Information Retrieval and Search Engine

Opas Wongtaweesap

Department of Computing

Faculty of Science, Silpakorn University

E-mail :: oatcomster@gmail.com

Outline

- Motivation
- Basic Concepts
- Past, Present, and Future
- The Retrieval Process
- IR Application

What is IR?

- You know Google?
- You know Yahoo?
- You know MSN Search?
- What goes on behind search engine???
- How do they work??????

Motivation

- Information Retrieval (IR) deal with
  - The *representation*, *storage*, *organization* of, and *access* to information items based on the user information need.
- Focus is on the Information need
  - One or several sentences including complex descriptions such as event, place, time, people, …
  - Translated into a query submitted to IR systems
    - A set of keywords (indexed terms or phrases) summarizes the information need
- Goal: retrieve information which might be useful or relevant to the user's information need

Relevance

- Much of IR depends upon idea that
  - Similar vocabulary -> relevant to same queries
- Usually look for documents matching query words
- “Similar” can be measured in many ways
  - String matching/comparison
  - Same vocabulary used
  - Probability that documents arise from same model
  - Same meaning of text

Information versus Data Retrieval

- Data Retrieval: determine a set of documents contain query keywords
  - Data corpus is well structured data
  - Exact match (no error): regular or relation algebra expression
  - Example: relational database systems
    - SELECT … FROM … WHERE …
- Information Retrieval: determine a set of documents which are relevant to the query
  - Data corpus is no always well structured
  - Proximity match (inaccurate and small error): documents that are probably match the user's information need
  - Interpret documents: extracting syntactic and semantic information

Keyword Search

- Simplest notion of relevance is that the query string appears verbatim in the document
- Slightly less strict notion is that the words in the query appear frequently in the document, in any order (*bag of words*)

Problems with Keywords

- May not retrieve relevant documents that include synonymous terms.
  - “restaurant” vs. “café”
- May retrieve irrelevant documents that include ambiguous terms.
  - “bat” (baseball vs. mammal)
  - “Apple” (company vs. fruit)
  - “bit” (unit of data vs. act of eating)

Intelligent IR

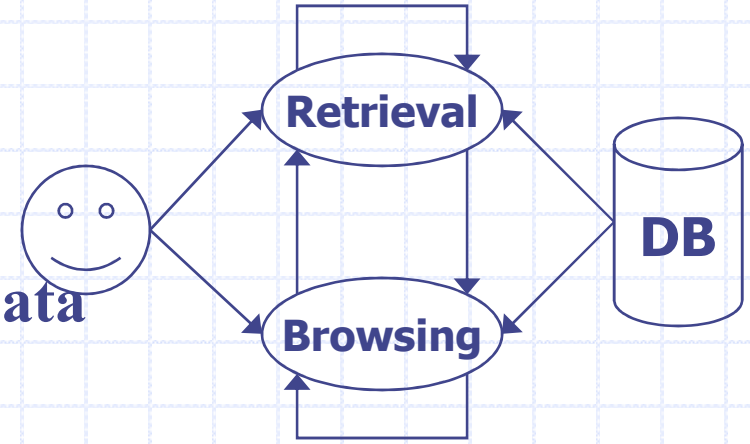
- *meaning* of the words used.
- *order* of words in the query.
- Adapting to the user based on direct or indirect feedback.
- *authority* of the source.

IR and the Web

- The beginning of the 1990s: the World Wide Web (WWW, Web, W3)
  - A world wide **hyperlinked** environment for **browsing** and finding information needs
  - The universal repository of human knowledge
  - Share ideas and information
  - Any user can create his own Web documents that are accessible to everybody
- New challenge:
  - Finding useful information on the Web: tedious and difficult
  - Navigate the vast **hyperspace**
- IR has gained a place with other technologies at the center of the Web

Basic Concept

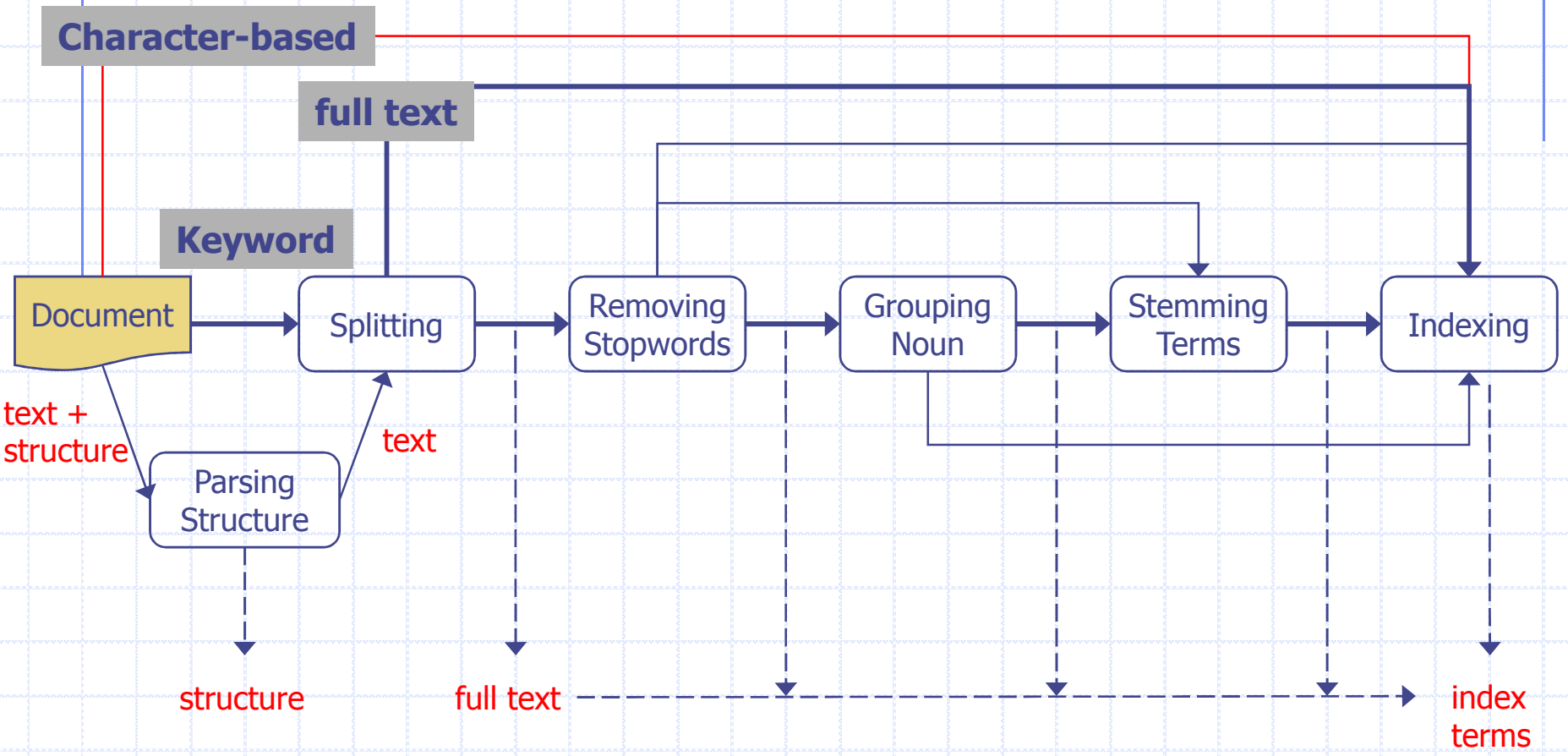
- The User Task
- Classic IR systems
  - Retrieving information or data
- Hypertext systems
  - Quick browsing
- Digital library and Web interface
  - Retrieving and browsing
  - Interacting with the user: user's relevance feedback



Logical View of the Document

- Represent a document in the computer
  - Keyword index: a set of terms (keywords or phrases)
  - Full text index: a full set of words
  - Character-based index: a full set of binary data (character)

Character, Full text, Keyword Index



Past, Present, and Future

- Developments
  - TOC: table of contents
  - Index: subjects, glossary, thesaurus
  - Library: hierarchical categorization of indexes
  - Automatic indexes
- Computer-centered view
  - Performance & Quality
    - efficient index and precise retrieval
- Human-centered view
  - Understanding user's information needs
    - Guide the user to find (browse) useful information
- An example: Web Portal Sites

TOC :Table of Contents

Table of Contents

W3C Candidate Recommendation

Index

Index

Inverted Files

- **Original text:**
  John Davenport, 52 years old, was appointed chief executive officer of this international telecommunications concern's U.S. subsidiary, Cable & Wireless North America Inc. Mr. Davenport, who succeeds John Zrno, is currently general manager of the group's operations in Bermuda.
- **One indexing result:**
  john davenport appoint chief executive officer international telecommunication concern subsidiary cable wireless north america davenport succeed john zrno current general manager group operation bermuda

Keyword ID

Document ID

	001	002	003	004	005	006	...
001	12	0	47	2	7	4	…
002	25	4	1	0	0	7	…
003	0	0	6	17	3	5	…
004	1	2	7	14	2	1	…
005	3	16	0	5	7	20	…
006	9	10	15	1	16	0	…
...	…	…	…	…	…	…	…

Numbers of each Keyword was found in each Document

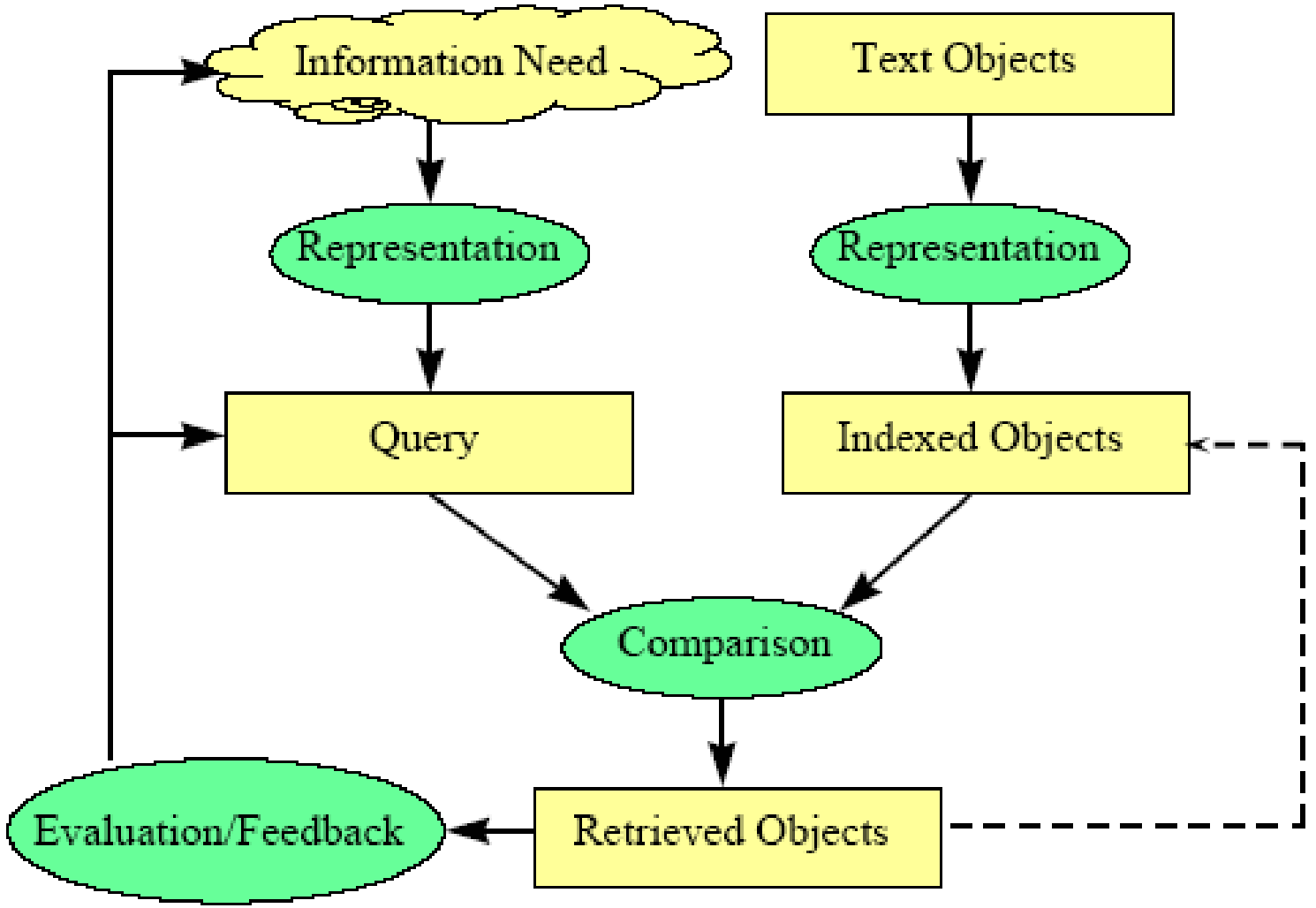
Issues of IR

- Issues of the Web and Digital Libraries
  - Retrieval of high quality: the right information
  - Quick response for unlimited users
  - User interaction
- Practical Issues
  - EC: electronic commerce
  - Security
  - Copyright and patent rights
  - Optical Character Recognition
  - Cross-language retrieval
  - …

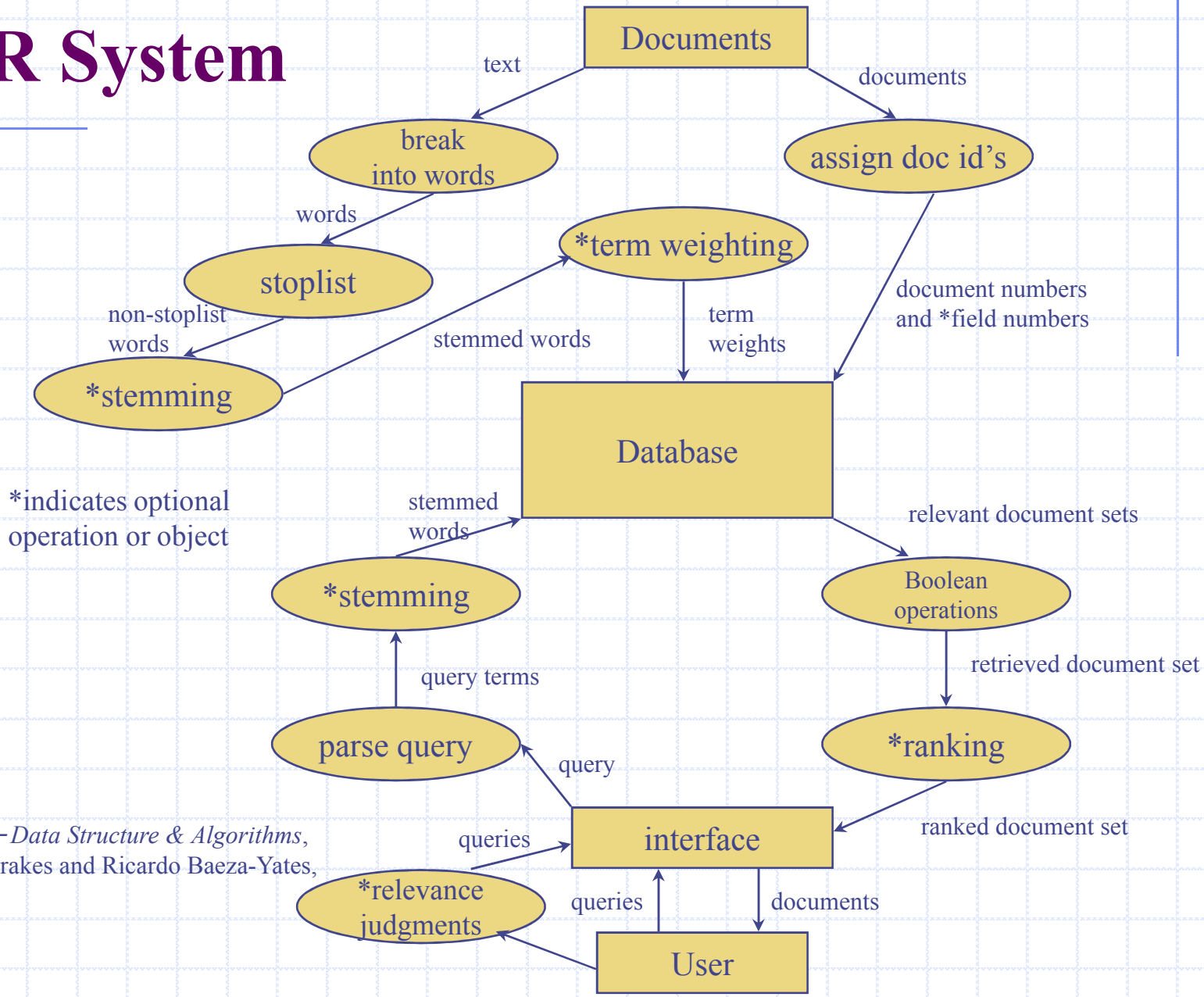
The Retrieval Process

- Document Operation
  - Document → Text Operation → Text Model (Logical View of Document)
- Index: Inverted File
  - Index once and Retrieve many time (for queries)
  - Time and storage space
- Query Operation
  - Regard the query as a short document text
  - Similar to text operation
  - Ranking: Relevance score (similarity)
  - Relevance Feedback

IR Processes



An IR System



From:
Information Retrieval – Data Structure & Algorithms,
Edited by William B. Frakes and Ricardo Baeza-Yates,
Prentice Hall, 1992

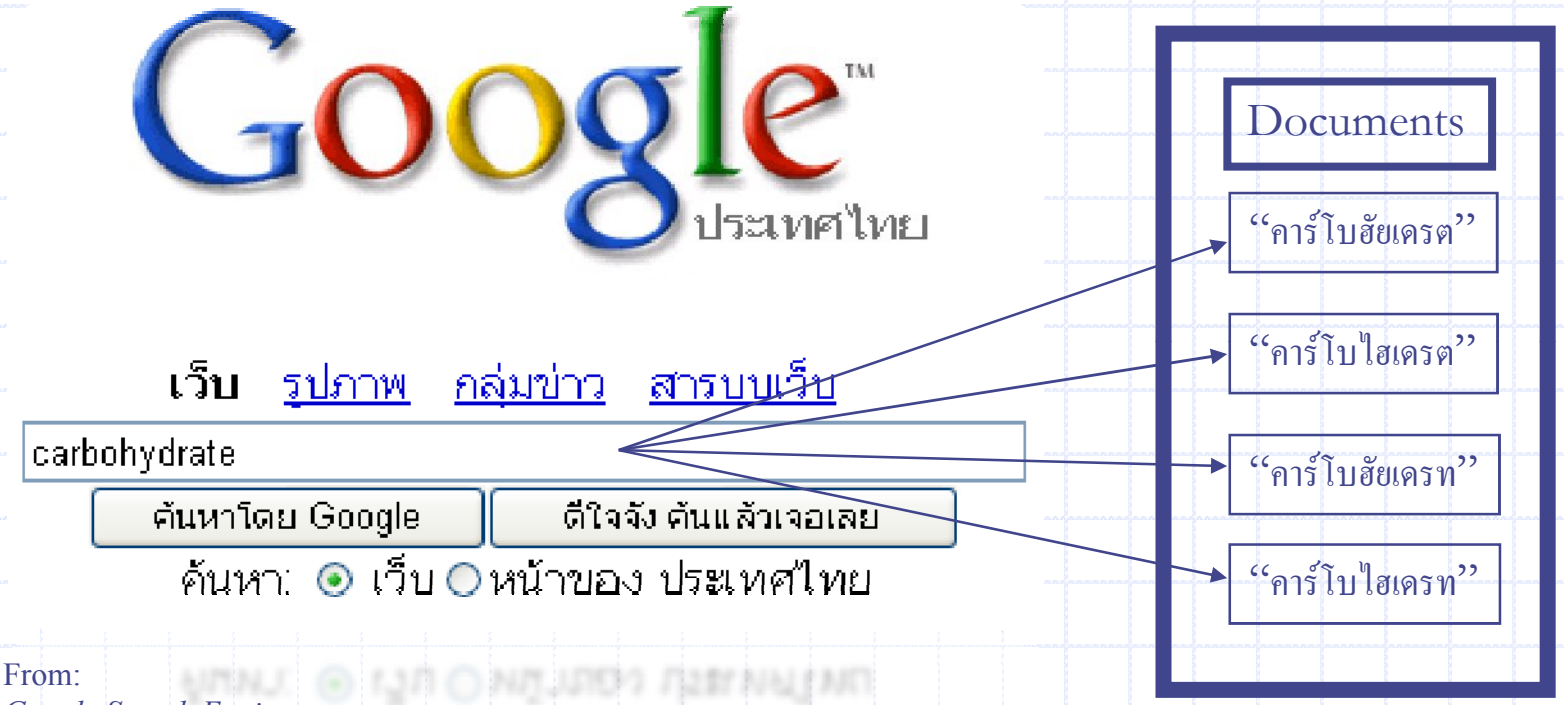
My IR Application – Senior Projects I

- Adaptive Testing System Development
- Tools for Important-Words Translation in Specific Document
- Information Retrieval System for Indexing Keywords and Multiple-Type Documents Inserted by User
- An Improvement of Information Retrieval System by Cross Language Retrieval Methodology

My IR Application – Senior Projects II

- Image Retrieval System Development by Indexing Keyword and Image Similarity Retrieval Techniques
- Online Transaction Management and Customer Relationship Management System for One Tambon One Product Case Study : Nakorn Pathom Province
- Tutorial System and Student's Programming Behavior Analysis System

Cross-Language Retrieval



From:
Google Search Engine,
http://www.google.co.th

Image Retrieval

เว็บ **รูปภาพ** แผนที่ Groups บล็อก แปลภาษา Gmail เพิ่มเติม ▼ **oatcomst**

Google รูปภาพ cat ค้นหาภาพ การค้นหารูปภาพขั้นสูง การตั้งค่า

SafeSearch ปานกลาง ▼

รูปภาพ แสดง: ขนาดใดก็ได้ ประเภทใดก็ได้ สีใดก็ได้ ▼ ผลการค้นหา 1 - 20

การค้นหาที่เกี่ยวข้อง : **kitten**

Cat
500x325-26k - jpg
61.19.69.9

อ้างอิง http://**cat**-lucky.
932x763-181k - jpg
littlecatzhome.net

Funny Picture : Baby and **cat**
299x400-28k
picture.forfun.us

Cat.jpg **Cat** image by
500x367-112k - jpg
blogger.sanook.com

++ Flying **Cat**
539x312-26k - jpg

Funny Picture : One Funny **Cat**

CAT
500x336-27k - jpg

ค่าสำคัญ: **cat** dog frier
450x401-37k - jpg

Image Retrieval System

- http://amazon.ece.utexas.edu/~qasim

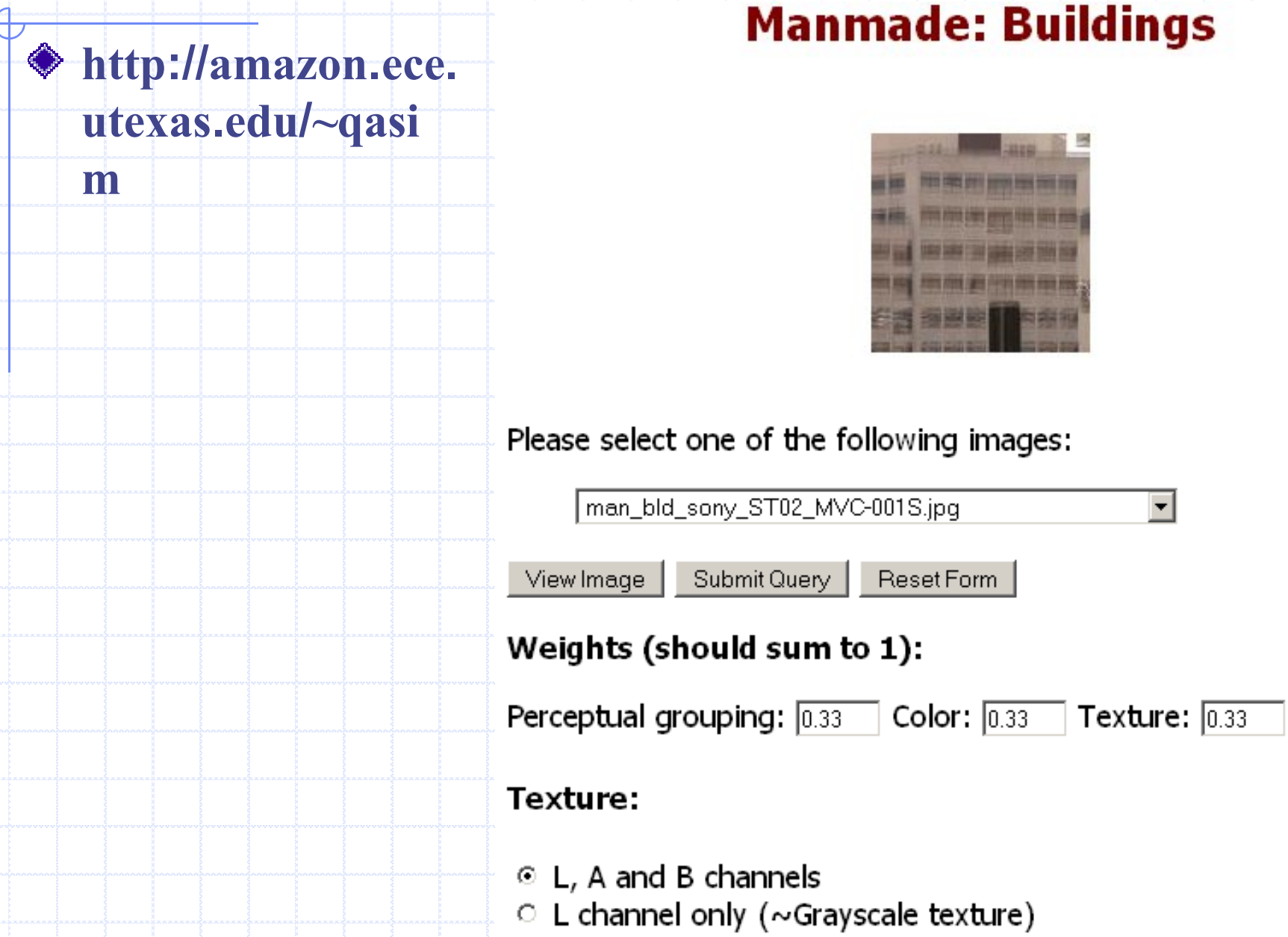


Image Retrieval System

Query Image

Relevance feedback type: Cluster. Weights: Perceptual Grouping = 0.33, Color = 0.33, Texture = 0.33, L, A, B channels.
For relevance feedback, please select the check boxes below each image, and then select the feedback type. Note that NS = "Not Sure".

Retrieved Images

Image Similarity

- http://www.myheritage.com/

Crime Retrieval System

Mirror Google

- **http://www.alltooflat.com /geeky/elgoog/m/index.cgi**

We're sorry!

Elgoog, the Google Mirror, is currenly undergoing a server upgrade.

All About Elgoog

elgooG 6002©

Blackle


Google
Google Search
I'm Feeling Lucky
Advanced Search
Preferences
Language Tools
Advertising Programs - Business Solutions - About Google
©2007 Google


Blackle™
a Heap Media site
Search
Google™ Custom Search
1,357,766.694 Watt hours saved

Combined Results of Search Engine

- http://gahooyoogle.com/
- http://www.twingine.no/

Google Books

- http://books.google.com/

ค้นหาหนังสือ

ค้นหาหนังสือขั้นสูง
วิธีใช้ Google Book Search

ค้นหาเนื้อหาเต็มของหนังสือและรายละเอียดใหม่ๆ

เกี่ยวกับ Google Book Search - ข้อมูลสำหรับผู้จัดพิมพ์ - Google หน้าแรก

Search Engine

- Search Engine คือ ระบบค้นหาข้อมูลบนอินเทอร์เน็ต แบ่งเป็น 2 ประเภทใหญ่ๆ ได้แก่ crawler-based search engines และ human-powered directories
- ตัวอย่าง Web ที่ให้บริการ
  - http://www.yahoo.com
  - http://www.alltheweb.com
  - http://www.google.com
- รายละเอียดเพิ่มเติมที่ http://searchenginewatch.com

Top Google Searches, 2008			
Fastest Rising (Global)		Fastest Rising (U.S.)	
1	sarah palin	1	obama
2	bejing 2008	2	facebook
3	facebook login	3	att
4	tuenti	4	iphone
5	heath ledger	5	youtube
6	obama	6	fox news
7	nasza klasa	7	palin
8	wer kennt wen	8	beijing 2008
9	euro 2008	9	david cook
10	jonas brothers	10	surf the channel

Top 10 Search Terms in 10 Categories, May 2009

Top 10 Search Terms by Category, Four Weeks Ending May 30, 2009 (%)			
IT and Internet		Automotive Manufacturers	
Search Term	Search Volume	Search Term	Search Volume
paypal	5.76	toyota	1.62
paypal.com	1.48	honda	1.44
www.paypal.com	0.86	ford	1.18
lady kathryn	0.85	harley davidson	1.14
people search	0.83	honda motorcycles	0.93
lite 1.4	0.62	nissan	0.89
paypal login	0.58	oreilly auto parts	0.88
pay pal	0.46	ford motor company	0.83
intelius	0.33	hyundai	0.78
experian	0.32	dodge	0.68

Movies		Net Communities and Chat	
Search Term	Search Volume	Search Term	Search Volume
netflix	2.60	myspace	4.83
imdb	1.43	facebook	4.49
netflix.com	0.78	myspace.com	2.30
blockbuster	0.46	youtube	1.81
fandango	0.44	facebook.com	1.28
star trek	0.44	facebook login	1.26
redbox	0.37	www.myspace.com	0.91
movies	0.36	www.facebook.com	0.57
new moon movie	0.27	my space	0.52
www.netflix.com	0.26	you tube	0.42

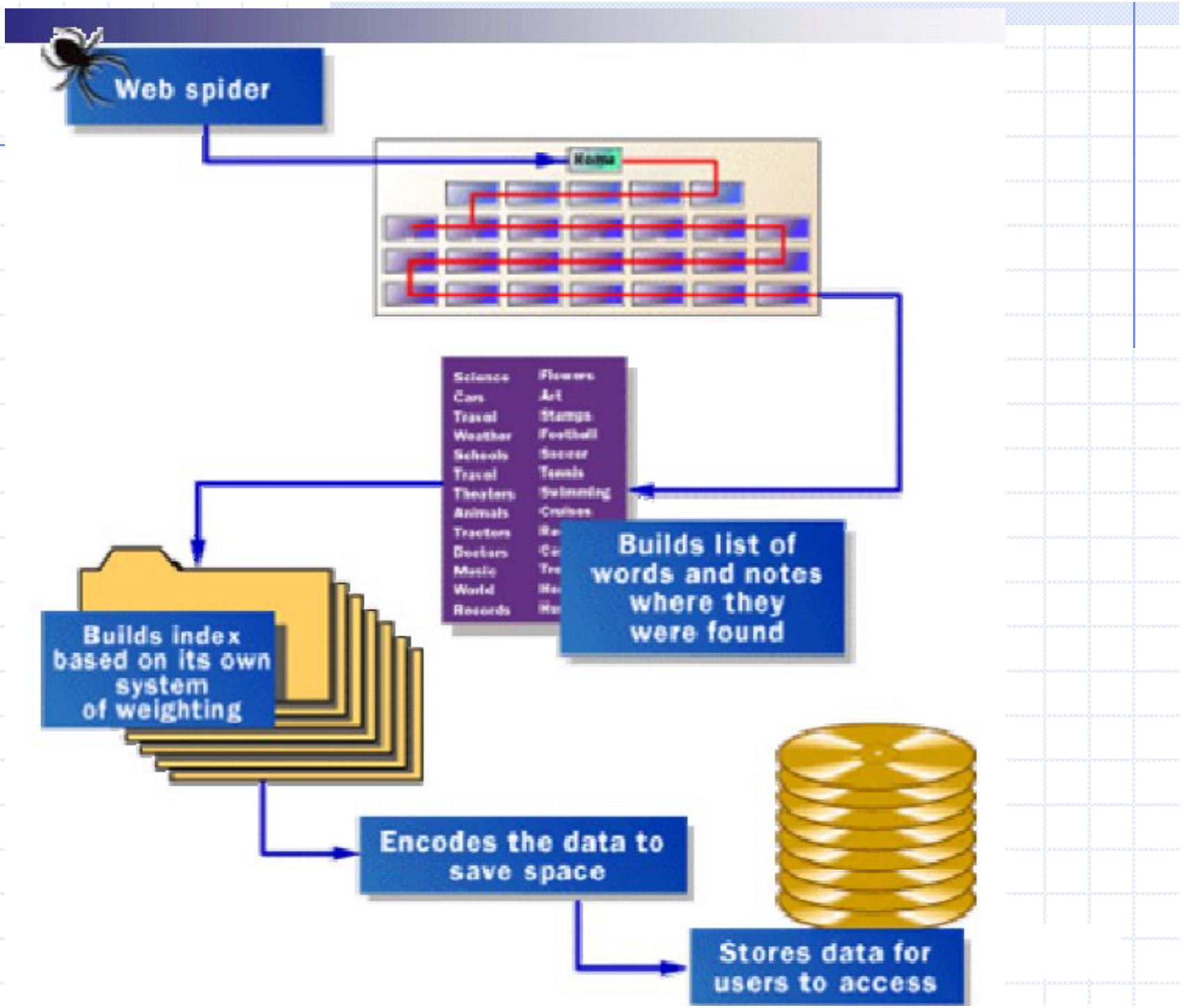
Top Social Networking Sites by Unique Visitors, December 2008			
Property	December 2007 (000)	December 2008 (000)	Change (%)
Total Internet audience	**183,619**	**190,650**	**4**
Social networking audience	**120,201**	**135,715**	**13**
MySpace.com	68,905	75,919	10
Facebook	34,658	54,552	57
Flickr	13,540	20,698	53
Classmates Online	10,002	16,553	66
MyLife.com**	N/A	15,018	N/A
Buzznet	4,973	9,781	97
AOL Community	40	9,208	22,701
Yahoo Buzz	4,864	8,724	79
AIM Profiles	2,587	8,618	233
Webs.com	N/A	8,053	N/A
Digg	6,026	6,844	14
LinkedIn	2,868	6,323	120
imeem	N/A	6,003	N/A
Tagged.com	1,156	5,778	400
Yahoo Groups	6,447	5,620	-13
Webshots	6,625	5,216	-21
DeviantART	4,102	4,905	20
Bebo	4,279	4,867	14
hi5	2,483	4,047	63
Windows Live Spaces	8,912	3,846	-57
Scribd.com	1,613	3,054	89
BlackPlanet.com	1,919	2,871	50
CafeMom.com	1,287	2,796	117
Sodahead.com	166	2,291	1,277

Notes:
1. ComScore audience measurement data report n media usage, visitor demographics, and online buying power for home, work, and university audiences across U.S. and worldwide Internet auciences.
2. Data excludes blogging sites.
**MyLife used to be known as Reunion.com

Source: comScore, 2009

Crawler-based search engines

- เป็นระบบที่สามารถที่จะจัดหาข้อมูลได้โดยอัตโนมัติ
- ระบบนี้อาศัย Web Crawler หรือ Spider ซึ่งเปรียบเสมือนกับหุ่นยนต์ที่คอยไปหาแหล่งข้อมูลต่างๆ เพื่อกลับนำข้อมูลมาเป็น index เก็บไว้ในฐานข้อมูล
- หาก Web มีการ update ในที่สุดตัวหุ่นยนต์จะค้นพบการเปลี่ยนแปลงและจัด index ใหม่ให้ทันสมัย เนื่องจากหุ่นยนต์ดังกล่าวจะถูกตั้งโปรแกรมให้กลับไปสำรวจอีกครั้งเมื่อถึงกำหนด


Web spider
Builds list of words and notes where they were found
Builds index based on its own system of weighting
Encodes the data to save space
Stores data for users to access

Human-powered directories

- เป็นระบบที่ดัชนีถูกจัดเรียงไว้แล้วโดยมนุษย์ ดัชนีจะถูกเปลี่ยนต่อเมื่อมีการจัดเรียงใหม่ ผู้ค้นสามารถเข้าไปค้นได้ตามดัชนีที่ถูกจัดเป็นหมวดหมู่เอาไว้ให้
- ระบบจะไม่สามารถตรวจสอบการเปลี่ยนแปลงข้อมูลของ Web ที่มีการเก็บดัชนีไว้แล้ว
- ปัจจุบันมีบาง Web Site ที่มีระบบ Hybrid Search Engines กล่าวคือใช้ระบบทั้งสองแบบเข้าด้วยกัน