี้อาจจะไม่ได้เป็นที่ยอมรับกันโดยสากล แต่จะใช้เฉพาะในบทความนี้ ข้อมูลจะถูกรับ จัดเก็บ และค้นคืนโดยส่วนที่ควบคุมได้ของระบบ นั่นคือข้อมูลไม่เกี่ยวข้องกับบุคคล ผู้ใช้ระบบสามารถที่จะเข้าถึงพวกมันได้อย่างเท่าเทียมกัน

ในทางตรงกันข้าม สารสนเทศเป็นชุดของข้อมูลซึ่งได้ถูกคัดหรือเลือกไว้ตามความต้องการสารสนเทศอันหนึ่งๆ แนวความคิดของสารสนเทศมีทั้งองค์ประกอบที่เป็นบุคคลและขึ้นกับเวลาซึ่งจะไม่มีในแนวความคิดของข้อมูล ยกตัวอย่างเช่น แม้ว่าระบบสามารถจัดหาสูตรอาหารให้กับผู้ใช้ได้ แต่สิ่งเหล่านั้นจะไม่ใช่สารสนเทศถ้าผู้ใช้รู้ข้อมูลเหล่านั้นแล้ว (ไม่มีประโยชน์ต่อผู้ใช้) หรือถ้ามันไม่ตรงต่อความต้องการสารสนเทศของผู้ใช้ ข้อมูลเหล่านี้กลายเป็นสิ่งรบกวน (Noise) ในระบบขัดขวางรบกวนการทำงานของผู้ใช้ อีกทั้งทำให้ระบบตอบสนองต่อความต้องการของผู้ใช้ล่าช้า

นอกเหนือจากข้อมูลและสารสนเทศแล้ว ยังมีคำอีกสามคำจัดเรียงลำดับชั้นตามความซับซ้อน คือสัญญาณ (Signal) ความรู้ (Knowledge) และความฉลาด (Wisdom) สัญญาณเป็นสิ่งที่มีความซับซ้อนน้อยกว่าข้อมูล สัญญาณจะถูกส่งจากที่หนึ่งไปยังอีกที่หนึ่ง ระหว่างการประมวลสารสนเทศ สัญญาณนี้อาจจะเป็นชุดของบิต รูปแบบของคลื่นแม่เหล็กไฟฟ้าหรือรูปแบบอื่นๆ โดยทั่วไปเรื่องเกี่ยวกับสัญญาณมักจะเป็นภาระหน้าที่ของวิศวกรสื่อสารที่จะทำให้การส่งสัญญาณจากจุดหนึ่งไปอีกจุดหนึ่งเป็นไปได้อย่างเชื่อถือได้ ซึ่งการส่งนี้ไม่เกี่ยวข้องกับเนื้อหา (Content) ที่ต้องการส่งสาขาวิชานี้ควรจะเรียกกันว่าเป็นทฤษฎีเกี่ยวกับการส่งสัญญาณ (Transmission Theory) แต่ Claude Shannon (1984) ได้เรียกไว้ในครั้งแรกว่าเป็น ทฤษฎีการสื่อสาร (Communication Theory) ปัจจุบันนี้นิยมเรียกกันว่าทฤษฎีสารสนเทศ (Information Theory) นักวิจัยในสาขาวิชานี้เน้นเรื่องคุณสมบัติทางสถิติของสัญญาณเพื่อที่ทำให้การส่งสัญญาณเป็นไปได้อย่างน่าเชื่อถือ การใช้คุณสมบัติเหล่านี้ทำให้เกิดการพัฒนารหัสส่งสัญญาณ ซึ่งสามารถตรวจสอบและแก้ไขความผิดพลาดระหว่างการส่งได้คำว่า สัญญาณรบกวน (Noise) ถูกใช้เพื่อแสดงความผิดพลาดของการส่งสัญญาณที่มีต่อสัญญาณดั้งเดิม สัญญาณที่ถูกรับอีกฝั่งหนึ่งประกอบไปด้วยสัญญาณเดิมรวมกับสัญญาณรบกวน (ซึ่งมักมีอยู่เพียงเล็กน้อย)

นอกเหนือจากสัญญาณ ข้อมูล และสารสนเทศ ยังมีความรู้ ความรู้ถูกสร้างขึ้นบน

สารสนเทศ โดยรวมสารสนเทศใหม่ๆ เข้ากับสารสนเทศเดิมซึ่งรับรู้กันอยู่แล้ว เป็นการรวมกันเพื่อที่จัดรูปแบบแสดงบางส่วนของความเป็นจริง ดังนั้นขณะที่สารสนเทศถูกทำให้เป็นท้องถิ่นในการตอบสนองต่อคำขอที่ระบุ ความรู้จะมีขอบเขตที่กว้างกว่าคนที่ทำงานทางด้านปัญญาประดิษฐ์ (Artificial Intelligence) จะพูดถึงฐานความรู้ (Knowledge Base) ฐานความรู้โดยทั่วไปจะถูกสร้างโดยความพยายามที่จะรวบรวมความจริง แนวความคิด กฎเกณฑ์ ซึ่งเป็นตัวแทนของความชำนาญในสาขาหนึ่งๆ มาเก็บในรูปแบบข้อมูลและอัลกอริธึม ระบบที่ถูกสร้างขึ้นบนฐานความรู้เหล่านี้เราเรียกว่าระบบผู้ชำนาญ (Expert Systems)

ท้ายที่สุดคำว่า ความฉลาด (Wisdom) เป็นสิ่งที่กว้างกว่าความรู้ เป็นการรวมความเป็นจริงที่รู้กันทั้งหมดและบริหารการใช้สารสนเทศซึ่งได้รับมา และความรู้ที่ได้ถูกพัฒนาไว้จะเกี่ยวข้องกับความสามารถที่จะตัดสินได้อย่างสมดุลในสถานการณ์ต่างๆ *(Lochen 1974; Debons, Home, and Croenweth 1988)* เท่าที่ทราบไม่มีความพยายามที่จะสร้างความฉลาดเข้าไปในระบบสารสนเทศใดๆ

ลูกโซ่จะเริ่มจากสัญญาณและจบลงที่ความฉลาด เนื่องจากสารสนเทศมีแง่มุมของบุคคล และระบบสารสนเทศทำงานกับข้อมูลเท่านั้น คำว่าการจัดเก็บและค้นคืนสารสนเทศจึงสะท้อนความหวังมากกว่าความเป็นจริง ผู้ใช้คนแรกสร้างและจัดเก็บข้อมูลที่เป็นตัวแทนสารสนเทศซึ่งเขาหวังที่จะรักษาไว้ ผู้ใช้คนที่สองค้นคืนข้อมูลเหล่านี้ ความหวังของผู้ใช้รายที่สองคือการค้นคืนสารสนเทศ นั่นคือข้อมูลซึ่งตรงตอบสนองความต้องการเฉพาะของเขา สารสนเทศที่ผู้ใช้คนที่สองค้นหาไม่จำเป็นต้องเป็นสารสนเทศที่ผู้ใช้คนแรกพยายามที่จะแสดงให้เห็นในข้อมูลนั้น ความรู้และความฉลาดเกี่ยวข้องกับผู้ใช้แต่ละคนอย่างมากจนกระทั่งมันอยู่เหนือขอบเขตของระบบค้นคืนสารสนเทศ (และบทความนี้) แม้กระนั้นการพัฒนาความรู้และความฉลาดขึ้นอยู่กับการเข้าถึงสารสนเทศที่ดี การพัฒนานี้ขึ้นอยู่กับระบบการจัดเก็บและค้นคืนสารสนเทศที่มีประสิทธิภาพและประสิทธิผล

ความต้องการสารสนเทศที่มีอยู่อย่างหลากหลายซึ่งระบบจะต้องตอบสนองทำให้การพัฒนาระบบเป็นสิ่งที่ท้าทายมากโดยเฉพาะอย่างยิ่งระบบที่ตั้งใจจะให้คนทั่วไปๆใช้ (ตรงข้ามกับระบบที่ออกแบบมาสำหรับกลุ่มผู้ใช้เล็กๆ กลุ่มหนึ่งที่มีพื้นฐานเหมือนๆ กัน)

5. เพื่อระบบค้นคืนสารสนเทศที่มีประสิทธิภาพ

ไม่ว่าจะเป็นการออกแบบระบบค้นคืนสารสนเทศใหม่ การประเมินระบบที่มีอยู่ หรือแม้แต่การพยายามใช้ระบบ สิ่งสำคัญที่เราจะต้องเข้าใจคือส่วนประกอบของระบบ และวิธีที่สามารถถูกนำมาใช้กับระบบได้ ประเภทของการทดสอบที่ถูกนำมาใช้ในการออกแบบระบบค้นคืนและการประเมินผล (ซึ่งสามารถใช้กับการออกแบบระบบทุกระบบ) จะรวมถึงตุ๊กตาการศึกษาการวิเคราะห์ การจำลอง การทดสอบในห้องทดลอง และท้ายที่สุดการทดสอบกับผู้ใช้ การทดสอบกับผู้ใช้เป็นประโยชน์อย่างมากต่อระบบค้นคืน เนื่องจากเรายังไม่สามารถตุ๊กตาแนวความคิดของประสิทธิภาพ (Concept of Effectiveness) ของผู้ใช้เข้าไปในการทำงานของระบบค้นคืนสารสนเทศอย่างมีประสิทธิภาพ

ระบบอาจจะทำงานมากมายหลายอย่างแต่เป้าหมายสูงสุดคือการพยายามที่จะจับคู่สารสนเทศกับความต้องการสารสนเทศโดยผ่านการจับคู่เอกสารและคำขอการพิจารณาตรวจสอบรูปแบบซึ่งเอกสารและคำขอสามารถ

จัดอยู่ได้จึงมีความสำคัญ เพื่อให้ได้การประมวลผลที่มีทั้งประสิทธิภาพและประสิทธิผล

งานวิจัยการค้นคืนสารสนเทศส่วนใหญ่เน้นไปในเอกสารที่เป็นข้อความ ดังนั้นแนวความคิดและโปรซีเยอร์อธิบายในงานวิจัยมักอยู่ในรูปแบบนั้น อย่างไรก็ตามข้อมูลที่เป็นภาพนิ่งและเสียงก็เป็นที่สนใจมากขึ้นและมีการศึกษากรรมวิธีที่จะสามารถระบุและค้นคืนข้อมูลเหล่านั้นได้

เอกสารจะถูกนำเข้ามายังระบบค้นคืนจากแหล่งภายนอกและมักมีแบบ (Forms) ที่ถูกกำหนดไว้ล่วงหน้า แบบเหล่านี้จะมีผลไปถึงแบบของคำขอและจะอยู่ภายใต้การควบคุมของนักออกแบบระบบ มีแบบของคำขอจำนวนมากที่ได้มีการทำการศึกษาไว้ ตั้งแต่แบบซึ่งมีประสิทธิผลในแง่ของการคำนวณการประมวลผลแต่ผู้ใช้อาจจะเข้าใจได้ยาก ไปจนถึงภาษาธรรมชาติซึ่งจะผลักภาระไปให้ระบบแทน แต่ว่าผู้ใช้จะรู้สึกง่ายในการระบุคำขอเนื่องจากเป็นสิ่งที่เขาคุ้นเคยอยู่แล้ว

วิธีของการจับคู่เอกสารกับคำขอมีความสัมพันธ์อย่างใกล้ชิดกับแบบของคำขอที่ใช้ ปัจจัยหลายชนิดมีผลต่อคุณภาพของการจับคู่

การวิเคราะห์เอกสาร (เพื่อเตรียมตัวต่อการจับคู่กันกับคำขอ) เกี่ยวข้องกับงานหลายขั้นตอน ขั้นตอนเหล่านี้โดยทั่วไปเริ่มที่ระดับตัวอักษร หาจำนวนของการปรากฏคำ หรือวลีเหล่านั้นในเอกสาร ข้อมูลเหล่านี้สามารถถูกนำไปใช้คำนวณคุณค่าของคำ (Term Value) เพื่อใช้สำหรับกระบวนการการจับคู่ คุณค่าของคำเป็นตัวเลขที่แสดงให้เห็นความสำคัญของคำในเอกสารนั้นๆ เทคนิคอื่นๆ ที่อาจจะนำมาใช้เพิ่ม เช่น ลักษณะทางภาษา ลักษณะการใช้ภาษาและฮิวริสติกต่างๆ เพื่อช่วยในการพิจารณา เนื้อเรื่องและนัยสำคัญของเอกสารรวมไปถึงความสัมพันธ์ระหว่างเอกสารกับคำขอด้วย เทคนิคเหล่านี้สามารถใช้ได้กับภาษามนุษย์หลายภาษาและเป็นที่สนใจเพิ่มมากขึ้น เนื่องจากสารสนเทศที่เป็นอิเลกทรอนิคส์ครอบคลุมไปทั่วโลก

ผู้ใช้ระบบเป็นศูนย์กลางของความสำเร็จของระบบค้นคืนสารสนเทศ แต่สารสนเทศเกี่ยวกับผู้ใช้แต่ละคนกลับถูกละเลยมาเป็นเวลานานไม่ค่อยมีผู้สนใจศึกษา ขณะนี้ผู้ใช้เริ่มเป็นองค์ประกอบที่สำคัญในการออกแบบระบบ ไม่ใช่แต่การให้ผู้ใช้ทดสอบระบบเท่านั้นที่มีความสำคัญขึ้น แต่วิธีต่างๆ ซึ่งนำความชอบส่วนตัว (Preference) และพื้นฐาน (Background) ของผู้ใช้เข้าไปในกระบวนการค้นคืนก็เป็นที่สนใจมากขึ้นด้วย

งานที่เกี่ยวข้องกับความสนใจและพื้นฐานของผู้ใช้ได้นำไปสู่แนวความคิดของการใช้จุดอ้างอิงหลายจุด (Multiple Viewpoint) แทนที่จะใช้คำขอเพียงอย่างเดียวควบคุม กระบวนการค้นคืน จุดอ้างอิงหลายจุดทำให้เกิดโครงสร้างการค้นคืนที่ซับซ้อนมากขึ้นกว่าโครงสร้างแบบเดิมซึ่งเป็นเพียงรายชื่อของเอกสารตัวแทนของโครงสร้างที่ซับซ้อนนี้ก่อให้เกิดแรงจูงใจอย่างมากในการพัฒนาการเชื่อมต่อ (Interface) ของระบบการค้นคืนด้วยภาพ *(Visual, ดู Korfhage 1997 บทที่ 7)*

การวัดผลแบบเดิมที่เรียกกันว่าความแม่นยำและความระลึก ทั้งสองตัวนี้แม้มีข้อบกพร่องค่อนข้างมากแต่ยังคงใช้ในการประเมินผลระบบค้นคืนส่วนใหญ่ วิธีการวัดผลแบบอื่นซึ่งมีความสัมพันธ์โดยตรงต่อความชอบของผู้ใช้หรือวิธีซึ่งพิจารณาถึงลำดับของเอกสารและผลทางด้านความสัมพันธ์ของเอกสารก็มีการใช้กันบ้าง นักวิจัยบางคนสนับสนุนให้มีการประเมินทางด้านคุณภาพของความพึงพอใจ

ของผู้ใช้แต่ผลกระทบของมันต่อการออกแบบและการใช้ระบบยังมีไม่มาก

ระบบค้นคืนสารสนเทศถ้าประเมินด้วยวิธีที่ยอมรับกันโดยทั่วไปจะพบว่ามันไม่มีประสิทธิภาพ เมื่อประสิทธิภาพของระบบได้ถูกวัดแล้ว ขั้นตอนที่ชัดเจนอันต่อมาก็คือการพยายามที่จะปรับปรุงประสิทธิภาพของมัน นักวิจัยได้พยายามที่จะทดลองอัลกอริธึมและโครงสร้างข้อมูลใหม่ๆ อยู่เสมอ อย่างไรก็ตาม การปรับปรุงอย่างมีนัยสำคัญในประสิทธิภาพของระบบค้นคืนสารสนเทศใดๆ สามารถทำได้โดยให้ผู้ใช้เข้ามาเกี่ยวข้องในกระบวนการโดยตรง กระบวนการที่ใช้กันส่วนใหญ่ถูกเรียกว่า การป้อนกลับความเกี่ยวข้อง (Relevance Feedback) ซึ่งผู้ใช้จะประเมินตัวอย่างของเอกสารที่ถูกค้นคืนออกมา และการประเมินนี้จะถูกนำไปใช้เพื่อแก้ไขกระบวนการการค้นคืน

ขณะที่มักใช้การจับคู่ของคำจากเอกสารกับคำที่อยู่ในคำขอในการทำการค้นคืน ก็มีเทคนิคอื่นๆ ที่ถูกศึกษาและนำเข้ามาใช้ในระบบบางระบบ ได้มีความพยายามจำนวนมากในการที่จะทำการวิเคราะห์ที่ลึกซึ้งของภาษาธรรมชาติของมนุษย์ที่ใช้ในเอกสารกับคำขอ และปรากฏว่าได้ผลลัพธ์ที่ดีในหลายๆ ระบบ บางระบบได้มีการใช้ เอกสารอ้างอิง (Bibliographic Citation) เพื่อที่จะเชื่อมเอกสารหนึ่งไปยังเอกสารที่คล้ายกัน มีการใช้การเชื่อมต่อของ Hypertext กันอย่างมากมายใน WWW ซึ่งชี้โดยตรงจากคำหรือแนวความคิดอันหนึ่งไปยังเอกสารหรือกลุ่มของเอกสาร เนื่องจากจำนวนของข้อมูลที่มากมายทำให้การกรองสารสนเทศ (Filtering) หรือการขุดข้อมูล (Data Mining) กลายมาเป็นส่วนที่สำคัญในการทำการค้นคืนให้มีประสิทธิภาพ ในแง่มุมของการค้นคืนสารสนเทศคำว่า การกรองสารสนเทศหมายถึงการเลือกเอกสารอย่างรวดเร็วและมีค่าใช้จ่ายต่ำ ทำการเลือกมันออกจากเอกสารจำนวนมหาศาลเพื่อที่จะไปประมวลผลให้ตอบสนองต่อความต้องการในลำดับต่อไป ขณะที่เอกสารที่เป็นสื่อประสม (Multimedia) ได้ถูกนำเข้ามาในระบบค้นคืนมากขึ้น ความสามารถที่จะประมวลภาพนิ่งและเสียงก็ได้กลายเป็นสิ่งที่สำคัญเช่นเดียวกับการประมวลตัวอักษร

ปัจจัยหลักอีกอันในการยอมรับต่อระบบของผู้ใช้คือ การเชื่อมต่อ ซึ่งผู้ใช้จะใช้ในการโต้ตอบกับระบบ การเชื่อมต่อแบบดั้งเดิมที่เป็นแบบตัวอักษรล้วนๆ ยังเป็นที่ใช้งานกันอยู่ทั่วไป ทั้งๆ ที่มันมีข้อบกพร่องมากมาย อย่างไรก็ตาม การเชื่อมต่อด้วยภาพ (Graphical Interface) ก็ได้ถูกพัฒนาโดยกลุ่มนักวิจัยอย่างต่อเนื่อง ปัจจุบันได้มีการศึกษาเกี่ยวกับยอมรับของผู้ใช้ต่อการเชื่อมต่อ เป็นจำนวนมาก

ในสมัยก่อนคอมพิวเตอร์มีข้อจำกัดมากมายไม่ว่าในแง่ของหน่วยความจำหรือความสามารถในการประมวลผล ทำให้ระบบค้นคืนในยุคก่อนมักจะแสดงเพียงแค่ตัวแทนบางส่วนของเอกสารและชี้ว่าเอกสารที่สมบูรณ์อยู่ที่ใด โดยที่ผู้ใช้จะต้องไปค้นหาเองต่อไป อย่างไรก็ตามด้วยความสามารถ ของสื่อเช่น ซีดีรอมที่สามารถบรรจุข้อมูลเป็นจำนวนมาก และเครือข่ายที่มีการเชื่อมต่อกันทั่วโลกอย่างอินเตอร์เน็ตทำให้ระบบใหม่ๆ จำนวนมากจะส่งเอกสารข้อความเต็ม (Fulltext) ให้กับผู้ใช้

ท้ายที่สุดระบบค้นคืนสารสนเทศที่สลับซับซ้อนได้ก่อให้เกิดปัญหาทางด้านจริยธรรมและนโยบายซึ่งมักถูกละเลยมาในอดีต เรื่องราวเหล่านี้รวมถึงลิขสิทธิ์ การละเมิดสิทธิส่วนบุคคลและความปลอดภัย ผู้มีอาชีพทางด้านสารสนเทศนายทุนและผู้ใช้ทุกคนทุกกลุ่มควรที่จะตระหนักในเรื่องเหล่านี้ และผลกระทบซึ่งอาจจะมีต่อการค้นคืน และควรจะศึกษาหาวิธีเพื่อประกันว่าทุกคนที่เกี่ยวข้องกับการสร้าง จัดเก็บ และค้นคืนสารสนเทศถูกปฏิบัติอย่างเท่าเทียมกัน

เอกสารอ้างอิง

Baker, Norman R. 1968. A descriptive model of library/ user/ funder behavior in a university environment. *Drexel Library Quarterly* 4:16-30.

Blair, David C., and M.E. Maron. 1985. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communicatons of the ACM 28,* no. 3 : 289-299.

Bottle, R.T. 1965. A user'r assessment of current awareness services. Journal of Documentation. 21 : 151-162.

Bourne, C.P., and D.F. Ford. 1964. Cost analysis and simulation procedures for the eveluation of large information systems. *American Documentation 15,* no. 2 : 142-149.

Charoenkitkarn, N. 1996. *The effect of markup-querying on search pattern performance in large-scale text retrieval.* Ph.D. Dissertation. Department of Mechanical and Industrial Engineering, University of Toronto.

Cooper, Michael D. 1972. A cost model for evaluating information retrieval systems *JASIS 23,* no. 5 : 306-312.

Debons, Anthony, Esther Horne, and Scott Croenweth. 1988. *Information Science, and integrated view.* Boston : G.K. Hall.

Dumsis, Susan. 1994. Evaluating interactive retrieval systems (panel Abstract). In *proceedings of the 17th Annual International ACM/ SIGIR Conference on Research and Development in Information Retrieval,* Dublin, Ireland, p. 361.

Eisenberh, Michael B., and Xiulan Hu. 1987. Dichotomous relevance judg ments and the evaluation of information systems. *In Proceedings of the 50th ASIS Annual Meeting,* Boston, pp. 66-70.

Frei, Hans-Peter, and Peter Schauble. 1991. Determining the effectiveness of retrieval algorithms. *Information Processing & Management 27,* no. 2/3 :153-164.

Frei, Hans-Peter, and M.F. Wyle. 1991. Retrieval algorithm effectiveness in a wide area network information filter. *In Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Design in Information Retrieval,* Chicago, pp. 114-122.

Hamilton, S., and N.L. Chervany. 1971. Evaluating Information system effectiveness. *MIS quarterly 5,* no. 4 : 649-652.

Harman, D. (Ed.) *Proceedings of the Fifth Text REtrieval Conference (TREC-5)* NIST Special Publication 500-238. Gaithersburg, Maryland : National Institute of Standards, November 1997.

Harman, D. (Ed.) *Proceedings of the Fourth Text REtrieval Conference (TREC-4).* NIST Special Publication 500-236. Gaithersburg, Marylnad : National Institute of Standards, October 1996.

Harman, D. (Ed.) *Proceedings of the Third Text REtrieval Conference (TREC-3).* NISt Special Publication 500-225. Gaithersburg, Maryland : National Institute of Standards, April 1995.

Harman, D. (Ed.) *Proceedings of the Second Text REtrieval Conference (TREC-2).* NIST Special Publication. 500-215. Gaithersburg, Maryland : National Institute of Standards, March 1994.

Harman, D. (Ed.) *Proceedings of the First Text REtrieval Conference (TREC-9).* NIST Special Publication 500-207. Gaithersbert, Maryland : National Institute of Standards, March 1993.

Heaps, H.S. 1971. Criteria for optimum effectiveness of information retrieval systems. *Information and Control* 18:156-167.

Kochen, Manfred. 1974. *Principles of information retrieval.* Los Angeles: Melville.

Korfhage, R.R. 1997. *Information Storage and Retrieval.* John Wiley & Sons, Inc.

Korfhage, Robert R., and Thomas G. DeLutis. 1969. A basis for time and cost evaluation of information systems. In *The Information Bazaar. Proceedings of the sixth Annual National Colloquium on Information Retrieval,* ed. Louise Schultz. Medical Documentation Service, The college of Physicians of Philadelphia, pp. 293-326.

Kraft, Donald H., and Abraham Bookstein. 1978. Evaluation of information retrieval systems : A decision theory approach. *JASIS 29,* no.1 : 31-40.

Lancaster, F. Wilfrid. 1971. The cost-effectiveness analysis of information retrieval and dissemination systems. *JASIS 22,* No.1 : 12-27.

Lancaster, F. Wilfrid, and W.D. Climenson. 1968. Evaluating the economic efficiency of a document retrieval system. *Journal of Documentation 24,* no.1 : 16-40.

Losee, Robert M., Jr. 1991. an analytic measure predicting information retrieval system performance. *Information processing & management 27,* no.1 :1-13.

McCain, Kate W., Howard D. White, and Belver C. Griffith. 1986. Text retrieval as a measure of system performance : MEDLINE and the medical behavioral sciences. In *Proceedings of the 49th ASIS Annual meeting*, pp. 199-203.

Meadow, Charles T. 1973. *Analysis of information systems*, 2d ed. New York: Wiley.

Nance, Richard E. 1967. Strategic simulation of a library / user / funder system. Ph.D. diss., Purdue University. West Lafayette, Indiana.

Shannon, Claud E. 1948. A Mathematical theory of communication. *Bell Systems Technical Journal* 27:379-423, 623-656.

Salton, G. 1989. Automatic text processing. The transformation, analysis and retrieval of information by computer. Addison-Wesley, Reading, Mass.

Salton, G., and McGill, M. 1983. Introduction to modern information retrieval McGraw-Hill.

Shaw, W.M., Jr. 1986. On the foundation of evaluation. *JASIS 37*, no.5: 346-348.

Sparck-Jones, K. 1981. Information retrieval experiment. London : Butterworths.

Tague, Jean M., and R. Schultz. 1988. Some measures and procedures for evaluation of the user interface in an IR system. *In Proceedings of the 11th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Greoble*, France, pp. 371-385.

Taylor, Robert S. 1986. *Value-added processes in information systems.* Norwood, New Jersey : Ablex.

van Rijsbergen, C.J. 1979. Information retrieval. London : Butterworths.

กนกวรรณ โสตถีวรกุล 2540 "*การค้นหาข้อมูลจากหนังสือพระคริสต์ธรรมคัมภีร์*" โครงการพิเศษตามหลักสูตรวิทยาศาสตรมหาบัณฑิต คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี กรุงเทพฯ.

สุวัฒน์ สถาพรพิริยะเดช 2540 "*ผลกระทบของลักษณะคำถามที่มีต่อประสิทธิภาพในการสืบคืนข้อมูลภาษาไทย*" วิทยานิพนธ์หลักสูตรวิทยาศาสตรมหาบัณฑิต คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยี พระจอมเกล้าธนบุรี กรุงเทพฯ