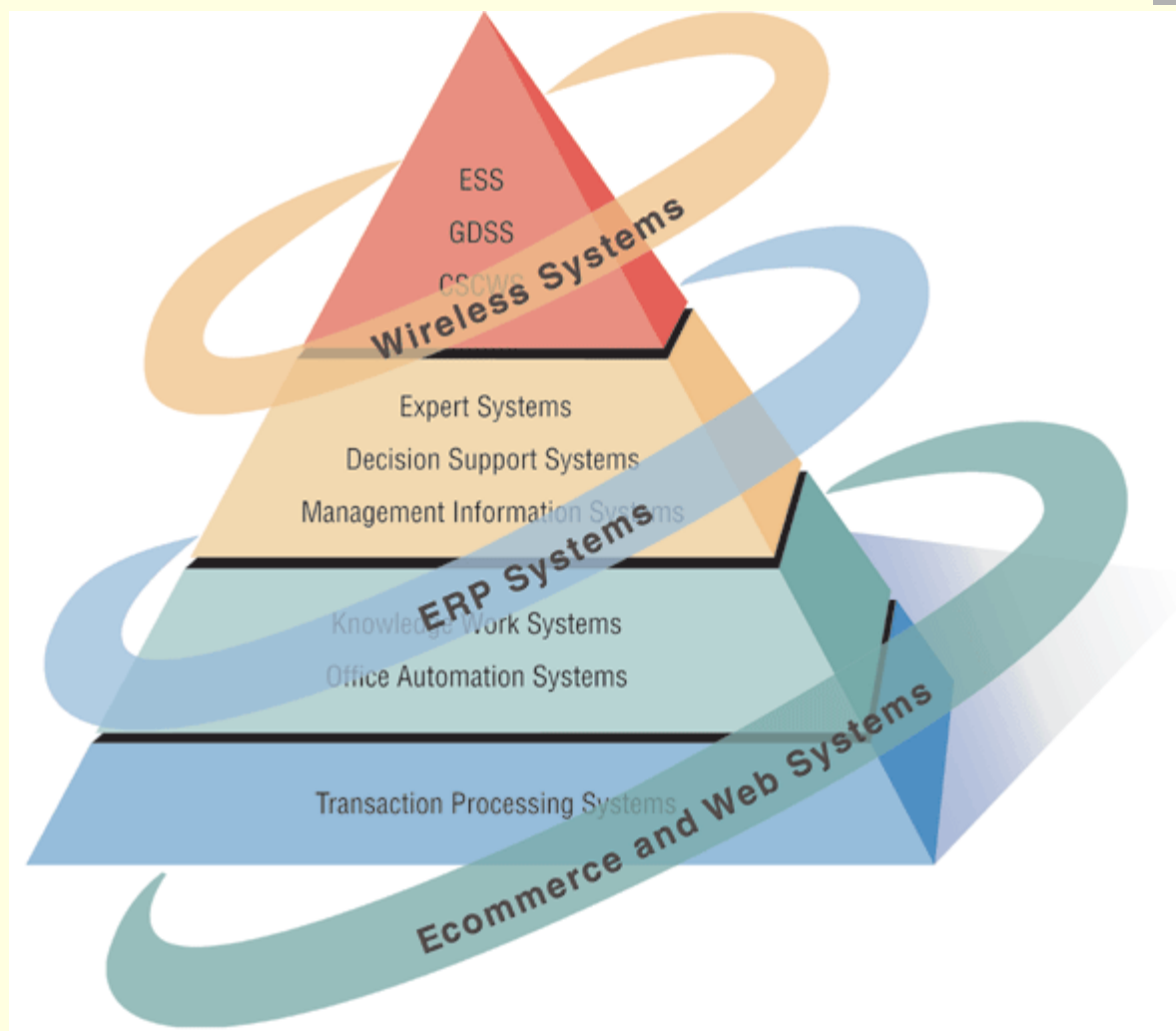


ครั้งที่ 2

ภาพรวมระบบคั่นคืนสารสนเทศ(ต่อ)

การนำเทคโนโลยีสมัยใหม่มาประยุกต์ใช้

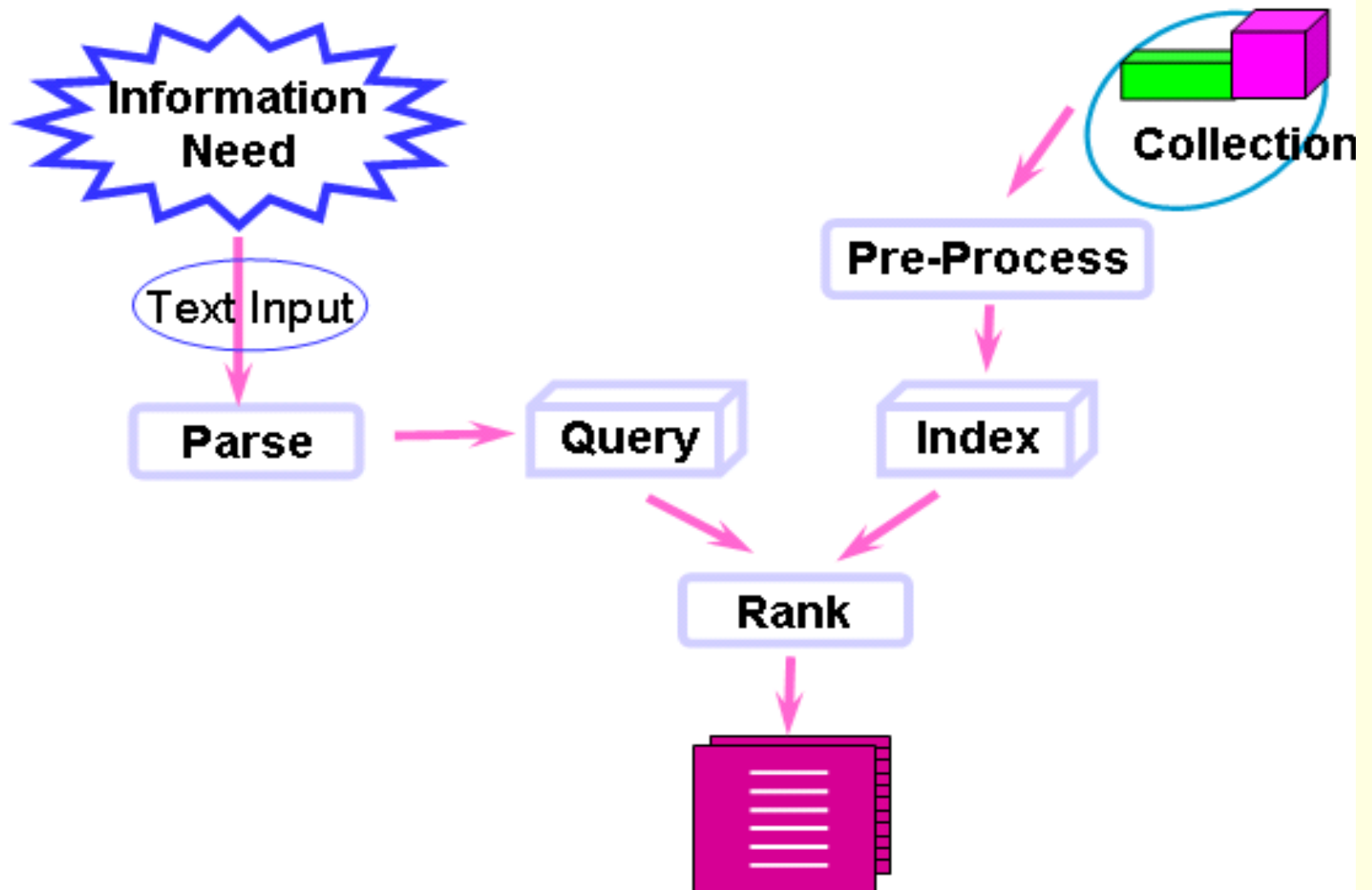


ระบบค้นคืนสารสนเทศ

(Information Retrieval System หรือ IR)

- ระบบที่จัดการประมวลผลสารสนเทศประเภทเอกสาร (Document) ในรูปแบบต่างๆ เช่น ข้อมูล, หนังสือ, วารสาร, บทความ เป็นต้น โดยเกี่ยวข้องในเรื่องการสร้างดัชนี, การเก็บบันทึก, การดึงเอกสาร โดยใช้กลยุทธ์และเทคนิคต่างๆ เพื่อแสดงรูปแบบของผลลัพธ์ให้กับผู้ใช้ซึ่งใช้เวลาและค่าใช้จ่ายในเกณฑ์ยอมรับได้

รูป : แสดงภาพรวมของระบบค้นคืนสารสนเทศ



ระบบค้นคืนสารสนเทศสามารถแบ่งได้ 3 ประเภทคือ

- ระบบค้นคืนที่ให้คำถาม — คำตอบ เป็นการบริการค้นคำตอบสำหรับคำถามที่ต้องการคำตอบ ที่เป็นข้อเท็จจริง
- ระบบค้นคืนที่ให้ข้อมูลเป็นตัวเลขหรือสัญลักษณ์ เป็นระบบจัดเก็บข้อมูลทางฟิสิกส์ เคมี สัมมะโนประชากร เป็นต้น
- ระบบค้นคืนข้อความจากวารสาร เป็นระบบที่จัดเก็บตัวเนื้อหาเอกสารและสามารถเรียกข้อความส่วนใดส่วนหนึ่งของเอกสารได้ เช่น ฐานข้อมูลทางกฎหมาย เป็นต้น

OSQA is the **free, open source** Q&A system

- it is a full-featured Q&A community.
- Users can ask and answer questions, comment and vote for the questions of others and their answers
- Both questions and answers can be revised and improved.

โปรแกรมตารางธาตุเคมี

Periodic Table of the Elements

Periods	I	II	III	IV	V	VI	VII	VIII			
I	H ¹								He ²		
II	Li ³	Be ⁴	B ⁵	C ⁶	N ⁷	O ⁸	F ⁹		Ne ¹⁰		
III	Na ¹¹	Mg ¹²	Al ¹³	Si ¹⁴	P ¹⁵	S ¹⁶	Cl ¹⁷		Ar ¹⁸		
IV	K ¹⁹	Ca ²⁰	Sc ²¹	Ti ²²	V ²³	Cr ²⁴	Mn ²⁵	Fe ²⁶	Co ²⁷	Ni ²⁸	
	Cu ²⁹	Zn ³⁰									Kr ³⁶
V	Rb ³⁷	Sr ³⁸									
	Ag ⁴⁷	Cd ⁴⁸									Xe ⁵⁴
VI	Cs ⁵⁵	Ba ⁵⁶									
	Au ⁷⁹	Hg ⁸⁰									Rn ⁸⁶
VII	Fr ⁸⁷	Ra ⁸⁸									
	Rg ¹¹¹	Cn ¹¹²									Uuo ¹¹⁸

*** Lantanoids**

La ⁵⁷	Ce ⁵⁸	Pr ⁵⁹	Nd ⁶⁰
------------------	------------------	------------------	------------------

**** Actinoids**

Ac ⁸⁹	Th ⁹⁰	Pa ⁹¹	U ⁹²	Np ⁹³	Pu ⁹⁴	Am ⁹⁵	Cm ⁹⁶	Bk ⁹⁷	Cf ⁹⁸	Es ⁹⁹	Fm ¹⁰⁰	Md ¹⁰¹	No ¹⁰²	Lr ¹⁰³
------------------	------------------	------------------	-----------------	------------------	------------------	------------------	------------------	------------------	------------------	------------------	-------------------	-------------------	-------------------	-------------------

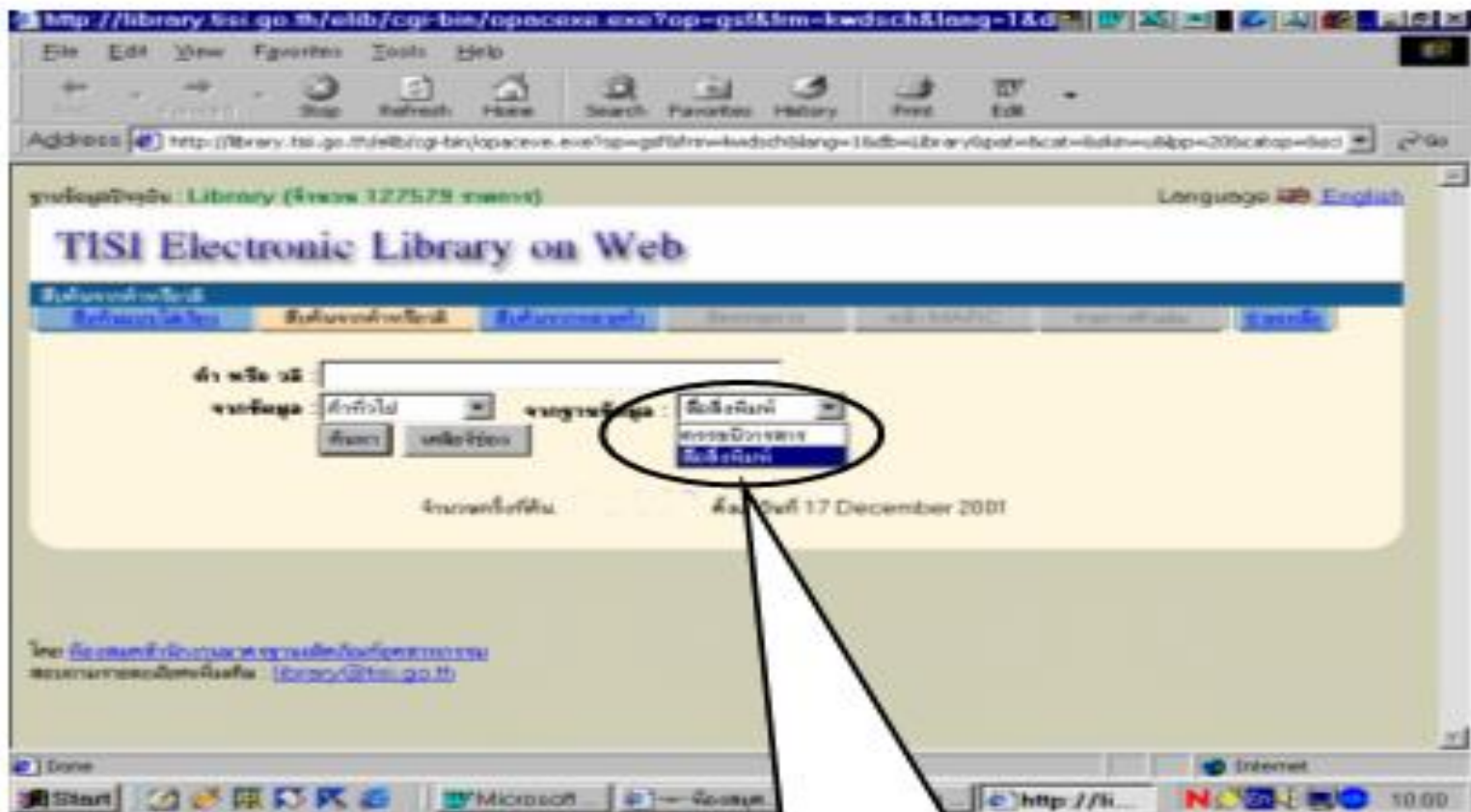
Color scheme:

s-block elements	p-block elements	d-block elements	f-block elements
------------------	------------------	------------------	------------------

To close the tooltip just click on X in the right upper corner

www.padlibrary.com

ระบบสืบค้นข้อมูลออนไลน์ (OPAC - Online Public Access Cataloging)



เลือกรูปแบบข้อมูลที่ต้องการสืบค้น
(สื่อสิ่งพิมพ์ / วรรณกรรม)

เครื่องมือช่วยค้นหาข้อมูลบนอินเทอร์เน็ตมี 5 ประเภท

ใหญ่ๆ คือ

(Bradley, 2002)

- Free text Search Engines
- Directory Search Engines
- Meta Search Engines
- Natural-language Search Engines
- Resource or Site-specific Search Engines

Free text Search Engines

<http://home.kku.ac.th/hslib/412141/internet/freetext.htm>

Google™
ประเทศไทย

เว็บ รูปภาพ กลุ่มข่าว สารบบเว็บ Desktop

ค้นหาโดย Google

ดีใจจัง ค้นแล้วเจอเลย

ค้นหา: ☒ เว็บ ☐ หน้าของ ประเทศไทย

ค้นหาแบบละเอียด
ปรับแต่งตัวเลือก
เครื่องมือเกี่ยวกับภาษา

altavista™

Web

Images

MP3/Audio

Video

Directory

News

FIND

Advanced Search

Settings

SEARCH: ☒ Worldwide or [Select a country](#)

RESULTS IN: ☒ All languages ☐ English

excite®

Join Now Sign In

Google Sponsored Links

Student Loans - Have a student loan or need one? Free resources and information.

Personalize: Settings Content Layout Colors & Themes Sign In

Web Search | Yellow Pages | White Pages

Search

Search Home | Submit Your Site

alltheweb

• • • find it all • • •

Web

News

Pictures

Video

Audio

SEARCH



sansarn.com
सरसार

เว็บ

ข่าวสาร.

สารบัญเว็บ

เว็บบอร์ด

เพิ่มเว็บ

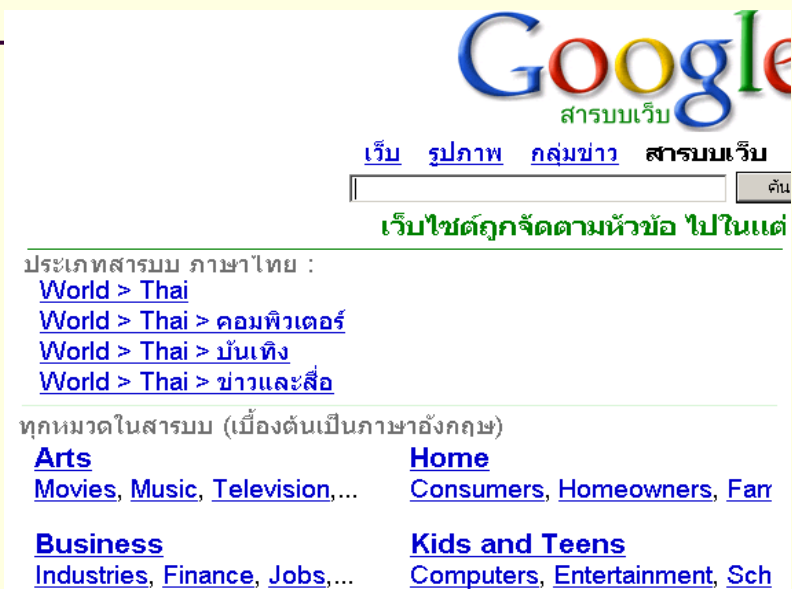
ค้นหา

ค้นหาคำใกล้เคียง

ค้นหามากกว่าหนึ่งคำ เลือก ☒ และ ☐ หรือ
เว้นวรรคระหว่างคำ ในกรณีที่ต้องการค้นหามากกว่าหนึ่งคำ เช่น "ค้นหา คำไทย"

[ติดต่อเว็บมาสเตอร์](#) | [ข้อมูลด้านเทคนิคของ"सरसार"](#)

Directory Search Engines



Meta Search Engines

The screenshot shows the Metacrawler website. At the top, the logo "metacrawler®" is displayed with the tagline "SEARCH THE SEARCH ENGINES!®". Below the logo, there are three tabs: "Web Search", "Yellow Pages", and "White Pages". The "Web Search" tab is active. Underneath, there are radio buttons for "Web Pages" (selected), "Images", "Audio", "Video", and "News". A search input field is present, followed by a "Search" button. To the right of the search field, there are links for "Advanced Web Search", "Preferences", and "Tools & Tips". Below the search area, there is a section titled "Metacrawler makes searching more of the Web easier by returning the best results from these leading engines:" followed by a list of engines: Google, Yahoo!, Ask Jeeves, About, LookSmart, Overture, and FindWhat. To the right of this list, there is a "MetaSpy®" section with the text "Ever wonder what the rest of the world is searching for?" and a small icon of a person. Below this, there is a "POPULAR SEARCHES" section with a list of search terms: "Chinese New Year", "Flowers", "Amazing Race", "Jennifer Aniston", "Minnie Driver", and "Misterio's Secret". To the right of the popular searches, there is a section titled "Try the all improved MetaCrawler Toolbar." with the text "MetaCrawler Toolbar now with:" followed by a list of features: "Personalized content ticker" and "Microsoft Word™ cursor search".

metacrawler®
SEARCH THE SEARCH ENGINES!®

Web Search **Yellow Pages** White Pages

☒ Web Pages ☐ Images ☐ Audio ☐ Video ☐ News

Search

☐ Exact Phrase

[Advanced Web Search](#)
[Preferences](#)
[Tools & Tips](#)

Metacrawler makes searching more of the Web easier by returning the best results from these leading engines:
Google • Yahoo! • Ask Jeeves
About • LookSmart • Overture
FindWhat

MetaSpy®
Ever wonder what the rest of the world is searching for?

POPULAR SEARCHES

- [Chinese New Year](#)
- [Flowers](#)
- [Amazing Race](#)
- [Jennifer Aniston](#)
- [Minnie Driver](#)
- [Misterio's Secret](#)

Try the all improved MetaCrawler Toolbar.
MetaCrawler Toolbar **now with:**

- Personalized content ticker
- Microsoft Word™ cursor search

Dogpile (<http://www.dogpile.com>)
Metacrawler (<http://www.metacrawler.com>)
ProFusion (<http://www.profusion.com>)
Search (<http://www.search.com>)
SurfWax (<http://www.surfwax.com>)
Ixquick (<http://www.ixquick.com>)

Natural-language Search Engines

Ask.com | MyJeeves^{BETA} Sign In | Your Settings | Help

Ask Jeeves[®] Ask.com

[Web](#) | [Pictures](#) | [News](#) | [Local](#) ^{NEW!} | [Products](#) | [More](#) »

what is thai language Search [Advanced](#)

Web Search: what is thai language 1-10 results out of 1,162,000

[Sponsored Web Results](#)

Thai Language
Our award-winning software will teach you the **Thai language**.
www.RosettaStone.com

Rapidly Learn Thai
Learn in 10 days or you don't!
www.pimsleurapproach.com

[Web Results](#)

Learning Thai Language the Easy Way!
Learn the **Thai language** the easy way.
 | www.learningthai.com/ | [Cached](#) | [Save](#)

คำถาม

คำตอบ

User-centric Math Information Retrieval System

Linked Expression:
 $a^2+b^2=c^2$

Filter by > **resource category**, experience, specificity

Tutorials, Slides, Problem Solution Set, Applets, Tools, Data, Algorithms

Pythagorean theorem - Wikipedia, the free encyclopedia

If we let c be the length of the hypotenuse and a and b be the lengths of the other two sides, the theorem can be expressed as the equation:

$$a^2 + b^2 = c^2$$

en.wikipedia.org/wiki/Pythagorean_theorem - 98k - [Cached](#) - [Similar pages](#)

Pythagorean Theorem and its many proofs from Interactive ...

Pythagoras' Theorem. 78 proofs of the **Pythagorean theorem**.

www.cut-the-knot.org/pythagoras/index.shtml - 157k - [Cached](#) - [Similar pages](#)

The Pythagorean Theorem Lesson

A lesson on the **pythagorean theorem** with the objective that the student discovers or figures out by him or herself the actual **theorem**.

www.arcytech.org/java/pythagoras/ - 5k - [Cached](#) - [Similar pages](#)

Definition...

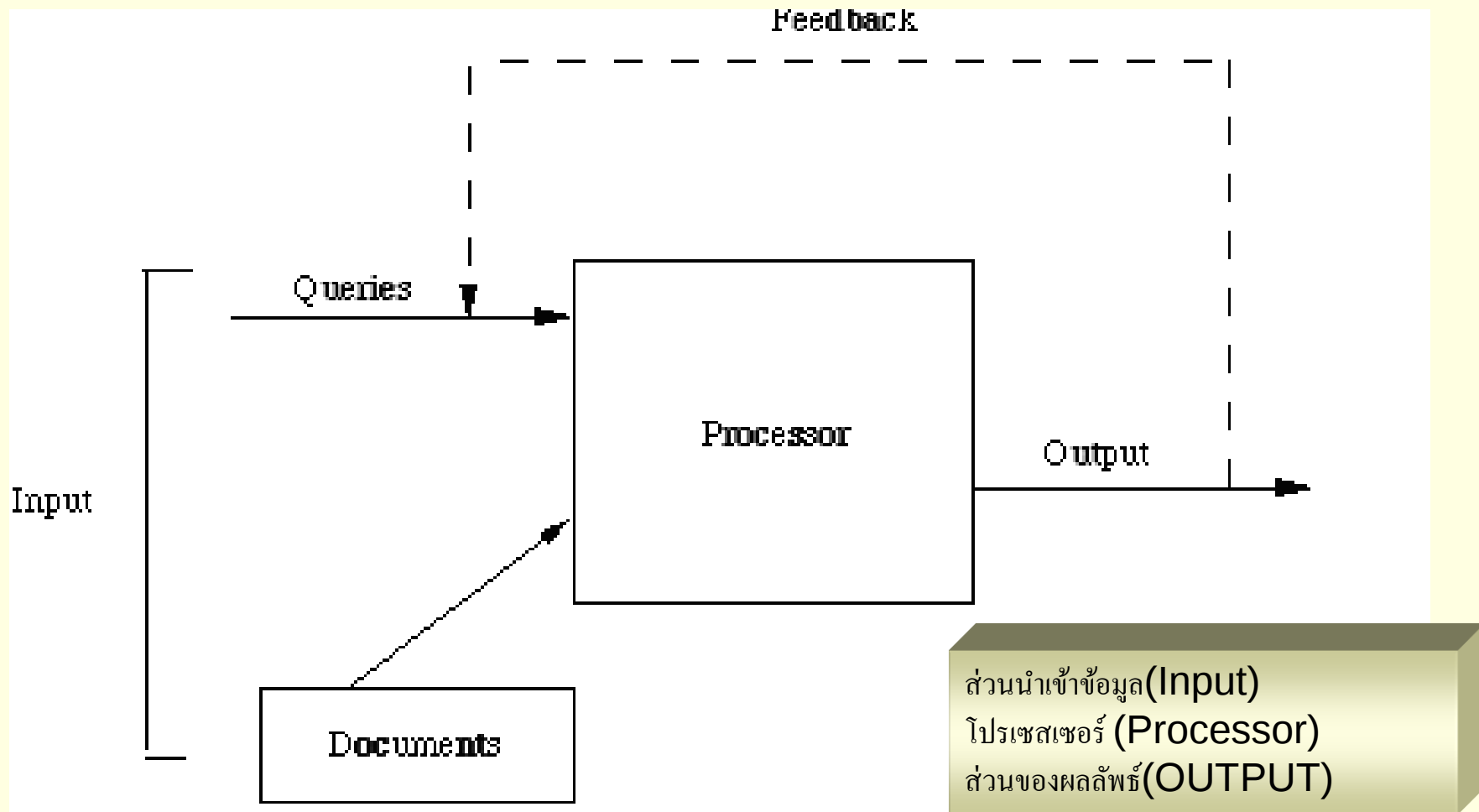
Proof

Tutorial

รูป แสดงความแตกต่างระหว่าง Database และ IR

	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. คลุมเครือ	Vague, imprecise information needs (often expressed in natural language)
Results we get	Exact. Always correct in a formal sense. แน่นนอน ถูกต้อง	Sometimes relevant, often not.
Interaction with system	หมาย One-shot queries.	Interaction is important. (relevance feedback)

รูป: แสดงส่วนประกอบของระบบค้นคืนสารสนเทศ



ส่วนนำเข้าข้อมูล(Input)

- เป็นส่วนของการป้อนข้อความ(query)จากผู้ใช้งานซึ่งเป็นภาษาธรรมชาติ หรืออาจเป็นการนำเข้าMetadata ซึ่งเป็นสารสนเทศเกี่ยวกับเอกสาร หรืออาจไม่เป็นส่วนหนึ่งของเอกสารก็ได้แต่เป็นข้อมูลเกี่ยวกับข้อมูล (data about data) หรือคำถาม

Descriptive metadata

- ผู้แต่ง(Author)
- ชื่อเรื่อง(Title)
- แหล่งที่มา (book, magazine)
- วันที่ (Date)
- ISBN
- สำนักพิมพ์(Publisher)
- ความยาว(Length)

Semantic metadata concerns the content

- บทคัดย่อ(Abstract)
- คำสำคัญ(Keywords)
- รหัสของหัวเรื่อง(Subject Codes)

เทอมของหัวเรื่อง (Subject terms)

สารสนเทศของเว็บ(Web Metadata)
-META tag in HTML

โปรเซสเซอร์ (Processor)

- เป็นส่วนของการประมวลผล ได้แก่ การจัดโครงสร้างของสารสนเทศในรูปแบบที่เหมาะสม

การสร้างตัวแทนเอกสาร

การแบ่งแยกกลุ่มของเอกสาร

การจัดเก็บสารสนเทศ

การดึงข้อมูลตามที่ใช้ต้องการ

การทำงานนั้นจะนำข้อคำถามไปเปรียบเทียบกับตัวแทนเอกสารที่มีอยู่ เพื่อดึงเอกสารที่ใกล้เคียงนำออกมาให้แก่ผู้สอบถาม

ส่วนของผลลัพธ์(OUTPUT)

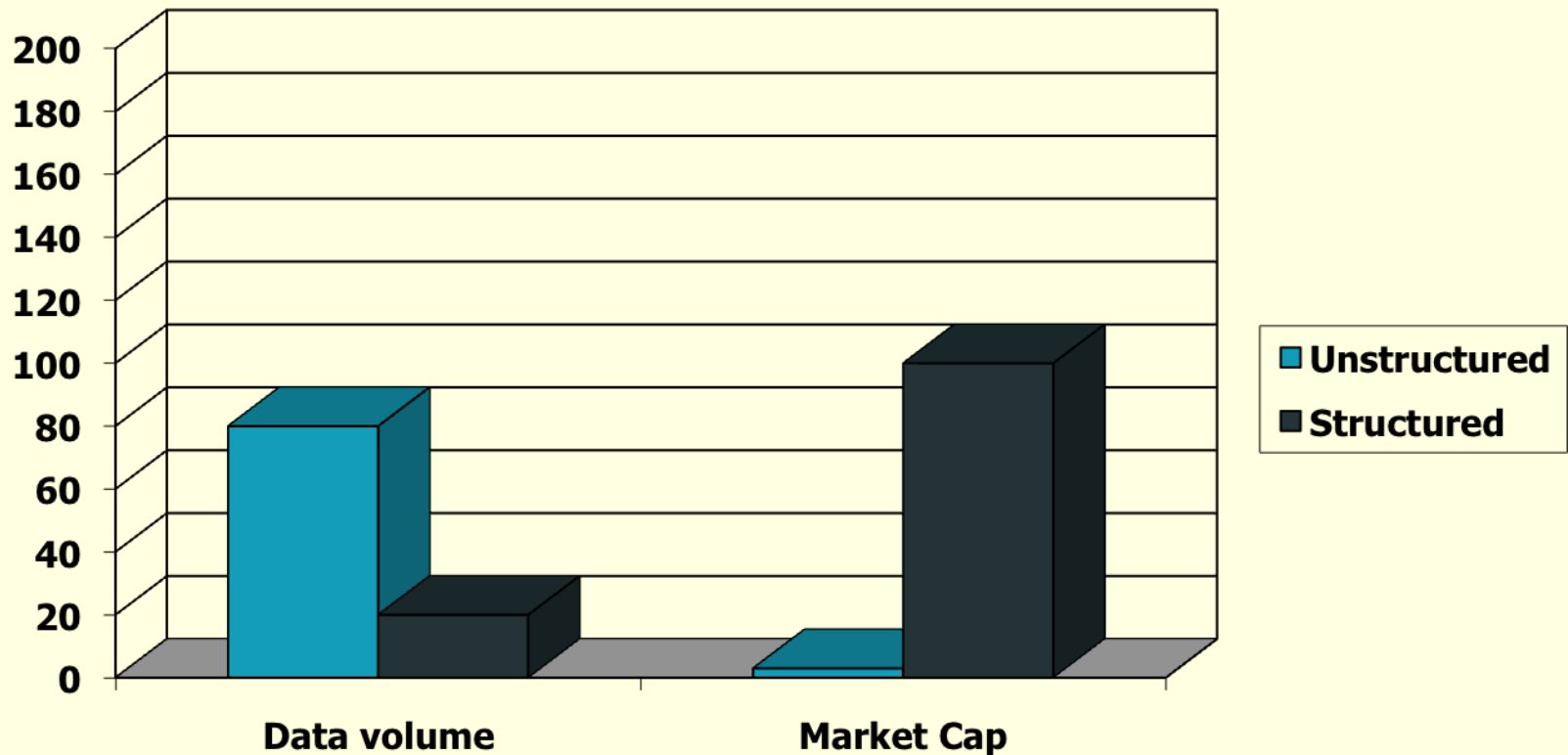
- ผลลัพธ์ที่ได้จากระบบเป็นข้อความสั้นๆ เช่น ชื่อ หนังสือ, หมายเลขเอกสาร, ชื่อผู้แต่ง, สำนักพิมพ์ เป็นต้น
- ผู้ใช้สามารถปรับปรุงข้อคำถามใหม่เพื่อให้ข้อคำถามนั้นสืบค้นสารสนเทศได้ตรงกับความต้องการมากที่สุด เป็นระบบตอบกลับ (feedback) ดังนั้นผลลัพธ์ที่ได้จึงขึ้นอยู่กับ ข้อคำถาม (Query)

ขั้นตอนในการสร้างระบบ IR

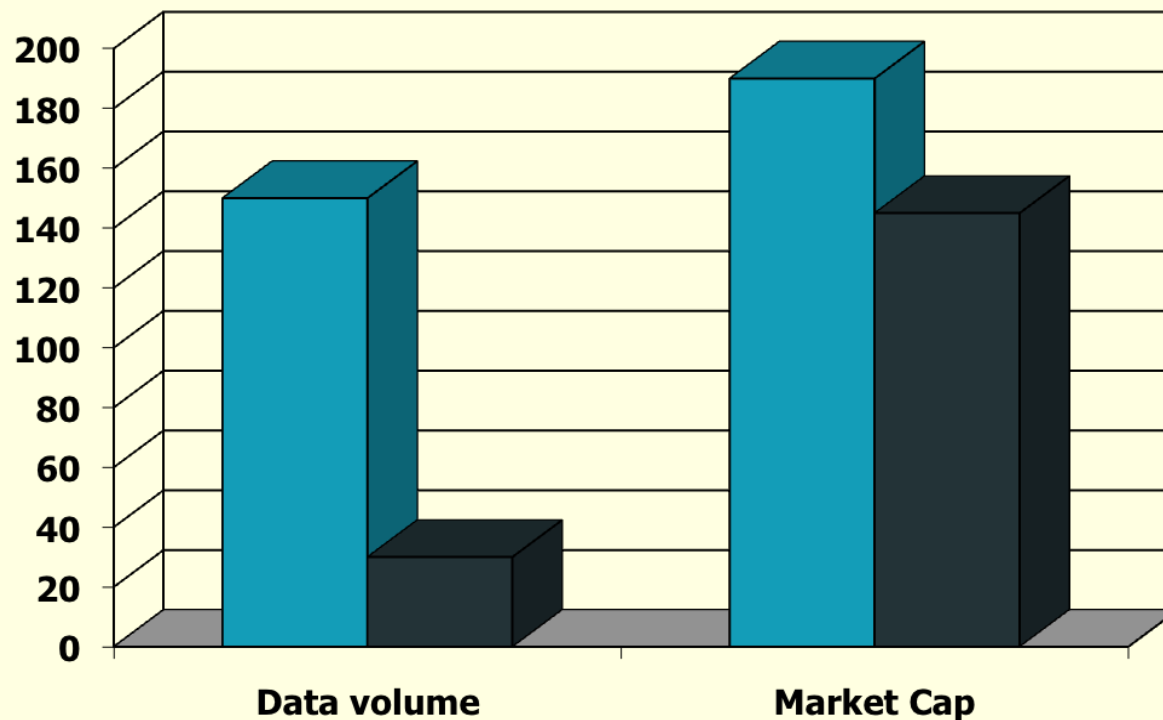
การสร้างระบบค้นคืนสารสนเทศ แบ่งออกเป็น 4 ขั้นตอนคือ

- การวิเคราะห์ข้อความ (Text Analysis)
- การจัดแบ่งกลุ่มข้อมูล (Classification)
- การเก็บบันทึกข้อมูลลงในแฟ้มข้อมูล
- การค้นคืนสารสนเทศ

Unstructured (text) vs. structured (database) data in 1996



Unstructured (text) vs. structured (database) data in 2009



Google™

YAHOO!®

■ Unstructured
■ Structured

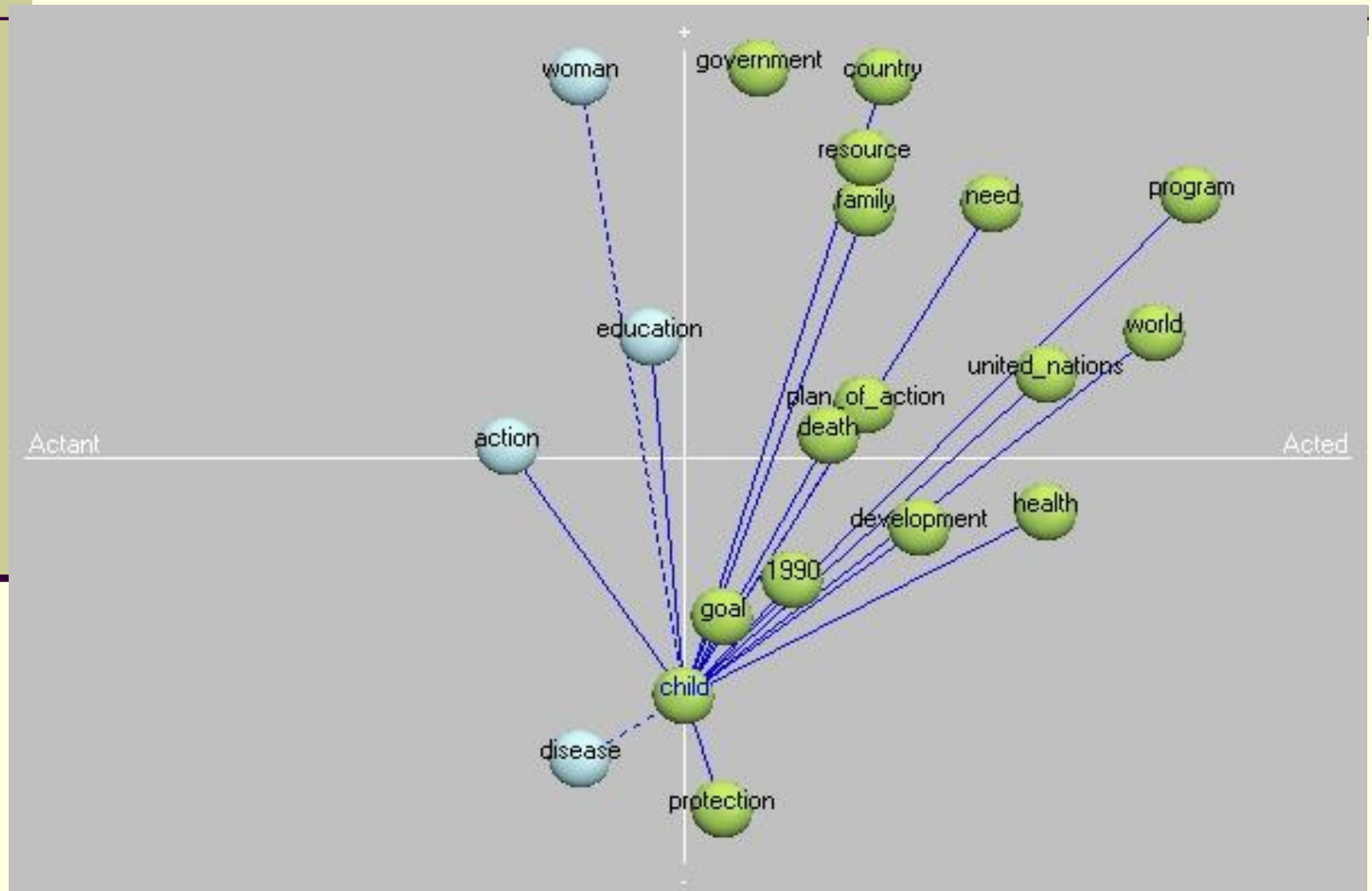
bing™

Ask™
.com

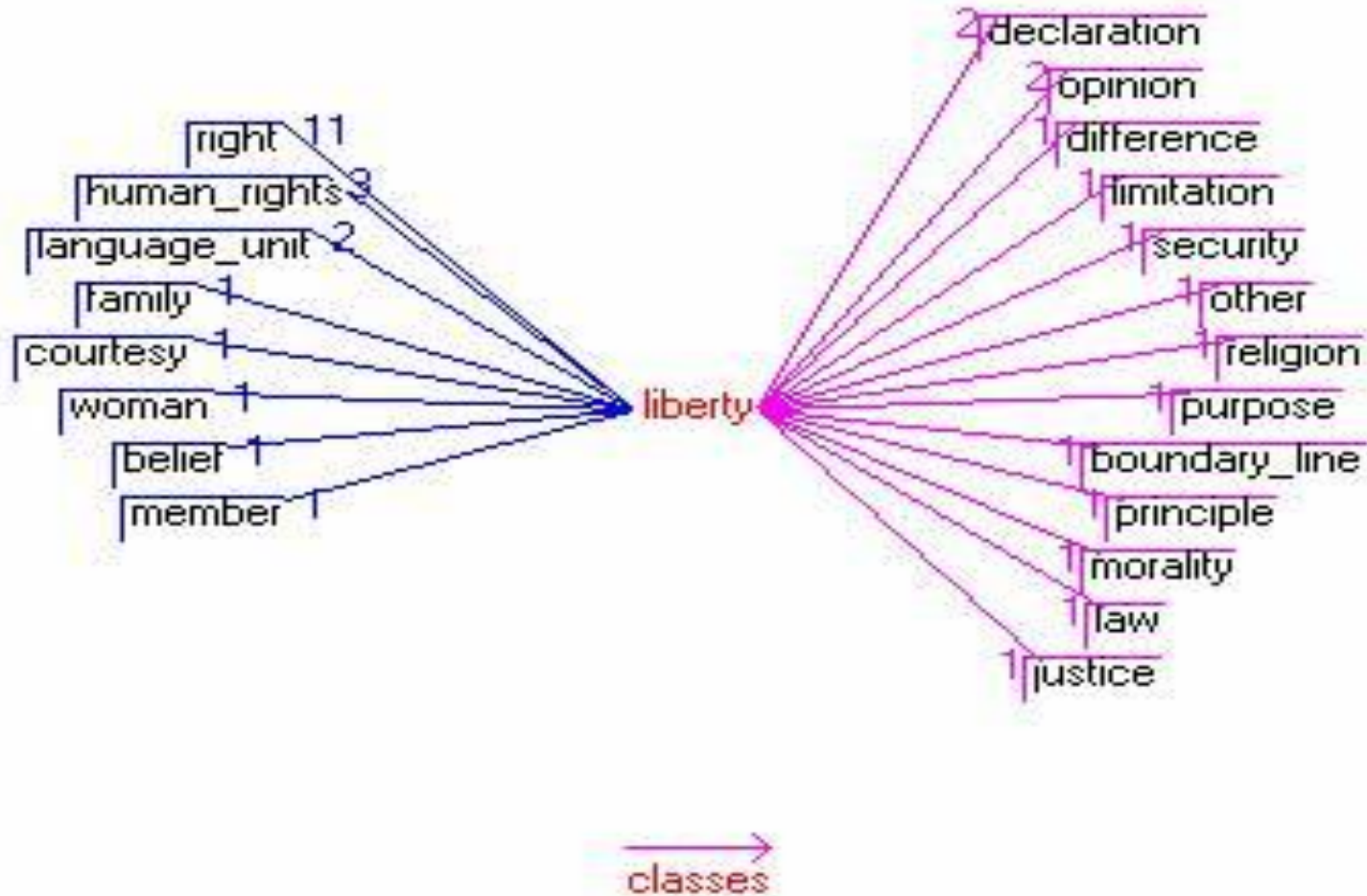
Information Retrieval



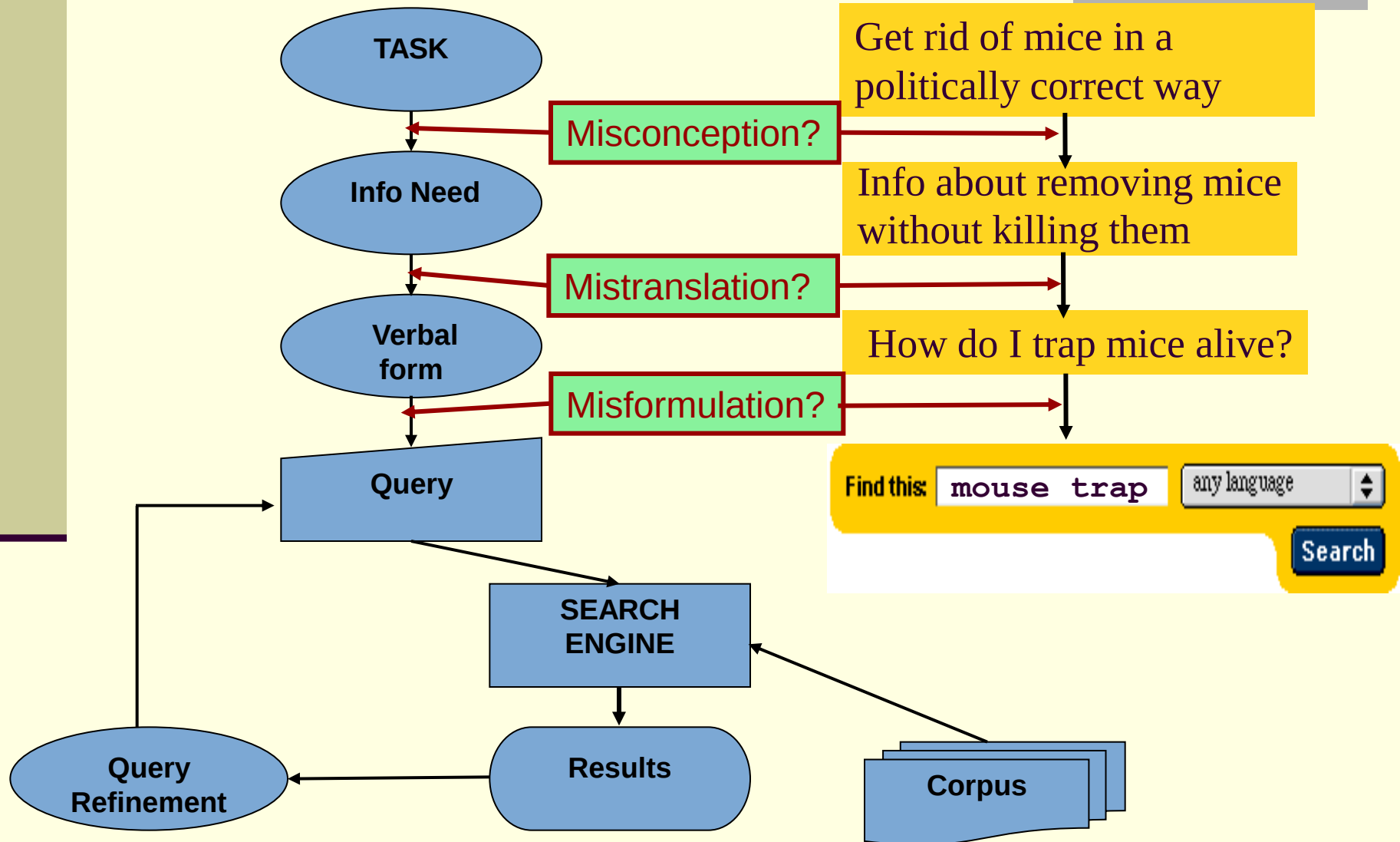
Text Analysis graph



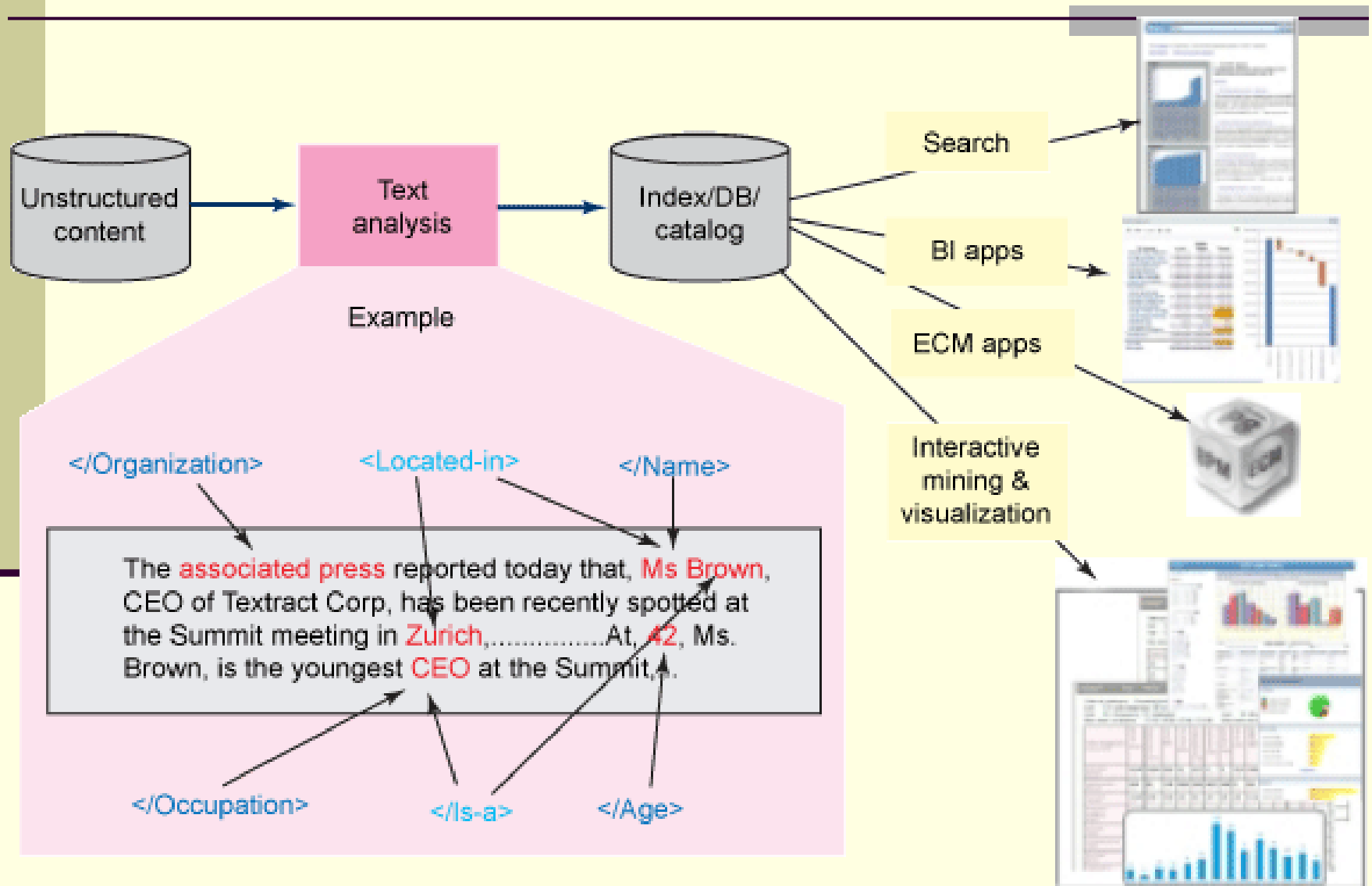
Text Analysis star graph



The classic search model



Text analysis enhances a wide variety of applications



OpenI Business Intelligence (BI)

- ซอฟต์แวร์ (**Software**) ที่นำข้อมูลที่มีอยู่เพื่อจัดทำรายงานในรูปแบบต่างๆ โดยทำหน้าที่ในการดึงข้อมูลจาก **Database** โดยตรงแล้วนำเสนอในรูปแบบของ **Report** ชนิดต่างๆที่เหมาะสมกับมุมมองในการวิเคราะห์ และตรงตามความต้องการของผู้ใช้งานในรูปแบบหลายมิติ (**Multidimensional Model**) ซึ่งจะ
ทำให้สามารถดูข้อมูลแบบเจาะลึก (**Drill-down**) ได้

Enterprise Content Management (ECM)

- ซอฟต์แวร์ใช้รวบรวมเนื้อหาที่มีปริมาณจำนวนมากมหาศาล และถูกจัดเก็บอยู่หลายรูปแบบ มาผ่านกระบวนการจัดการเนื้อหาจัดหมวดหมู่, จัดเรียง, การประสานข้อมูลชนิดต่างๆ เพื่อให้ข้อมูลมีความสัมพันธ์ ทำให้ง่ายต่อการนำมาใช้งาน เป็นการจัดการเนื้อหาในระดับองค์กร โดยจัดเก็บข้อมูลแบบไม่มีโครงสร้าง(UnStructure Data)
 - จากไมโครซอฟต์ออฟฟิศ(MS office, exels, Power Point)
 - ไฟล์เว็บ เช่น .html, XML
 - ไฟล์ที่เป็นเอกสารอิเล็กทรอนิกส์ ได้แก่ .PDF
 - ไฟล์ภาพถ่าย เช่น .JPG, .GIF, .RAW,
 - ไฟล์วิดีโอ หรือไฟล์เสียง ได้แก่ .avi, .dat, .VOB
 - ไฟล์ข้อมูล ไม่ว่าจะเป็นไฟล์ข้อความ, .DBF, .MDB, .GDB, .DB)
 - ไฟล์บีบ(Zip files)

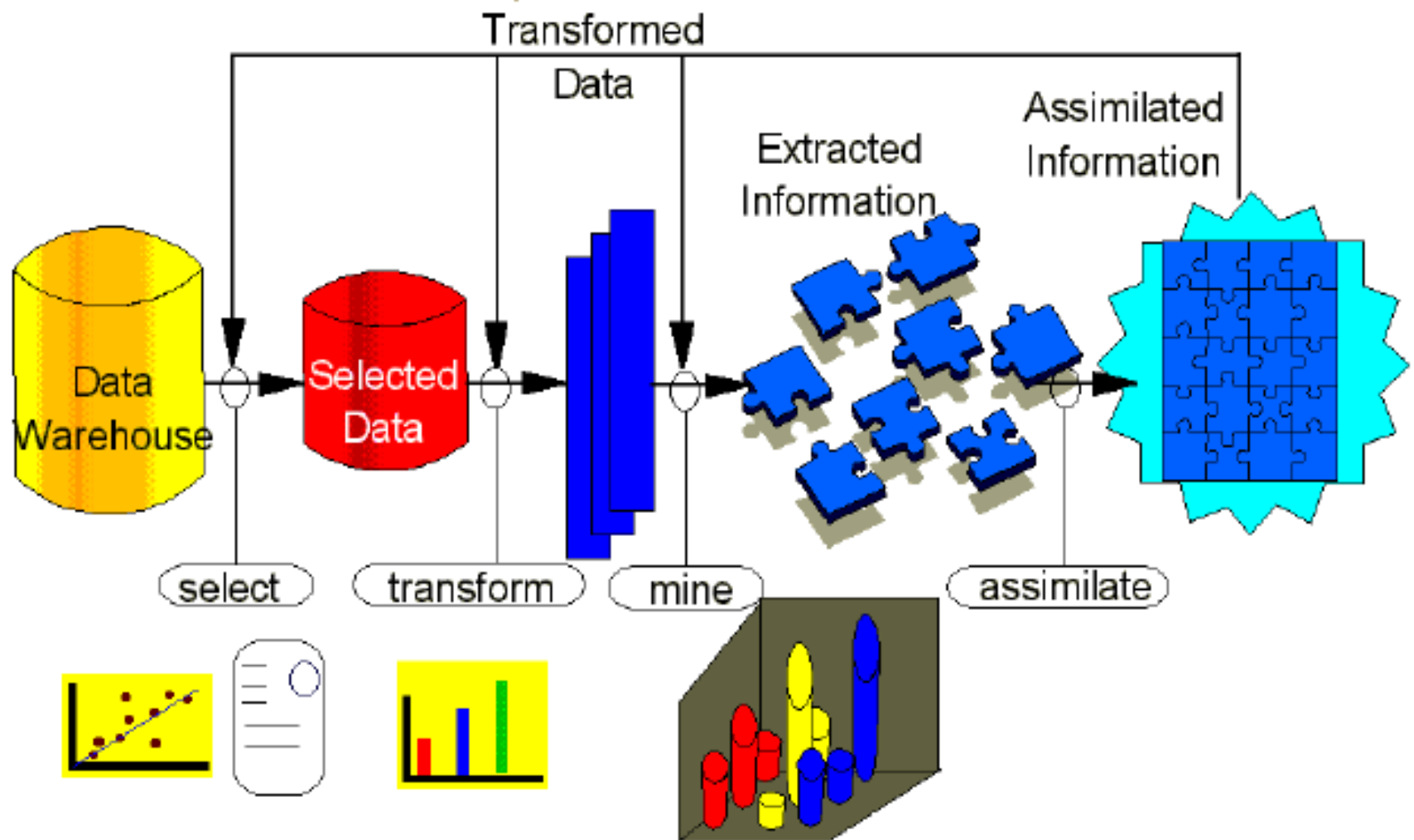


Data mining

- กระบวนการค้นหาสารสนเทศทางธุรกิจที่มีมูลค่าต่อองค์กร
ในฐานะข้อมูลขนาดใหญ่ คลังข้อมูล หรือตลาดข้อมูล
- สามารถทำนายแนวโน้มและพฤติกรรมต่างๆ โดยอัตโนมัติ



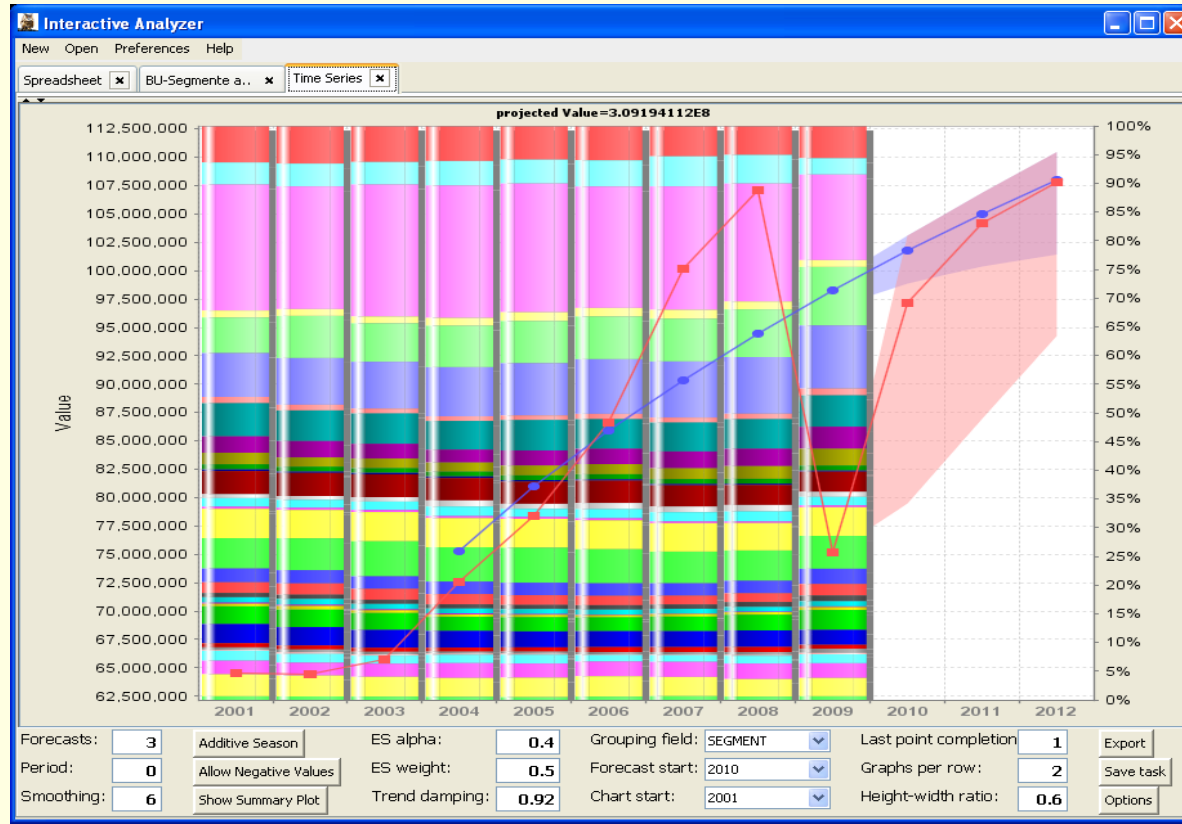
Data mining



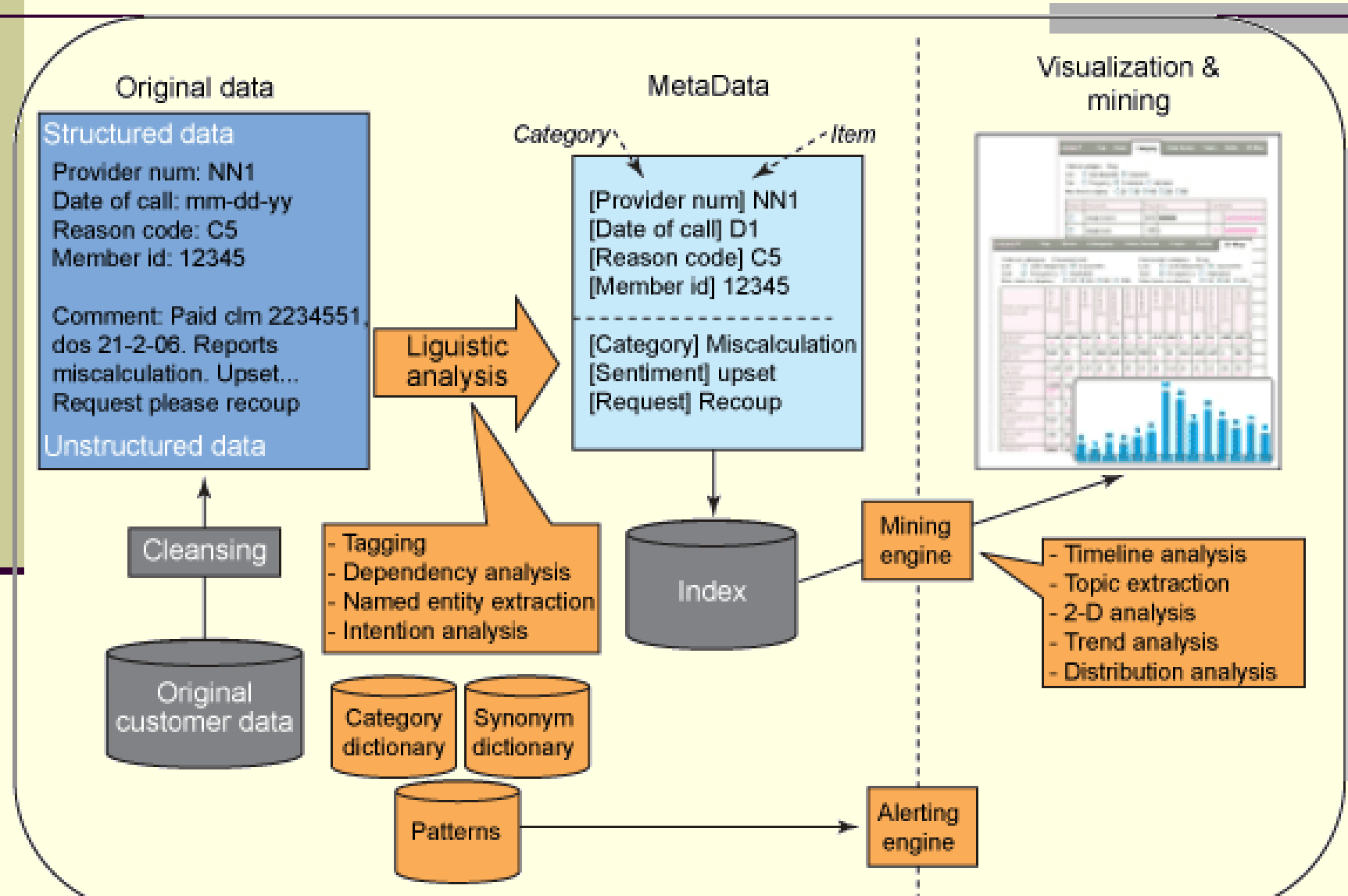
Data Visualization

- เป็นการนำเสนอข้อมูลโดยใช้เทคโนโลยีต่างๆ เช่น รูปภาพ ตาราง วิดีโอ กราฟ ภาพเคลื่อนไหว สื่อประสม
- สามารถมองภาพในหลายมิติในเวลาเดียวกัน

- สินค้า
- พนักงานขาย
- ส่วนแบ่งตลาด



Analytics Edition Architecture

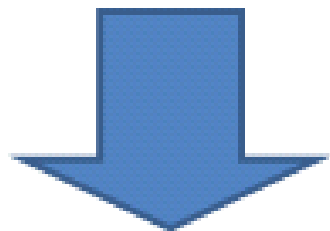


Data Cleansing

- **Data cleansing** คือขั้นตอนในการรับประกันว่าข้อมูลนั้นๆมีความชัดเจนและถูกต้อง ซึ่งในขั้นตอนการทำนั้นข้อมูลที่บันทึกไว้จะถูกตรวจสอบความถูกต้องหรือลบข้อมูลบางส่วนที่ไม่จำเป็นออกไป
- การทำ **Data cleansing** สามารถทำได้ทั้งภายในข้อมูลชุดเดียวหรือระหว่างข้อมูลหลายชุดซึ่งสามารถนำมารวมกันได้ หรือเป็นข้อมูลที่ต้องมีการใช้งานร่วมกัน

Database cleaning

#	First Name	Last Name	Age	Gender	Mobile	Telephone 1	Address
1	David	Ishikawa	26	Male		02-211-0063	3-12 soi 5 Rachtevi
2	Bill	Chringsam	43	Male	08-9999-8765		354 soi 4 Nana Sukumvit
3	Nancy	Namachue	29	Female	08-1111-3456	02-310-8633	Sukumvit Klongtoey Bangkok
4	Mie	Nakanishi	26	Female	0-1234-9965	02-260-5544	
5	David	Ishika	26	Male	08-7511-1234		



- Check same person or not (If we find a duplicate name, combine as one)
- Update the telephone number (mobile & Fix)
- Update the new address

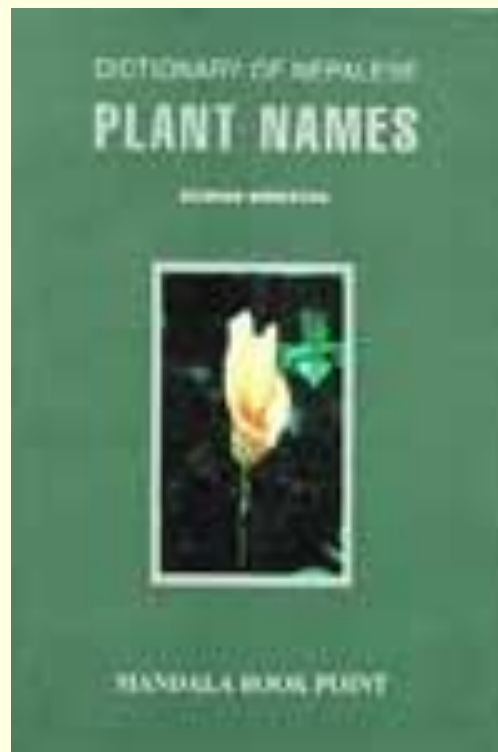
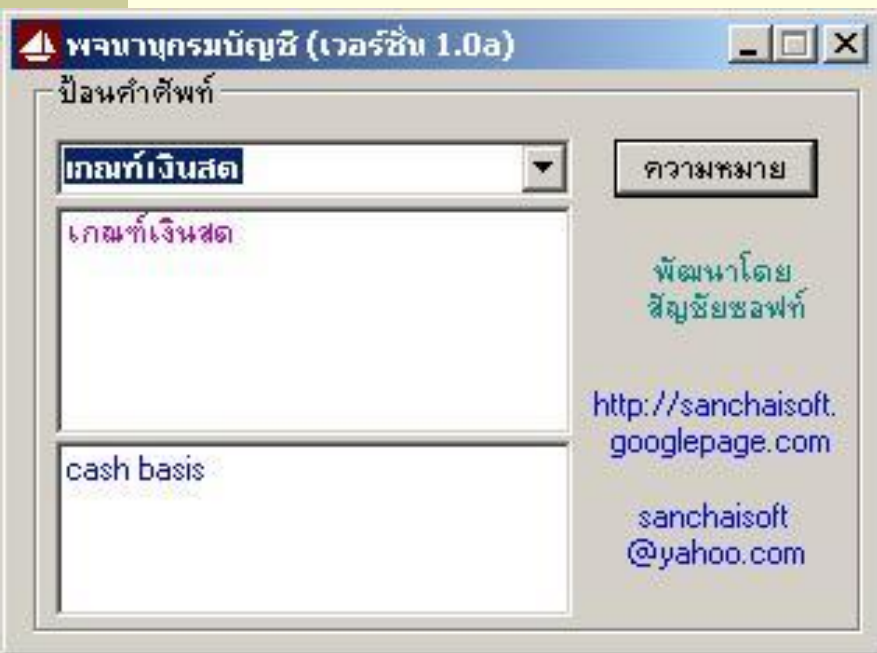
#	First Name	Last Name	Age	Gender	Mobile	Telephone 1	Address
1	David	Ishikawa	26	Male	08-7511-1234	02-211-0063	3-12 soi 5 Rachtevi
2	Bill	Chringsam	43	Male	08-9999-8765	02-416-0011	354 soi 4 Nana Sukumvit
3	Nancy	Namachue	29	Female	08-1111-3456	02-310-8633	5-10 A Mansion 104 soi 4 Sukumvit ...
4	Mie	Nakanishi	26	Female	08-1234-9965	02-260-5544	71 soi Ratprao ...

Linguistic Analysis

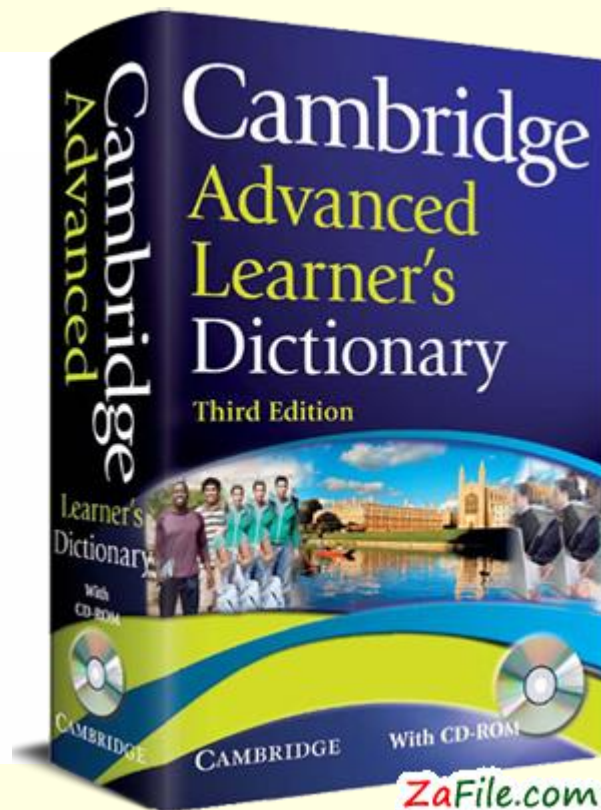
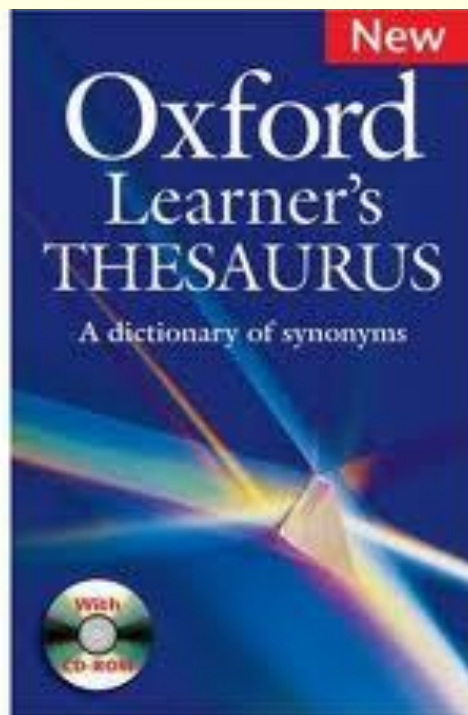
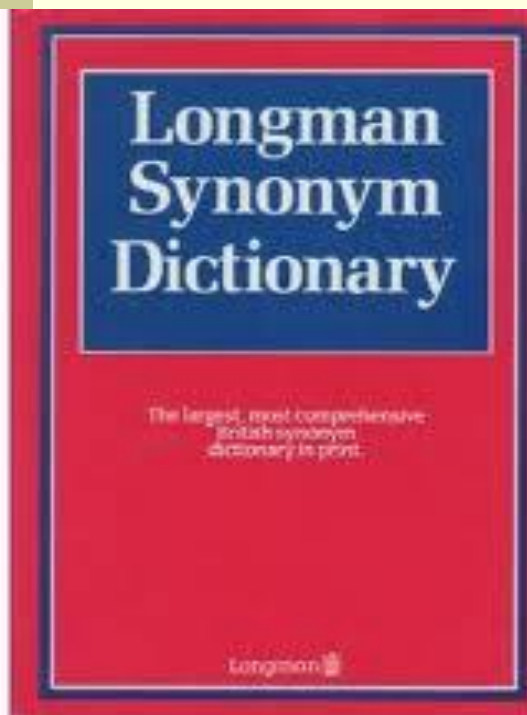
■ ขจัดคำกำกวมที่ใช้ในภาษา

- Tagging คือ index หรือ keyword ของเนื้อหา
- dependency analysis วิเคราะห์คำที่ขึ้นต่อกัน
- name entity extraction แยกเ็นดีที่ชื่อ
- intention analysis วิเคราะห์เจตจำนงค์ ความมุ่งมั่น

Category dictionary



Synonym dictionary





View text



File path :

Article.doc

Open

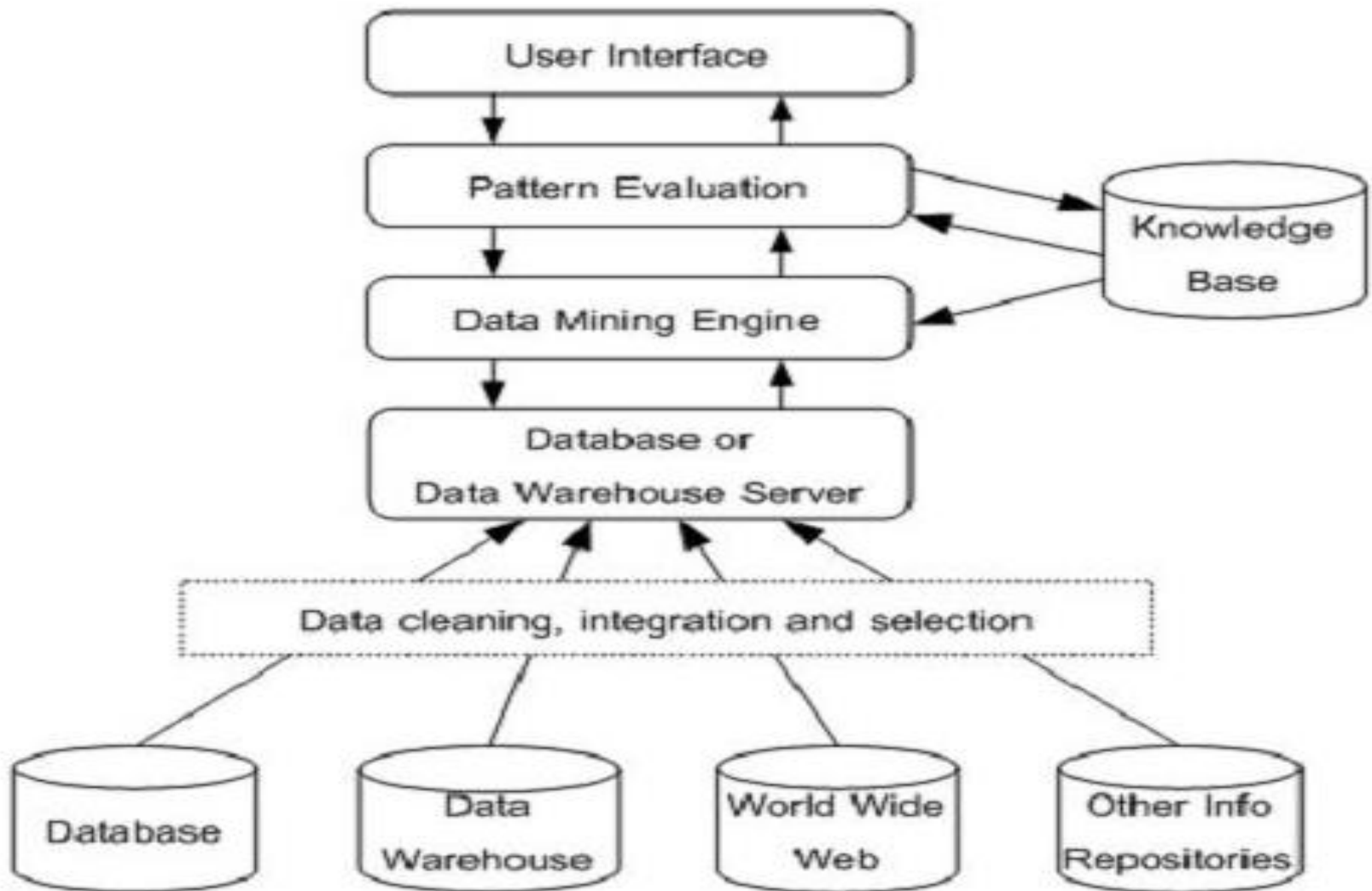
Introduction

The most popular file format for rich text representation is *.doc file format created by Microsoft. OLE (Object Linking and Embedding) is the most popular way to work with MS Word documents programmatically, but this way has several disadvantages, e.g. low speed, need of MS Office to be installed and inconvenient document model.

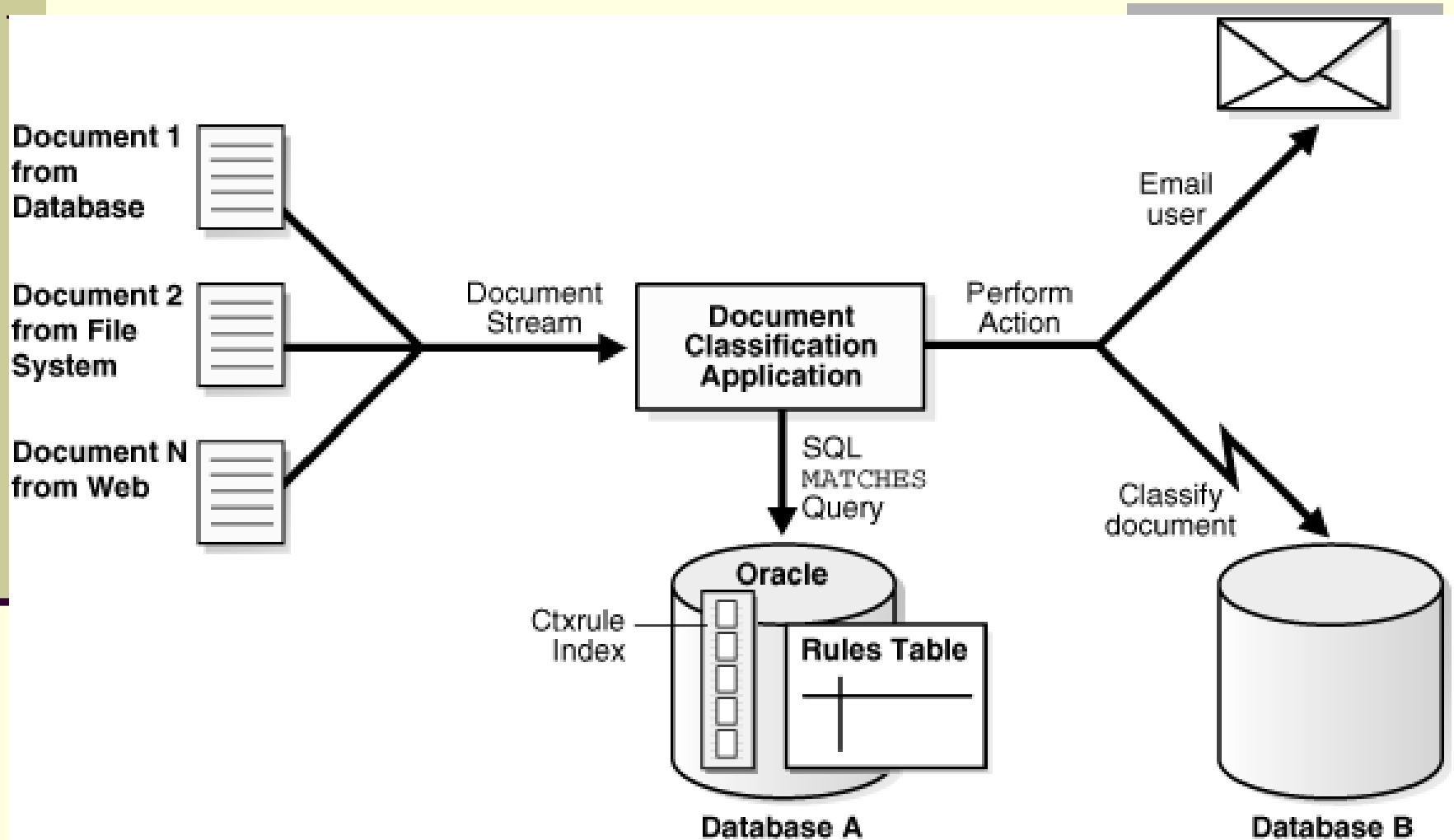
Another way to manipulate MS Word files is to read and write them directly. This is one way to solve the described above disadvantages. Direct manipulation provides much more speed and allows use of your own document model. However the most significant difficulty here is binary file format knowledge.

In this article we'll take a brief look into MS Word binary file format and present a simple way to obtain document text from *.doc file. The complete source code for this article can be found here: [Cellbi.GetDocText](#). You may also take a look at the [Cellbi.DocFramework](#) project we are currently working on.

Mining engine

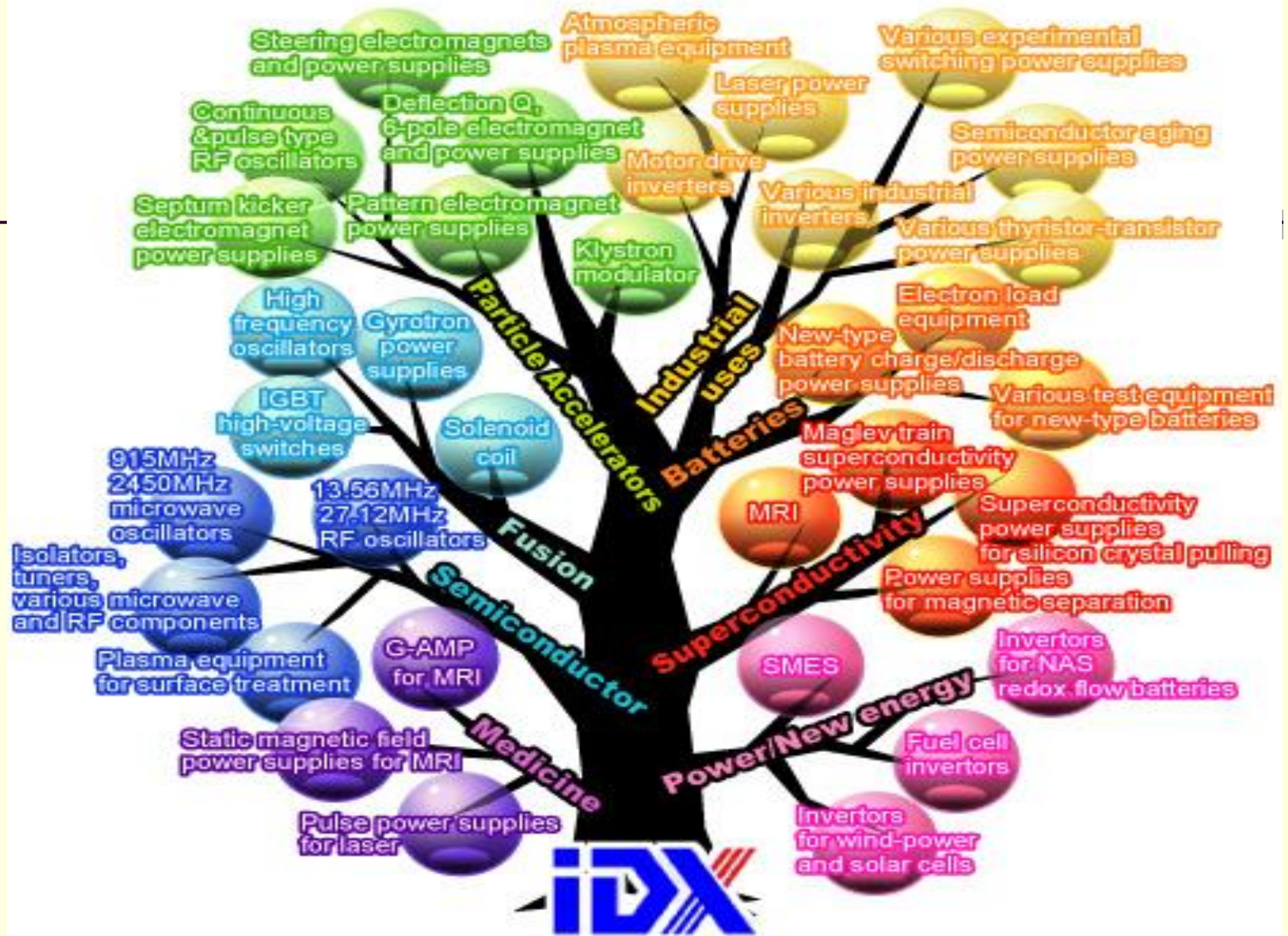


Overview of a Document Classification Application



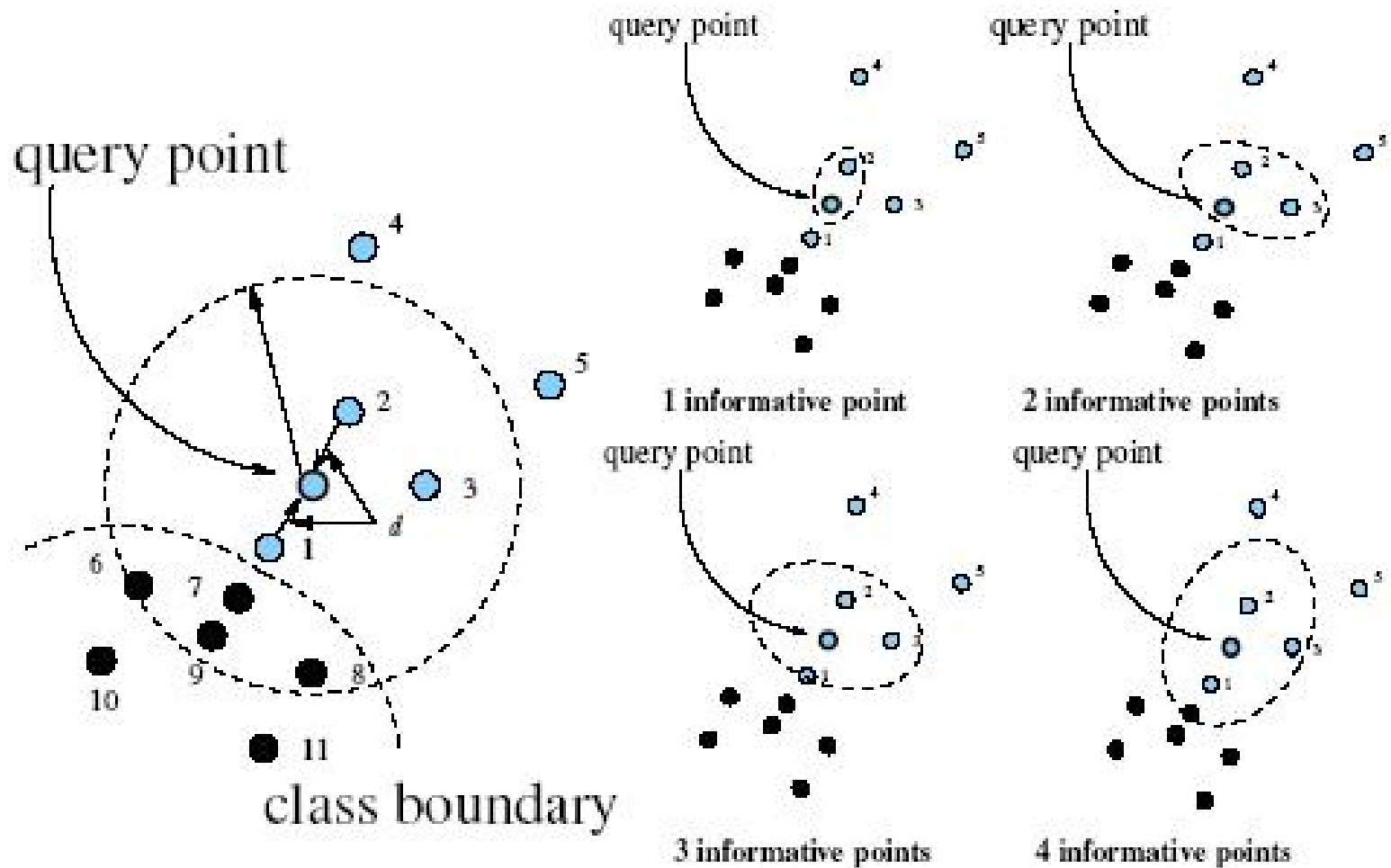
Tree of classification of the order Coleoptera (to family level) on the site on the Minnesota State University (USA)



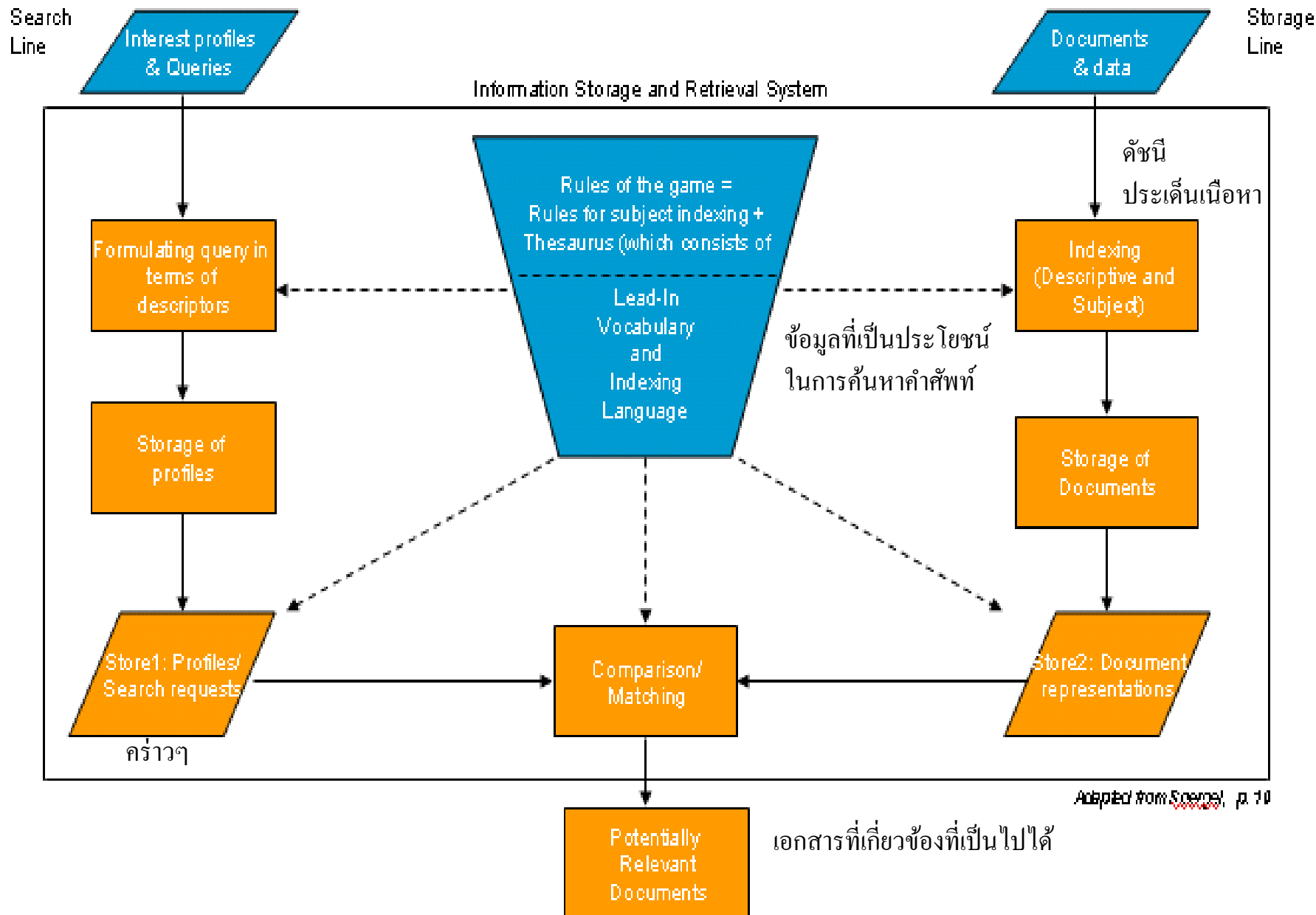


IDX Power Electronics Tree

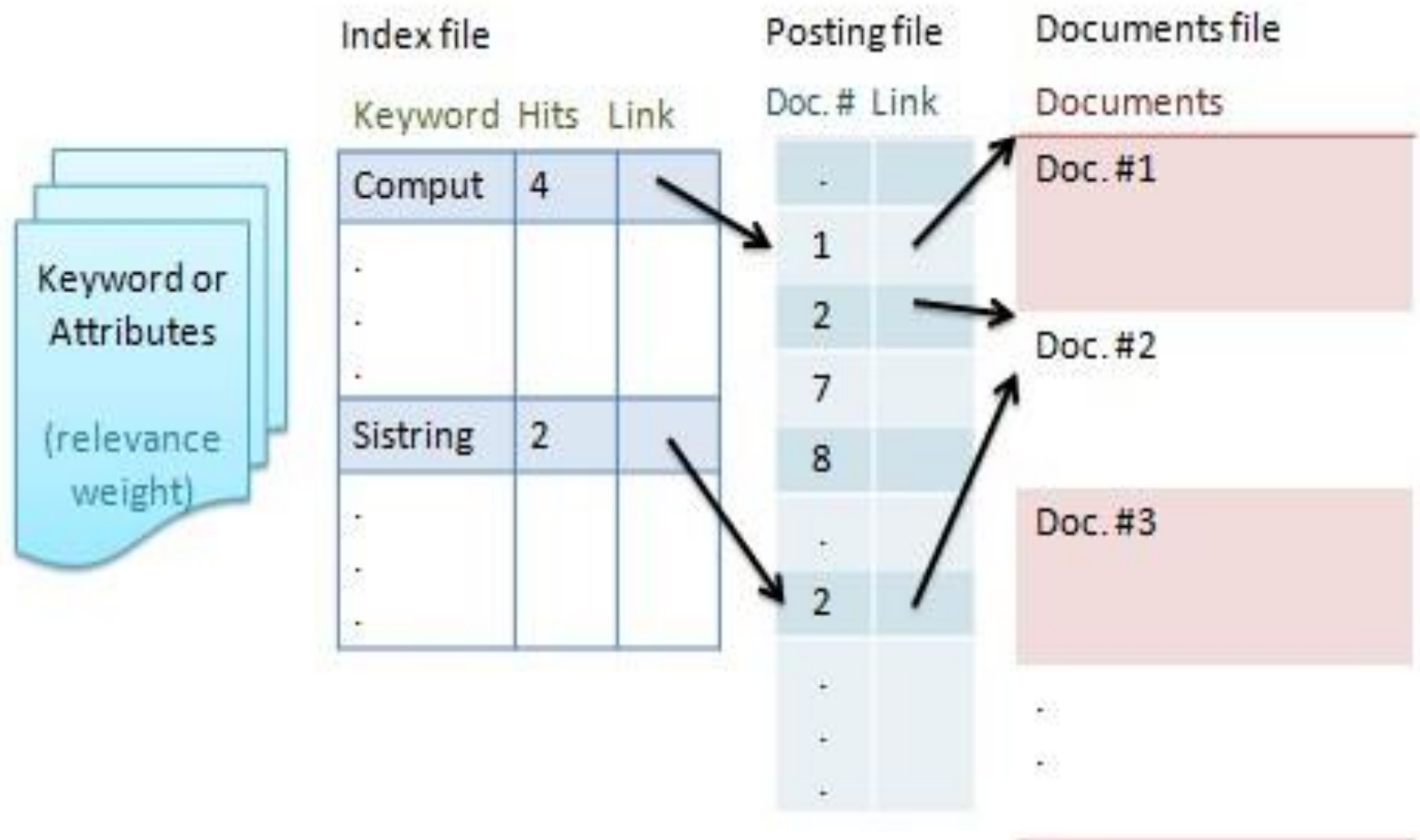
Text Classification.



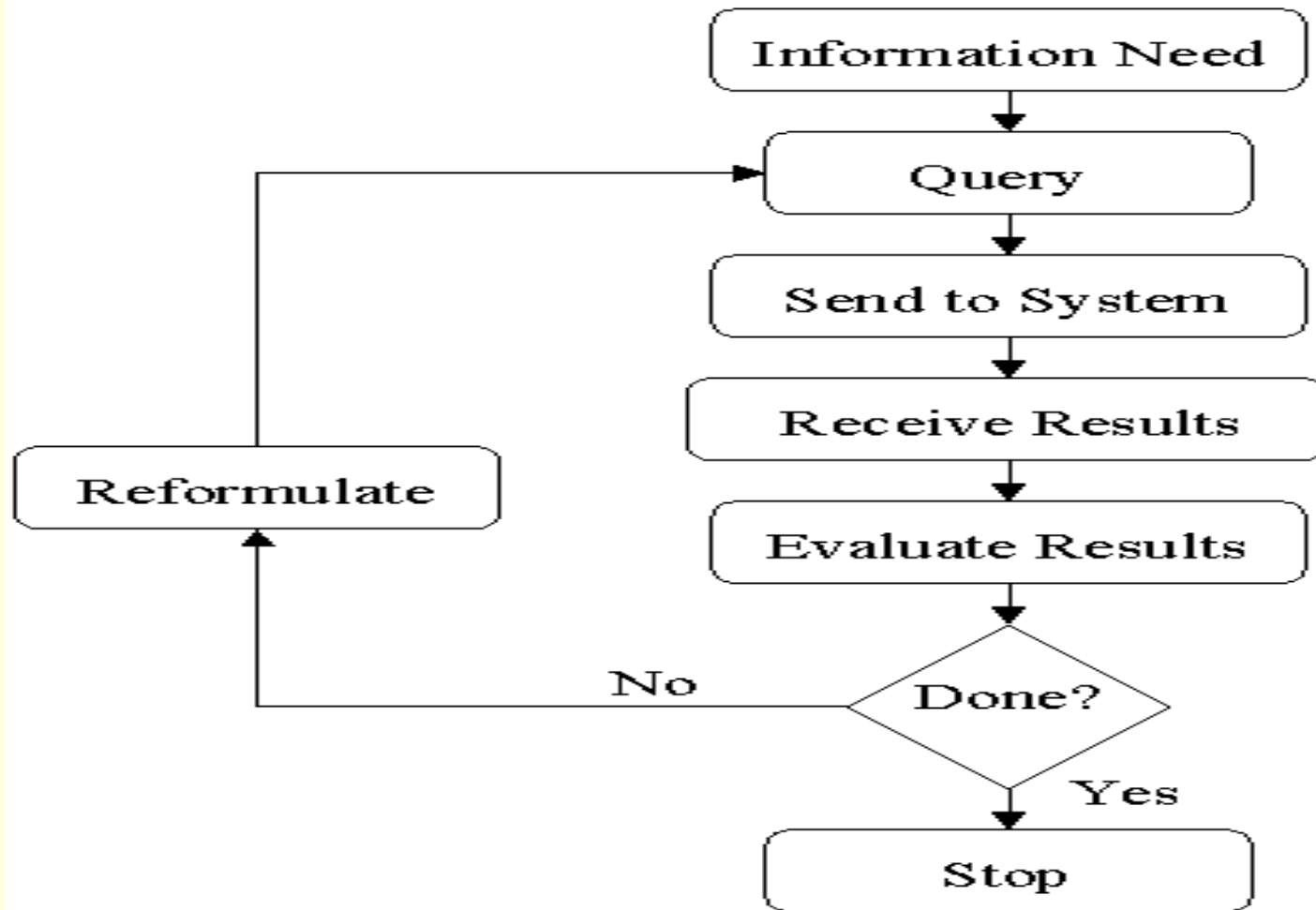
รูป : แสดงโครงสร้างของระบบค้นคืนสารสนเทศ



Inverted file



រូប៖ The Standard Retrieval Interaction Model



การประเมินผลระบบค้นคืนสารสนเทศ (Evaluation of IR System)

พัฒนาประสิทธิภาพ
(Efficiency)

วัดจาก

- เนื้อหาในการจัดเก็บในหน่วยความจำ
- **CPU Time**

ประสิทธิผล

(Effectiveness)

วัดจาก

- ค่าใช้จ่าย
- ต้นทุนในการสร้างระบบ
- **Recall**
- **Precision**

ประสิทธิภาพของระบบ

Precision = จำนวนของเอกสารที่เกี่ยวข้องที่ถูกลีงออกมา /
จำนวนทั้งหมดของเอกสารที่ถูกลีงออกมา

Recall = จำนวนของเอกสารที่เกี่ยวข้องที่ถูกลีงออกมา /
จำนวนของเอกสารทั้งหมดที่เกี่ยวข้อง

ประสิทธิผลของระบบ

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

Relevant(เกี่ยวข้อง)		Non-relevant(ไม่เกี่ยวข้อง)	
MISSED	HIT	WASTED	PASSED
Not Retrieved	Retrieved (ดึงออกมา)		Not Retrieved

$\text{Recall} = \text{HIT} / \text{Relevant}$
เป็นการวัดความสามารถของระบบในการดึงเอกสารที่เกี่ยวข้องออกมา

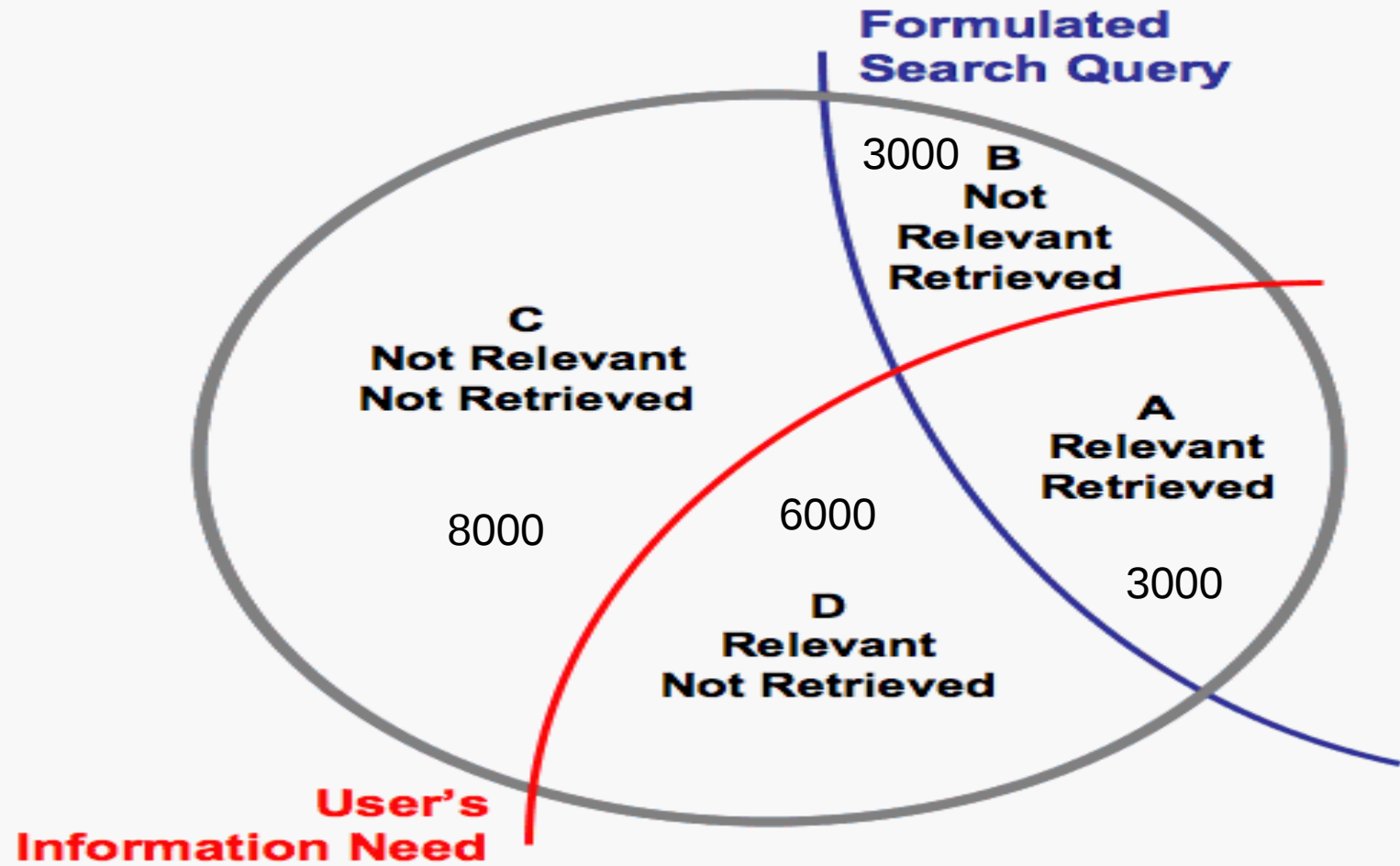
$\text{Precision} = \text{HIT} / \text{Retrieved}$
เป็นการวัดความสามารถของระบบในการจัดเอกสารที่ไม่เกี่ยวข้องออกไป

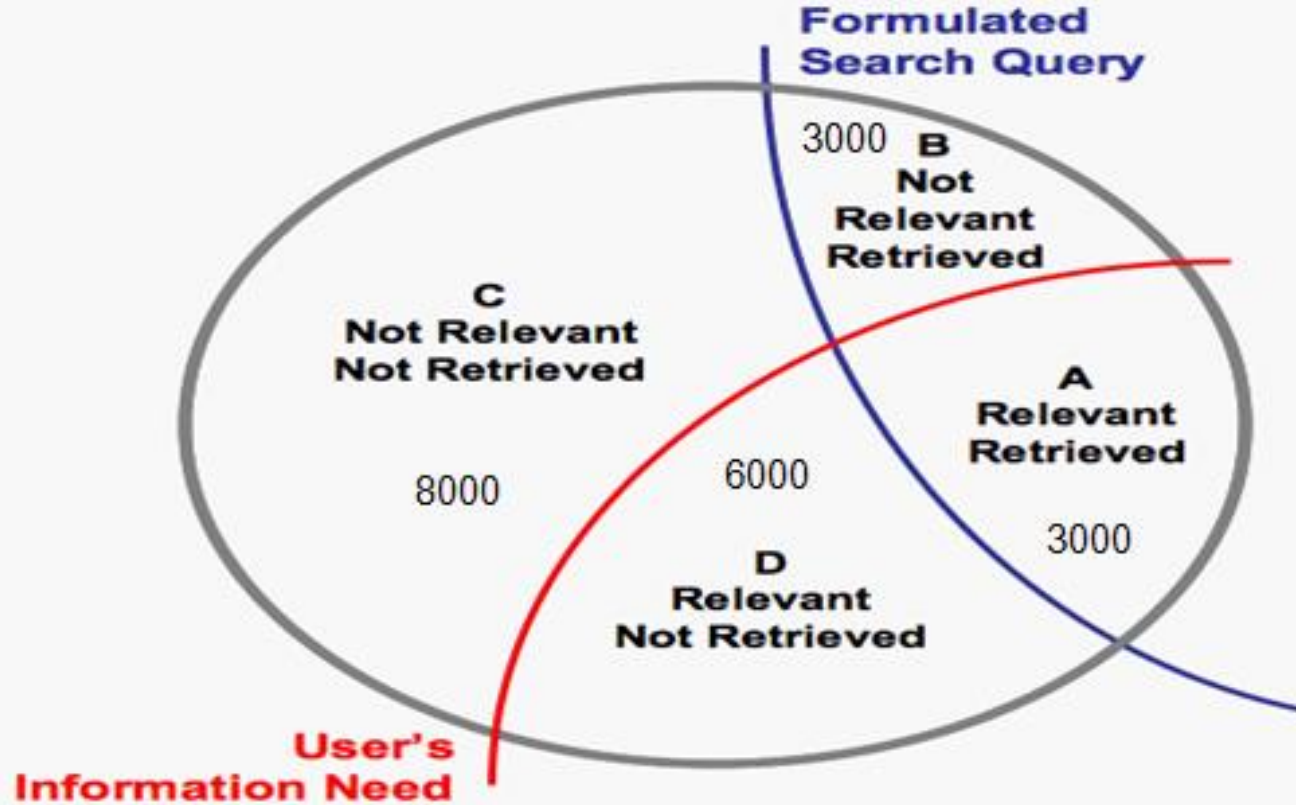
การหาประสิทธิภาพของระบบค้นคืนสารสนเทศ

	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 = $B/(B+D)$

Recall & Precision =? Effectiveness?





$$\text{Recall} = A/(A+D)$$

Proportion of documents relevant to a search question that are retrieved by a given search formulation.

$$\text{Precision} = A/(A+B)$$

Proportion of documents retrieved by a given search formulation that is relevant to the search question.

<http://choo.fis.utoronto.ca/FIS/Courses/LIS1325/RecallPrecision3.html>

How to improve PRECISION

- **NARROW** the search:
- Condense FACETS with fewer synonyms or more specific terms ย่อแง่มุมให้เป็นคำเฉพาะ
- Add a new FACET
- Search in *fewer* SUBJECT FIELDS (Title, Abstract, Descriptors)
- Use *more restrictive* WORD OPERATORS

How to improve RECALL

- **BROADEN** the search:
- Expand FACETS with more search terms
- Drop an existing FACET
- Search in *more* SUBJECT FIELDS (Title, Abstract, Descriptors)
- Use *less restrictive* WORD OPERATORS

วิวัฒนาการของระบบคั่นคืนสารสนเทศ

วิวัฒนาการของระบบค้นคืนสารสนเทศ

คศ. 1960-70

- เริ่มต้นในการสำรวจระบบค้นคืนสารสนเทศสำหรับข้อความที่มีขนาดเล็กที่เป็นบทความย่อทางด้านวิทยาศาสตร์กฎหมาย หรือ เป็นเอกสารทางด้านธุรกิจ มีการพัฒนา บูลีนพื้นฐาน และ

Vector Space Model

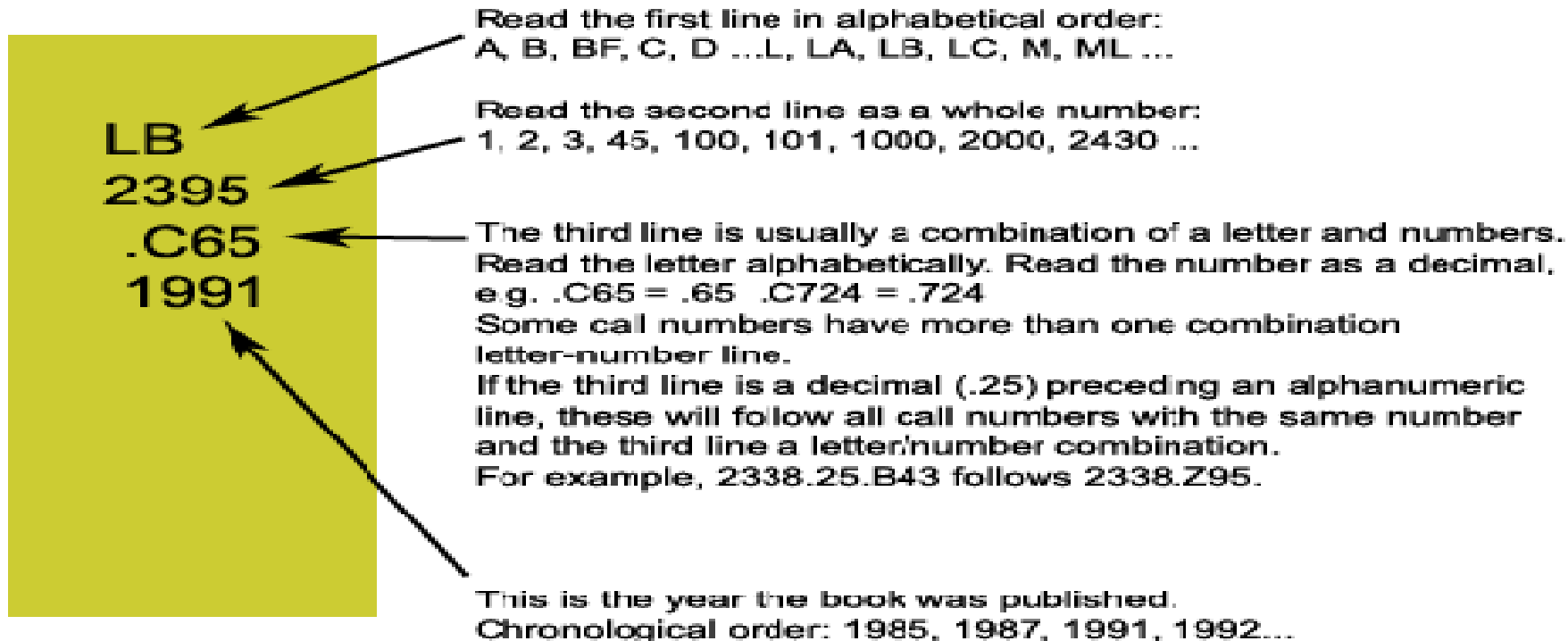
สำหรับการค้นคืนสารสนเทศ

คศ. 1980

- ระบบจะมีการสืบค้นจากฐานข้อมูลเอกสารขนาดใหญ่ที่มีการปฏิบัติงานในหลายๆบริษัท เช่น **Lexis-Nexis/MEDLINE**

พัฒนาการของการจัดเก็บและค้นคืนสารสนเทศ ทศวรรษ 1960

- ระยะแรกเริ่ม เป็นการพัฒนาระบบ และเครื่องมือจัดเก็บสื่อบันทึกความรู้และค้นหาทรัพยากรสารสนเทศในลักษณะสิ่งพิมพ์ เช่นระบบจัดหมวดหมู่ การจัดทำบรรณานุกรมและสารบัญแบบด้วยระบบมือ



พัฒนาการของการจัดเก็บและค้นคืนสารสนเทศ ทศวรรษ 1960

- เทคโนโลยีระยะแรก มีการนำไมโครกราฟิก (ถ่ายภาพย่อส่วนลงฟิล์ม) มาใช้จัดเก็บและ ค้นคืนสารสนเทศ โดยมีการใช้ระบบคอมพิวเตอร์ทำดัชนีและ บัตรรายการ

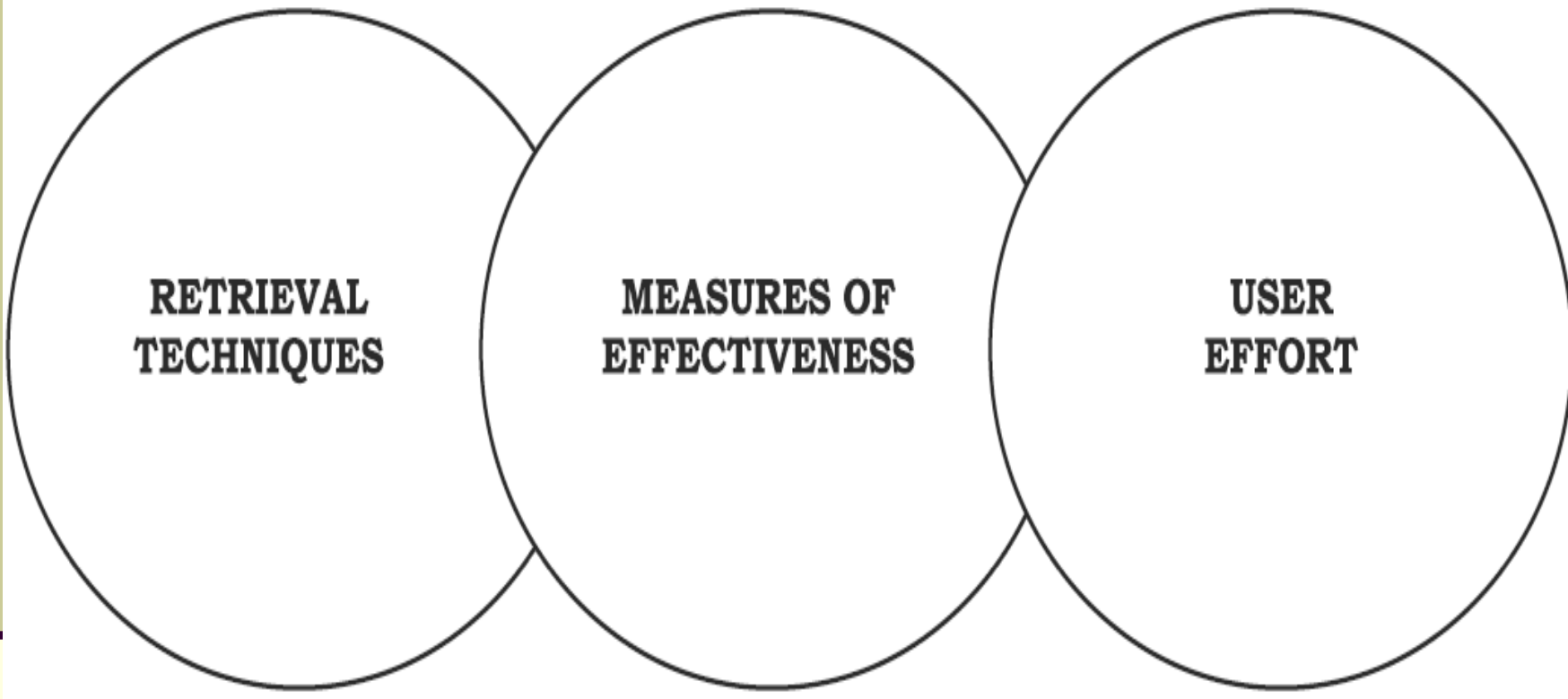


ระยะที่ 2 พัฒนาการของการจัดเก็บและค้นคืนสารสนเทศ ตั้งแต่ คศ.1960

- พัฒนาการด้านการวิจัยและพัฒนาระบบจัดเก็บและค้นคืนสารสนเทศที่สำคัญ
โครงการต่าง ๆ อาทิ
 - การวิจัยกลุ่มแคเรนฟีลด์ เมคคาร์ที สมาร์ท
 - การ วิจัยของซาระเชวิก และกลุ่มเทอร์ค เกิดกลุ่มความร่วมมือระหว่างหน่วยงาน
บริการสารสนเทศ รวมถึงบริการค้นคืนสารสนเทศเชิงพาณิชย์
 - การจัดเก็บ และค้นคืนสารสนเทศบน อินเทอร์เน็ต



PERCEPTION OF



**RETRIEVAL
TECHNIQUES**

**MEASURES OF
EFFECTIVENESS**

**USER
EFFORT**

LEXIS-NEXIS ACADEMIC UNIVERSE

วิวัฒนาการของระบบค้นคืนสารสนเทศ

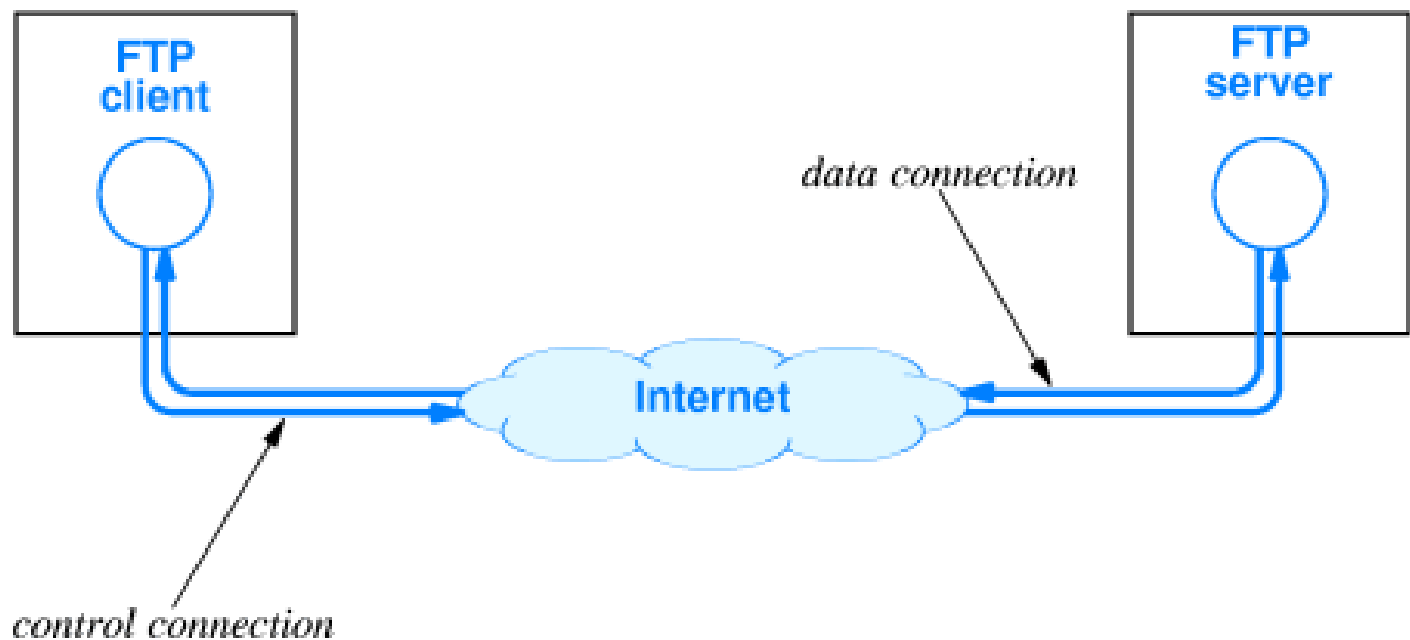
คศ. 1990

มีการสืบค้น FTP และ World Wide Web บนอินเทอร์เน็ต
เช่น Archie ,WAIS ,Lycos ,
Yahoo ,Altavista

คศ. 2000 จนถึงปัจจุบัน

มีการเชื่อมโยงโดยวิเคราะห์จาก Web Search เช่น Google
นอกจากนี้มีการสืบค้นสื่อผสม
(multimedia) ไม่ว่าจะเป็นภาพ
เสียง เพลง วิดีโอ การสรุปเอกสาร เป็น
ต้น

FTP(Fle Transfer Protocol)



เป็นการถ่ายโอนแฟ้มข้อมูลระหว่างเครื่องคอมพิวเตอร์ 2 เครื่อง ซึ่งอยู่บนเครือข่ายอินเทอร์เน็ต
ในระบบเครือข่ายอินเทอร์เน็ต มีเครื่องคอมพิวเตอร์ที่ให้บริการการถ่ายโอนแฟ้มข้อมูลเป็นสาธารณะ
มีอยู่เป็นจำนวนมาก เรียกเครื่องคอมพิวเตอร์ที่ให้บริการถ่ายโอนแฟ้มข้อมูลนี้ว่า **FTP Server**

Search engine

Archie Query Form

Search for:

Database: ☒ Worldwide Anonymous FTP ☐ Polish Web Index

Search Type: ☒ Sub String ☐ Exact ☐ Regular Expression

Case: ☒ Insensitive ☐ Sensitive

Do you want to look up strings only (no sites returned):
☒ NO ☐ YES

Output Format For Web Index Search: ☐ Keywords Only ☒ Excerpts Only ☐ Links Only

Search Reset

What is WAIS? A World Wide Information Server (WAIS) is a program for finding documents on the Internet. WAIS is rather primitive in its search capabilities.

ACTIVATION: THINKING ABOUT SOFTWARE

Internet.com You are in the World Wide Web (WWW) Home

4.0 International | Local Business | Community | News | Search | Contact Us | About Us

Webmaster: Please email: webmaster@internet.com or call: 1-800-444-4444. We have a new paper to learn how to use the Internet. Please click on the link to download the paper.

Internet.com (Webopedia) The #1 online encyclopedia dedicated to computer technology

Enter a word for a definition: or choose a computer category:

MENU

- Home
- Terms of the Day
- New Tutorials
- Personalization
- Hot Links
- Quick Reference
- Did You Know?
- Categories
- Tech Support
- Technology Jobs

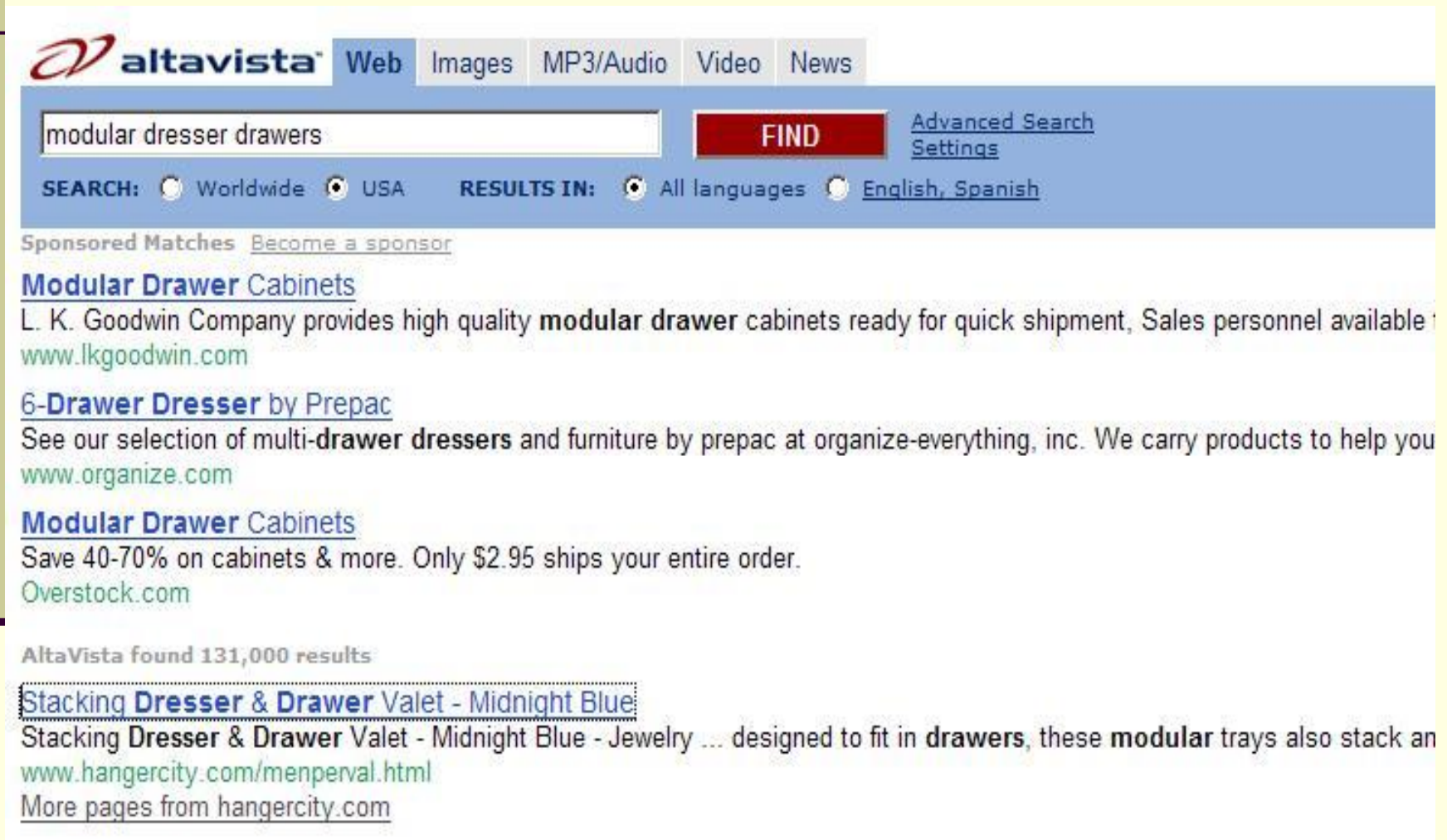
WAIS

Short for Wide Area Information Server, and pronounced wyeis, a program for finding documents on the Internet. WAIS is rather primitive in its search capabilities.

internet.com VIDEO

Log on to the Internet

Altavista



The screenshot displays the Altavista search engine interface. At the top, the Altavista logo is followed by navigation tabs for Web, Images, MP3/Audio, Video, and News. A search bar contains the text 'modular dresser drawers', and a red 'FIND' button is to its right. Links for 'Advanced Search' and 'Settings' are also present. Below the search bar, there are radio buttons for 'Worldwide' and 'USA' (selected), and 'RESULTS IN:' with radio buttons for 'All languages' and 'English, Spanish' (selected).

Sponsored Matches [Become a sponsor](#)

Modular Drawer Cabinets
L. K. Goodwin Company provides high quality **modular drawer** cabinets ready for quick shipment, Sales personnel available
www.lkgoodwin.com

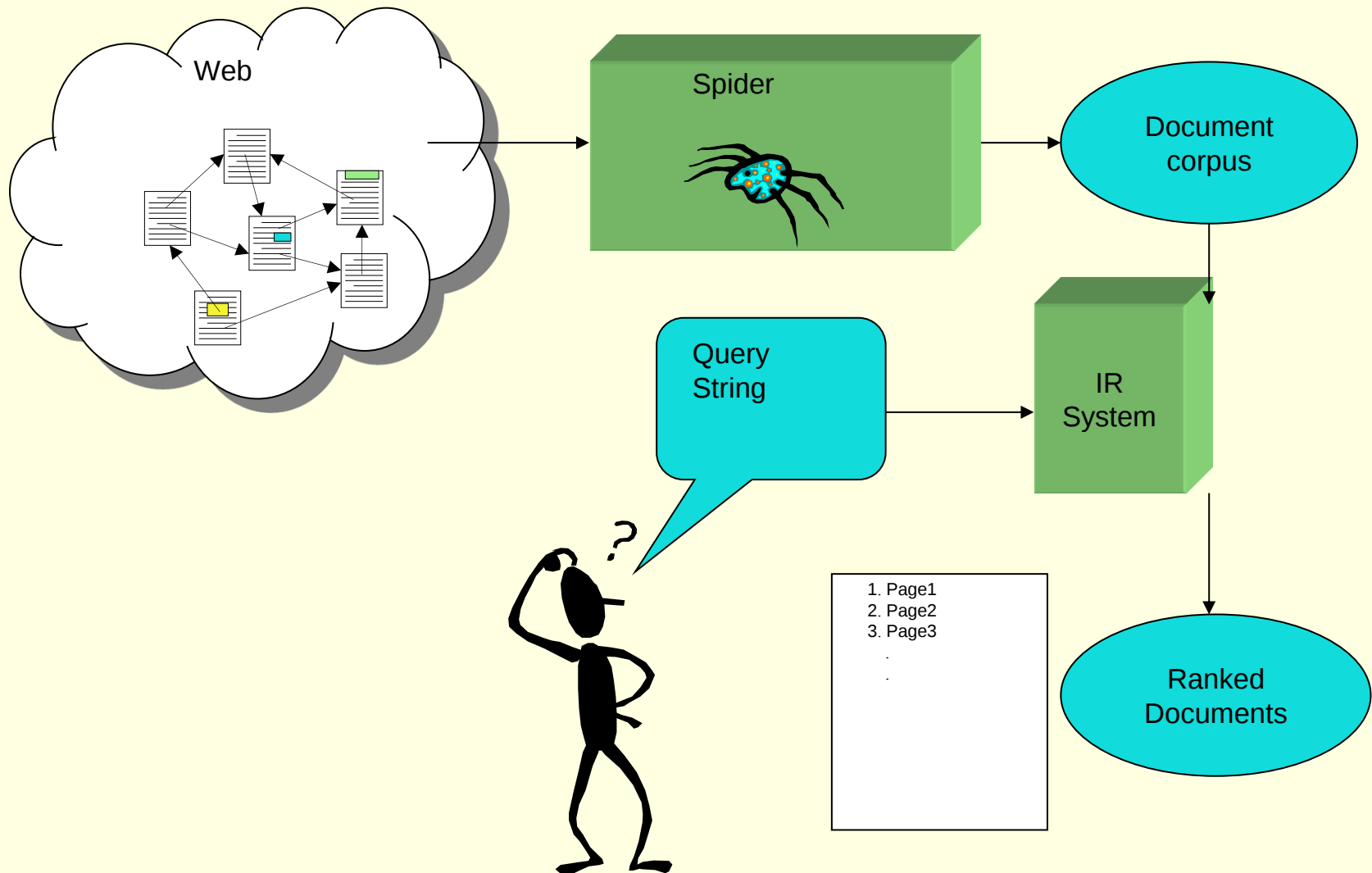
6-Drawer Dresser by Prepac
See our selection of multi-**drawer dressers** and furniture by prepac at organize-everything, inc. We carry products to help you
www.organize.com

Modular Drawer Cabinets
Save 40-70% on cabinets & more. Only \$2.95 ships your entire order.
Overstock.com

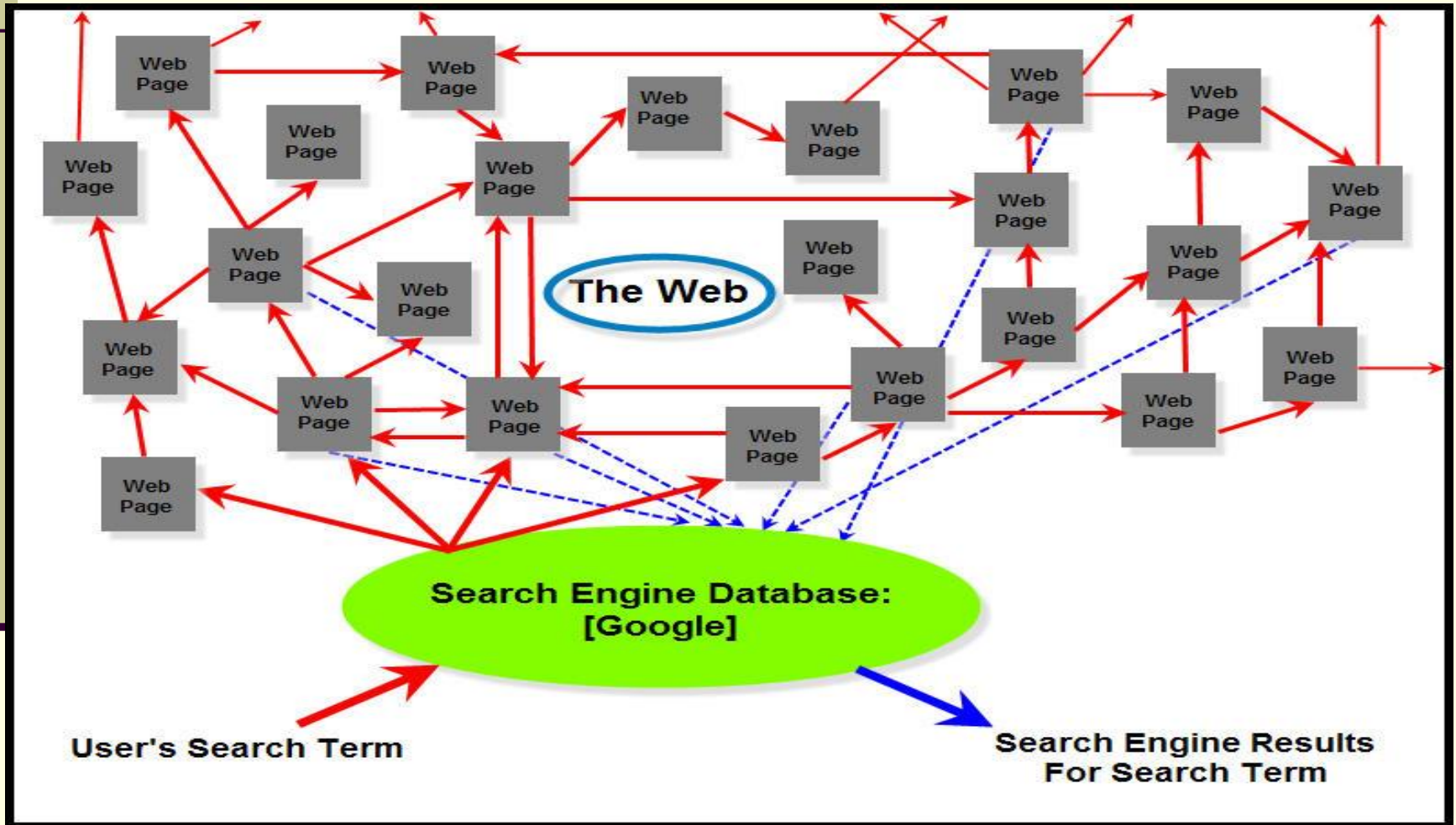
AltaVista found 131,000 results

Stacking Dresser & Drawer Valet - Midnight Blue
Stacking **Dresser & Drawer Valet - Midnight Blue - Jewelry** ... designed to fit in **drawers**, these **modular** trays also stack an
www.hangercity.com/menperval.html
[More pages from hangercity.com](http://www.hangercity.com)

แสดง Web Search System



Spider



บันทึกข้อมูลหน้าเพจต่างๆ จากทั่วโลกเอามาจัดเก็บให้เป็นระเบียบและมีการจัดอันดับตามความเหมาะสมของเนื้อหานั้นๆ (SEO)

สิ่งที่คนส่วนมากค้นหาใน web

โดยการศึกษาจาก Spink et al., Oct 98

■ 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%

Sexual/pornography related Web searching studies

Table 1 provides an overview of the nine studies conducted from 1997 to 2005 that examined the level of sexual or pornographic Web queries, including the single Web search engines Excite, AlltheWeb, Alta Vista, Vivisimo, and Dogpile. No studies are publicly available that examined Web queries from Google, MSN, AOL or Yahoo!. Table 2 shows the distribution of Web queries across topic categories for the nine Web search engine studies. Table 3 shows the top 20 query terms across the nine studies.

Table 1: Basic data for the nine studies conducted from 1997 to 2005									
Study Number	1	2	3	4	5	6	7	8	9
Web search engine	Excite (Jansen, Spink and Saracevic, 2000)	Excite (Wolfram, Spink, Jansen and Saracevic, 2001)	AlltheWeb (Spink, Ozmutlu, Ozmutlu and Jansen, 2002)	Excite (Spink, Jansen, Wolfram and Saracevic, 2002)	AlltheWeb (Spink, Ozmutlu, Ozmutlu and Jansen, 2002)	Alta Vista (Jansen, Spink and Pedersen, 2005)	Vivisimo (1) (Koshman, Spink and Jansen, in press)	Vivisimo (2) (Koshman, Spink and Jansen, in press)	Dogpile (Jansen, Spink, Blakely and Koshman, in press)
Region	U.S.	U.S.	Europe	U.S.	Europe	U.S.	U.S.	U.S.	U.S.
Date collected	Tues 16 September 1997	Wed 1 December 1999	Tues 6 February 2001	Mon 30 April 2001	Tues 28 May 2002	Sun 8 September 2002	28 March — 4 April 2004	25 April — 4 May 2004	Fri 6 May 2005
Sessions	211,063	325,711	153,297	262,025	345,093	369,350	279,766	275,456	534,507
Queries	1,025,908	1,025,910	451,551	1,025,910	957,303	1,073,388	1,082,431	927,303	1,523,793
Terms	1,277,763	1,500,500	1,350,619	1,538,120	2,225,141	1,073,388	3,394,989	2,905,592	4,250,656

Table 2: Distribution of Web query topic categories for each Web search engine study

Categories/Study	1 1997 Excite (2,414 queries)	2 1999 Excite (2,539 queries)	3 2001 Excite (2,453 queries)	4 2001 AlltheWeb (2,503 queries)	5 2002 AltaVista (2,603 queries)	6 2002 AlltheWeb (2,535 queries)	7 2004 (1) Vivisimo (3,600 queries)	8 2004 (2) Vivisimo (1,368 queries)	9 2005 Dogpile (2,500 queries)
People, places or things	6.7%	20.3%	19.7%	22.5%	49.3%	41.5%	15%	27.9%	16.0%
Commerce, travel, employment, or economy	13.3%	24.5%	24.7%	12.3%	12.5%	12.7%	21%	14.6%	30.4%
Computers or Internet	12.5%	10.9%	9.7%	21.8%	12.4%	16.3%	13%	8.5%	5.7%
Health or sciences	9.5%	7.8%	7.5%	7.8%	7.5%	4.5%	6%		8.9%
Education or humanities	5.6%	5.3%	4.6%	2.9%	5.0%	2.3%	5%	11.6%	5.6%
Entertainment or recreation	19.9%	7.5%	6.7%	9.1%	4.6%	9.5%	2%	11.4%	7.0%
ลามก Sex and pornography	16.8%	7.5%	8.6%	10.8%	3.3%	4.5%	4%	3.6%	3.8%
Society, culture, ethnicity, or religion	5.7%	4.2%	3.9%	4.8%	3.1%	4.5%	9%	9.7%	4.7%
Government or legal	3.4%	1.6%	2.0%	2.7%	1.6%	2.6%	3%	0.5%	3.6%

Table 3: Most frequent Web query terms

	Excite 1997	Excite 1999	Excite 2001	AlltheWeb 2001	AltaVista 2002	AlltheWeb 2002	Vivisimo 2004	Vivisimo 2004	Dogpile 2005
1	and	and	free	free	free	free	download	free	of
2	of	of	sex	download	sex	sex	new	download	the
3	sex	free	pictures	sex	pictures	download	software	new	in
4	free	the	new	pictures	new	software	windows	2004	and
5	the	sex	nude	uk	nude	uk	sex	pictures	free
6	nude	in	pics	nude	music	windows	school	how	for
7	pictures	christmas	school	hotel	school	new	history	software	a
8	in	nude	music	mp3	how	hotel	online	windows	to
9	university	for	download	new	lyrics	mp3	video	history	girls
10	pics	pictures	university	windows	home	video	what	sex	sex
11	chat	new	state	pics	pics	crack	music	school	on
12	for	pics	lyrics	software	download	nude	world	online	how
13	adult	to	home	lyrics	online	pictures	lyrics	music	nude
14	women	music	games	crack	american	web	hotel	video	lyrics
15	new	university	software	music	state	home	map	world	music
16	xxxx	a	american	video	country	world	system	hotel	new
17	girls	games	stories	map	university	online	home	map	pictures
18	music	porn	porn	university	car	Internet	mp3	uk	mp3

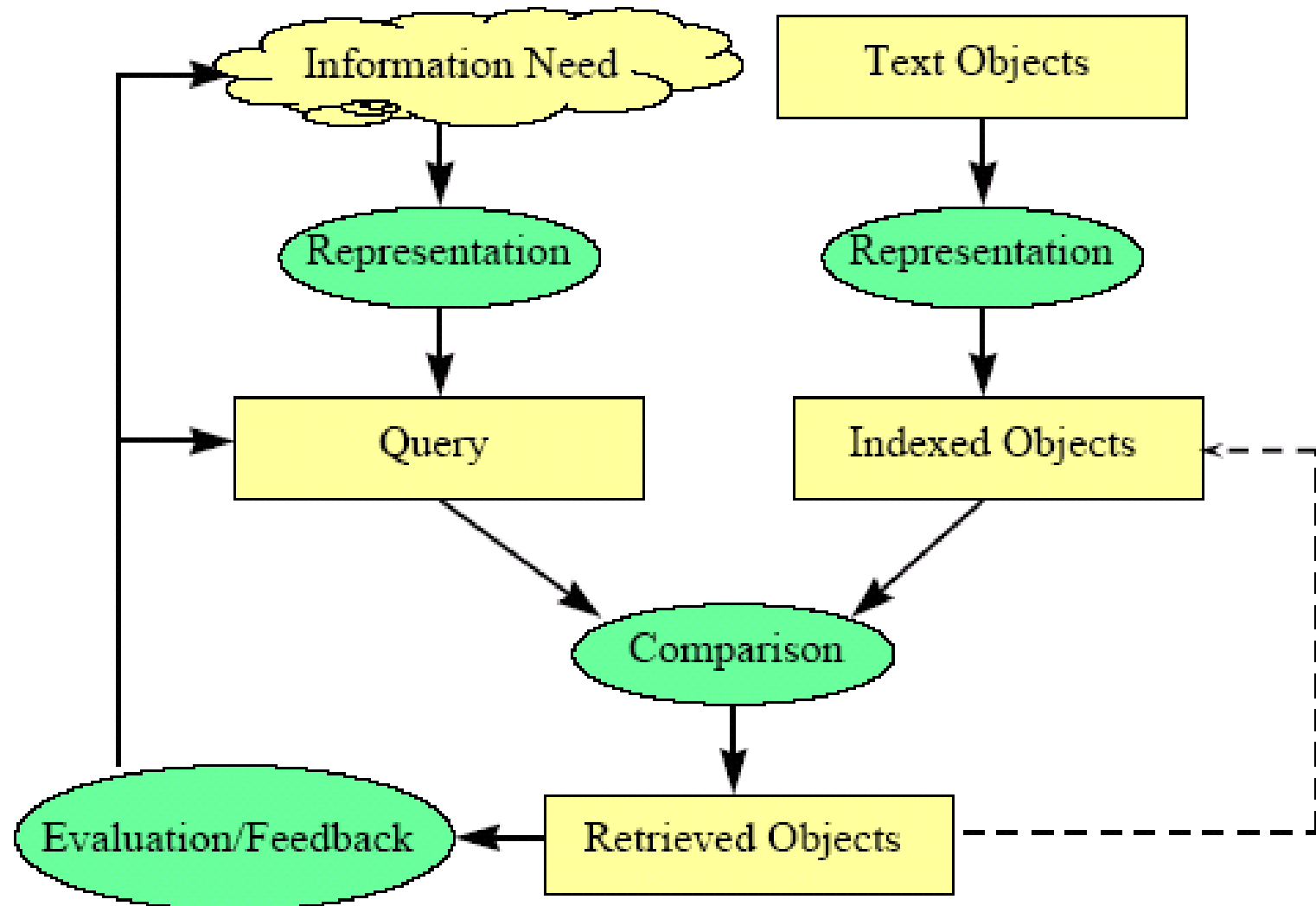
Web Search Market Share and Volume

(Top web search engine's share of searches, volume, performance and sponsored search referrals as a % of all referrals*)

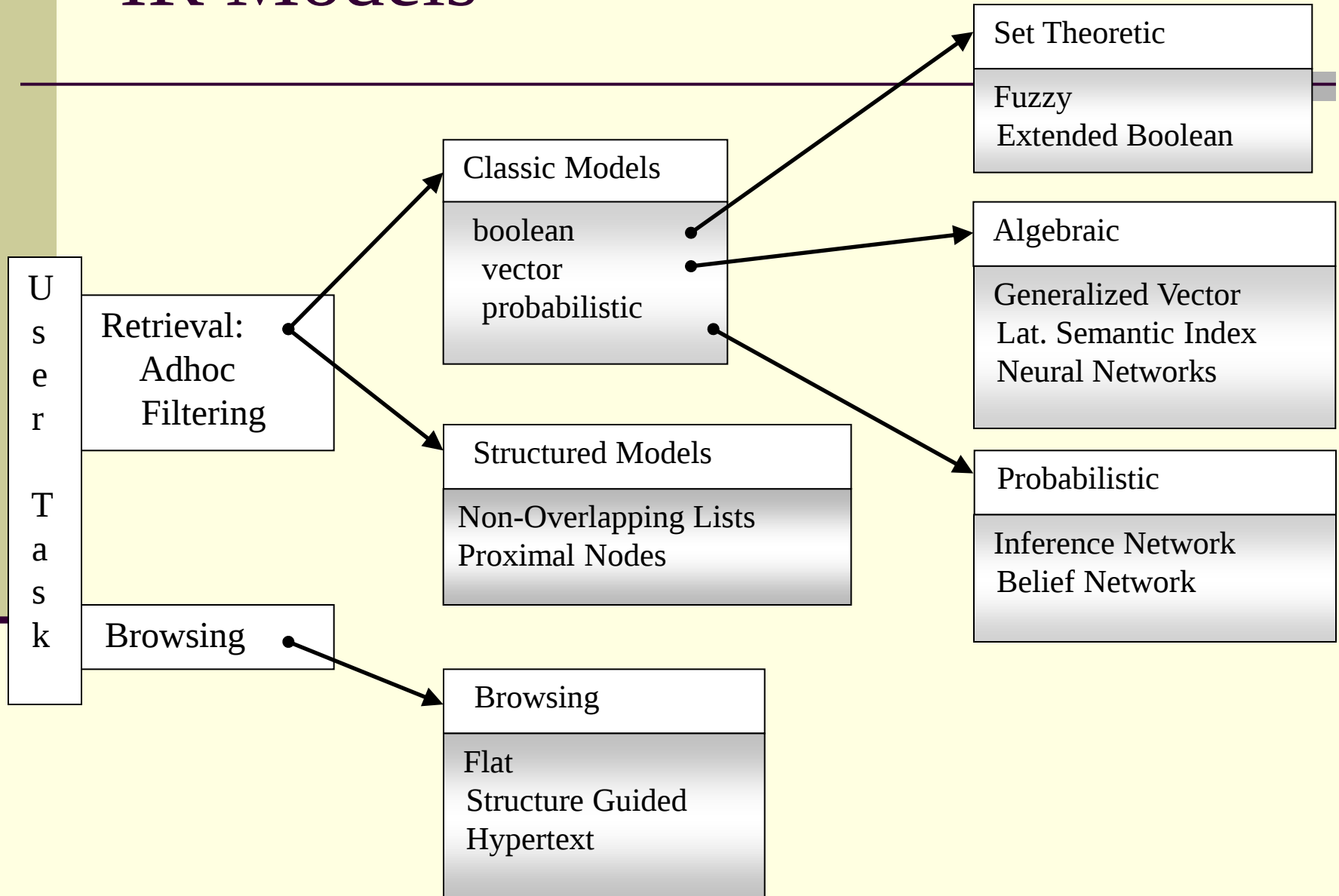
Engine	Market Share			Share Change		Query Volume (millions)			Volume Change		Sponsored Referrals		
	Jul-10	Jun-10	Jul-09	M-O-M	Y-O-Y	Jul-10	Jun-10	Jul-09	M-O-M	Y-O-Y	Jul-10	Jun-10	Jul-09
Google	72.9%	73.8%	72.5%	-1.1%	0.6%	10,461	10,373	10,261	0.8%	1.9%	8.4%	7.8%	8.1%
Yahoo!	13.1%	12.7%	15.7%	3.0%	-16.6%	1,873	1,782	2,221	5.1%	-15.7%	11.5%	11.9%	9.4%
MSFT	10.6%	10.3%	8.7%	3.1%	21.1%	1,518	1,443	1,237	5.1%	22.6%	8.7%	8.8%	7.5%
Ask	2.9%	2.7%	2.5%	7.2%	16.4%	417	381	353	9.4%	17.9%	10.8%	10.7%	7.7%
DL	0.5%	0.6%	0.6%	-9.6%	-13.2%	78	85	89	-7.8%	-12.0%	1.7%	1.8%	2.7%
Total	100.0%	100.0%	100.0%	n/a	n/a	14,347	14,064	14,163	2.0%	1.3%	n/a	n/a	n/a

Data does not include vertical search queries.

Simple model of IR



IR Models



IR Model สามารถจำแนกเป็น 4 ชนิด

Set Theoretic Models

ใช้บูลีนโมเดล ซึ่งอยู่บนพื้นฐานของ
แนวความคิดเชิงตรรกะหรือพีชคณิตบูลีน

Algebraic Models

เอกสารแต่ละอันจะถูกนำเสนอแสดงโดย
Vector หรือชุดของค่าที่มีการจัด
เรียงลำดับ

Probabilistic Model

แบบความน่าจะเป็นโดยฟังก์ชันตัดสินใจที่ถูก
ใช้เป็นแบบความน่าจะเป็น

Hybrid Model

Boolean+Vector

Set Theoretic Models

- ใช้บูลีน โมเดล ซึ่งอยู่บนพื้นฐานของแนวความคิดเชิงตรรกะหรือพีชคณิตบูลีน (Boolean Algebra) กับคำที่ถูกรวมกันโดยตัวเชื่อมทางตรรกะหรือพีชคณิต และ (AND) หรือ (OR) ไม่ (NOT)
- ตัวแบบการค้น พิจารณาจาก การจับคู่หรือเปรียบเทียบ ระหว่าง ข้อความ และ ดัชนี แทนสารสนเทศที่สะสมไว้ให้สามารถค้นคืนสารสนเทศที่เข้าเรื่องหรือตรงกับความต้องการ

ตัวแบบการค้น

- **Boolean model** เป็นการจับคู่แบบตรงกันระหว่างศัพท์ดัชนีและคำค้น
- **Retrieval Status Value (RSV)** เป็นมาตรการเกี่ยวกับการคิวรีเอกสารที่มีความ **Similality**
 - **RSV** จะมีค่าเท่ากับ 1 เมื่อคิวรีที่แสดงมีค่าเป็น **True** มีการค้นพบเอกสารที่เกี่ยวข้องตามที่คิวรี
 - **RVS** มีค่าเป็น 0 เมื่อคิวรีที่แสดงเป็นค่าอื่น

Set Theoretic Models

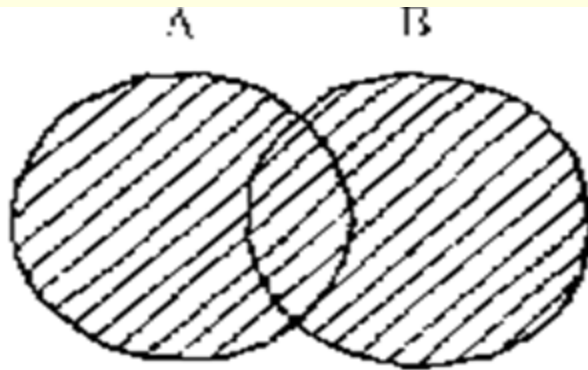
ข้อดีของ Boolean Model

- เข้าใจง่าย เวลา Query กำหนดความต้องการได้ว่าจะเอาคำไหนซึ่งเป็นลักษณะง่าย ๆ ที่มีรูปแบบไม่ซับซ้อน
- เป็นการหาข้อมูลในลักษณะ ใช่ / ไม่ใช่ เจอ / ไม่เจอ

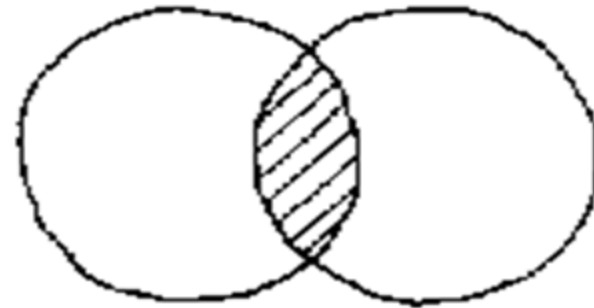
ข้อเสียของ Boolean Model

- ไม่มีการจัดลำดับของเอกสาร (Ranking) และไม่มีการเปรียบเทียบ (Relevance)
- ใช้ Query ที่มีความซับซ้อนไม่ได้
- ควบคุมจำนวนเอกสารยาก
- มีความลำบากในการยอมรับเนื่องจากตรงหรือสำคัญเป็นความต้องการของ User

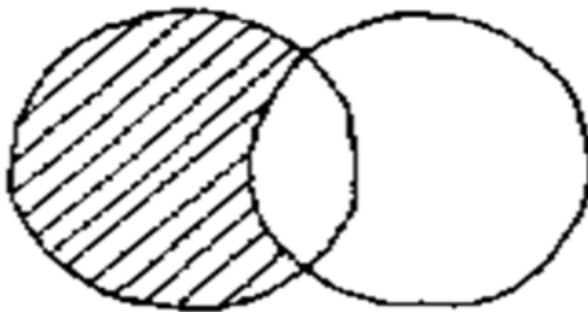
Venn diagram



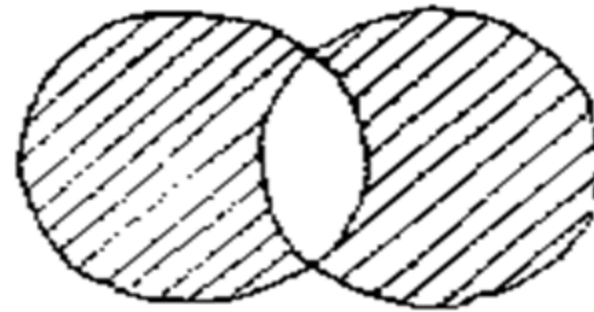
Union $A \cup B$



Intersection $A \cap B$



Difference $A - B$



Symmetric difference $A \Delta B$

SQL

The screenshot displays the FlySpeed SQL Query application window, titled "FlySpeed SQL Query - [Order Products*]". The interface includes a menu bar (Connection, Query, Edit, Tools, View, Window, Help), a toolbar, and a status bar at the bottom showing "Order Details" and "Northwind".

Left Panel (Northwind): A tree view showing the database structure. Under "Tables", the following tables are listed: Employees, Customers, Categories, Order Details, Orders, Products, Shippers, Suppliers, and Table1. Under "Views", several views are listed, including "Ten Most Expensive Products", "Sales by Category", "Quarterly Orders", "Products by Category", "Products Above Average Price", "Product Sales for 1997", "Orders Qry", "Order Subtotals", and "Order Details Extended".

Query Design (Main): The "Main" tab shows a query design grid with three tables: Products, Order Details, and Orders. The relationships are as follows:

- Products (1) to Order Details (∞) on ProductID.
- Order Details (∞) to Orders (1) on OrderID.

The fields included in the query are:

- Products: ProductName (checked).
- Order Details: Quantity (checked).
- Orders: OrderDate (checked).

Query Results Table:

Output	Expression	Aggregate	Alias	Sort Type	Sort Order	Grouping	Criteria
☒	Products.ProductName			Ascending	1	☐	
☒	[Order Details].Quantity					☐	> 1000
☒	Orders.OrderDate					☐	

SQL Code:

```
Select
Products.ProductName,
[Order Details].Quantity,
Orders.OrderDate
From
(Products Inner Join
[Order Details] On Products.ProductID = [Order Details].Pr
Inner Join
Orders On Orders.OrderID = [Order Details].OrderID
```

Algebraic Models

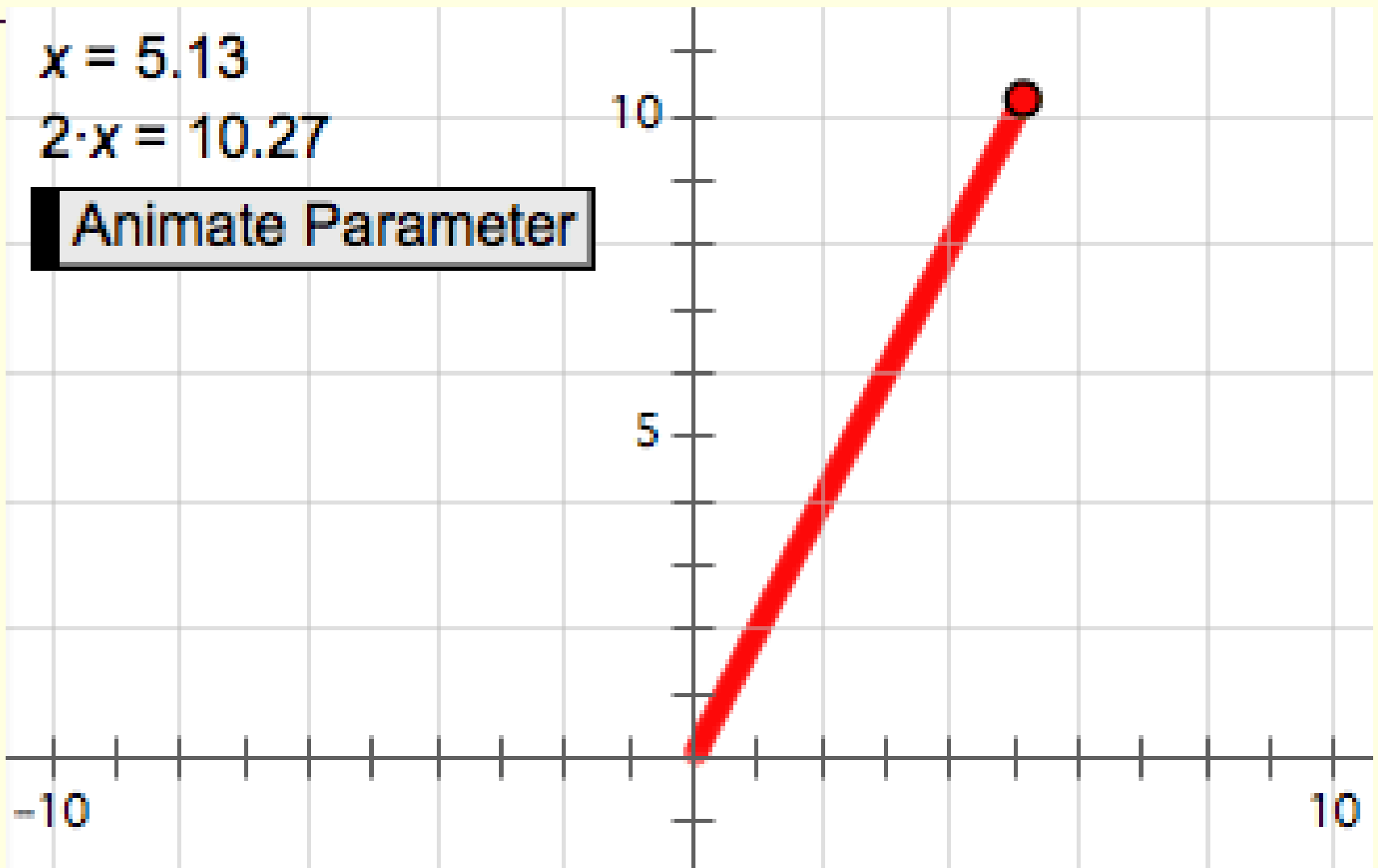
- แทนเอกสารและข้อความในรูปเวกเตอร์ (**Vector**) โดยกำหนดค่าน้ำหนักของคำ ด้วยความถี่ของคำที่ปรากฏในเอกสาร และคำที่ปรากฏในฐานข้อมูล
- วิธีการประเมินขึ้นอยู่กับ **Vector 0-1**
 - ซึ่งแต่ละองค์ประกอบเป็น **0** ถ้าคำนั้น ๆ ไม่ปรากฏ
 - หรือเป็น **1** ถ้าคำนั้น ๆ ปรากฏในเอกสารตามที่ควรี
 - และการประเมินอีกทางหนึ่งอยู่บนพื้นฐานของ **Vector** น้ำหนักซึ่งองค์ประกอบของมันเป็นน้ำหนักหรือค่าที่ถูกกำหนดให้แต่ละคำในเอกสาร

Algebraic Models

$$x = 5.13$$

$$2 \cdot x = 10.27$$

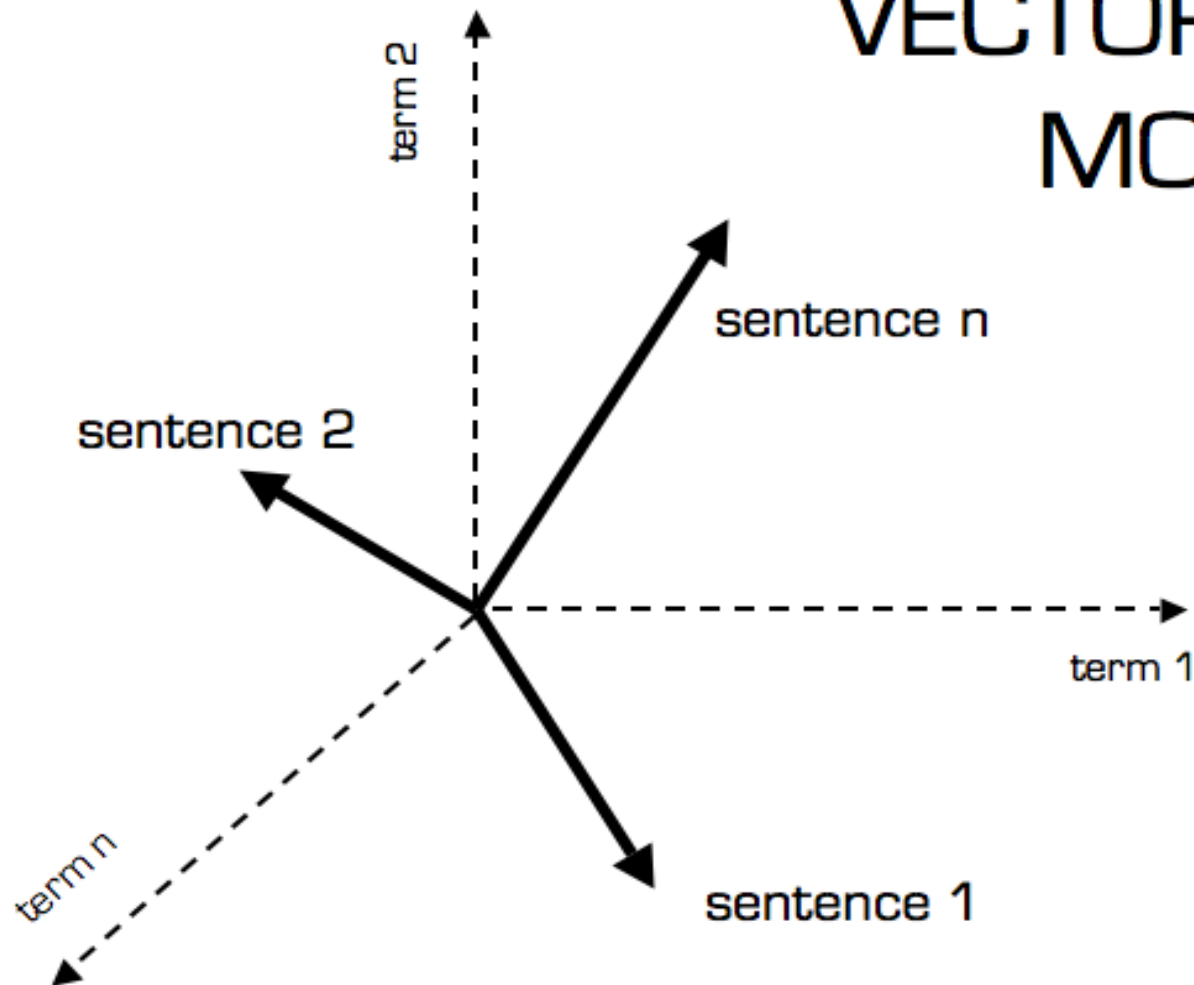
Animate Parameter



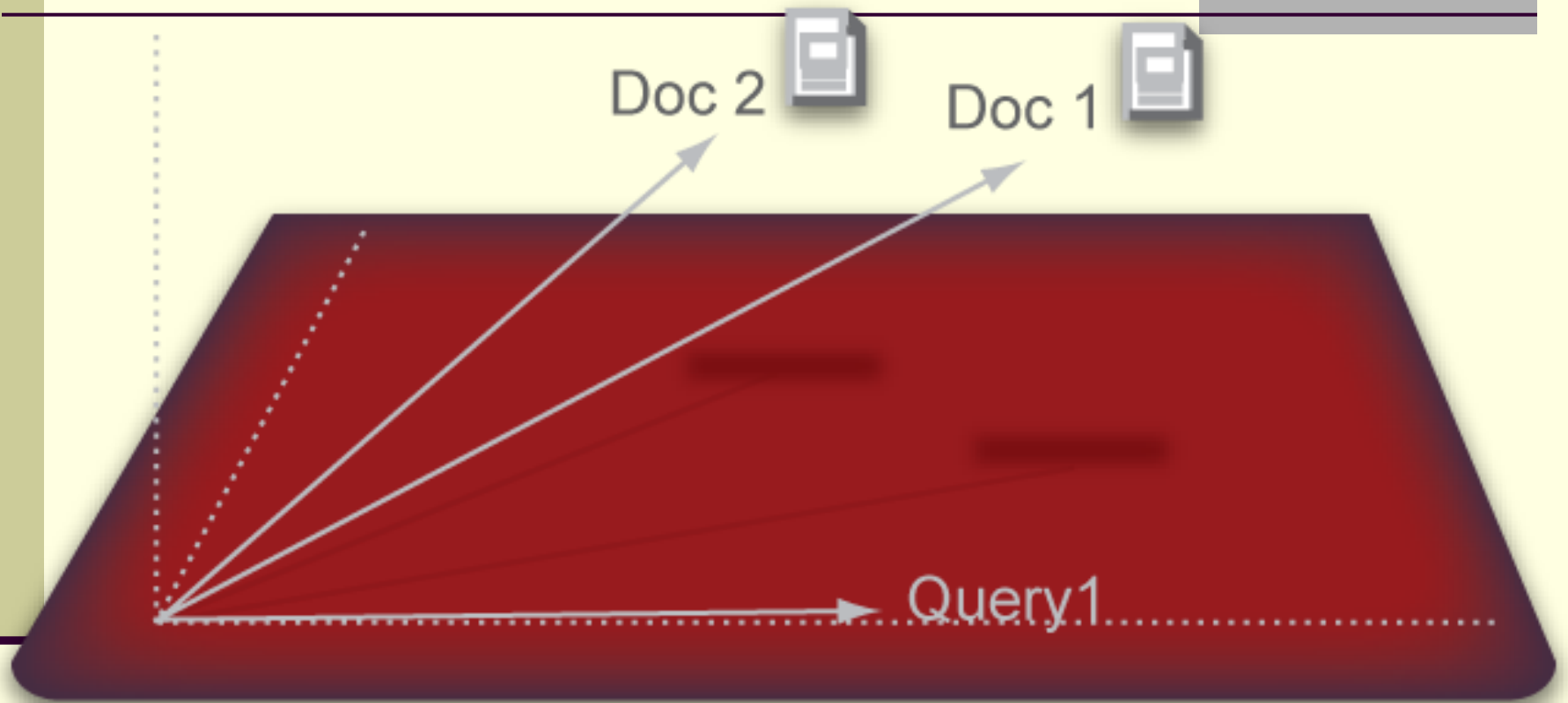
รูปแบบการทำงานของ Vector Model

- ให้ความสำคัญความถี่ของคำที่ปรากฏอยู่ในเอกสารและความถี่มีผลต่อการให้น้ำหนักของคำ ได้แก่
 - **Term Frequency** คือการใช้ความถี่ของคำ เช่นเจอ **1** ครั้ง เรียกว่า **Term** ทั้งนี้ขึ้นอยู่กับจำนวนคำของเอกสาร โดย **Term** จะแทนคำศัพท์ของแต่ละคำ
 - **Term Weight** (น้ำหนักของคำ) ความถี่ของคำ ๆ หนึ่งที่พบในทุก ๆ เอกสาร
- สามารถจัดอันดับของเอกสารโดยใช้เกณฑ์ความสำคัญของคำและการ **Match** กันของคำ

VECTOR SPACE MODEL



Vector space Model



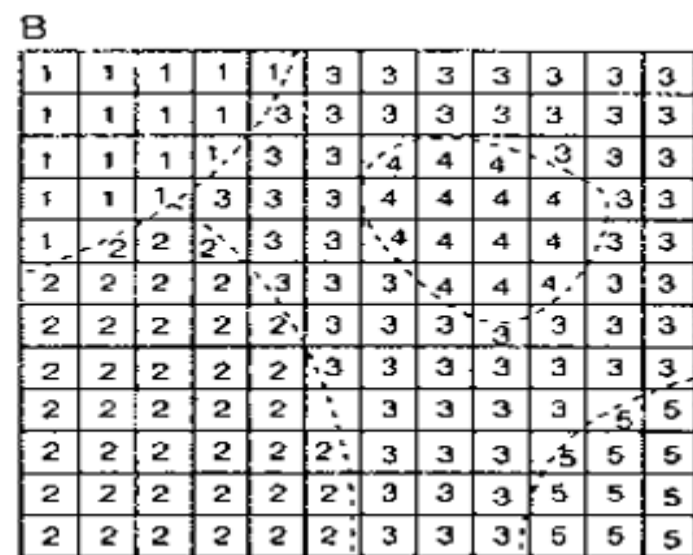
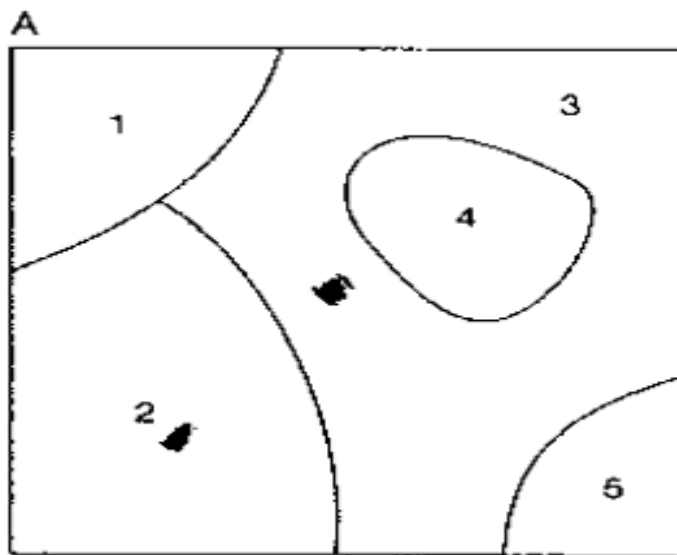
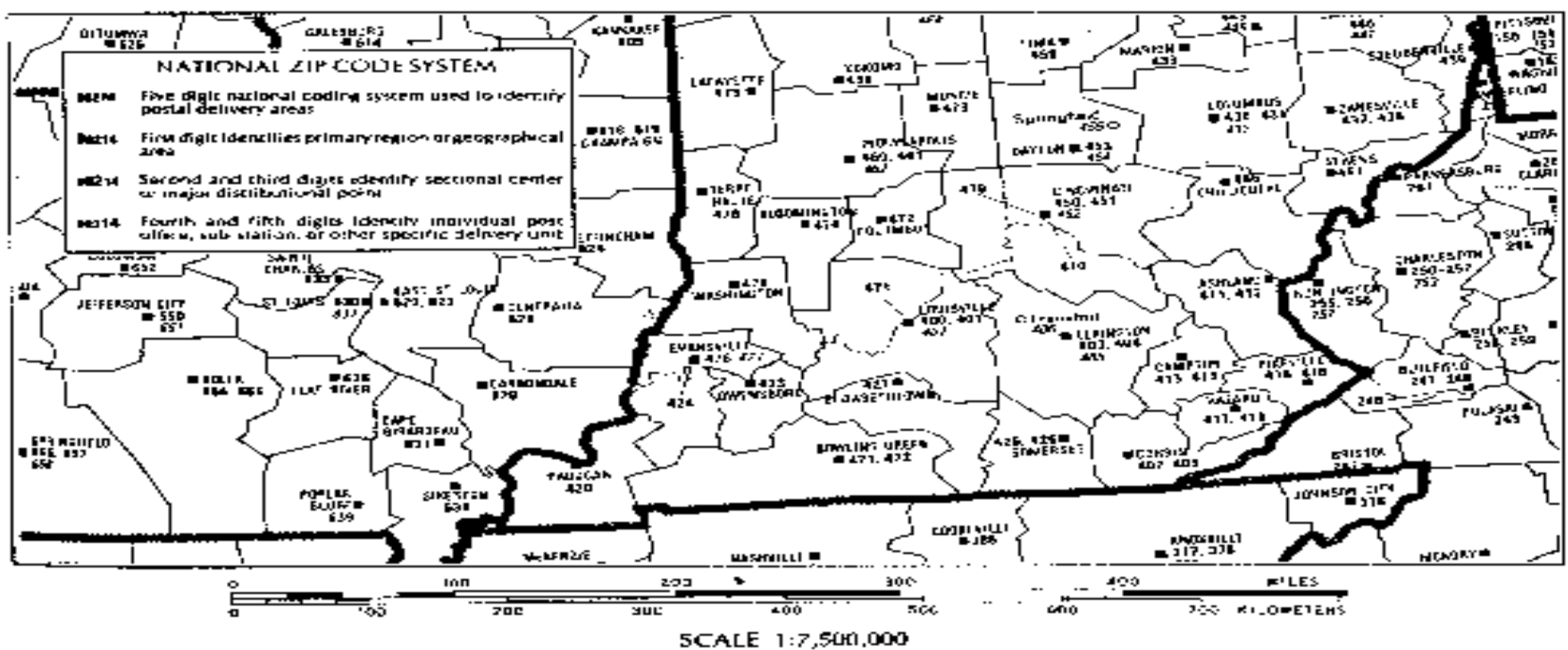
Vector Space Model

ข้อดี

ใช้คณิตศาสตร์เรียบง่ายในการคิด
มีการพิจารณาจากความถี่ของคำ
และสามารถจัด **Ranking**
ของเอกสารได้ สามารถใช้กับเอกสารที่
มีข้อมูลมาก ๆ ได้ดี

ข้อเสีย

- ไม่สนใจความหมายของคำ , วลี ,
โครงสร้างของคำ , คำที่มีความหมาย
เหมือนกัน (**Synonymy**)
- สืบค้นใส่เงื่อนไขแบบ **Boolean**
Model ไม่ได้



Comparison Between the Vector Model (A) and the Raster Data Model (B)

Probabilistic Model

- Probabilistic Model เป็นการจัดลำดับเอกสารตามความน่าจะเป็นด้านความเข้าเรื่องของแต่ละเอกสารกับข้อคำถามตามค่าที่กำหนด โดยเรียงลำดับจากมากไปหาน้อย
- ความน่าจะเป็นที่เอกสารได้จากการค้นคืนต้องสนองต่อคำขอ และความน่าจะเป็นต้องรวมกันเป็น 1 เหมือนการคำนวณของฟังก์ชันสมาชิกตามคอมพลิเมนต์ของเซต
- ข้อดี มีวิธีการคำนวณความน่าจะเป็นที่ได้รับการยอมรับโดยคำนวณความน่าจะเป็นจากข้อมูลความถี่ของคำ

Best Match Searching using the Probabilistic Model

- The probabilistic model query-document matching score $ms(j)$ for document j can be determined using:

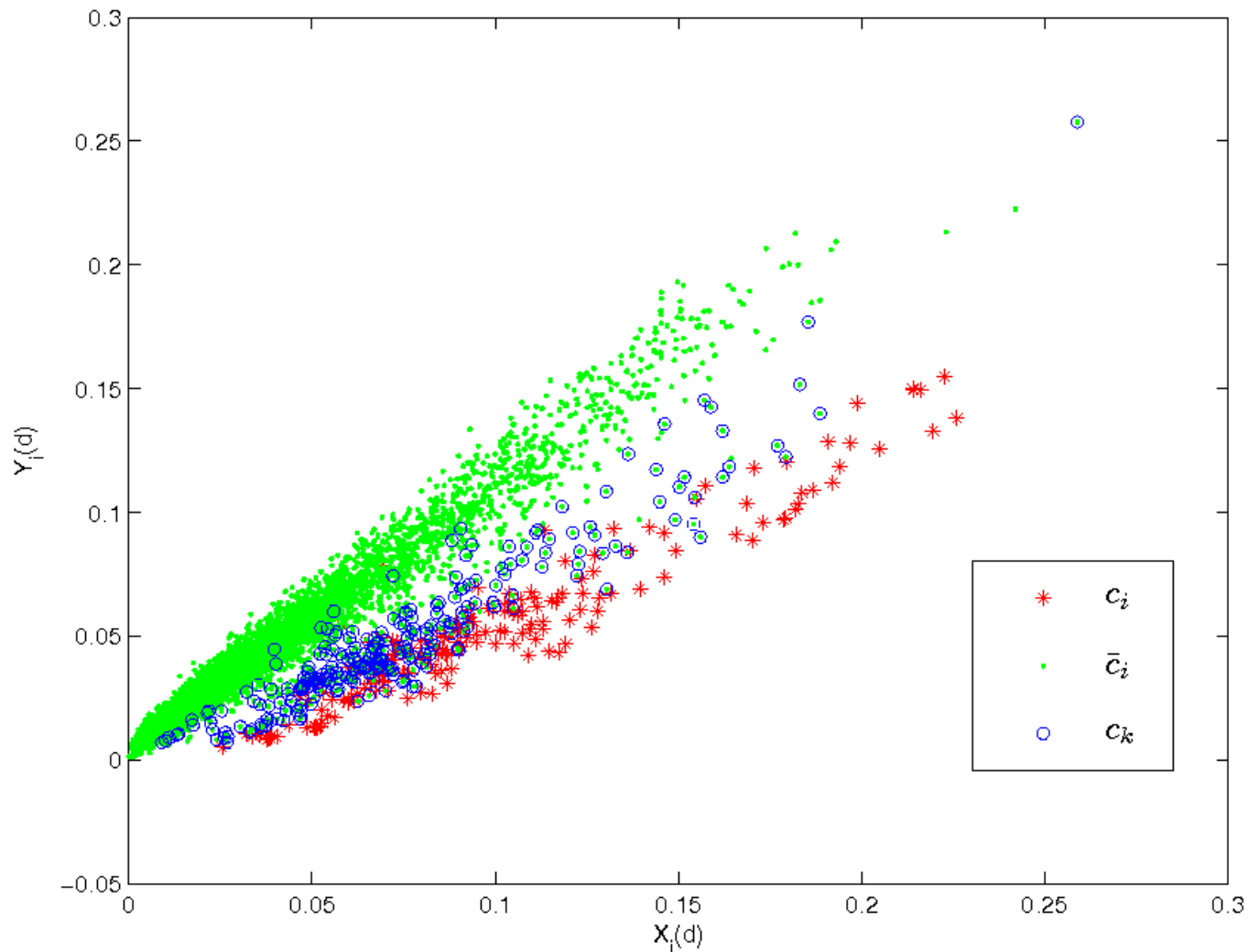
$$ms(j) = \sum_{i=0}^{I-1} w(i, j)$$

- where $w(i, j)$ is a term weighting scheme.

Visualization and Classification of Documents

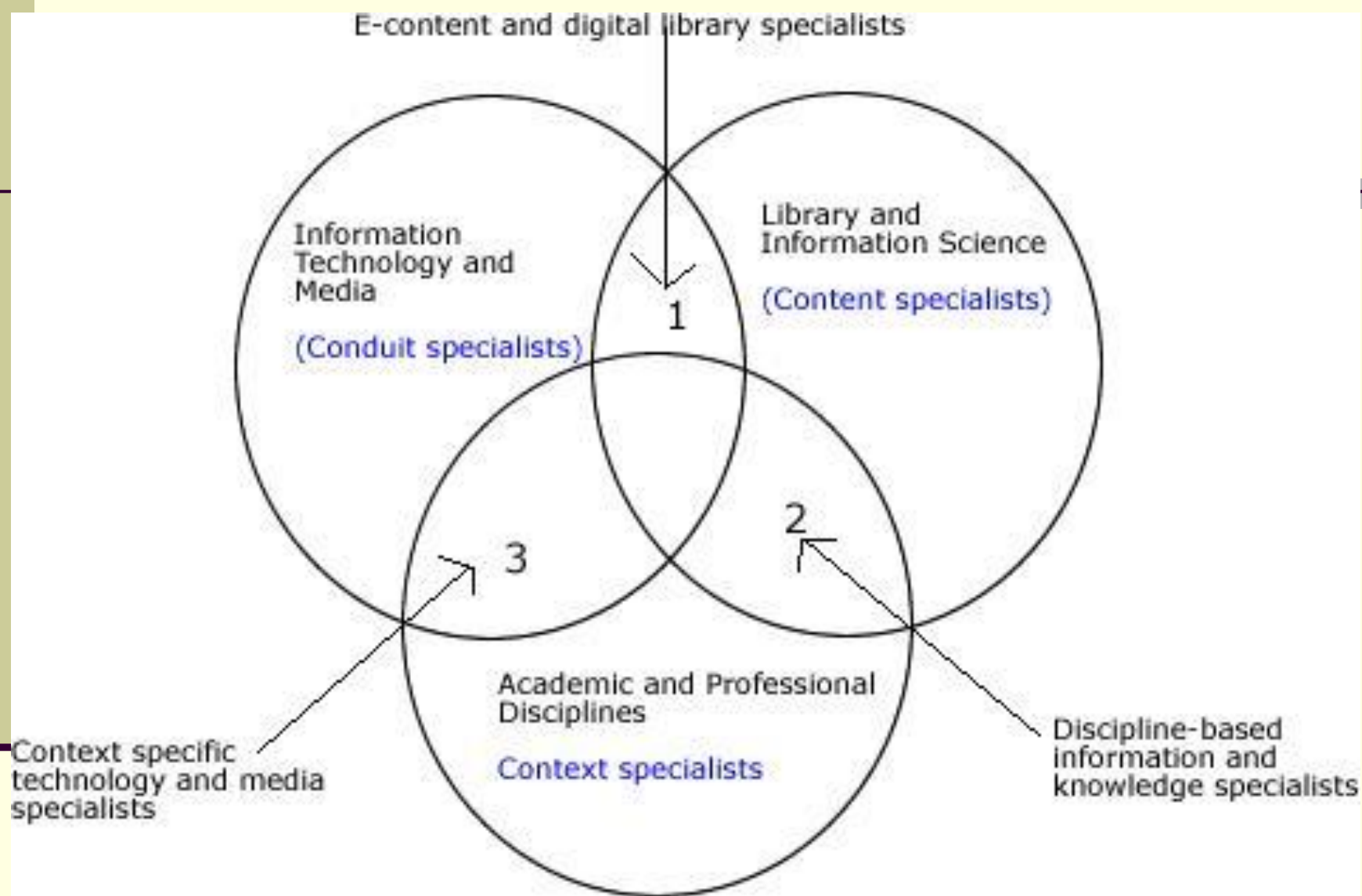
A New Probabilistic Model to Automated Text Classification

<http://www.ieee-tcdl.org/Bulletin/v2n2/dinunzio/dinunzio.html>



Hybrid Model

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



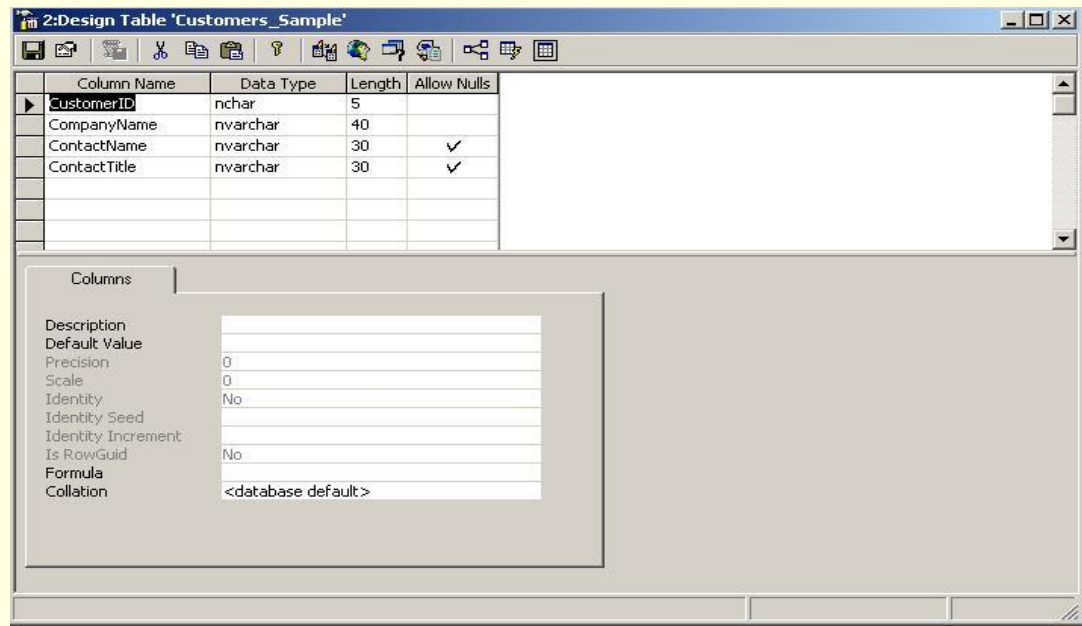
Hybrid professionals (after Corrall and Lester, 1996)

ตาราง: แสดงการเปรียบเทียบโมเดลระบบค้นคืนสารสนเทศในรูปแบบต่างๆ

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

Flat-file database

- ฐานข้อมูลรายแฟ้มหมายถึง โปรแกรมฐานข้อมูลที่ใช้เก็บและเรียกหาข้อมูล ออกมาใช้ได้ที่ละแฟ้ม เป็นสำคัญ ตรงข้ามกับฐานข้อมูลสัมพันธ์ (relational database) ซึ่งจะช่วยให้เรียกหาใช้ได้ มากกว่าครั้งละหนึ่งแฟ้ม



Stem

Search Bible Text

Search in: KJV (King James Version)

Normal Search Proximity Search Character String Search RegEx Search

Search Text
Specify the words, phrase, or boolean expression to search for:
love View Word List...

Search Method
☒ Find verses with ALL of the words specified
☐ Find verses with ANY of the words specified
☐ Find verses with the exact PHRASE specified
☐ Find verses matching the BOOLEAN EXPRESSION specified

Search Options
☐ Match case ☐ Words of Christ only
☒ Find all word forms (English stemming) ☐ Include footnotes in search, if any
☐ Find close word matches (fuzzy matches) ☐ Consonant search (for Hebrew texts)
☐ "Substring" search (partial word matches, same as *word*)

Search Range:
Genesis - Revelation
All Old Testament New Testament Book Custom

Cancel Help Do Search

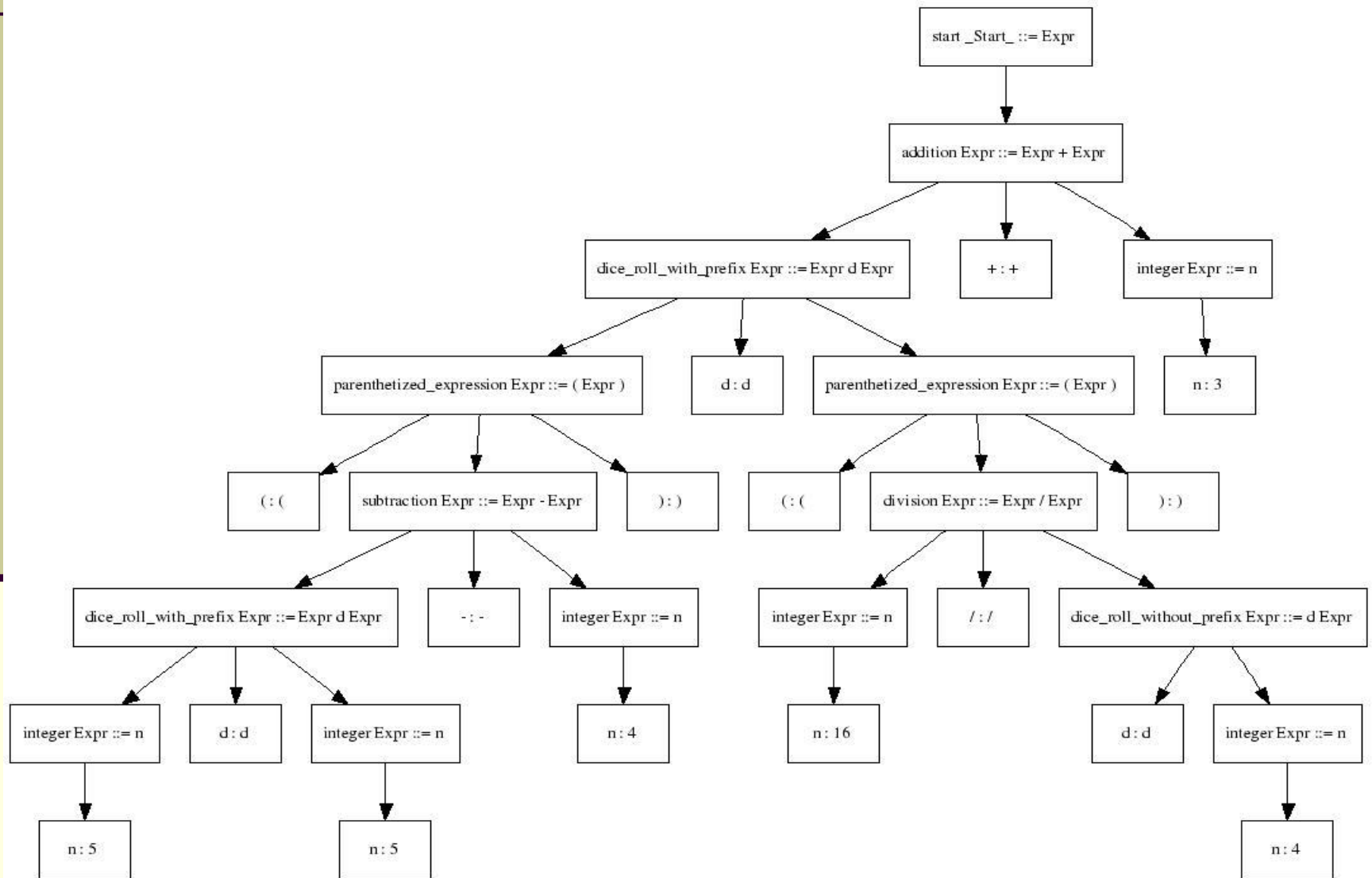
Search results for: boat

1 result found.

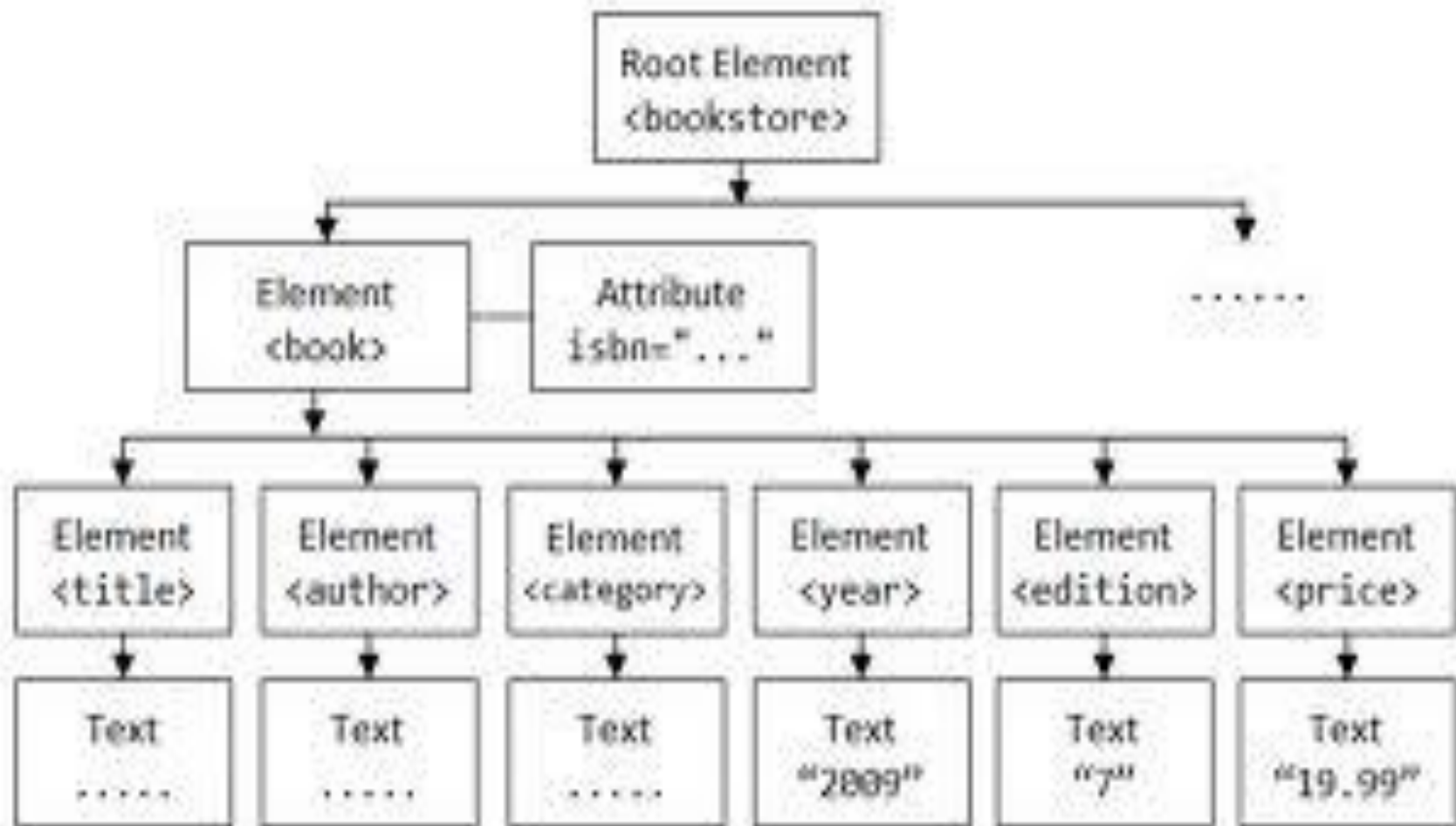
1. [Norfolk Broads Boat Hire](http://www.aboatsite.co.uk/)

... Boat Hire Luxurious breaks on the water Norfolk Broads boating holidays Our Luxury Boats enquiries Norfolk Broads boat hire prices Offers Norfolk Broads visitors guide Posh Boats new
Terms matched: 1 - Score: 422 - 9k - URL: <http://www.aboatsite.co.uk/>

Parse tree



Parse



Signature file

The screenshot shows the 'Options' dialog box with the 'Signatures' tab selected. The 'Signature settings' section has two checkboxes: 'Add signatures to all outgoing messages' (unchecked) and 'Don't add signatures to Replies and Forwards' (checked). The 'Signatures' section shows a list with 'Signature #1' and 'Default signature'. To the right of the list are 'New', 'Remove', and 'Rename' buttons. The 'Edit Signature' section has a 'Text' radio button selected, a text area, and a 'File' radio button with a text box containing 'C:\signatures\signature1.htm'. To the right of the text area are 'Set as Default' and 'Advanced...' buttons. To the right of the text box is a 'Browse...' button. At the bottom are 'OK', 'Cancel', and 'Apply' buttons.

Spelling Security Connection Maintenance
General Read Receipts Send Compose Signatures

Signature settings

☐ Add signatures to all outgoing messages
☒ Don't add signatures to Replies and Forwards

Signatures

Signature #1	Default signature

New
Remove
Rename

Edit Signature

☒ Text ☐ File

Set as Default
Advanced...

C:\signatures\signature1.htm Browse...

OK Cancel Apply

This screenshot is identical to the one on the left, but with red annotations. Red circles and arrows highlight specific elements: 1. The 'Signatures' tab label. 2. The 'Add signatures to all outgoing messages' checkbox. 3. The 'New' button in the 'Signatures' list. 4. The 'Text' radio button in the 'Edit Signature' section. 5. The text area in the 'Edit Signature' section, which contains the text '#5 see below for details'. 6. The 'Apply' button at the bottom right.

Options

Spelling Security Connection Maintenance
General Read Receipts Send Compose Signatures

Signature settings

☒ Add signatures to all outgoing messages
☒ Don't add signatures to Replies and Forwards

Signatures

Signature #1	Default signature

New
Remove
Rename

Edit Signature

☒ Text ☐ File

Set as Default
Advanced...

#5 see below for details

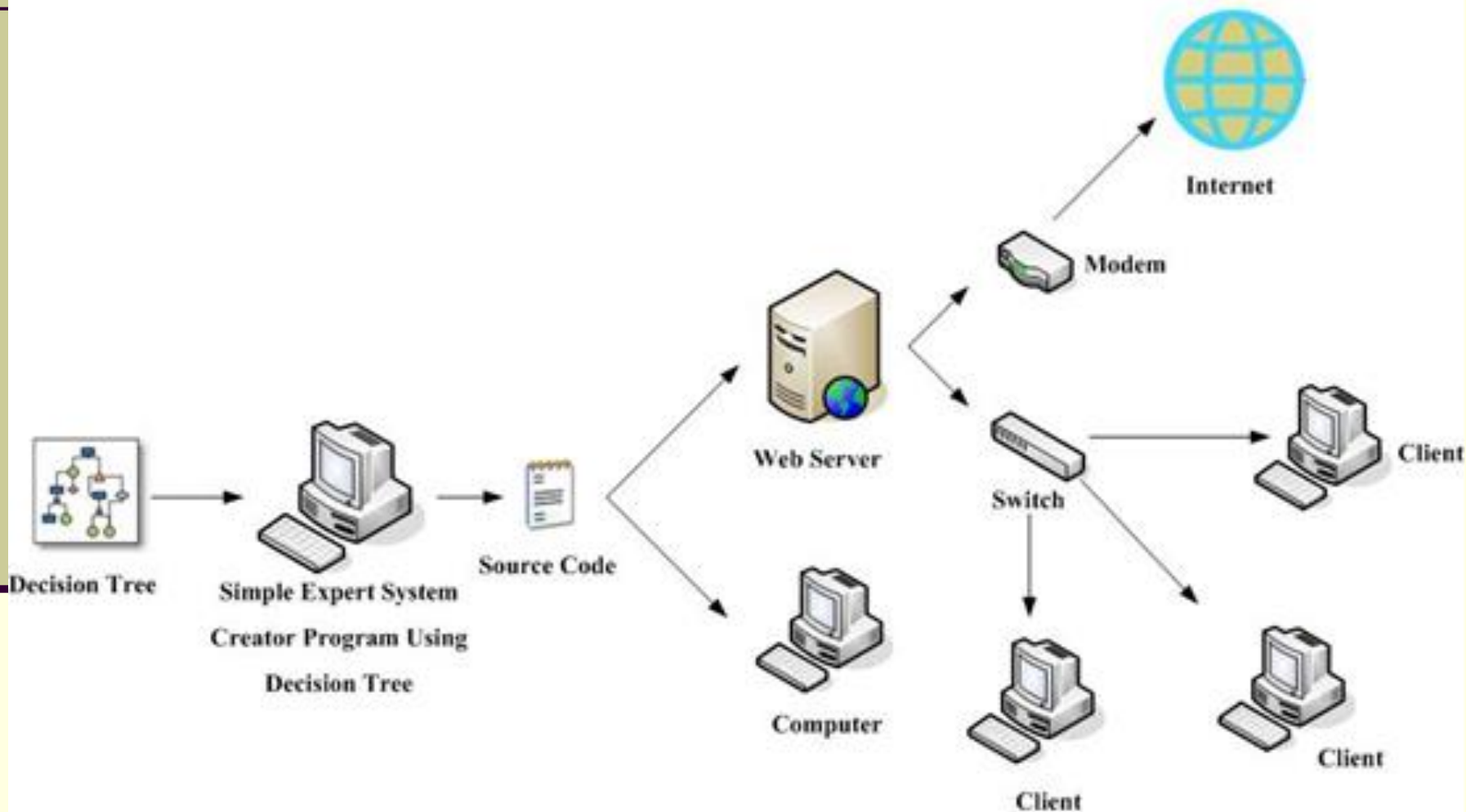
C:\signatures\signature1.htm Browse...

OK Cancel Apply

เทคโนโลยีที่สนับสนุน => ปัญญาประดิษฐ์

- ระบบผู้เชี่ยวชาญ
- โปรแกรมตัวแทนอัจฉริยะ
- การประมวลผลภาษาธรรมชาติ
 - การใช้ภาษาธรรมชาติสำหรับข้อคำถาม ได้แก่ **Ask**
 - การค้นคืนสารสนเทศหลายภาษา หรือ การค้นคืนสารสนเทศข้ามภาษา

ระบบผู้เชี่ยวชาญ



โปรแกรมตัวแทนอัจฉริยะ

- ซอฟต์แวร์ที่มีการปัญญาประดิษฐ์เพื่อช่วยเหลือผู้ใช้
- มนุษย์โดยสามารถเรียนรู้ได้ด้วยตัวเองจากประสบการณ์และข้อมูลที่เกี่ยวข้องในอดีต และเพื่อพัฒนาอัลกอริทึมทาง **Computational Intelligence** เพื่อช่วยสนับสนุนการตัดสินใจของมนุษย์ หรือแม้กระทั่งทำงานแทนที่มนุษย์
 - การพัฒนาโปรแกรมแปลงอักษรเบรลล์
 - การวิเคราะห์การใช้เทคโนโลยีคอมพิวเตอร์สำหรับนักเรียนพิการทางการมองเห็น กรณีเรียนร่วมระดับมัธยมศึกษา กรุงเทพมหานคร

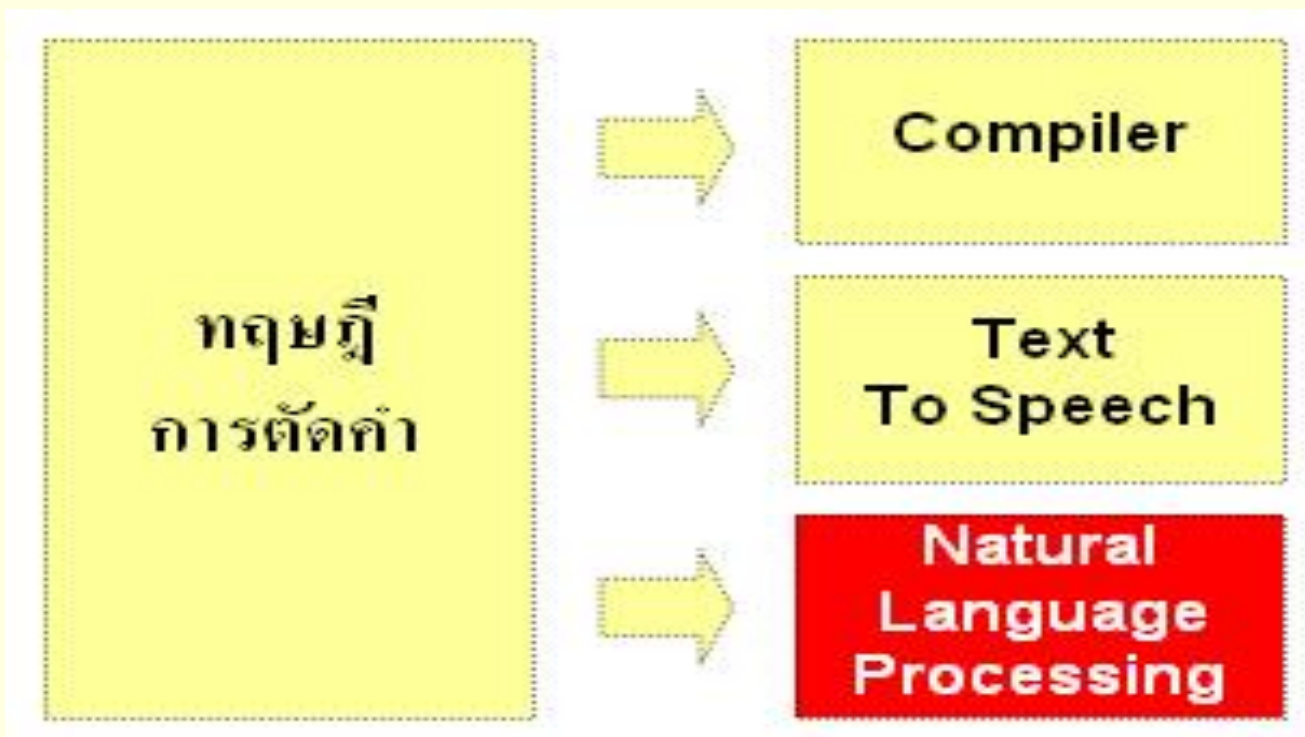
โปรแกรมแปลงอักษรเบรลล์

ตัวเลข				
⠠	เครื่องหมายนำเลข			
1	2	3	4	5
⠠⠠	⠠⠠	⠠⠠	⠠⠠	⠠⠠
6	7	8	9	0
⠠⠠	⠠⠠	⠠⠠	⠠⠠	⠠⠠



การประมวลผลภาษาธรรมชาติ

- การรับอินพุตจากการพิมพ์ผ่านคีย์บอร์ด หรือเสียงพูดผ่านไมโครโฟน หรือจากการอ่านเข้าระบบผ่านแฟ้มข้อมูลเวิร์ดโปรเซสซึ่ง แล้วจัดการตัดทอนคัดเอาเฉพาะที่มีความหมายสมบูรณ์เข้าเก็บในคอมพิวเตอร์



NLP องค์ประกอบที่สำคัญดังนี้

- การวิเคราะห์ในเชิงโครงสร้าง (**Syntactic Analysis**) จะเป็นการตรวจสอบโครงสร้างทางไวยากรณ์เกี่ยวกับการวางตำแหน่งของคำนาม กริยา คำบุพบท ฯลฯ ที่รวมเป็นประโยค
- การวิเคราะห์ในเชิงความหมาย (**Semantic Analysis**) จะเป็นการแยกแยะความถูกต้องทางความหมายของประโยค ประโยคที่ถูกต้องแบบตามโครงสร้างไวยากรณ์จะมีความหมายอย่างไรอย่างหนึ่งแน่นอน แต่บางครั้งความหมายที่ได้เป็นความหมายที่กำกวม หรืออาจไร้ความหมายเป็นไปไม่ได้

Content scanning

- การใช้คอมพิวเตอร์มาช่วยแสดงหรือตีความเนื้อหาของข้อความ
 - เครื่องอ่านและแปลข้อความของการเดินเรือทะเล
 - เครื่องอ่านและแปลเทเล็กซ์การโอนเงินระหว่างประเทศ
- ลักษณะของข้อความที่ใช้เป็นแบบสั้น ๆ คำที่ใช้มักจะเป็นคำย่อและไม่มีรูปไวยากรณ์ที่ดีนัก
- การทำงานของระบบจะทำในระดับความหมาย (semantic level)
การวิเคราะห์คำหรือข้อความโดยอาศัยพจนานุกรมที่เก็บคำศัพท์ไว้ในหน่วยความจำคอมพิวเตอร์ แล้วทำงานร่วมกับกฎตามหลักของภาษาเพื่อตีความบอกรูขี้นกและความสัมพันธ์ของข้อมูล

Capture documents by scanning or uploading documents

Content Central - Document Management Software

http://ContentCentral/

content central v5 Search Folders Capture Search All Catalogs Admin Options Logout Help

☐ Search by Packet

Full-Text Search

Jones Engineering

☐ Search all versions

Search Clear

Catalogs

☒ Accounts Payable

☐ Human Resources

Document Types

☒ Check

☒ Invoice

☒ Packing List

☒ Purchase Order

☒ Purchase Req

☐ Quote

Search

AdEMERO

Check File Modify View

Drag a column header here to group by that column.

		*DocName	Vendor Name	PO Number	*DocType	*Catalog
☒		Check 2482 - Jones Engineering.pdf	Jones Engineering	7312	Check	Accounts Payable
☐		Jones Engineering - Invoice 4802.pdf	Jones Engineering	7312	Invoice	Accounts Payable
☐		Jones Engineering - Packing List 4802.pdf	Jones Engineering	7312	Packing List	Accounts Payable
☐		Purchase Requisition 7312 - Jones Engineering.pdf	Jones Engineering	7312	Purchase Req	Accounts Payable
☐		Purchase Order 7312 - Jones Engineering.pdf	Jones Engineering	7312	Purchase Order	Accounts Payable

Results: 5

Southside Manufacturing, Inc.
1100 Commercial Blvd.
Hollywood, FL 33020
754-890-5001

First Bank of Bal Harbour
1505 Branch St.
Bal Harbour, FL 33154
305-320-1232

2482

DATE March 22, 2009

PAY TO THE ORDER OF Jones Engineering \$ xx5,002.22

Five-Thousand Two Dollars and Twenty-Two Cents DOLLARS

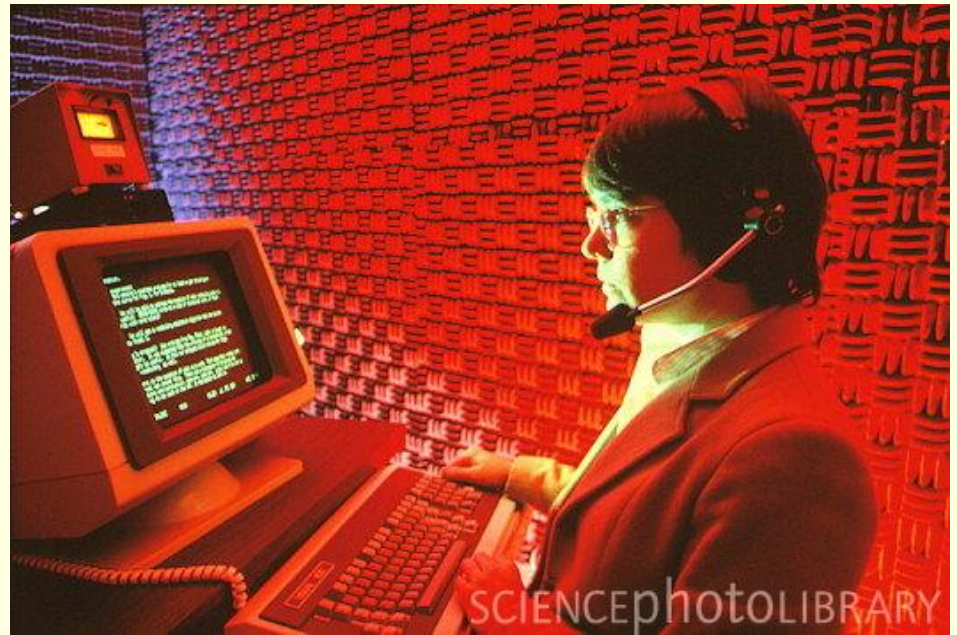
MEMO

James Smith

⑈001000⑈ ⑆123456789⑆ 123 12345 5⑈

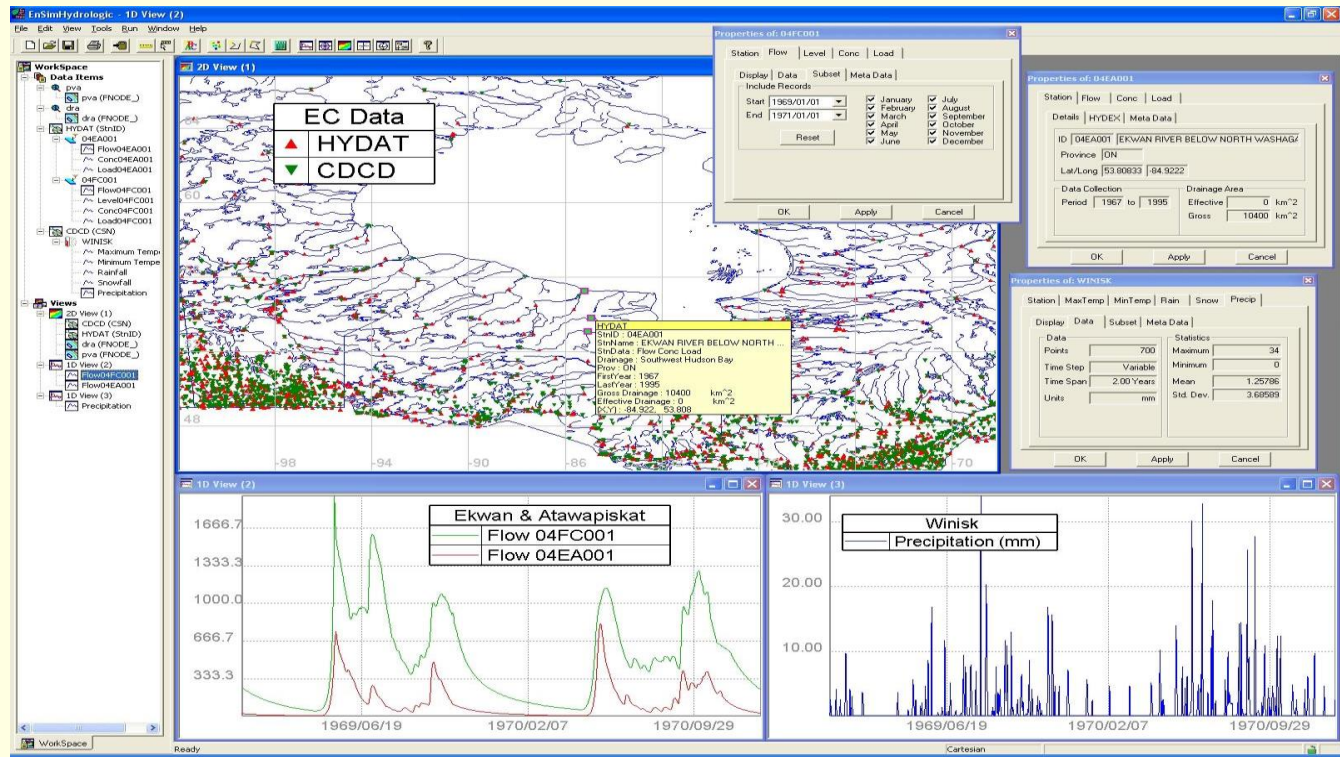
Talkwriter

- เครื่องที่รับเสียงพูดของมนุษย์ แล้วมีการถ่ายเปลี่ยน
สัญญาณเสียงให้อยู่ในรูปสัญญาณที่เครื่องสามารถอ่านได้
หรือในรูปของข้อความแบบเวิร์ดโปรเซสซิ่ง หรือในรูปของ
คำสั่งเพื่อจะส่งต่อไปใช้งานฐานข้อมูล

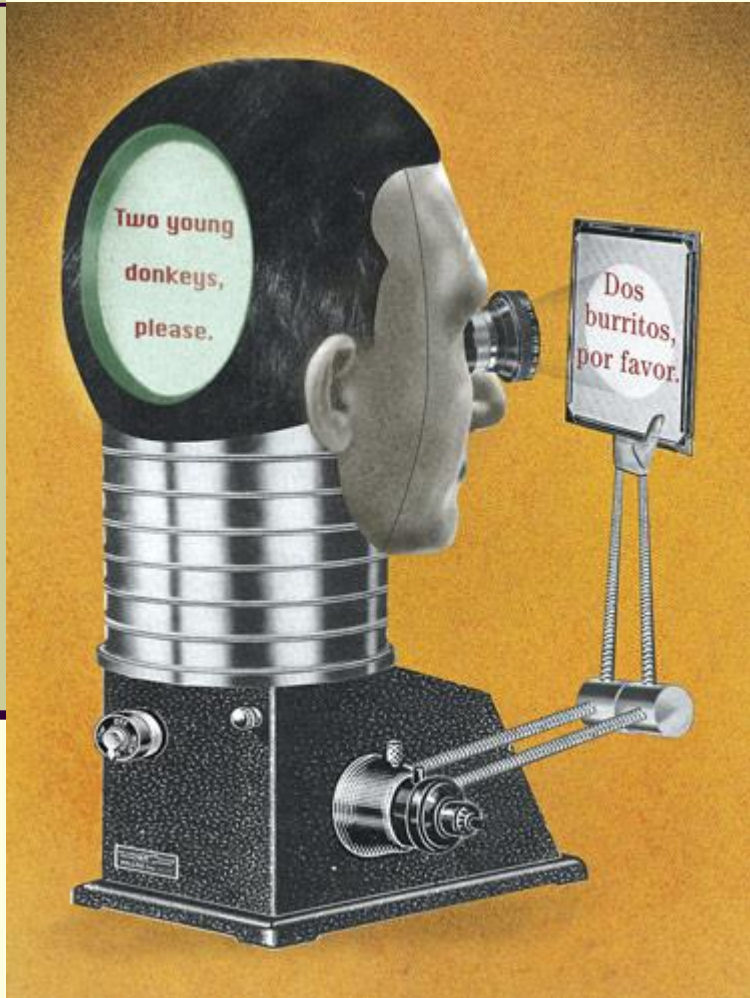


Database interface

- การใช้ภาษาธรรมชาติ มาสั่งงานเรียกใช้
- เป็นลักษณะของการถามหาข้อมูลด้วยภาษาที่ใช้ในชีวิตประจำวันจากภาษาธรรมชาติที่ใช้ ก็จะมีการแปลตีความ และสร้างเป็นคำสั่งเพื่อเรียกใช้ฐานข้อมูล



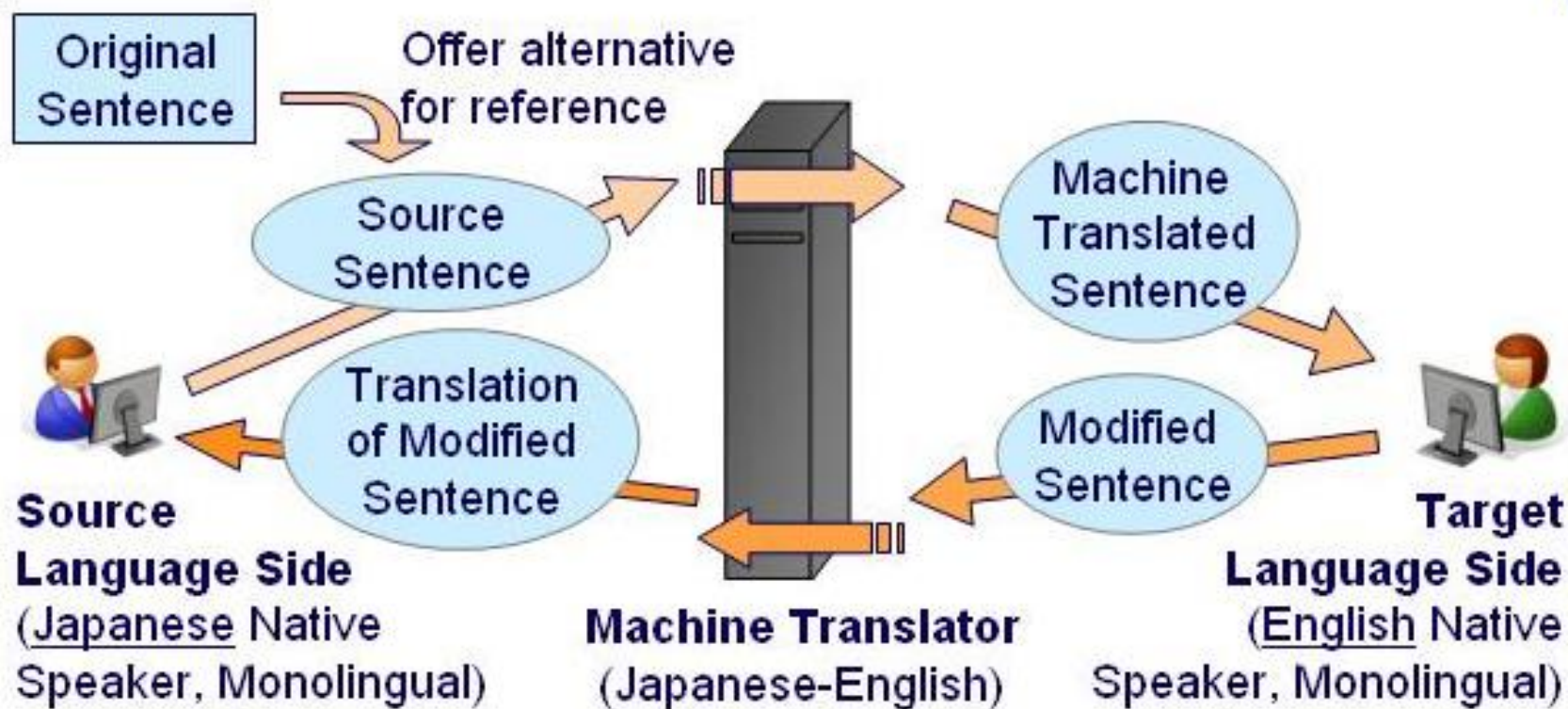
Machine translation



- เป็นการใช้อุปกรณ์คอมพิวเตอร์มาเพื่อการแปลภาษาระหว่างภาษาของมนุษย์
 - การแปลภาษาระหว่างภาษาไทยกับภาษาอังกฤษ
- การรวมข่ายงานของภาษาธรรมชาติ ทั้งในแง่ของการอินเทอร์เน็ตด้วยเสียงด้วย **database interface** และ **text editing** เข้าด้วยกัน
 - ช่วยอำนวยความสะดวกและเร่งการติดต่อแลกเปลี่ยนข้อมูลให้เป็นไปอย่างรวดเร็ว

Machine translation

มีการเก็บคำศัพท์พร้อมวิธีการใช้คำนั้นภายในหน่วยความจำของคอมพิวเตอร์ไว้ มีขนาดใหญ่มากจริง ๆ ดังนั้นจึงต้องมีการแปลเฉพาะสาขาในระยະนี้ แต่ถ้ำเทคโนโลยีของคอมพิวเตอร์ขยายไปอีก การแปลมากสาขาก็อาจเป็นไปได้ในอนาคต



การวิจัยและการพัฒนา

- แบ่งออกได้เป็น 2 ประเภท คือ
 - เชิงเทคนิค การจับคู่ได้มีประสิทธิภาพขึ้น การทำตรรชนีหรือสาระสังเขปอัตโนมัติ
 - เชิงผู้ใช้และการใช้ เน้นการใช้ที่ตรงตามความต้องการ

ผู้ใช้และการใช้สารสนเทศ (1)

- **ผู้ใช้** คือ ผู้มีความต้องการสารสนเทศ แสวงหา ค้นหา หรือ ใช้สารสนเทศ/ ตัวแทนสารสนเทศ
- **ความต้องการสารสนเทศของผู้ใช้** — ภาวะที่บุคคลตระหนักถึงช่องว่างทางความรู้ หรือการขาดสารสนเทศ ซึ่งผลักดันให้เกิดการแสวงหาสารสนเทศ
- ความต้องการสารสนเทศ** — มีลักษณะเป็นพลวัต จำแนกได้ตามเงื่อนไขของเวลา ตามขอบเขตและลักษณะความต้องการ

ตัวอย่าง ห้องสมุดดิจิทัล

- องค์ประกอบสำคัญ คือ
 - ทรัพยากรในรูปแบบดิจิทัล
 - กระบวนการจัดเก็บ
 - บริการสารสนเทศ
 - ผู้ใช้
 - เทคโนโลยี





Innovation, Demand and Investment in Telehealth

February 2008



From Computer Desktop Encyclopedia
Reproduced with permission.
© 1999 NavoMedia, Inc.



ห้องสมุดดิจิทัลที่แตกต่างจากเว็บไซต์ทั่วไป

- จัดการตามมาตรฐานห้องสมุดดิจิทัลที่เป็นสากล
- จัดหมวดหมู่และการทำรายการที่เป็นมาตรฐาน
- ค้นทำได้อย่างมีมาตรฐาน (หัวเรื่อง ชื่อเรื่อง ชื่อผู้แต่ง คำสำคัญ)
- นำเสนอผลการค้นอย่างมีมาตรฐาน (ภาพในลักษณะต่าง ๆ ฯลฯ)
- ข้อมูลมีคุณภาพ: ความถูกต้อง ความน่าเชื่อถือ ฯลฯ
- ข้อมูลและแหล่งยั่งยืน
- มีการเพิ่มเติมเนื้อหาอย่างต่อเนื่อง