บทที่ 2
การค้นคืนข้อมูล

2.1 ภาพรวมของระบบค้นคืนข้อมูล

ระบบค้นคืนข้อมูล เกี่ยวข้องกับการ นำเสนอข้อมูล การจัดเก็บ การจัดการ และ การเข้าถึงข้อมูล การนำเสนอและการจัดการกับข้อมูล จำเป็นต้องให้ผู้ใช้เข้าถึงข้อมูลที่สนใจได้ง่าย แต่สิ่งนี้ความเป็นจริงแล้วไม่ใช้สิ่งง่าย เพราะ พฤติกรรมความต้องการของผู้ใช้นั้นย่อมมีความแตกต่างกันไป เป็นเพราะข้อมูลที่ผู้ใช้ในการค้นหาข้อมูลนั้นส่วนใหญ่จะเป็นภาษาไม่มีระเบียบทางด้านโครงสร้างของภาษา ดังนั้นต้องมีระบบค้นคืนที่สามารถแปลงความต้องการของผู้ใช้ที่อยู่ในรูปภาษาที่ไม่มีระเบียบทางด้านโครงสร้างให้เป็นภาษาที่มีโครงสร้างเพื่อใช้ในการค้นคืนข้อมูล

2.2 ความแตกต่างระหว่าง Information Retrieval กับ Data Retrieval

ในการค้นคืนข้อมูลแบบ Data Retrieval นั้นผู้ใช้จำเป็นต้องระบุ คำสำคัญ(key word) เพื่อใช้ในการค้นหา แต่ การค้นคืนข้อมูลแบบ information retrieval นั้นสามารถใช้ข้อมูลที่ไม่เป็นโครงสร้างและไม่จำเป็นต้องเป็นคำสำคัญ ดังนั้นในการค้นคืนข้อมูลแบบ Data Retrieval นั้นผู้ใช้จำเป็นต้องเข้าใจวิธีการสร้างคำสำคัญเพราะไม่เช่นนั้นแล้วการค้นคืนข้อมูลจะไม่สามารถกระทำได้เพราะ ข้อมูลที่ใช้ในการค้นหานั้นต้องมีความสัมพันธ์แบบตรงกันทุกประการ ยกตัวอย่างเช่น ในระบบการค้นคืนข้อมูลจำเพาะ ถ้าผู้ใช้ต้องการหาข้อมูลเกี่ยวกับ Internet ผู้ใช้อาจใช้คำว่า ‘Internet’ เป็นข้อมูลในการค้นหา แต่ผู้ใช้ไม่อาจรู้ได้เลยว่าคำว่า ‘Internet’ เป็นคำสำคัญหรือไม่ ถ้าคำว่า ‘Internet’ ไม่ใช่คำสำคัญ ก็ไม่สามารถใช้คำว่า ‘Internet’ ในการค้นหาได้

ถึงแม้ว่าการค้นคืนข้อมูลแบบ Data Retrieval จะใช้ได้ดีในระบบฐานข้อมูลแต่ก็มีจุดเสียที่ไม่สามารถค้นหาโดยใช้หัวเรื่อง(subject) หรือหัวข้อ(title) ได้ ดังนั้นในการแก้ไขปัญหานี้ จึงเกิดระบบค้นคืนข้อมูลที่สามารถสรุปความหมายของเอกสารนั้น ๆ เพื่อระบุหัวเรื่อง หรือหัวข้อของเอกสารนั้น ๆ ได้ โดยจะนำเอาข้อมูลที่ได้จากการตีความหมายมาเป็นคิดความสัมพันธ์กับข้อมูลการค้นหาของผู้ใช้

โดยสรุปแล้วระบบการค้นคืนข้อมูลแบบ Information Retrieval นั้นจะเน้นความสัมพันธ์กับผู้ใช้เป็นหลัก โดยที่ระบบต้องสามารถค้นคืนเอกสารที่มีความสัมพันธ์กับข้อมูลของผู้ใช้ได้ทั้งหมด และต้องไม่นำเอกสารที่ไม่มีความสัมพันธ์กับข้อมูลของผู้ใช้มาแสดง หรือ มาแสดงน้อยที่สุด

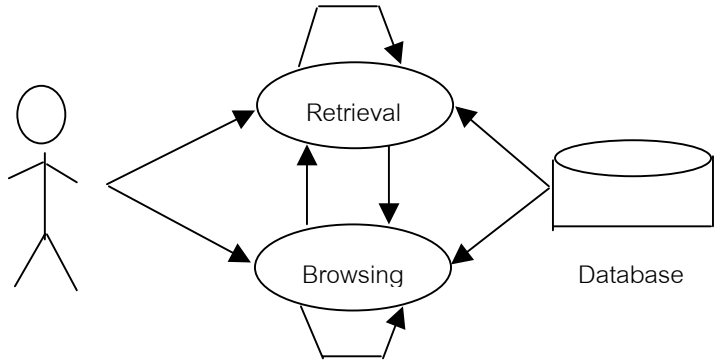
2.3 แนวคิดพื้นฐานของระบบค้นคืนข้อมูล

สิ่งที่มีผลต่อการค้นหาข้อมูลที่สัมพันธ์กันจะขึ้นอยู่กับปัจจัย 2 สิ่งด้วยกันนั่นคือ งานที่เกิดจากผู้ใช้ และ มุมมองในการค้นคืนเอกสาร(logical view of document)

2.3.1 งานที่เกิดจากผู้ใช้

จะแบ่งงานของผู้ใช้ออกเป็นได้ 2 แบบใหญ่ ๆ ด้วยกันคือ

1. การค้นคืนข้อมูล
2. การ browse ข้อมูล



รูปที่ 2-1 งานของผู้ใช้ที่มีต่อระบบ

จากรูป 2-1 จะเห็นได้ว่างานของผู้ใช้นั้นแบ่งเป็น 2 ประเภทด้วยกัน คือการค้นคืนข้อมูล และการ browse ข้อมูล ซึ่ง

การค้นคืนข้อมูล(retrieval) งานของผู้ใช้ที่ต้องแปลความต้องการของผู้ใช้ให้เป็นภาษาที่สามารถใช้ในการค้นคืนข้อมูลได้ทั้งในรูปแบบการค้นคืนข้อมูลแบบ Information Retrieval และการค้นคืนข้อมูลแบบ Data Retrieval เราจะเรียกสิ่งนี้ว่าเป็นงานของการค้นคืน

การ browse ข้อมูล คือการค้นคืนข้อมูลที่มีโต้ตอบระหว่างผู้ใช้กับระบบ ซึ่งปัจจุบันเป็นส่วนจำเป็นมากขึ้นซึ่งอาจเนื่องมาจากการค้นคืนข้อมูลแบบปัจจุบัน จะเป็นการค้นคืนโดยให้ข้อมูลแบบกว้าง ๆ ดังการค้นหาข้อมูล url ของ web site ดังนั้นการโต้ตอบกับผู้ใช้จะสามารถอำนวยความสะดวกกับผู้ใช้ได้มากขึ้น

2.3.2 มุมมองในการค้นคืนเอกสาร

มุมมองในการค้นคืนเอกสารนั้นมีด้วยกันหลายวิธี อย่างแรกเป็นวิธีการสร้าง index term โดยการนำคำที่อยู่ในเอกสารที่มีค่าความถี่สูง ออกมาเป็น index term หรือ อาจเป็นการระบุโดยคน ซึ่งผลที่ได้ยังเป็นผลที่มีประสิทธิภาพในการค้นคืนที่ต่ำอยู่ ต่อมา ประสิทธิภาพของเครื่องคอมพิวเตอร์มีมากขึ้น ดังนั้นมุมมองในการค้นคืนเอกสาร จะนำเอาข้อความทั้งหมดที่มีอยู่ในเอกสารนั้นมาเป็นมุมมองในการค้ฯคืน แต่ข้อเสียคือ เนื้อที่ในการจัดเก็บ เวลาในการค้นหา และที่สำคัญคือ คำบางคำจะมีทุก ๆ เอกสารเช่น article ‘the’

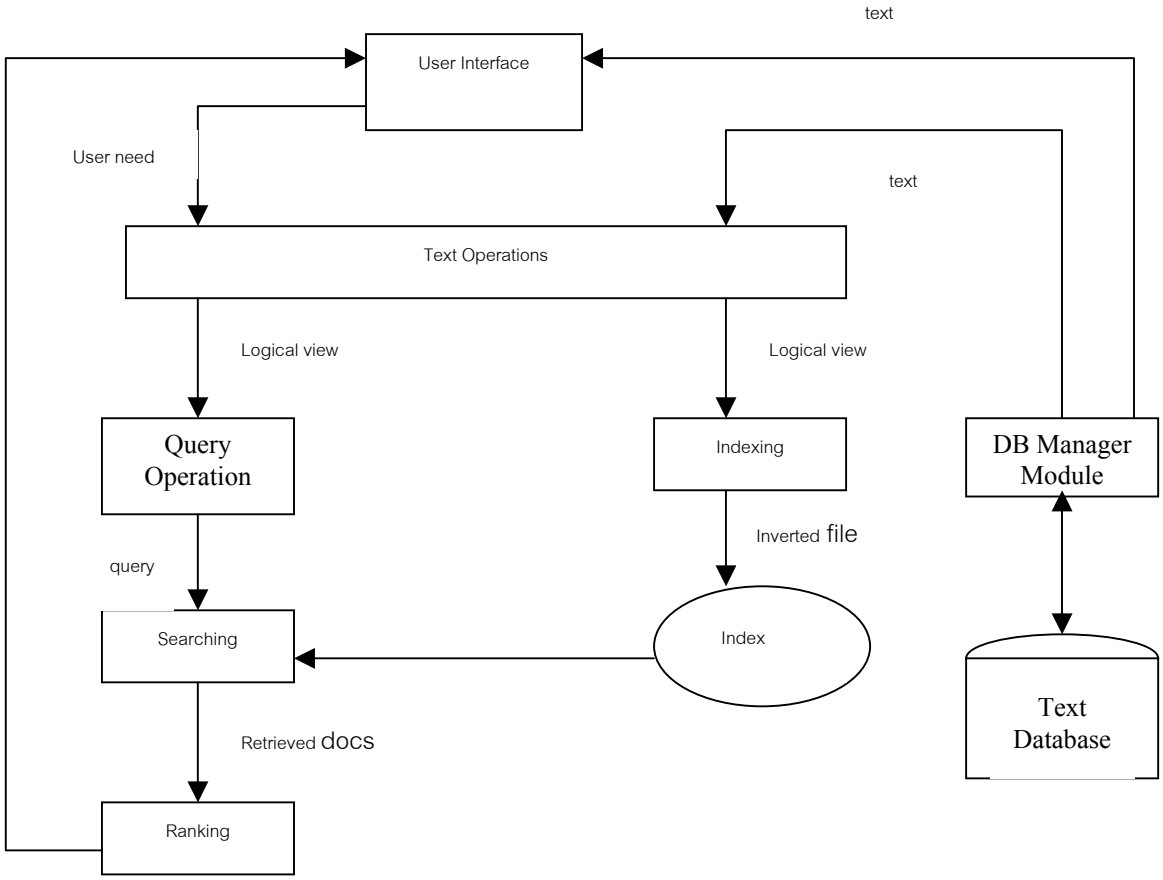
ดังนั้นในมุมมองสมัยใหม่ จะพยายามลดจำนวน keyword โดยใช้วิธีการกำจัดคำ(stop word) ในที่นี้หมายถึงการกำจัด คำจำพวก article หรือ คำเชื่อมประโยค with และอื่น ๆ วิธี stemming คือการลดขนาดของคำเช่น normally ไปเป็น normal เป็นต้น วิธีการการทำหนดกลุ่มของคำนาม โดยการกำจัดคำขยายกริยา กำจัดคำขยายคำนาม เป็นต้น นอกจากนี้แล้วจำเป็นต้องแบ่งแยกโครงสร้างของเอกสาร ได้แก่ บท หัวข้อ หัวข้อย่อย ในการทำกระบวนการเช่นนี้ไม่สามารถทำได้โดยคนทั้งหมด จึงได้เกิดรูปแบบที่เรียกว่า การทำ Automatic indexing

Automatic Indexing

การจัดทำ automatic indexing คือการโปรแกรมคอมพิวเตอร์โดยให้โปรแกรมนั้นสามารถที่จะรับอินพุทของระบบเป็นเท็กซ์ไฟล์ (text file) และสามารถที่จะให้ out put ของระบบออกมาเป็น index term ได้ ซึ่ง index term นั้นจะเป็นตัวแทนของเอกสารที่จะบอกได้ว่าเอกสารนั้นกล่าวถึงหรือเกี่ยวข้องสัมพันธ์กับเรื่องใด

โดยส่วนมากแล้วการจัดทำอินเด็กซ์อัตโนมัตินั้นจะจัดทำเพื่อที่จะนำไปใช้เกี่ยวกับการจัดเก็บข้อมูลและค้นคืนข้อมูล (information retrieval) ให้ได้เร็วที่สุด และเป็นการจัดเก็บข้อมูลให้เป็นหมวดหมู่ที่มีประสิทธิภาพ เช่นในระบบของห้องสมุดสมัยใหม่ ก็จะทำการจัดเก็บข้อมูลพร้อมทั้ง index term ของหนังสือหรือเอกสารนั้น ๆ เมื่อมีผู้ใช้ต้องการเข้ามาค้นหาข้อมูลต่าง ๆ ก็ทำการใส่ข้อมูลในการค้นหา และระบบจะทำการเปรียบเทียบ ข้อมูลที่ผู้ใช้ใส่เข้ามากับตัว index term ที่ได้มีการจัดเก็บไว้แล้ว เพื่อที่จะดึงข้อมูลที่ตรงกับสิ่งที่ผู้ใช้ต้องการได้

2.4 กระบวนการค้นคืนข้อมูลเบื้องต้น



รูปที่ 2-2 กระบวนการค้นคืนเบื้องต้น

จากรูป 2-2 ได้อธิบายระบบค้นคืนสารสนเทศเบื้องต้นที่เป็นเพียงระบบที่ยังไม่ได้มีโครงสร้างอื่นมาขยายประสิทธิภาพในการค้นคืนข้อมูลโดยรูปข้างต้นนี้จะเป็นพื้นฐานในการนำเอาไปพัฒนาระบบอื่นๆ ที่มีประสิทธิภาพได้ โดยวิทยานิพนธ์ฉบับนี้จะนำเอาโครงสร้างนำไปพัฒนาระบบค้นคืนสารสนเทศที่ใช้ทฤษฎี ฟัซซี่ มาขยายประสิทธิภาพในการค้นคืนซึ่งจะได้อธิบายไว้ในบทที่ 7