An Overview of
Web Retrieval and Mining

Instructor: Pu-Jen Cheng (鄭卜壬)

Department of Computer Science & Information Engineering

National Taiwan University

World-Wide Web

- **Initiated at CERN (the European Organization for Nuclear Research)**
  - **By Tim Berners-Lee (1980)**

提姆・柏納斯李(Tim Berners-Lee)
http://www.eroach.net/revolution/0.htm

發明WWW的柏納斯李，他大學時研讀物理，至歐洲粒子物理實驗室(CERN，European Laboratory for Particle Physics)工作後在高能物理社群中發展出WWW的雛形。網路的發展，並不只是由技術人員主導，而是包含了許多領域的共同參與。

- **Mosaic (1993)**
  - **A hypertext GUI for the X-window system**
  - **CERN HTTPD: server of hypertext documents**

1994: the Landmark Year

- **Foundation of the “Mosaic Communications Corporation"**
- **First World-wide Web conference**

  http://www.iw3c2.org/conferences/
- **MIT and CERN agreed to set up the World-wide Web Consortium (W3C).**

Crawling and Indexing

- **Crawler/Spiders/Web robots/Bots**
- **Purpose of crawling and indexing**
  - **Quick fetching of large number of Web pages into a local repository**
  - **Indexing based on keywords**
  - **Ordering responses to maximize user's chances of the first few responses satisfying his information need.**
- **Earliest search engine: Lycos (Jan 1994)**
- **Followed by….**
  - **Alta Vista (1995), HotBot and Inktomi, Excite**

Topic Directories

- **Yahoo! directory**
  - **To locate useful Web sites**
  - **Jerry Yang & David Filo (Ph.D. students at Stanford University), 1994**
- **Efforts for organizing knowledge into ontologies**
  - **Centralized: (Yahoo!)**
  - **Decentralized: About.COM and the Open Directory**

Hyperlink Analysis

- **Take advantage of the structure of the Web graph.**
  - **Indicators of prestige of a page (e.g. citations)**
  - **HITS (Kleinberg 1998) & PageRank (Page 1998)**
- **Bibliometry**
  - **Bibliographic citation graph of academic papers**
- **Topic distillation**
  - **The process of finding quality documents on a query topic**
  - **Adapting to idioms of Web authorship and linking styles**

Paid Placement Ranking

- **Goto.com (→ Overture.com → Yahoo!)**
  - **Search ranking depended on how much you paid**
  - **Auction for keywords: *casino* was expensive!**
  - **Goto/Overture's annual revenues were nearing $1 billion**
- **Result: Google added paid-placement "ads" to the side, independent of search results**
  - **Yahoo follows suit, acquiring Overture (for paid placement) and Inktomi (for search)**

Web 2.0

- **Web 2.0**
  - **Human-centered, social value, participation & co-creation**
- **Explosive growth of multimedia data**
- **Huge incentives in search business**
- **Applications**

flickr

YouTube™ Broadcast Yourself

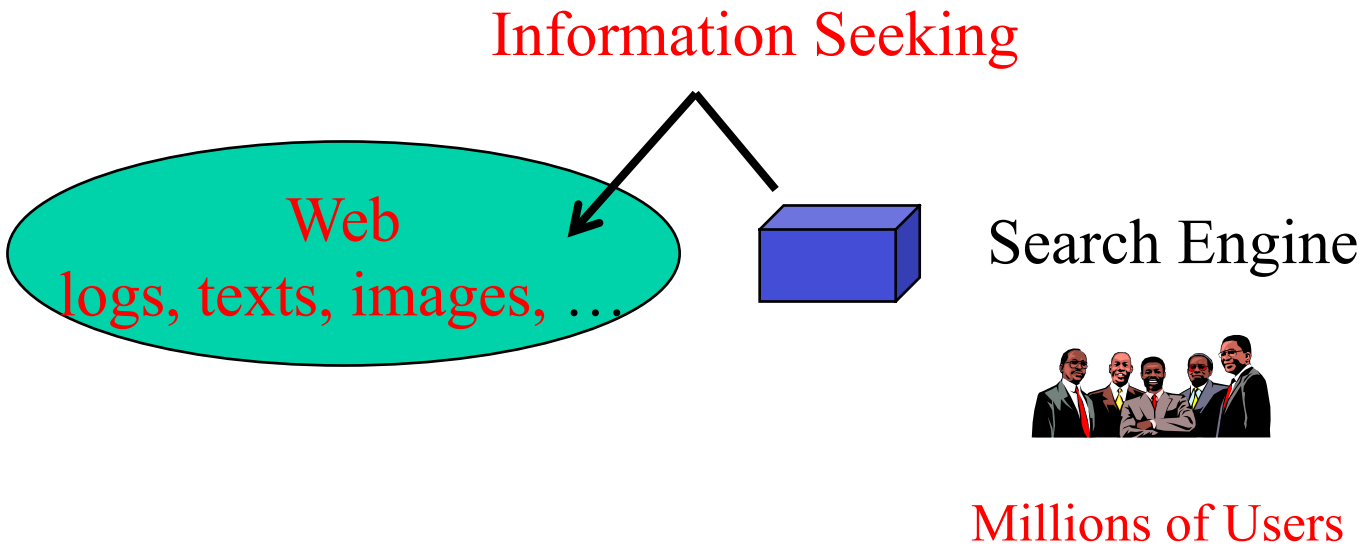
facebook

twitter

The Problem of

Information Overload

Web Retrieval/Search



Without Search Engines the Web Wouldn't Scale

- No incentive in creating content unless it can be easily found
- The web is both a technology artifact and a social environment
  - The Web has become the "new normal" in the American way of life
  - Those who don't go online constitute an ever-shrinking minority.

    [Pew Foundation report, January 2005]
- Search engines make aggregation of interest possible:
  - Create incentives for very specialized niche players
    - Economical – specialized stores, providers, etc
    - Social – narrow interests, specialized communities, etc
- The acceptance of search interaction makes "unlimited selection" stores possible
- Search turned out to be the best mechanism for advertising on the web

Information Retrieval (IR)

- **Information retrieval is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him. It is the finding or discovery process with respect to stored information.**

- Introduced by Calvin Mooers in 1951

Mooers, C. N. (1951). Zatocoding applied to mechanical organization of knowledge. *American Documentation, 2*, 20-32.

Information Retrieval (cont.)

- **Information retrieval (IR) deals with the representation, storage, organization of, and access to information items.**
- **Converting information need**

  to information items

  – **Information need**

    full description, keyword-based query

  – **Information item**

    text documents (often unstructured), Web pages (semi -structured), images, audios, videos, ….

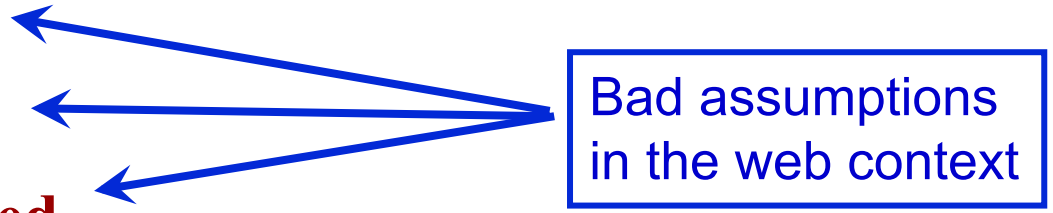
Classical IR vs. Web IR

Basic Assumptions of Classical IR

- **Corpus:** Fixed document collection
- **Goal:** Retrieve documents with information content that is relevant to user's information need

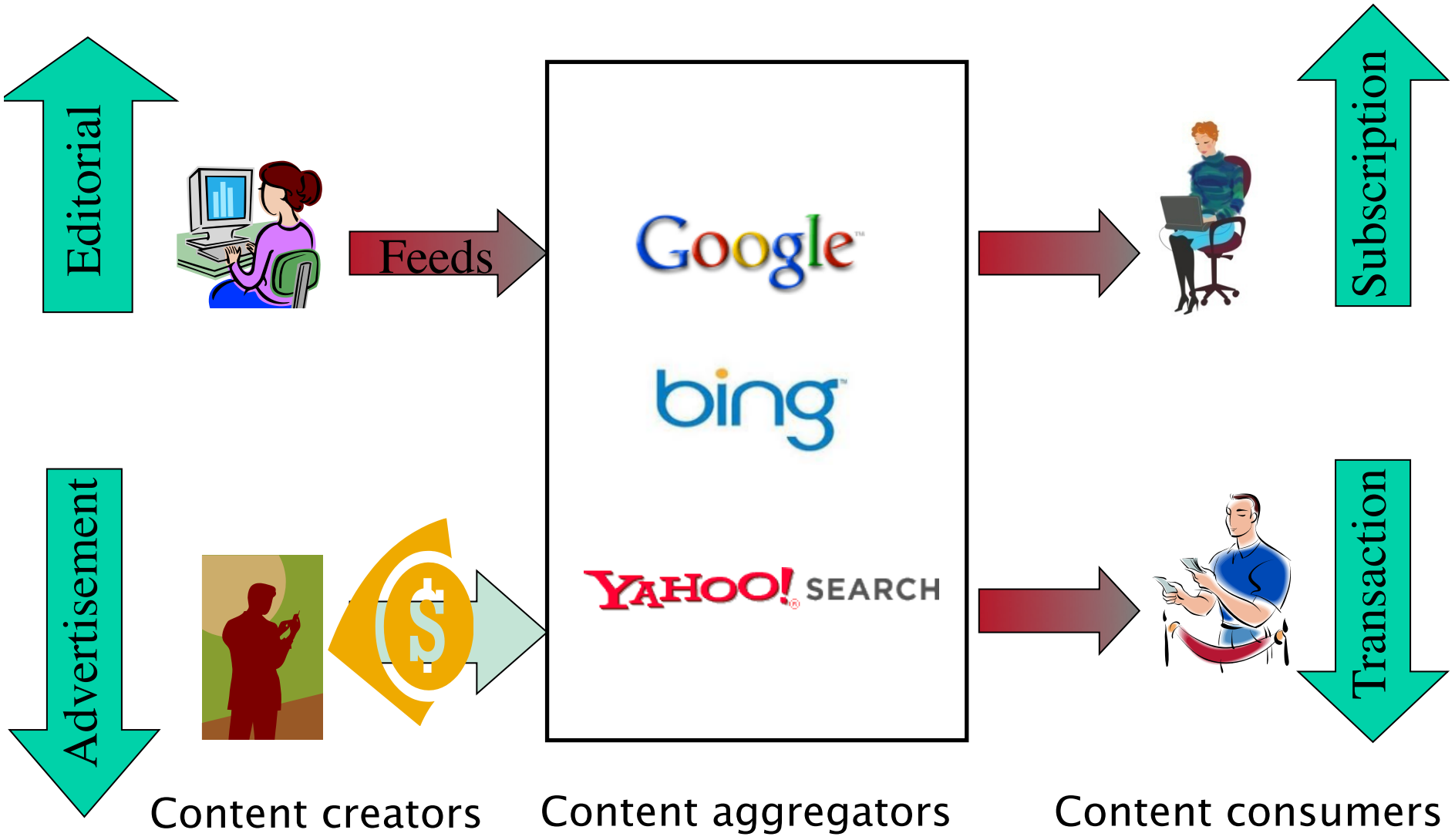
Classic IR Goal

- **Classic Relevance**
  - **For each query Q and stored document D in a given corpus assume there exists relevance Score(Q, D)**
    - **Score is average over users U and contexts C**
  - **Optimize Score(Q, D) as opposed to Score(Q, D, U, C)**
  - **That is, usually:**
    - **Context ignored**
    - **Individuals ignored**
    - **Corpus predetermined**



Web IR

The Coarse-level Dynamics



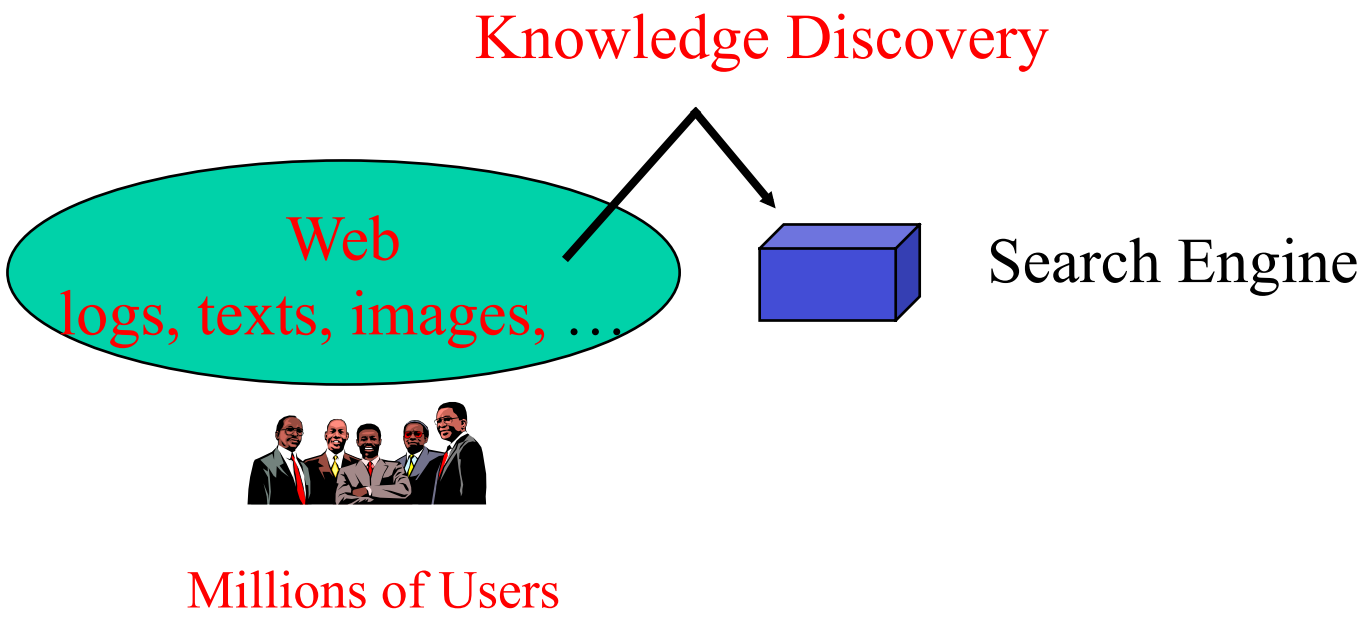
Web IR Goal

- **Topical Relevance vs. User Relevance**
  - **Optimize Score(Q, D, U, C) as opposed to Score(Q, D)**
  - **That is, usually:**
    - **Context: query session, clickthrough, location, time, …**
    - **Individuals: personalization, user-centered search, …**
    - **Corpus: dynamic, refresh rate, …**

The Niche of

Information Overload

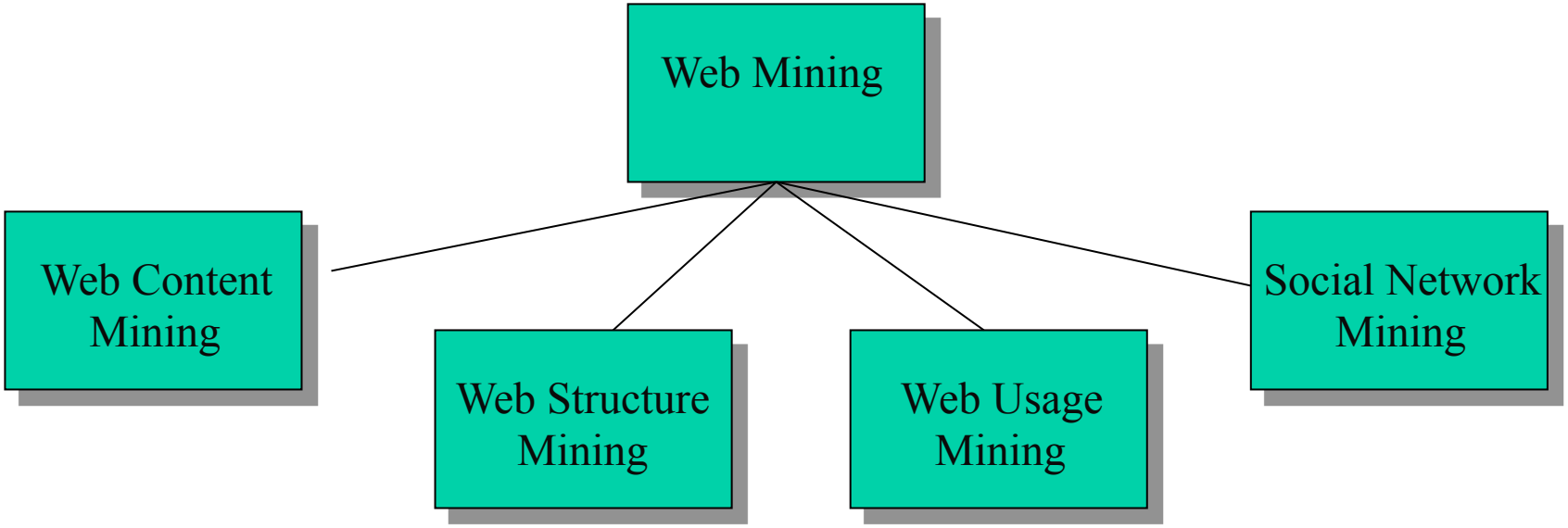
Web Mining


Knowledge Discovery
Web
logs, texts, images, …
Search Engine
Millions of Users

Web Mining [Srivastava'01]

- **Web Mining**
  - **Discovery of interesting patterns from Web content, structure and usage data.**
  - **A combination of WWW and Data Mining areas (Viewpoint of data mining)**
- **Typical Source of Data**
  - **Page/multimedia content**
  - **Intra-page and inter-page structure**
  - **Server access logs, registration information, demographics, past history, etc.**
- **Different Approaches**
  - **Database/Data Mining approach**
  - **Agent-based approach (or AI approach)**
  - **Information Retrieval/Web Search approach**
  - **Information Extraction/Natural Language Processing approach**

Taxonomy of Web Mining [R. Cooley]



- Web Content Mining (web page/text, search-result page, multimedia, tags, …)
- Web Usage Mining (query log analysis, user gap, community, …)
- Web Structure Mining (hyperlink, anchor text, web site, …)
- Social Network Mining (blog, wikipedia, email, instant messaging, …)

AltaVista前20大查詢語彙及比例

Document Query Logs

查詢語彙	查詢次數	比例
sex	1551477	0.27%
applet	1169031	0.20%
porno	712790	0.12%
mp3	613902	0.11%
chat	406014	0.07%
warez	398953	0.07%
yahoo	377025	0.07%
playboy	356556	0.06%
xxx	324923	0.06%
hotmail	321267	0.06%
[non-ASCII query]	263760	0.05%
pamela anderson	256559	0.04%
p****	234037	0.04%
sexo	226705	0.04%
porn	212161	0.04%
nude	190641	0.03%
lolita	179629	0.03%
games	166781	0.03%
spice girls	162272	0.03%
beastiality	152143	0.03%
註：總查詢次數：575,244,993		

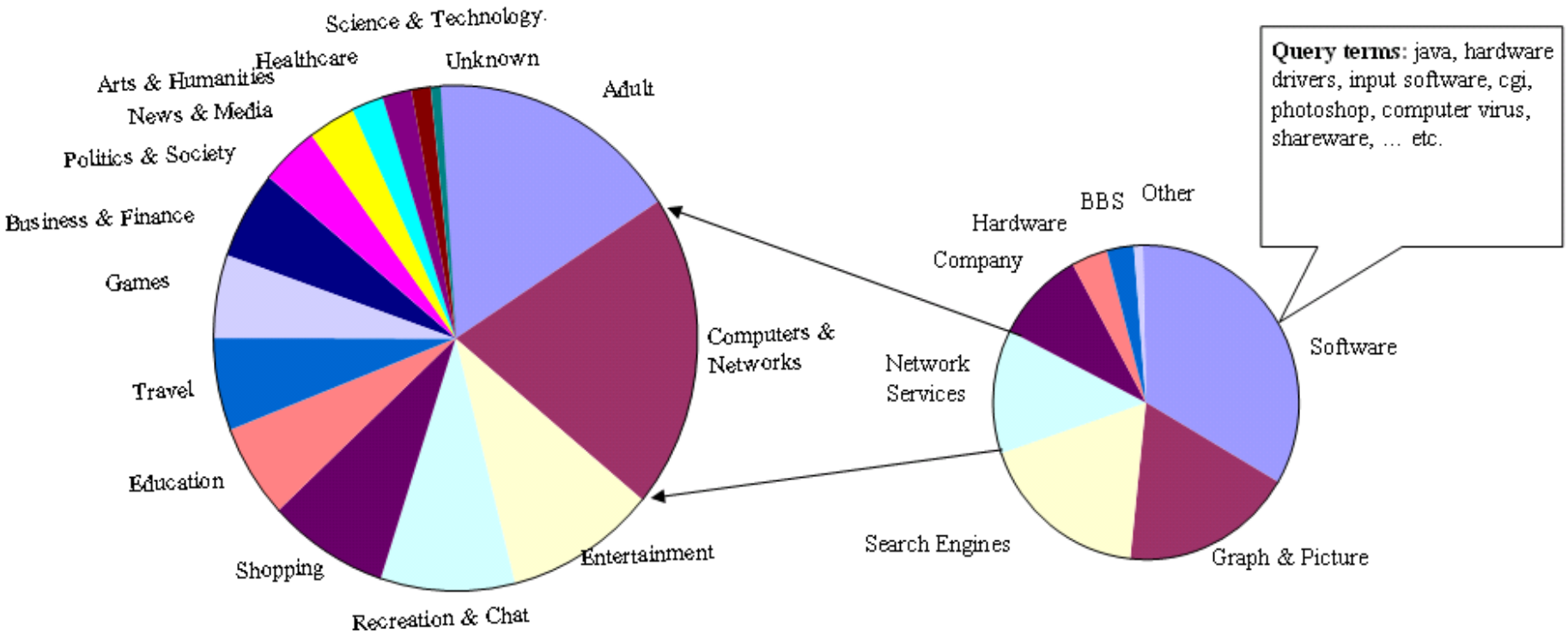
查詢語彙	查詢次數	比例
MP3	42561	1.95%
色情	24970	1.14%
情色	24363	1.12%
sex	20182	0.92%
模擬器	15071	0.69%
icq	13899	0.64%
同志	13622	0.62%
貼圖	12210	0.56%
桌面	12092	0.55%
桌面王	11680	0.53%
寫真集	11640	0.53%
蕃薯藤	10000	0.46%
情色文學	9817	0.45%
寫真	9530	0.44%
奇摩	9328	0.43%
bbs	8613	0.39%
kimo	8166	0.37%
104	7943	0.36%
小說	7456	0.34%
歌詞	7217	0.33%
註：總查詢次數：2,183,506		

Image Query Log

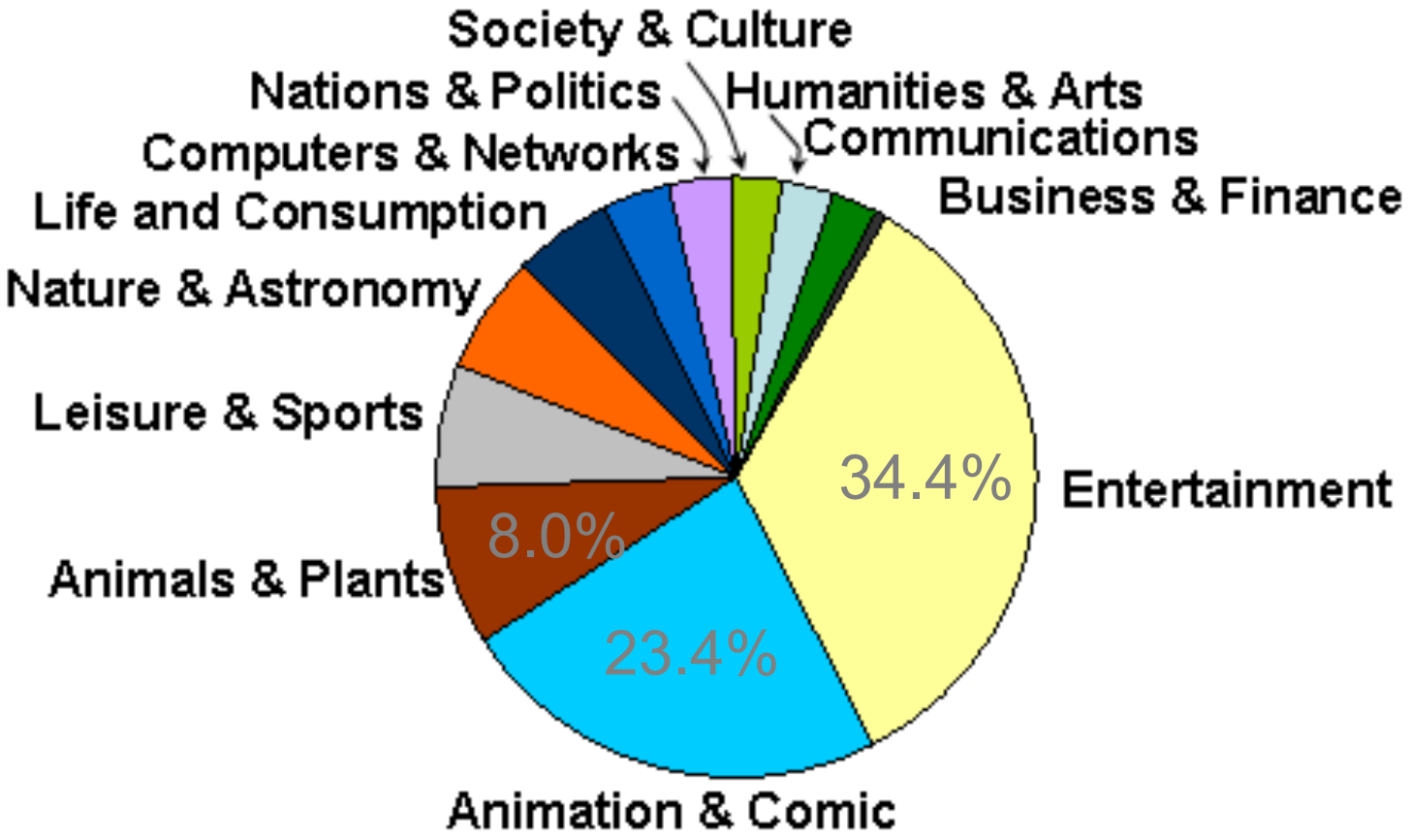
PCHome 2002/01~2002/03

一月熱門查詢詞彙		二月熱門查詢詞彙		三月熱門查詢詞彙	
Qoo	40943	pucca	8079	交大水果妹妹	23608
2002 車展	16179	Qoo	4953	網路美女	16662
美女	8571	大頭狗	4655	水果妹	15485
金城武	6627	鹽水蜂炮	4529	交大水果妹	9277
S.H.E	4577	可愛的圖片	4177	薰衣草	8234
賤兔	3967	神隱少女	2819	蔡依林	5966
孫燕姿	3453	關穎珊	2696	賤兔	5659
怪獸電力公司	3334	哈姆太郎	2531	背景	5409
流氓兔	3113	李英愛	1703	荷莉貝莉	5119
宋慧喬	2564	後藤希美子	1352	丁文琪	4854
背景底圖	2395	許紹洋	1246	背景底圖	4610
西瓜熊	1923	怪獸電力公司	1144	丹佐華盛頓	3568
璩美鳳	1859	櫻花	1135	水果妹妹	3489
BMW	1581	賤兔	1105	美麗境界	3309
S.H.E	1576	松島菜菜子	1096	許慧欣	3244
周杰倫	1263			孫燕姿	2874
深田恭子	1249			周杰倫	2613
天心	1171			哈姆太郎	2594
喬丹	1148			大頭狗	2136

Common Interests in Web Pages

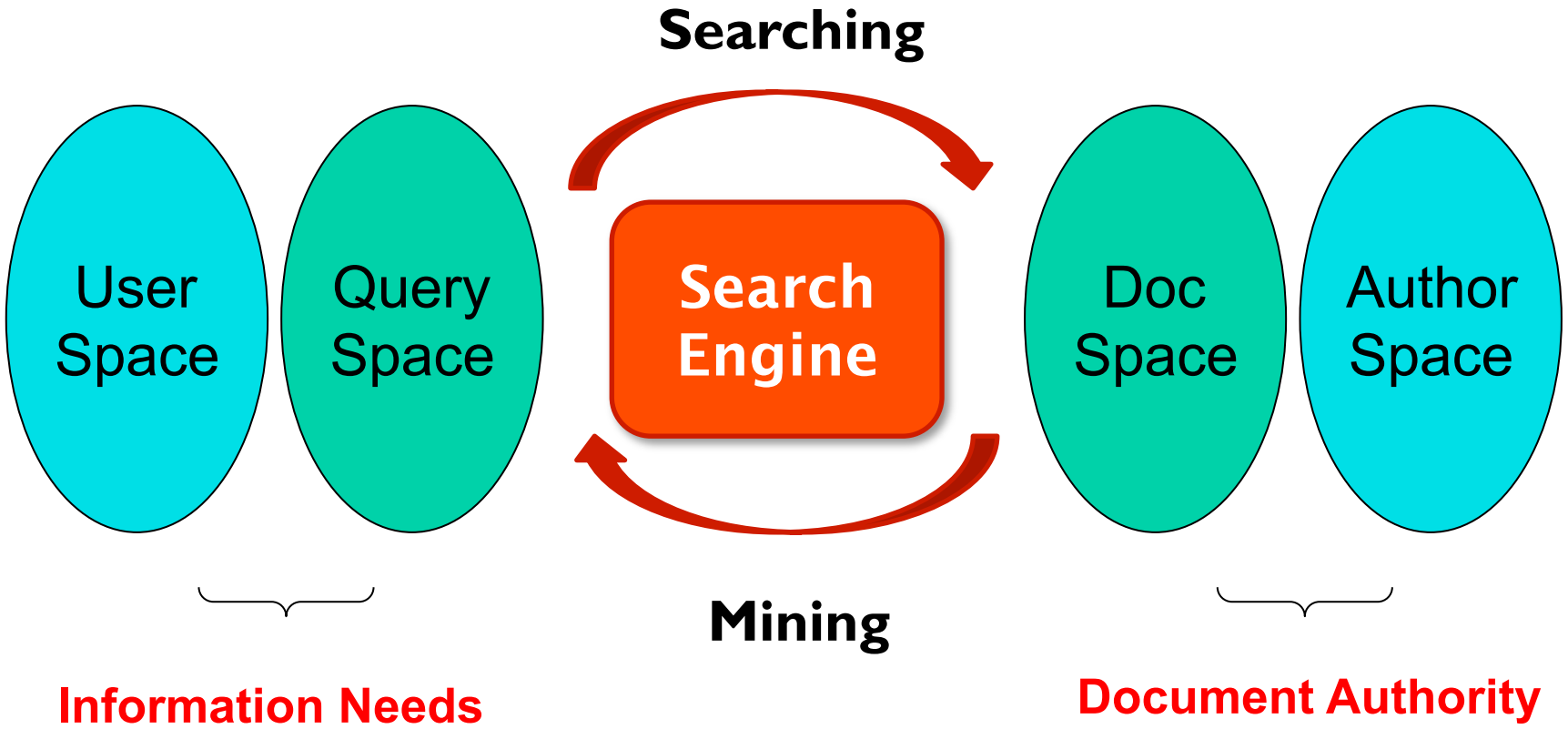


Common Interests in Web Images

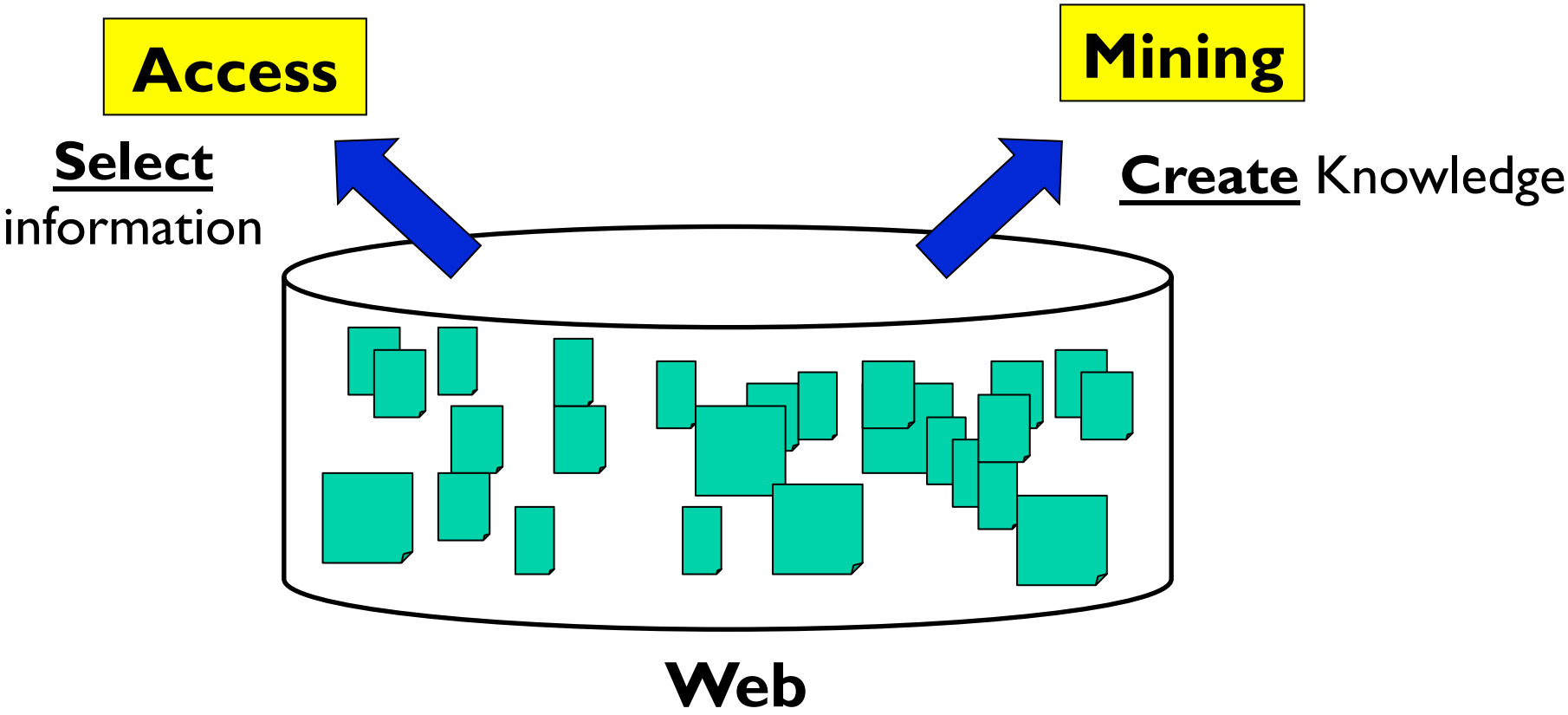


Course Summary

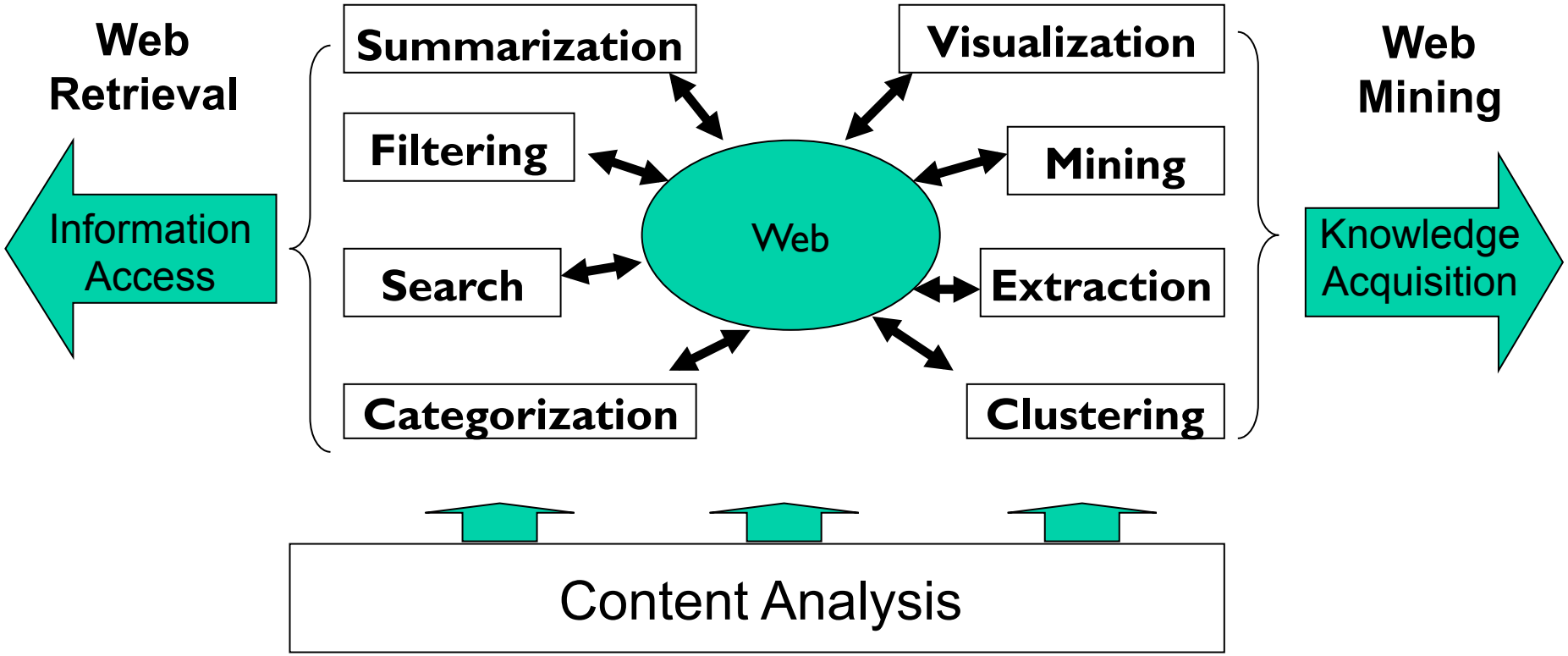
Web Mining & Information Retrieval



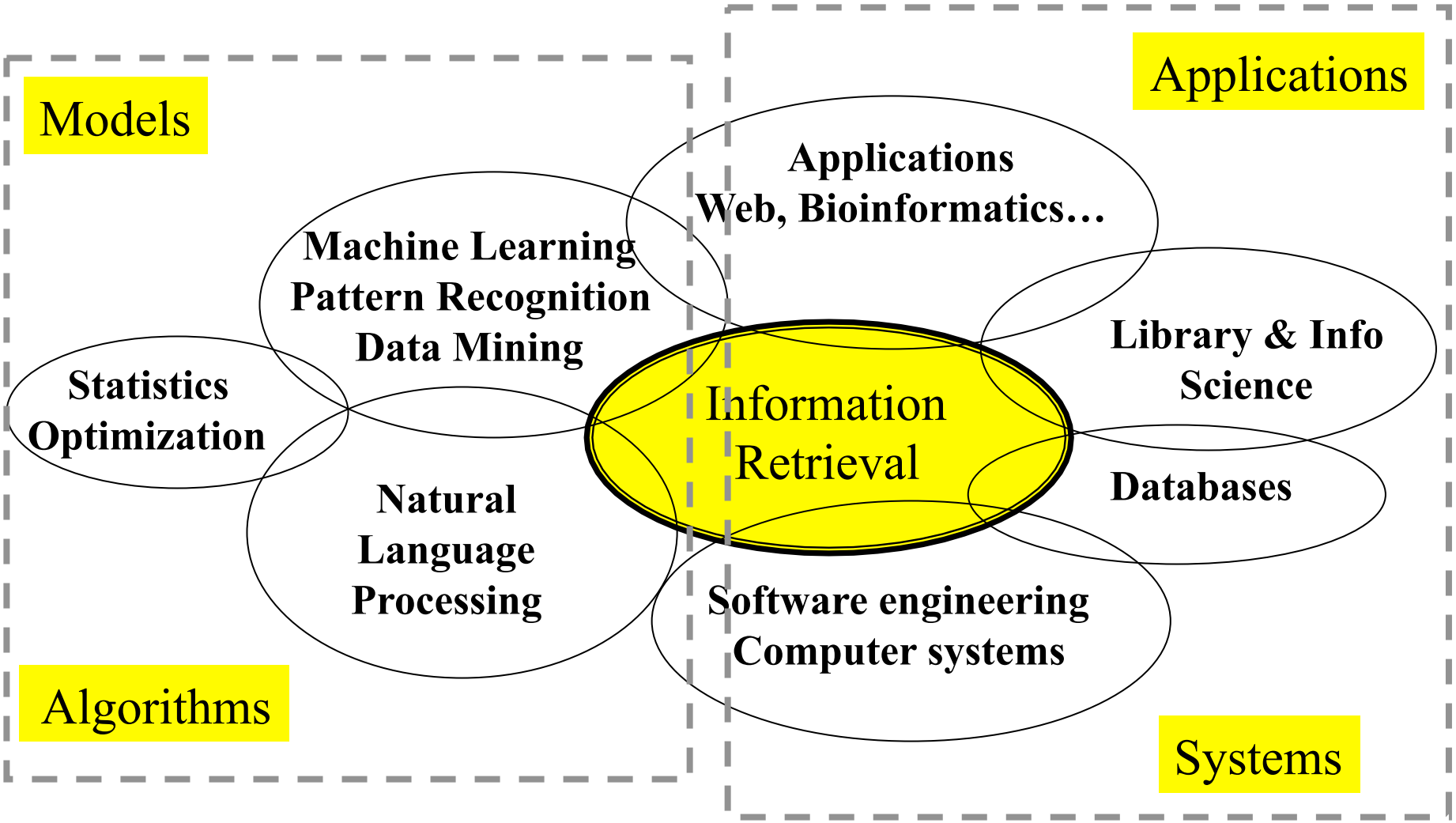
The Two Topics of this Course



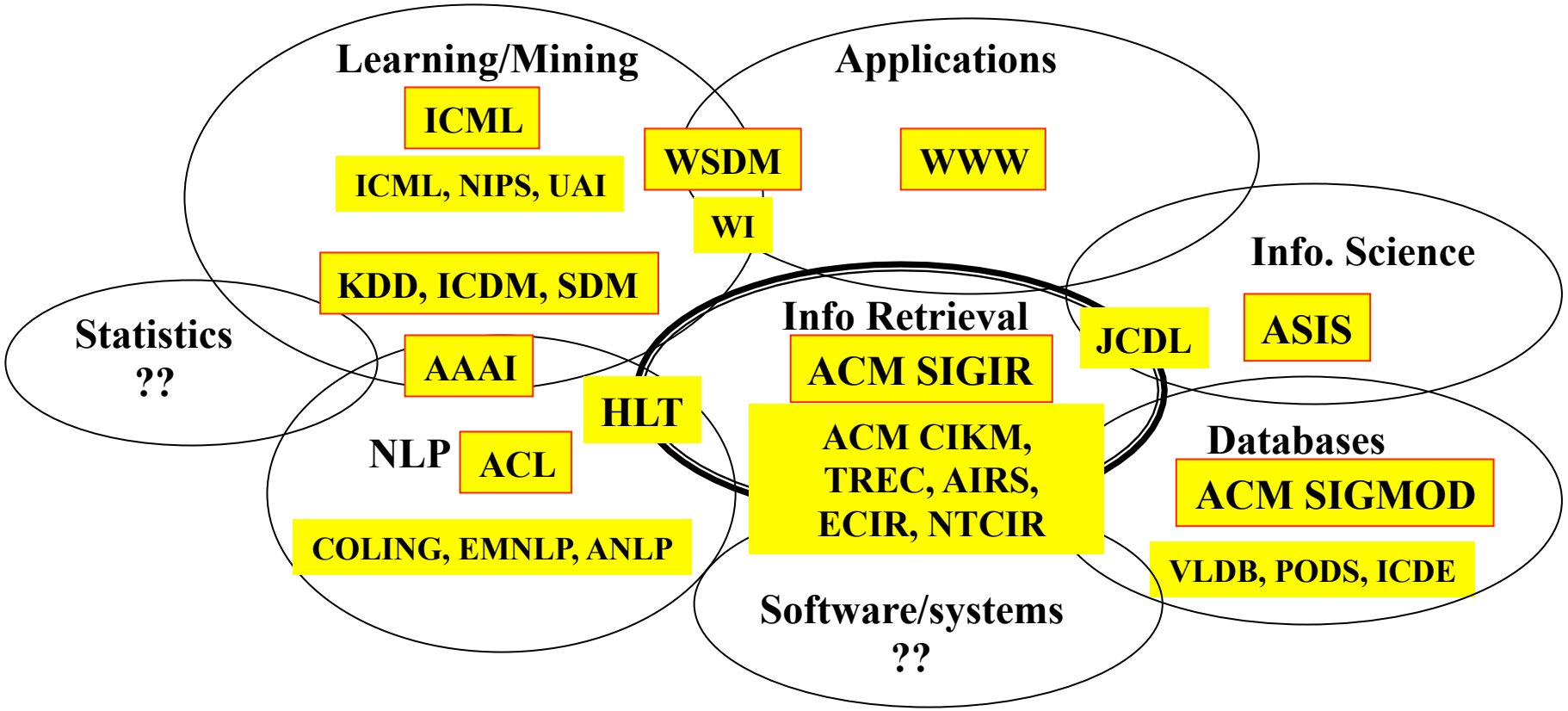
Relevant Technologies



Related Areas



What to Read?


Learning/Mining
ICML
ICML, NIPS, UAI
KDD, ICDM, SDM
AAAI
NLP
ACL
COLING, EMNLP, ANLP
Statistics
??
WSDM
WI
HLT
Applications
WWW
Info Retrieval
ACM SIGIR
ACM CIKM,
TREC, AIRS,
ECIR, NTCIR
JCDL
Info. Science
ASIS
Databases
ACM SIGMOD
VLDB, PODS, ICDE
Software/systems
??

922 U3640

Web Retrieval and Mining

(Spring 2013)

Goal & Design

- Introduce "Web Retrieval/Search" and "Web Mining"
- Prepare students for doing research/development in related fields
- Targeted at (senior) undergraduate students and graduate students with computer science background

Schedule

- **Part I: Web Information Retrieval**
  - Retrieval Model
  - User Interaction
  - Evaluation
  - Link Analysis
- **Part II: Web Mining**
  - Classification
  - Clustering
- **Part III: SIG Study (tentative)**
  - Multimedia/Multilingual Information Retrieval
  - User Behavior Analysis
  - NLP & ML for IR
  - Information Extraction & Filtering
  - Advertisement

Some Relevant NTU CISE Courses

- Information Retrieval
- Natural Language Processing
- Machine Learning
- Data Mining
- Social Network
- Statistical Artificial Intelligence

Format

- **Handwritten Assignments (individual work)**
- **2 Programming Assignments (individual work)**
  - **Programming + Report**
- **Midterm Exam**
- **Final Project**
  - **Team work (2~4 people, which depends on # of students)**
  - **Programming**
  - **Presentation & report**

    (including idea, literature review, method & experiment)

Grading

- **Assignments: 50%**
- **Midterm Exam: 20%**
- **Term Project: 30%**

Readings

- ***Introduction to Information Retrieval*****,** *by Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze*, (Selected Chapters) **Available online!**
- ***Modern Information Retrieval,*** *by Ricardo Baeza-Yates, Berthier Ribeiro -Neto.* (Selected Chapters)
- ***Search Engines: Information Retrieval in Practice*****,** *by W. Bruce Croft, Donald Metzler, Trevor Strohman***.** (Selected Chapters)
- ***Mining the Web: Discovering Knowledge from Hypertext Data*****,** *by Soumen Chakrabarti, Morgan Kaufmann.* (Selected Chapters)
- ***Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*****,** *by Bing Liu, Springer, 2006***.** (Selected Chapters) **Available online!**
- **Additional readings will be available online**

Questions?

http://www.csie.ntu.edu.tw/~pjcheng/course/wm2013

Office hours:

Pu-Jen Cheng: Tuesday 9:00-12:00am, R323

Good Luck!