บทที่ 1
ภาพรวมของระบบค้นคืนสารสนเทศ

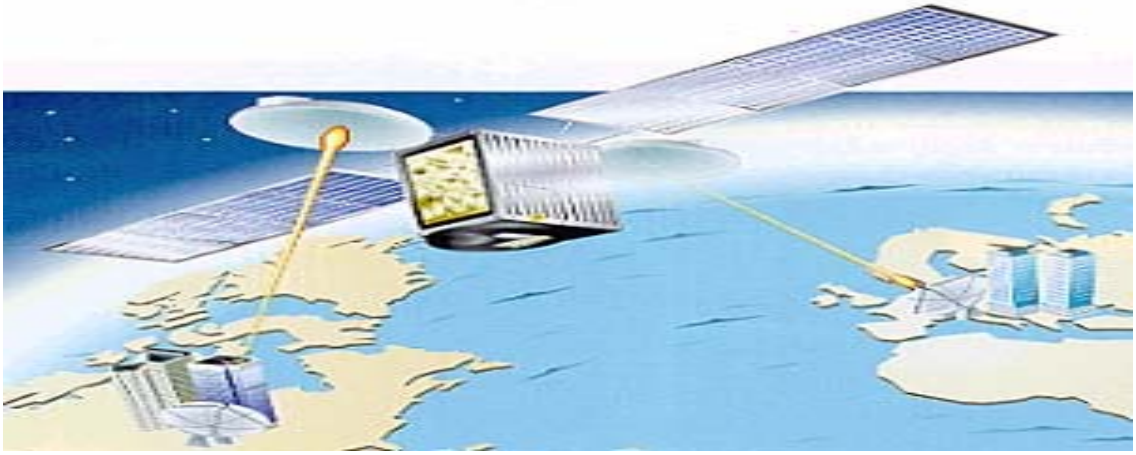
วัตถุประสงค์

1. เพื่อให้นักศึกษาทราบประเภทของระบบสารสนเทศ
2. เพื่อให้นักศึกษาทราบถึงขั้นตอนในการสร้างระบบค้นคืนสารสนเทศ
3. เพื่อให้นักศึกษาทราบถึงการประมวลผลระบบค้นคืนสารสนเทศ
4. เพื่อให้นักศึกษาทราบถึงวิวัฒนาการของระบบค้นคืนสารสนเทศ
5. เพื่อให้นักศึกษาทราบถึง Information Retrieval Model

สารบัญ

1.1 บทนำ

ในปัจจุบันนี้โลกเข้าสู่ยุคของข้อมูลข่าวสาร สารสนเทศ(Information) เป็นสิ่งที่มีความจำเป็นที่ทางภาครัฐและภาคเอกชนต้องอาศัยสารสนเทศเข้ามาช่วยในการทำงาน, การบริหารจัดการ หรือช่วยสนับสนุนการตัดสินใจ เทคโนโลยีในปัจจุบันมุ่งเน้นในเรื่องของอุปกรณ์ ,เครื่องมือหรือ ผลิตภัณฑ์ที่มีพื้นฐานทางคอมพิวเตอร์เป็นองค์ประกอบ ซึ่งการพัฒนาเทคโนโลยีนี่เองที่ก่อให้เกิดความก้าวหน้าในโลกในด้านต่างๆไม่ว่าจะเป็นทางด้านอุตสาหกรรม , โทรคมนาคม , ดาวเทียม , การผลิต ,การบริหาร , การแพทย์ แม้กระทั่งในวงการศึกษา



สารสนเทศต่างๆได้มาจากการเก็บรวบรวมข้อมูลทั้งแหล่งปฐมภูมิจากแหล่งกำเนิดข้อมูลโดยตรง และแหล่งทุติยภูมิ จากแหล่ง ที่ได้จัดเก็บรวบรวมข้อมูลอยู่แล้ว โดยผู้ใช้จะนำข้อมูลต่างๆเหล่านี้ป้อนเข้าสู่ระบบคอมพิวเตอร์โดยอุปกรณ์นำเข้าที่เหมาะสมขึ้นอยู่กับชนิดของข้อมูล ซึ่งอาจเป็นข้อความ , ตัวเลข , ภาพนิ่ง ภาพเคลื่อนไหว , เสียงหรือสื่อผสม ต่อจากนั้นจะนำข้อมูลเหล่านี้ไปผ่านกระบวนการประมวลผล(process) ซึ่งเป็นการปฏิบัติการใดๆกับข้อมูลอาจประกอบด้วย การตรวจสอบข้อมูล (Verifying) , การจำแนกข้อมูล (Classifying) , การจัดเรียง (Sorting) , การสรุปผล (Summarizing) , การคิดคำนวณ (Calculating) , การดึงข้อมูลมาใช้งาน (Retrieving) , การสำรองข้อมูล (Backup) , การเผยแพร่และการติดต่อสื่อสาร (Communicating) และอื่นๆ ผลลัพธ์ของการทำงานจะออกมาในรูปแบบที่เป็นประโยชน์ต่อผู้ใช้ซึ่งอาจอยู่ในรูปของข้อความ , รายงาน , กราฟ , รูปภาพ หรือในรูปของอุปกรณ์หลายสื่อ (Multimedia) โดยมีภาพและเสียงประกอบเป็นภาพสองหรือสามมิติเป็นต้น

เทคโนโลยีสารสนเทศและการสื่อสาร หรือ ICT ซึ่งย่อมาจาก Information and Communication Technology หมายถึงการนำเทคโนโลยีสารสนเทศ (IT) และ เทคโนโลยีการสื่อสาร (CT) นำมาสร้างสารสนเทศและเชื่อมต่อระบบย่อยๆขององค์กรเข้าด้วยกัน โดยมี

เทคโนโลยี (Technology) หมายถึง การใช้เครื่องมือให้เหมาะสมกับสถานการณ์ในการแก้ปัญหา ผู้ที่นำเอาเทคโนโลยีมาใช้เรียกว่า นักเทคโนโลยี(Technologist)

สารสนเทศ(Information) หมายถึง ข่าวสารที่ได้จากการนำ ข้อมูลดิบ (raw data) มาประมวลผลอย่างใดอย่างหนึ่ง ซึ่งข่าวสารที่ได้ออกมาจะอยู่ในรูปที่สามารถนำไปใช้งานได้ทันที

เทคโนโลยีสารสนเทศ (Information Technology) หมายถึงกระบวนการต่างๆและระบบงานที่ช่วยให้ได้สารสนเทศที่ต้องการโดยจะรวมถึง

(1) เครื่องมือและอุปกรณ์ต่างๆหมายถึง เครื่องคอมพิวเตอร์ เครื่องใช้สำนักงาน อุปกรณ์โทรคมนาคมต่างๆรวมทั้งซอฟต์แวร์ทั้งระบบสำเร็จรูปและพัฒนาขึ้นโดยเฉพาะด้าน

(2) กระบวนการในการนำอุปกรณ์เครื่องมือต่างๆ มาใช้งาน โดย รวบรวมเก็บข้อมูล จัดเก็บประมวลผล และแสดงผลลัพธ์เป็นสารสนเทศในรูปแบบต่างๆ ที่สามารถนำไปใช้ประโยชน์ได้ต่อไป

1.2 ประเภทของเทคโนโลยีสารสนเทศ

เทคโนโลยีของระบบสารสนเทศในปัจจุบัน ประกอบด้วย

(1) ระบบประมวลผลข้อมูล(Data Processing System หรือ DP) หรือบางครั้งเรียกว่าระบบประมวลผลรายการ(Transaction Processing System หรือ TPS) เป็นการนำคอมพิวเตอร์มาใช้ในการจัดข้อมูลพื้นฐาน โดยเน้นที่การประมวลผลรายการประจำวัน และการเก็บรักษาข้อมูล

(2) ระบบสารสนเทศเพื่อการบริหาร(Management Information System หรือ MIS) เป็นระบบบริหารที่ให้สารสนเทศที่ผู้บริหารต้องการ ระบบนี้ผลิตรายงานสรุปสารสนเทศซึ่งรวบรวมจากฐานข้อมูลขององค์กรเป็นคาบระยะเวลา หรือเป็นรายงานตามความต้องการ หรือเป็นรายงานในเหตุการณ์ที่เป็นสภาวการณ์ผิดปกติ หรืออาจเป็นรายงานพยากรณ์ รายงานเหล่านี้เป็นสารสนเทศที่สำคัญสำหรับผู้บริหารในการควบคุม สั่งการ การปฏิบัติการ และวางแผนในองค์กรได้ถูกต้อง

(3) ระบบสนับสนุนการตัดสินใจ(Decision Support System หรือ DSS) เป็นระบบสารสนเทศสนับสนุนการตัดสินใจของผู้บริหารอย่างมีประสิทธิภาพ โดยระบบนี้จะมีการสร้างแบบจำลอง(Model)สถานการณ์ปัจจุบันของการดำเนินงานในองค์กรขึ้นมา เพื่อให้ผู้บริหารได้

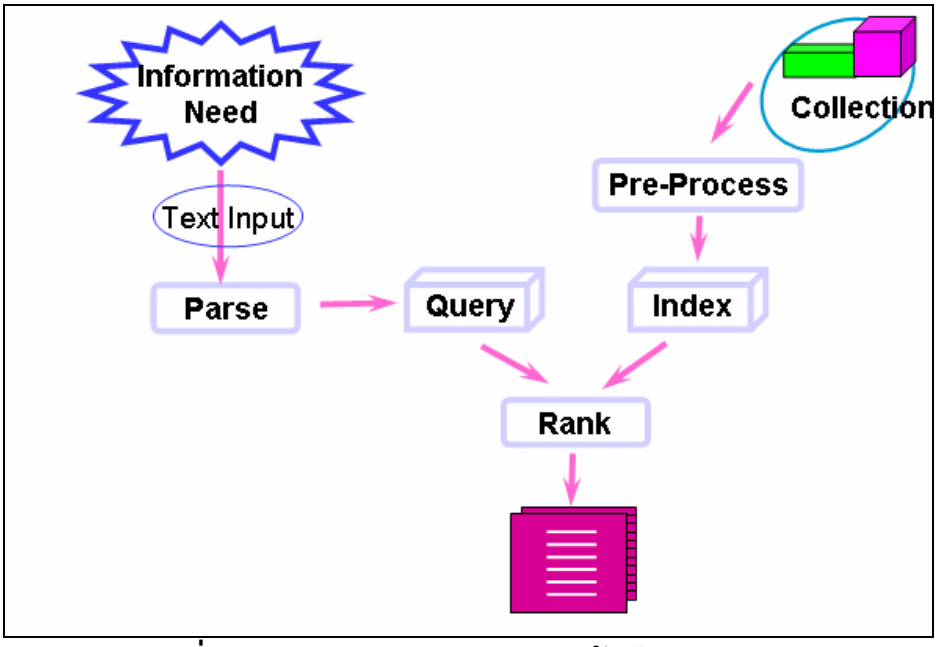
(4) ระบบสารสนเทศเพื่อผู้บริหารระดับสูง(Executive Information System หรือ EIS) เป็นระบบสารสนเทศที่เน้นการให้สารสนเทศที่สำคัญต่อการบริหารแก่ผู้บริหารระดับสูงสุด เพื่อใช้ในการตรวจสอบ ควบคุมการดำเนินงานในองค์กร รวมทั้งการตัดสินใจ การใช้งานจะเป็นภาพกราฟิก ที่ผู้ใช้งานไม่จำเป็นต้องมีทักษะสูง ผลลัพธ์จะอยู่ในรูปแบบที่เข้าใจได้ง่ายในรูปของกราฟิก ภาพ เสียง หรือ ระบบมัลติมิเดีย

(5) ระบบผู้เชี่ยวชาญ(Expert System หรือ ES) เป็นระบบคอมพิวเตอร์ที่ช่วยในการตัดสินใจโดยใช้วิธีเดียวกับมนุษย์ที่มีความเชี่ยวชาญ ระบบนี้เกี่ยวข้องกับการจัดการความรู้ (Knowledge) มากกว่าสารสนเทศ โดยใช้หลักการทำงานด้วยระบบ ปัญญาประดิษฐ์ (Artificial Intelligence) ซึ่งเป็นระบบถาม-ตอบ (Question-Answering System หรือ QA) จัดการเกี่ยวกับสารสนเทศที่เป็นลักษณะแบบธรรมชาติ เป็นความจริง (facts) ที่ให้ความรู้แก่ผู้ใช้ โดยข้อมูลนำเข้า(INPUT) และข้อมูลผลลัพธ์(OUTPUT) เป็นภาษาธรรมชาติ

(6) ระบบบริหารฐานข้อมูล (Data Base Management System หรือ DBMS)เป็นระบบที่จัดการเกี่ยวกับฐานข้อมูล เช่น การเก็บบันทึก การบำรุงรักษา และการค้นคืนข้อมูลต่างๆ รูปแบบการจัดเก็บ ไม่ใช่ภาษาธรรมชาติ อยู่ในรูปแบบตัวเลข หรือตาราง ประกอบด้วยเรคอร์ดข้อมูล แต่ละฟิลด์มีลักษณะเฉพาะ ผลลัพธ์ เป็นเรคอร์ดที่ต้องการ ที่ถูกจัดตามรูปแบบต่างๆ การค้นหา ต้องระบุค่าที่เป็นเอกลักษณ์ สารสนเทศที่ดึงออกมาจะเป็นรายการข้อมูลตามที่ผู้ใช้ระบุ

(7) ระบบค้นคืนสารสนเทศ(Information Retrieval System หรือ IR) เป็นระบบที่จัดการประมวลผลสารสนเทศประเภทเอกสาร(Document) ในรูปแบบต่างๆ เช่น หนังสือ , วารสาร , บทความ เป็นต้น โดยเกี่ยวข้องในเรื่องการแสดงรูปแบบ ,การเก็บบันทึก ,การดึงเอกสาร ตามรูปที่ 1.1 :แสดงภาพรวมของระบบสารสนเทศ ซึ่งปัญหาในปัจจุบันคือ เอกสารมีจำนวนมาก การค้นหาเอกสารหรือข้อมูลจากแหล่งข้อมูลที่มีอยู่ที่ไม่สามารถกระทำได้อย่างถูกต้องและรวดเร็ว วิธีการของระบบค้นคืนสารสนเทศคือ จะไม่อ่านเอกสารทั้งหมดเพื่อดึงเอกเอกสารที่ต้องการออกมา แต่จะใช้ลักษณะเด่นของเนื้อหาของเอกสารเป็นตัวแทนของเอกสาร ที่สามารถแยกแยะเอกสารที่เกี่ยวข้อง(relevant) กับข้อถามหรือสิ่งที่เราต้องการออกจากเอกสารที่ไม่เกี่ยวข้อง (Non-relevant) ซึ่งการดึงเอกสารที่เราต้องการนั้นยากที่จะระบุเอกสารหรือสารสนเทศที่ต้องการแน่นอนได้ บ่อยครั้งที่ผลลัพธ์หรือเอกสารที่ดึงออกมานั้น ไม่

และถ้ากล่าวถึง Google , Yahoo หรือ MSN search ก็คงจะนึกออกได้ไม่ยากว่า IR นั้นเป็นตัวแทน, หน่วยเก็บข้อมูล รวมถึงโครงสร้างและการเข้าถึงสารสนเทศต่างๆที่ผู้ใช้ต้องการ



รูปที่ 1.1 : แสดงภาพรวมของระบบค้นคืนสารสนเทศ

สรุปความแตกต่างระหว่างประเภทของสารสนเทศต่างๆนั้น จะเห็นได้ว่า ระบบบริหารฐานข้อมูล(DBMS) และ ระบบสารสนเทศเพื่อการบริหาร(MIS) มีการนำข้อมูลมาจัดเป็นโครงสร้าง ในรูปแบบตารางตัวเลขต่างๆ การค้นหาสามารถเข้าถึงข้อมูลที่ถูกจัดเก็บได้โดยตรง ส่วน ระบบค้นคืนสารสนเทศ(IR) และระบบถามตอบ(QA) หรือ ปัญญาประดิษฐ์ (AI) เป็นระบบที่เกี่ยวข้องกับข้อมูลที่เป็นความจริง(facts) มีการโต้ตอบกับผู้ใช้ด้วยภาษาธรรมชาติ ผลลัพธ์จะเกิดจากการค้นหาเป็นการอนุมาน หรือ เปรียบเทียบความใกล้เคียงกับข้อคำถาม (Query Language) มากที่สุด ดังรูปที่ 1.2 :ความแตกต่างระหว่าง Database และ IR

ตารางที่ 1.1 : แสดงความแตกต่างระหว่างการค้นคืนข้อมูลและการค้นคืนสารสนเทศ

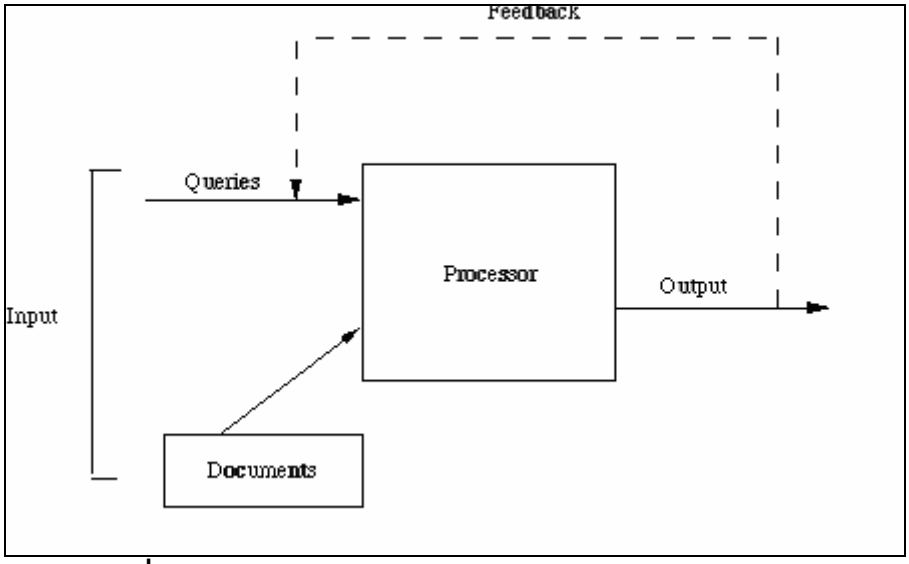
	Data Retrieval (DR)	*Information Retrieval* (IR)
Matching	Exact match	Partial match, best match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Classification	Monothetic	Polythetic
Query language	Artificial	Natural
Query specification	Complete	Incomplete
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive

การค้นคืนข้อมูล(Data retrieval) นั้นเอกสารจะถูกบรรจุเป็นเซตของคำสำคัญ (keyword) ซึ่งมีความหมาย สำหรับการสืบค้นสารสนเทศ (Information retrieval) นั้น สารสนเทศเป็นหัวข้อหรือประธานของประโยค(subject) ซึ่งมีความหมายเพียงคร่าวๆ สำหรับ ระบบค้นคืนสารสนเทศ(IR system) นั้นเป็นระบบที่ทำการแปลเนื้อความของข่าวสาร การจัดตำแหน่ง(ranking) ของความสัมพันธ์กันเป็นสิ่งที่สำคัญที่สุด

	Databases	IR
What we're retrieving	Structured data. Clear semantics based on a formal model.	Mostly unstructured. Free text with some metadata.
Queries we're posing	Formally (mathematically) defined queries. Unambiguous.	Vague, imprecise information needs (often expressed in natural language)
Results we get	Exact. Always correct in a formal sense.	Sometimes relevant, often not.
Interaction with system	One-shot queries.	Interaction is important. (relevance feedback)

รูปที่1.2 : แสดงความแตกต่างระหว่าง Database และ IR

1.3 ส่วนประกอบของระบบ IR



รูปที่ 1.3 : แสดงส่วนประกอบของระบบค้นคืนสารสนเทศ

ส่วนประกอบของระบบค้นคืนสารสนเทศแบ่งได้เป็น 3 ส่วนได้แก่

(1) ส่วนนำเข้าข้อมูล(Input) เป็นส่วนของการป้อนข้อคำถาม(query)จากผู้ใช้ซึ่งเป็นภาษาธรรมชาติ หรืออาจเป็นการนำเข้าMetadata ซึ่งเป็นสารสนเทศเกี่ยวกับเอกสารหรืออาจไม่เป็นส่วนหนึ่งของเอกสารก็ได้แต่เป็นข้อมูลเกี่ยวกับข้อมูล (data about data) อทิเช่น

(1.1) Descriptive metadata เป็นการนำเข้าสารสนเทศที่เป็นความหมายของเอกสารที่อยู่ภายนอก เช่น ผู้แต่ง(Author), ชื่อเรื่อง(Title), แหล่งที่มา (Source : book, magazine, newspaper, journal) ,วันที่ (Date) ,ISBN ,สำนักพิมพ์(Publisher),ความยาว (Length)

(1.2)Semantic metadata concerns the content: เป็นการนำเข้าเนื้อความที่มีความหมาย เช่น บทคัดย่อ(Abstract), คำสำคัญ(Keywords) ,รหัสของหัวเรื่อง(Subject Codes) ซึ่งอาจเป็น Library of Congress หรือ Dewey Decimal หรือ UMLS (Unified Medical Language System) ก็ได้

(1.3) เทอมของหัวเรื่อง (Subject terms) ซึ่งอาจมาจาก ontologies พิเศษเป็นลำดับขั้นของเทอมมาตรฐาน(hierarchical taxonomies of standardized semantic terms)

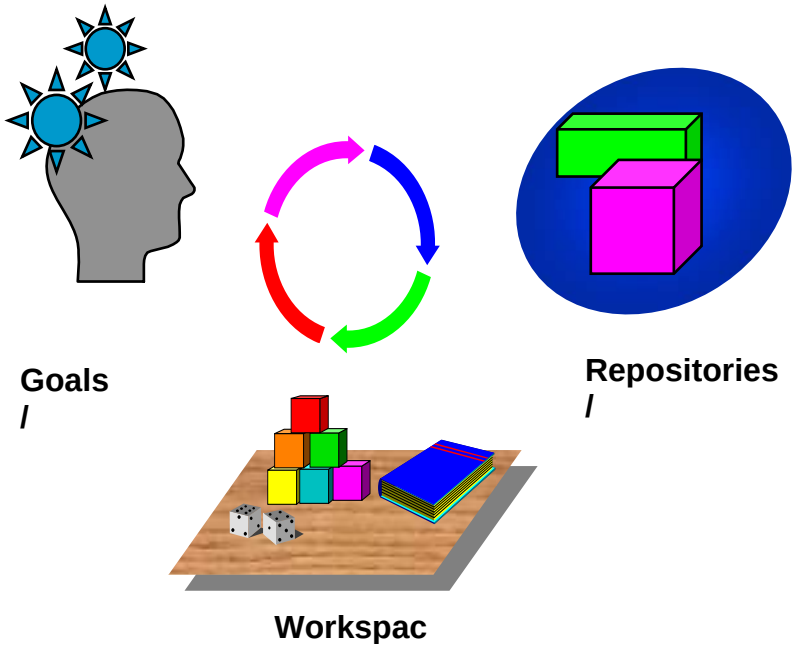
(1.4) อาจเป็นสารสนเทศของเวป(Web Metadata) ก็ได้ เช่น META tag in HTML ดังนี้

```
<META NAME="keywords"
CONTENT="pets, cats, dogs">
```

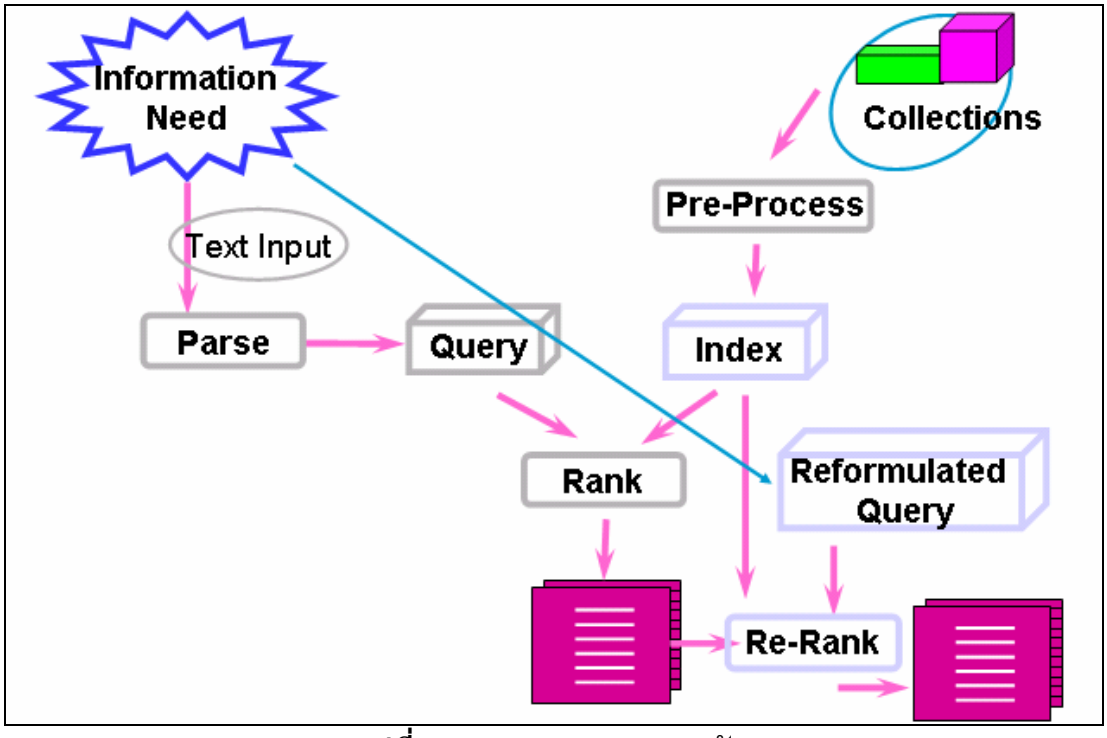
ระบบค้นคืนสารสนเทศจะนำสารสนเทศเหล่านี้ผ่านการประมวลผลแบบเชื่อมโดยตรงกับระบบคอมพิวเตอร์ ซึ่งผู้ใช้จะมีการโต้ตอบหรือปฏิสัมพันธ์กับระบบโดยตรง

(2) โปรเซสเซอร์ (Processor) เป็นส่วนของการประมวลผล ได้แก่ การจัดโครงสร้างของสารสนเทศในรูปแบบที่เหมาะสม อันประกอบด้วย การสร้างตัวแทนเอกสาร ,การแบ่งแยกกลุ่มของเอกสาร ,การจัดเก็บสารสนเทศ ,การดึงข้อมูลตามที่ผู้ใช้ต้องการ การทำงานนั้นจะนำข้อคำถามไปเปรียบเทียบกับตัวแทนเอกสารที่มีอยู่ เพื่อดึงเอกสารที่ใกล้เคียงนำออกมาให้แก่ผู้สอบถาม

(3) ส่วนของผลลัพธ์(OUTPUT) ผลลัพธ์ที่ได้จากระบบเป็นข้อความสั้นๆ เช่น ชื่อหนังสือ, หมายเลขเอกสาร, ชื่อผู้แต่ง, สำนักพิมพ์ เป็นต้น ผู้ใช้สามารถพิจารณาจากข้อมูลต่างๆที่ได้จากระบบถ้าเอกสารที่ได้มีจำนวนมากเกินไปหรือไม่ใกล้เคียงกับสิ่งที่ต้องการ ผู้ใช้สามารถปรับปรุงข้อคำถามใหม่เพื่อให้ข้อคำถามนั้นสืบค้นสารสนเทศได้ตรงกับความต้องการมากที่สุด เป็นระบบตอบกลับ(feedback) ดังนั้นผลลัพธ์ที่ได้จึงขึ้นอยู่กับ ข้อคำถาม (Query) ของผู้ใช้ตามรูปที่ 1.4 : แสดงกระบวนการกระทำซ้ำของระบบค้นคืนสารสนเทศ และ รูปที่ 1.5: แสดงระบบตอบกลับ



รูปที่ 1.4 : แสดงกระบวนการกระทำซ้ำของระบบค้นคืนสารสนเทศ



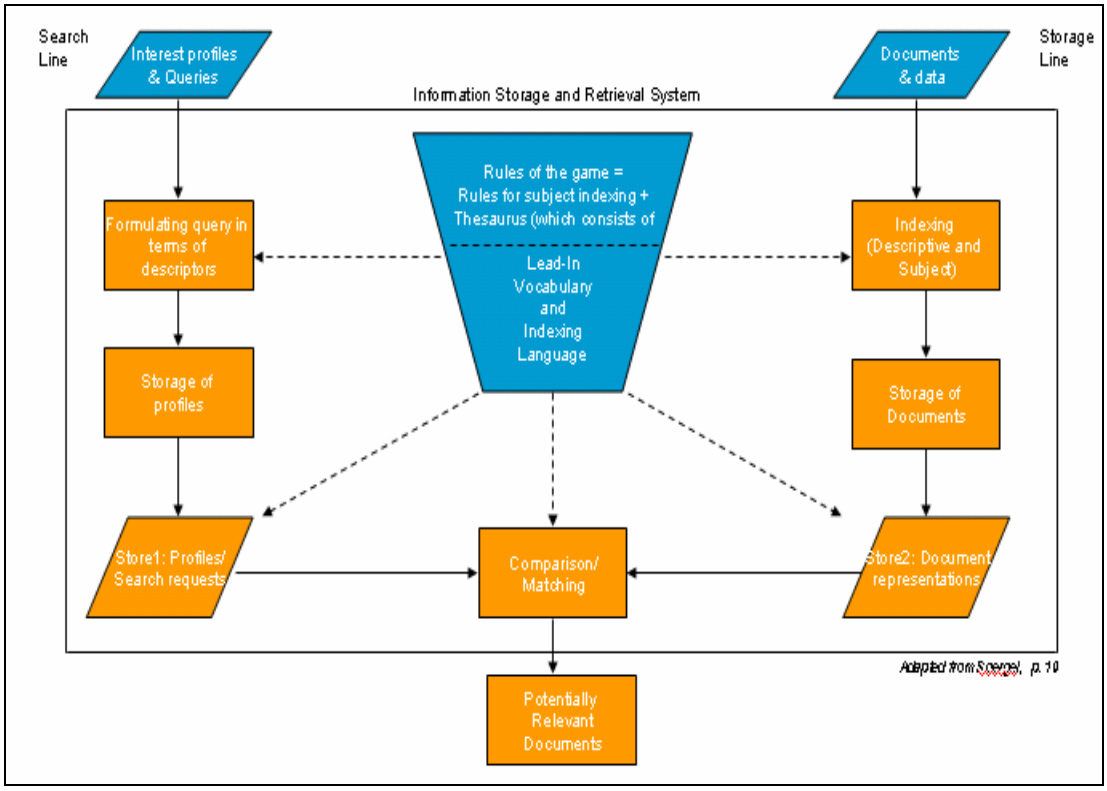
รูปที่ 1.5 :แสดงระบบตอบกลับ

1.4 ขั้นตอนในการสร้างระบบ IR

การสร้างระบบค้นคืนสารสนเทศ แบ่งออกเป็น 4 ขั้นตอนคือ

1. การวิเคราะห์ข้อความ (Text Analysis)
2. การจัดแบ่งกลุ่มข้อมุล (Classification)
3. การเก็บบันทึกข้อมูลลงในแฟ้มข้อมูล
4. การค้นคืนสารสนเทศ

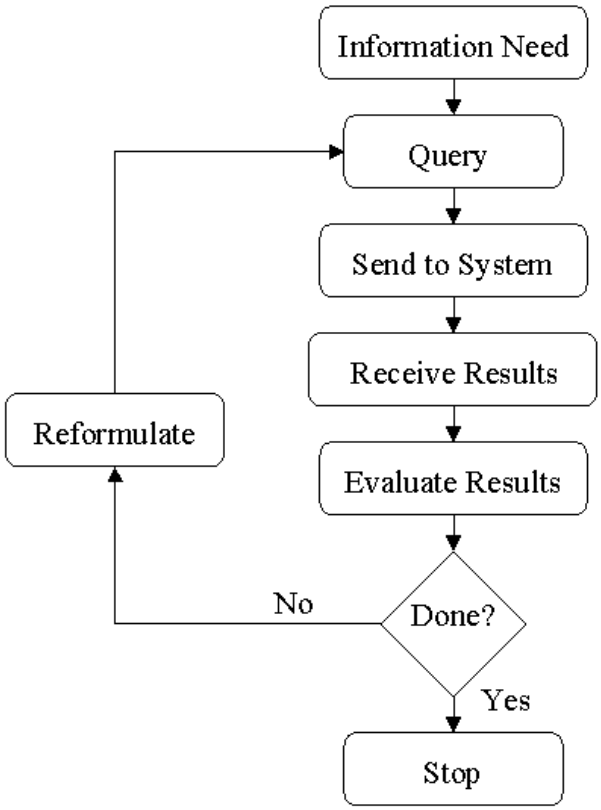
ปัญหาของระบบค้นคืนสารสนเทศนั้นคือการนำข้อความทั้งหมดในเอกสารไปเก็บในระบบ จะใช้เวลาและเนื้อที่ในหน่วยความจำมาก จึงมีการวิเคราะห์ข้อความ(Text Analysis) เพื่อแก้ปัญหาเพื่อลดเวลาและค่าใช้จ่ายลง การวิเคราะห์ข้อความของเอกสารมีวัตถุประสงค์เพื่อหาตัวแทนเอกสารที่เหมาะสม นำตัวแทนของเอกสารหรือดรรชนี(index)ที่ได้มาจัดเก็บแทนข้อความฉบับสมบูรณ์ ในการสืบค้นเอกสารจะมีการเปรียบเทียบกับตัวแทนเอกสารเหล่านี้กับข้อคำถามของผู้ใช้ เพื่อดึงสารสนเทศที่ต้องการออกมาทำให้เกิดความรวดเร็วในการประมวลผล วิธีการในการหาตัวแทนเอกสารนั้นอาจอาศัยมนุษย์ ซึ่งต้องใช้วิจารณญาณของผู้เชี่ยวชาญในสาขานั้นๆ และอาศัยหลักทางด้านความหมายของคำ หรืออาจกระทำด้วยระบบ



รูปที่ 1.6 : แสดงโครงสร้างของระบบค้นคืนสารสนเทศ

1.5 การประเมินผลระบบค้นคืนสารสนเทศ (Evaluation of IR System)

การพัฒนาและวิจัยทางด้าน IR นั้นมุ่งเน้น พัฒนาประสิทธิภาพ(Efficiency) และประสิทธิผล (Effectiveness) ของระบบค้นคืนสารสนเทศ ซึ่งงานวิจัยหลายๆท่านจะพยายามหาเทคนิคหรือวิธีการใหม่ๆ อย่างต่อเนื่องเพราะในยุคโลกาภิวัฒน์นี้มีเอกสารเกิดขึ้นในแต่ละวันมากมาย ไม่ว่าเป็นหนังสือ บทความ อีกทั้งภาพและสื่อผสม จึงมีการประเมินผลระบบค้นคืนสารสนเทศ เพื่อพยายามที่จะปรับปรุงให้ระบบมีประสิทธิภาพและประสิทธิผลมากที่สุด



รูปที่ 1.7 : The Standard Retrieval Interaction Model

จากรูปที่ 1.7 แสดงถึงโมเดลในการปฏิสัมพันธ์ในการสืบค้นมาตรฐาน จะเห็นได้ว่าผู้ใช้จะร้องขอสารสนเทศที่ต้องการโดยมีการส่งเข้าไปยังระบบ ระบบจะส่งผลลัพธ์จากการประมวลผลกลับมายังให้ผู้ใช้ ผลลัพธ์ที่ได้จะถูกประเมินว่าเป็นสารสนเทศที่ต้องการใช่หรือไม่ ในกรณีที่ได้รับข้อมูลข่าวสารที่ไม่ตรงกับความต้องการจะทำการคำนวณใหม่โดยป้อนข้อสอบถามใหม่ที่ปรับปรุงจากเดิมเป็นการทำงานวนรอบ จนกระทั่งได้รับสารสนเทศที่ต้องการ

ประสิทธิภาพของระบบวัดจาก เนื้อที่ในการจัดเก็บในหน่วยความจำ , CPU Time

ประสิทธิผลของระบบวัดจาก ค่าใช้จ่าย ,ต้นทุนในการสร้างระบบ, Recall ,Precision

โดย

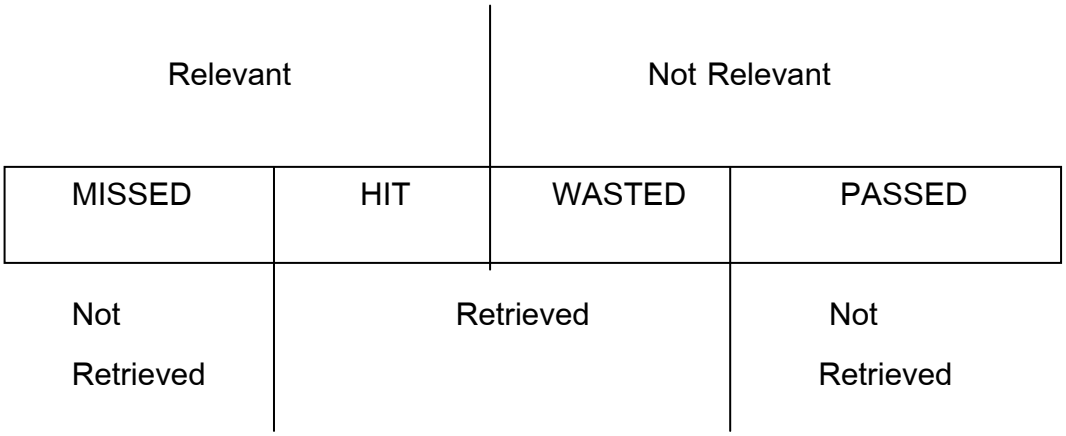
Precision = จำนวนของเอกสารที่เกี่ยวข้องที่ถูกดึงออกมา /
จำนวนทั้งหมดของเอกสารที่ถูกดึงออกมา

Recall = จำนวนของเอกสารที่เกี่ยวข้องที่ถูกดึงออกมา /
จำนวนของเอกสารทั้งหมดที่เกี่ยวข้อง

การดึงเอกสารนั้นขึ้นอยู่กับข้อคำถาม เมื่อผู้ใช้ป้อนคำถามระบบจะแบ่งกลุ่มของเอกสารออกเป็น 2 ส่วนคือ เอกสารที่ถูกดึงออกมา(Retrieved) และ เอกสารที่ไม่ถูกดึงออกมา(Not Retrieved) ซึ่งเอกสารต่างๆใน 2 กลุ่มนี้ อาจมีทั้งเอกสารที่เกี่ยวข้อง(Relevant) และไม่เกี่ยวข้อง (Non-Relevant) กับสิ่งที่ต้องการก็ได้

กำหนดให้

HIT : เอกสารที่เกี่ยวข้องที่ถูกดึงออกมา
WASTED : เอกสารที่ไม่เกี่ยวข้องที่ถูกถึงออกมา
MISSED :เอกสารที่เกี่ยวข้องที่ไม่ถูกดึงออกมา
PASSED : เอกสารที่ไม่เกี่ยวข้องที่ไม่ถูกดึงออกมา



Recall = HIT / Relevant
เป็นการวัดความสามารถของระบบในการดึงเอกสารที่เกี่ยวข้องออกมา
Precision = HIT / Retrieved
เป็นการวัดความสามารถของระบบในการขจัดเอกสารที่ไม่เกี่ยวข้องออกไป

ระบบ IR ที่มีประสิทธิผลดี ย่อมมี Recall ที่ได้สมดุลย์กับ Precision

ตัวอย่างที่ 1.1 แสดงการหาประสิทธิภาพของระบบค้นคืนสารสนเทศ

	Relevant	Not Relevant
Retrieved	A	B
Not Retrieved	C	D

- Relevant = A+C
- Retrieved = A+B
- Collections size = A+B+C+D
- Precision = A/(A+B)
- Recall = A/(A+C)
- Miss = C/(A+C)
- False alarm = B/(B+D)

1.6 วิวัฒนาการของระบบค้นคืนสารสนเทศ

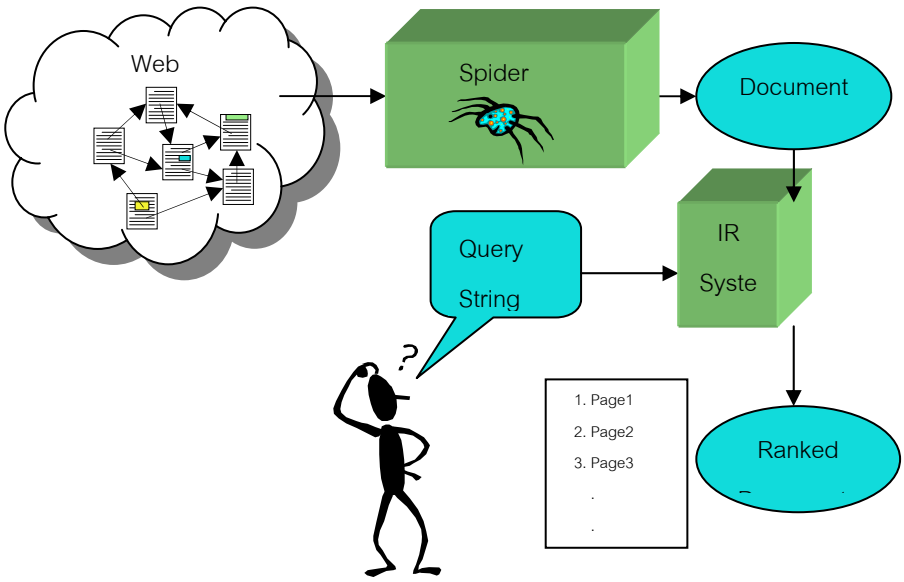
วิวัฒนาการของระบบค้นคืนสารสนเทศตั้งแต่อดีตจนถึงปัจจุบันนั้นสามารถสรุปได้ดังนี้

ปี คศ. 1960-70 เริ่มต้นในการสำรวจระบบค้นคืนสารสนเทศสำหรับข้อความที่มีขนาดเล็กที่เป็นบทคัดย่อทางด้านวิทยาศาสตร์ กฎหมาย หรือ เป็นเอกสารทางด้านธุรกิจ มีการพัฒนา บูลีนพื้นฐาน และ Vevtor Space Model สำหรับการค้นคืนสารสนเทศ

ปี คศ. 1980 นั้นระบบจะมีการสืบค้นจากฐานข้อมูลเอกสารขนาดใหญ่ที่มีการปฏิบัติงานในหลายๆบริษัท เช่น Lexis-Nexis/MEDLINE

ปี คศ. 1990 มีการสืบค้น FTP และ World Wide Web บนเตอร์เน็ตเช่น Archie ,WAIS ,Lycos , Yahoo ,Altavista เป็นต้น

ปี คศ. 2000 จนถึงปัจจุบัน มีการเชื่อมโยงโดยวิเคราะห์จาก Web Search เช่น Google นอกจากนี้มีการสืบค้นสื่อผสม (multimedia) ไม่ว่าเป็นภาพ เสียง เพลง วีดีโอ การสรุปเอกสาร เป็นต้น



รูปที่ 1.8 : แสดง Web Search System

สิ่งที่คนส่วนมากค้นหาใน web โดยการศึกษาจาก Spink et al., Oct 98 (www.shef.ac.uk/~is/publications/infres/paper53.html) สำรวจจาก Excite จาก 13 คำถาม โดยสำรวจจากประชากร 316 คน ได้แก่

- Genealogy/Public Figure: 12%
- Computer related: 12%
- Business: 12%
- Entertainment: 8%
- Medical: 8%
- Politics & Government 7%
- News 7%
- Hobbies 6%
- General info/surfing 6%
- Science 6%
- Travel 5%
- Arts/education/shopping/images 14%

มีการสำรวจผู้ใช้ web จากข้อคำถามทั้งหมด 50,000 ข้อคำถาม ที่เข้าใช้ excite ในปี คศ. 1997 สรุปได้ว่า เทอม(terms) ที่ใช้บ่อยมากๆ ได้แก่

จำนวน	เทอม	จำนวน	เทอม
4660	sex	1163	games
1151	mp3	1140	weather
1127	www.yahoo.com	1110	maps
1036	yahoo.com	983	ebay
980	recipes	1223	hotmail
3129	yahoo	2191	internal site admin check from kho
1520	chat	1498	porn
1315	horoscopes	1284	pokemon
1283	SiteScope test		

จากการสำรวจ web pages โดยการวิเคราะห์เทอมต่างๆทั้งหมดจำนวน 31,928,892 terms จากการวิเคราะห์ 49,602,191 web pages (***Source: http:/ /elib. cs. berkeley. edu/ docfreq/index.html)*** สามารถสรุปได้ว่าคำที่มีจำนวนมากได้แก่คำดังต่อไปนี้

จำนวน	เทอม	จำนวน	เทอม
65002930	the	62789720	a
60857930	to	57248022	of
54078359	and	52928506	in
50686940	s	49986064	for
45999001	on	42205245	this
41203451	is	39779377	by
35439894	with	35284151	or
34446866	at	33528897	all
31583607	are	30998255	from
30755410	e	30080013	you
29669506	be	29417504	that
28542378	not	28162417	an

28110383	as	28076530	home
27650474	it	27572533	i
24548796	have	24420453	if
24376758	new	24171603	t
23951805	your	23875218	page
22292805	about	22265579	com
22107392	information		

ในปัจจุบันมีงานวิจัยทางระบบค้นคืนสารสนเทศในการปรับปรุงเทคนิคและวิธีการเพื่อเพิ่มประสิทธิภาพและประสิทธิผลในการค้นคืนสารสนเทศ เพื่อให้สามารถรองรับกับปริมาณของเอกสารในปัจจุบันที่เพิ่มจำนวนมากขึ้นอย่างมากมายมหาศาล สำหรับประวัติของงานวิจัยทางระบบค้นคืนสารสนเทศตั้งแต่อดีตนั้น มีความเป็นมาดังนี้

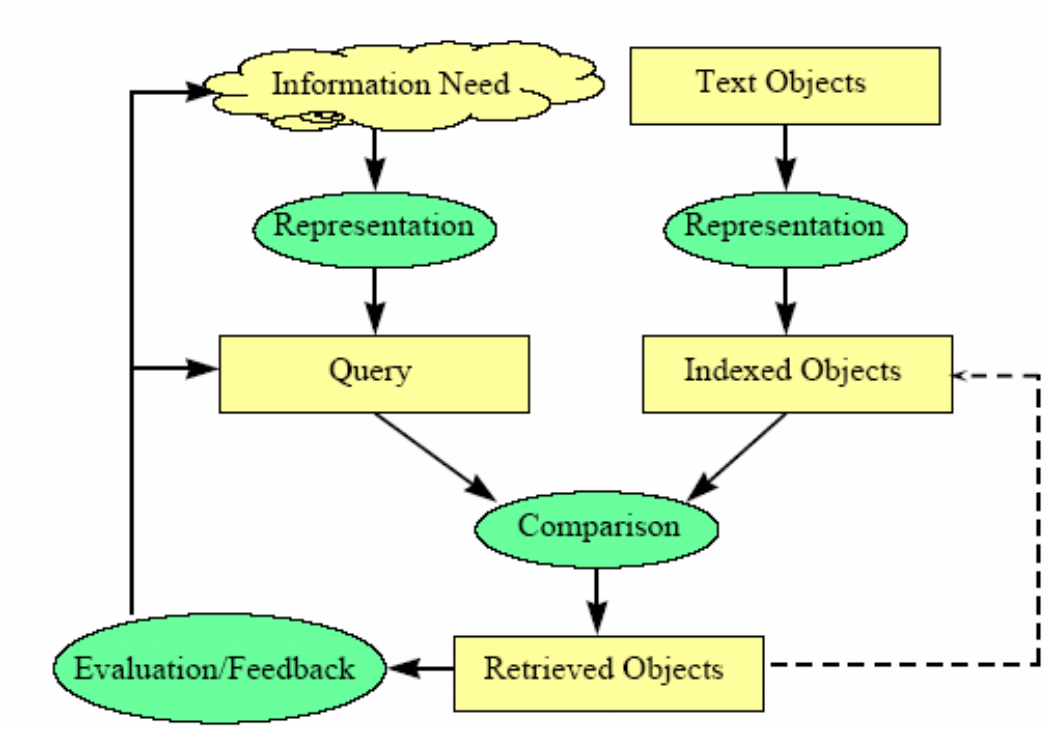
- 1958 Statistic Language Properties (Luhn)
- 1960 Probabilistic Indexing (Maron & Kuhns)
- 1961 Term association and clustering (Doyle)
- 1965 Vector Space Model (Salton)
- 1968 Query expansion (Roccio, Salton)
- 1972 Statistical Weighting (Sparck-Jones)
- 1975 2-Poisson Model (Harter, Bookstein, Swanson)
- 1976 Relevance Weighting (Robertson, Sparck-Jones)
- 1980 Fuzzy sets (Bookstein)
- 1981 Probability without training (Croft)
- 1983 Linear Regression (Fox)
- 1983 Probabilistic Dependence (Salton, Yu)
- 1985 Generalized Vector Space Model (Wong, Rhagavan)
- 1987 Fuzzy logic and RUBRIC/TOPIC (Tong, et al.)
- 1990 Latent Semantic Indexing (Dumais, Deerwester)
- 1991 Polynomial & Logistic Regression (Cooper, Gey, Fuhr)
- 1992 TREC (Harman)

- 1992 Inference networks (Turtle, Croft)
- 1994 Neural networks (Kwok)

ในปัจจุบันมีการวิจัยอย่างต่อเนื่องซึ่งแปรเปลี่ยนไปตามยุคสมัย ในความต้องการสารสนเทศของผู้ใช้ โดยมีเป้าหมายเพื่อให้ผู้ใช้สามารถสืบค้นได้อย่างรวดเร็ว โดยพัฒนาระบบค้นคืนสารสนเทศให้มีประสิทธิภาพและประสิทธิผลให้มากที่สุด

1.7 Information Retrieval Models

IR model ได้จำแนกคุณลักษณะเฉพาะออกเป็น 4 พารามิเตอร์ คือ การสร้างตัวแทนเอกสารหรือข้อสอบถาม(representations for documents and queries) , การใช้กลยุทธ์ในการจับคู่จากการวัดความคล้ายคลึงกันของเอกสารจากข้อคำถามของผู้ใช้ (matching strategies for assessing the relevance of documents to a user query) , วิธีการในการจัดลำดับผลลัพธ์จากข้อสอบถาม(methods for ranking query output) และ กลไกสำหรับการได้มาของระบบตอบกลับ(Mechanisms for acquiring user-relevance feedback.) ตามรูปที่ 1.9 แสดงถึง Simple model of IR



รูปที่ 1.9 : Simple model of IR

IR Model สามารถจำแนกเป็น 4 ชนิด ได้แก่

1. Set Theoretic Models ใช้บูลีนโมเดล ซึ่งอยู่บนพื้นฐานของแนวความคิดเชิงตรรกะหรือพีชคณิตบูลีน (Boolean Algebra) กับคำที่ถูกรวมกันโดยตัวเชื่อมทางตรรกะหรือพีชคณิต และ (AND) หรือ (OR) และไม่ (NOT) ตรรกะเหล่านี้รวมกัน

- การใช้ AND กำหนดว่าทั้งสองคำที่ถูกเชื่อมอยู่ในเอกสารที่ถูกค้นคืน
- การใช้ OR กำหนดว่าอย่างน้อยที่สุดหนึ่งในสองคำนั้นต้องอยู่
- การใช้ NOT กำหนดว่าคำที่ระบุต้องไม่ปรากฏในเอกสารที่ค้นคืนได้ แต่ Operator NOT จะทำการค้นคืนทุก ๆ เอกสาร ที่ไม่บรรจุคำที่กำหนด จึงมีความเสี่ยงที่จะได้เกือบทุกเอกสารในฐานข้อมูล วิธีการหนึ่งที่จะแก้ปัญหาคือการกำจัดการใช้ NOT และจะใช้ในสถานการณ์ที่ใช้ในชุดของเอกสารที่เล็ก ๆ น้อยเท่านั้น

Retrieval Status Value (RSV) เป็นมาตรการเกี่ยวกับการคิวรี่เอกสารที่มีความ Similality ซึ่งใน Boolean Model ค่า RSV จะมีค่าเท่ากับ 1 เมื่อคิวรี่ที่แสดงมีค่าเป็น True และ RVS มีค่าเป็น 0 เมื่อคิวรี่ที่แสดงเป็นค่าอื่น ซึ่งเอกสารทั้งหมดที่ค่า RSV มีค่าเท่ากับ 1 คือมีการค้นพบเอกสารที่เกี่ยวข้องตามที่คิวรี่

ในคำขอบูลีนการพิจารณาว่าคำไหนควรจะถูกพิจารณาก่อนหรือหลังจะถูกกำหนดอย่างน้อยที่สุดบางส่วนโดย Operator ทางตรรกะ

ปัญหาของ Boolean Model

(1) คำขอล้วน ๆ ไม่มีทางที่จะให้น้ำหนักของคำได้อย่างง่าย ๆ คำ ๆ หนึ่งจะปรากฏหรือไม่ปรากฏเท่านั้น ผู้ใช้ควบคุมความสำคัญของคำที่มีต่อคำขอได้น้อยมากการขาดกลไกลการให้น้ำหนักส่งผลให้คำขอที่ไม่ดีที่สุด

(2) บูลีนอาจทำให้การค้นคืนที่ผิดอันเนื่องมาจากคิวรี่ที่ระบุผิด ซึ่งอาจจะเป็นปัญหาของการแปลความหมายที่ผิดของการเชื่อมต่อบูลีน

(3) ระบบการค้นคืนแบบบูลีนอยู่ที่ลำดับของการจัดเรียงของตัวเชื่อมตรรกะ มีมาตรฐานสองอันที่ถูกใช้กันอยู่ ทั้งคู่จะพิจารณาวงเล็บที่รวมกลุ่มของคำด้วยกันเป็นอันดับแรกการรวมภายในวงเล็บ

ข้อดีของ Boolean Model

(1) Boolean Model ได้รับความนิยมมากเพราะว่าเข้าใจง่าย เวลา Query กำหนดความต้องการได้ว่าจะเอาคำไหนซึ่งเป็นลักษณะง่าย ๆ ที่มีรูปแบบไม่ซับซ้อน

(2) เป็นการหาข้อมูลในลักษณะ ใช่ / ไม่ใช่ เจอ / ไม่เจอ

ข้อเสียของ Boolean Model

(1) ไม่มีการจัดลำดับของเอกสาร (Ranking) และ ไม่มีการเปรียบเทียบ (Relevance)

(3) ใช้ Query ที่มีความซับซ้อนไม่ได้และไม่มีความยืดหยุ่น เพราะว่ามี Expression แค่ AND OR NOT

(4) ควบคุมจำนวนเอกสารยาก

(5) มีความลำบากในการยอมรับเนื่องจากตรงหรือสำคัญเป็นความต้องการของ User แม้ว่าการคิวรี่แบบ Boolean จะมีปัญหาต่าง ๆ แต่ระบบการค้นคืนแบบ Boolean เป็นระบบที่มีความนิยมใช้กันอย่างสูงและค่อนข้างที่จะมีประสิทธิภาพ และมีการใช้งานที่ง่าย

Fuzzy Set Model ในทฤษฎีของ Set ปกติ set มีขอบแหลม (Sharp Edges)คือแต่ละตัวจะอยู่หรือไม่อยู่ใน Set ซึ่งในทฤษฎีของ Fuzzy Set แต่ละตัว (สมาชิก) จะมีระดับสมาชิก (Membership Grade) ติดตัวอยู่ตามที่ Set กำหนดซึ่งค่านี้จะแสดงกำลังหรือระดับของความเชื่อในสมาชิกของ Set ค่าสมาชิกมักถูกกำหนดเป็นค่าในช่วง 0.0 ถึง 1.0

ปัญหาของ Fuzzy Set คือระบบไม่สามารถบอกได้อย่างแม่นยำว่า เอกสารอันที่ได้จะตอบสนองต่อความต้องการสารสนเทศ ซึ่งความไม่แน่นอนจะถูกเข้ามาในการประเมินแบบ Fuzzy Set ซึ่งการค้นคืนแบบนี้มีความสัมพันธ์กับการค้นคืนที่มีการคำนวณการประเมินของความเกี่ยวข้อง (Relevance)

2. Algebraic Models

Vector Space Model เอกสารแต่ละอันจะถูกนำแสดงโดย Vector หรือชุดของคำที่มีการจัดเรียงลำดับ วิธีการประเมินขึ้นอยู่กับ Vector 0-1 ซึ่งแต่ละองค์ประกอบเป็น 0 ถ้าคำนั้น ๆ ไม่ปรากฏ หรือเป็น 1 ถ้าคำนั้น ๆ ปรากฏในเอกสารตามที่คิวรี่ และการประเมินอีกทางหนึ่งอยู่บนพื้นฐานของ Vector น้ำหนักซึ่งองค์ประกอบของมันเป็นน้ำหนักหรือค่าที่ถูกกำหนดให้แต่ละคำในเอกสาร

ความสำเร็จในการใช้ Vector space Model อยู่ที่ความเข้ากันได้ดีในเชิงมิติ (Dimensional Compatibility) คือการเปรียบเทียบของเอกสารสองอัน (หรือเอกสารกับคำขอ) จะอยู่บนพื้นฐานของการเปรียบเทียบคำที่เหมือนกันในแต่ละเอกสารเสมอ

การกำหนดน้ำหนักให้กับคำ ใน Vector เป็นกระบวนการที่ซับซ้อน น้ำหนักสามารถถูกกำหนดโดยอัตโนมัติในเอกสารขึ้นอยู่กับการนับความถี่ของคำ คำที่ปรากฏยิ่งบ่อยในเอกสารจะมีความสำคัญต่อเอกสารนั้นมาก (ข้อยกเว้นที่คือคำที่เป็นคำเชื่อม เช่น The , a , and , of)

รูปแบบการทำงานของ Vector Model

(1) ให้ความสำคัญความถี่ของคำที่ปรากฏอยู่ในเอกสารและความถี่มีผลต่อการให้ค่าน้ำหนักของคำ ได้แก่

- Term Frequency คือการใช้ความถี่ของคำ เช่นเจอ 1 ครั้ง เรียกว่า Term ทั้งนี้ขึ้นอยู่กับจำนวนคำของเอกสาร โดย Term จะแทนคำศัพท์ของแต่ละคำ
- Term Weight (น้ำหนักของคำ) ความถี่ของคำ ๆ หนึ่งที่พบในทุก ๆ เอกสาร

(2) สามารถจัดอันดับของเอกสารโดยใช้เกณฑ์ความสำคัญของคำและการ Match กันของคำ

ข้อดีของ Vector Space Model

(1) ใช้คณิตศาสตร์เรียบง่ายในการคิด มีการพิจารณาจากความถี่ของคำ และสามารถจัด Ranking ของเอกสารได้ สามารถใช้กับเอกสารที่มีข้อมูลมาก ๆ ได้ดี

ข้อเสียของ Vector Space Model

(1) ไม่สนใจความหมายของคำ , วลี , โครงสร้างของคำ , คำที่มีความหมายเหมือนกัน (Synonymy)

(2) สืบค้นใส่เงื่อนไขแบบ Boolean Model ไม่ได้

3. Probabilistic Model

Probabilistic Model แบบความน่าจะเป็นจะคล้ายกับคำขอแบบ Fuzzy set Model แต่มีข้อกำหนดเพิ่มเติมว่าฟังก์ชันสมาชิกหรือฟังก์ชันตัดสินที่ถูกใช้เป็นแบบน่าจะเป็น ในระบบการค้นคืนแบบน่าจะเป็น Set ที่ถูกค้นคืนจากคำขอใด ๆ ถูกสมมติว่าจะต้องประกอบด้วยเอกสารซึ่งสนองต่อคำขอด้วยความน่าจะเป็นที่สูงกว่าค่าที่ตั้งไว้ความน่าจะเป็นที่เอกสารได้จากการค้นคืนต้องสนองต่อคำขอและความน่าจะเป็นไม่ว่าต้องรวมกันเป็น 1 เหมือนการคำนวณของฟังก์ชันสมาชิกตามคอมพลีเมนต์ของเซต

ซึ่งข้อดีของคำขอแบบความน่าจะเป็นคือมีวิธีการคำนวณความน่าจะเป็นที่ได้รับการยอมรับโดยคำนวณความน่าจะเป็นจากข้อมูลความถี่ของคำ ข้อมูลความถี่ของคำจะสามารถนำมาใช้คะแนนความน่าจะเป็นที่เอกสารซึ่งบรรจุ Set ของคำจะสนองต่อคำขอ

4. Hybrid Model

Boolean Model มีข้อเสียคือการไม่รวมน้ำหนักของคำ Vector Space Model มีข้อเสียของการที่ไม่สามารถทำการเชื่อมต่อทางตรรกะได้โดยง่าย จึงได้มีความพยายามที่จะทำการเอาข้อดีของทั้งสองมารวมกัน จึงได้จึงได้มี Extended Boolean Model

ตารางที่ 1.2 : แสดงการเปรียบเทียบโมเดลระบบค้นคืนสารสนเทศในรูปแบบต่าง ๆ

Conceptual Model	File Structure	Query Operation	Term Operation	Document Operation
Boolean	Flat File	Feed Back	Stem (แกนคำ)	Parse
Extended Boolean	Inverted File	Parse (การแบ่ง)	Weight	Display
Probabilistic	Signature	Boolean	Thesaurus	Cluster
Vector space	Graphs		Truncation (บางส่วนของคำ)	Sort
	Hashing		Filed Mark	

สำหรับบทนี้นักศึกษาพอจะมองเห็นความสำคัญของระบบค้นคืนสารสนเทศแล้วว่า ระบบค้นคืนสารสนเทศ(Information Retrieval System หรือ IR) เป็นระบบที่จัดการประมวลผลสารสนเทศประเภทเอกสาร(Document) ในรูปแบบต่างๆ เช่น หนังสือ , วารสาร , บทความ เป็นต้น โดยเกี่ยวข้องในเรื่องการแสดงรูปแบบ ,การเก็บบันทึก ,การดึงเอกสาร มีการวิเคราะห์ข้อความของเอกสารเพื่อหาตัวแทนเอกสารและจัดเก็บข้อความเป็นตัวแทนของเอกสารหรือดรรชนี(index)แทนข้อความฉบับสมบูรณ์ นำตัวแทนเอกสารมาจัดแบ่งกลุ่ม (Classification) ข้อมูลที่ได้จัดแบ่งกลุ่มว่า คลัสเตอร์ (Cluster) ซึ่งการจัดแบ่งกลุ่มนี้ทำให้สามารถดึงเอกสารที่ต้องการได้รวดเร็วขึ้น และดึงกลุ่มของเอกสารที่เกี่ยวข้องได้มากขึ้น

นอกจากนี้นักศึกษายังได้ทราบถึงโมเดลของระบบค้นคืนสารสนเทศในรูปแบบต่างๆ ซึ่งจะกล่าวในรายละเอียดในบทต่อๆไป

แบบฝึกหัด

1. ระบบค้นคืนสารสนเทศ แตกต่างจากระบบบริหารฐานข้อมูลอย่างไร
2. ท่านคิดว่าทำไมต้องมีการพัฒนาประสิทธิภาพและประสิทธิผลของระบบค้นคืนสารสนเทศ จงให้เหตุผล
3. จงอธิบายความแตกต่างระหว่างการค้นคืนข้อมูลและการค้นคืนสารสนเทศ
4. จงอธิบายส่วนประกอบของระบบค้นคืนสารสนเทศมาพอเข้าใจ
5. จงอธิบายขั้นตอนในการสร้างระบบค้นคืนสารสนเทศมาพอเข้าใจ
6. การวัดประสิทธิภาพและประสิทธิผลของระบบค้นคืนสารสนเทศวัดจากสิ่งใดบ้าง
7. จงอธิบายถึง Precision และ Recall มาพอเข้าใจ
8. จงอธิบายถึงระบบตอบกลับ (feedback) มาพอเข้าใจ
9. จงอธิบายถึง Information Retrieval Model มาพอเข้าใจ

บรรณานุกรม

สุมาลี เมืองไพศาล ,"การจัดการข้อมูล และการเรียกใช้ข้อมูล" ,สำนักพิมพ์มหาวิทยาลัยรามคำแหง ,CS337 ,2535

ดร.ชูชาติ หฤไชยะศักดิ์ ,"Information Retrieval and Search engine" , ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

Sirak Kaewjamnong, "Text Properties and Languages" , ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยศิลปากร ,http://www.cs.su.ac.th/~sirak/517632/IR%20Model.ppt,

Sirak Kaewjamnong, "Basic Implementation and Evaluations", , ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยศิลปากร, http://www.cs.su.ac.th/~sirak/517632/IR%20Model.ppt

Prof. Ray Larson & Prof. Marc Davis , **"Information Organization and Retrieval"**, UC Berkeley SIMS , http://www.sims.berkeley.edu/academics/courses/is202/f04/