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 (European Organization for Nuclear Research)**
  - **By Tim Berners-Lee (1989)**

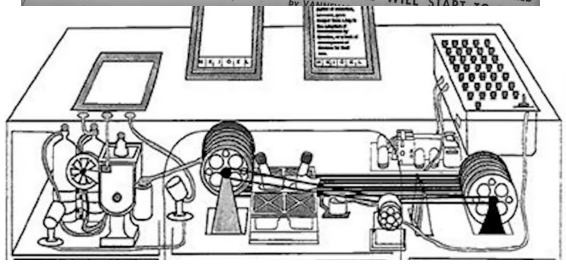
Tim Berners-Lee

1945: Vannevar Bush

As We May Think

1965: Ted Nelson

'hypertext' : non-sequential writing



- **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 her 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!**
- **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

What's going on?

- More searches from mobile devices
- More data from users
- More integrated services for users
- From passive to active services
- More semantics from learning
- More AI-driven generation

Google

information retrieval

全部 圖片 影片 書籍 購物 新聞 網頁 更多 工具

维基百科
https://zh.wikipedia.org › zh-tw › 信息檢索

資訊檢索- 維基百科，自由的百科全書

資訊檢索（英語：Information Retrieval）是從資訊資源集合獲得與資訊需求相關的資訊資源的活動。搜尋可以基於全文或其他基於內容的索引。 自動資訊檢索系統用於減少所謂 ...

HackMD
https://hackmd.io › ...

資訊檢索導論(Introduction to Information Retrieval) Autumn ...

2024年4月19日 — 資訊檢索導論(Introduction to Information Retrieval) Autumn 2021 · 成績計算 · Overview · Boolean Queries · Phrase Queries · Tolerant Retrieval.

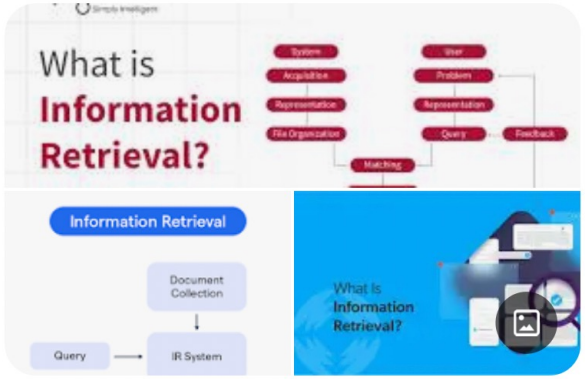
國立臺灣大學資訊工程學系
https://www.csie.ntu.edu.tw › intro_irlab PDF

網路探勘與資訊檢索實驗室

實驗室簡介: 網路探勘與資訊檢索實驗室是一個充滿創新與合作精神的研究環境，致力於探. 索網路資料的潛力以及資訊檢索的最新技術。「網路探勘」探討如何從大規模網路.

2 頁

資訊檢索



資訊檢索是從資訊資源集合獲得與資訊需求相關的資訊資源的活動。搜尋可以基於全文或其他基於內容的索引。自動資訊檢索系統用於減少所謂的「資訊超載」。許多大學和公共圖書館使用IR系統... More

Source: 維基百科 ›

Wikipedia
https://en.wikipedia.org › wiki › Information_retrieval

Information retrieval - Wikipedia

Information retrieval (IR) in computing and information science is the task of identifying and retrieving information system resources that are relevant to an information need. The...

預估閱讀時間: 10 分鐘

其他人也問了以下問題

What is information retrieval (IR)?

Information Retrieval is believed to be the **dominant form of Information access**. The IR system assists the users in finding the information they require but it does...

What is Information Retrieval? - ...
geeksforgeeks.org

What is the evaluation of an information retrieval system?

The evaluation of an information retrieval system' is **the process of assessing how well a system meets the information needs of its users**. In general, measurement...

Information retrieval - Wikipedia
en.wikipedia.org

Where inform

This is t
the follc
Mann
Hinrich
Informa

Introdu
nlp.s

意見反應

维基百科
https://zh.wikipedia.org › wiki › 信息檢索

信息檢索 - 维基百科，自由的百科全书

資訊檢索（英語： Information Retrieval ）是从信息资源集合获得与信息需求相关的信息资源的活动。 搜索可以基于 全文 或其他基于内容的索引。 自动信息检索系统用于减少所谓的“ 資訊超載...

預估閱讀時間: 3 分鐘

在資訊檢索中，查詢不會唯一地標識集合中的單個對象。 相反可以有不止一個對象匹配查詢，它們可能具有不同程度的相關性。

資訊檢索基本上是一個**資訊檢索系統** (**Information Retrieval System**)的最重要功能，IR是指因應使用者對資訊的需求提供查尋的方法與查尋過程而言。

資訊檢索 (Information Retrieval,IR)這門學問過去主要應用在**圖書館**的**資料檢索系統**中。 隨著網際網路的蓬勃發展，人們在網路上透過搜尋引擎進行文本或是多媒體的搜索需求也與日俱增。

探索其他


資訊


自然語言處理

資訊系統


資訊學


網路搜尋引擎

資料: 維基百科 · zh.wikipedia.org · pedia.cloud.edu.tw · ithelp.ithome.com.tw
源自維基百科 CC-BY-SA 授權

意見反應

相關搜尋

資訊系統與資訊檢索

现在

information retrieval

Information Retrieval

What is Information Retrieval? A...
zilliz

What is Information Retrieval? - Elastic
elastic

What is information retrieval? - IBM
ibm

+5 来源

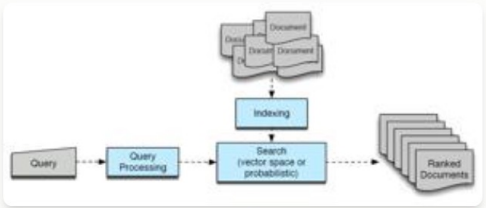
Perplexity

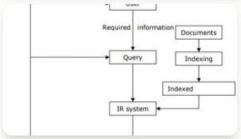
Information retrieval (IR) is a crucial field in computer science and information science focused on the efficient extraction of relevant information from large collections of unstructured or semi-structured data. This process is vital for various applications, including web search engines, digital libraries, and information management systems.

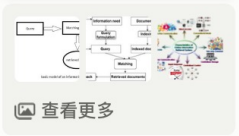
Definition and Importance

IR involves searching for, locating, and presenting information that matches a user's query or information need. It operates on the principle that user queries are processed to identify relevant terms, which are then used to rank documents based on their relevance. The significance of IR can be summarized by the following points:

- **Efficient Information Access:** Enables quick retrieval of relevant data from vast datasets.
- **Personalization:** Adapts search results based on user preferences and previous interactions.
- **Scalability:** Capable of managing large volumes of diverse content, including text and







搜索视频

PRO 生成图片

From World-Wide Web to Now

- **1989 Initiated at CERN**
- **1990 Machine Learning (Data-driven)**
- **1993 Mosaic (First GUI-based Brower)**

  Data Mining

Tim Berners-Lee

- **1994 Mosaic Communications Corporation (Commercialized)**

  W3C, WWW (Standard)

  Lycos Search Engine

  Yahoo! Directory

- **1997 Deep Blue (Computer > Human: Chess (10^{123}))**
- **1998 Web Structure Analysis (PageRank)**

  Pay-for-placement (Goto.com)

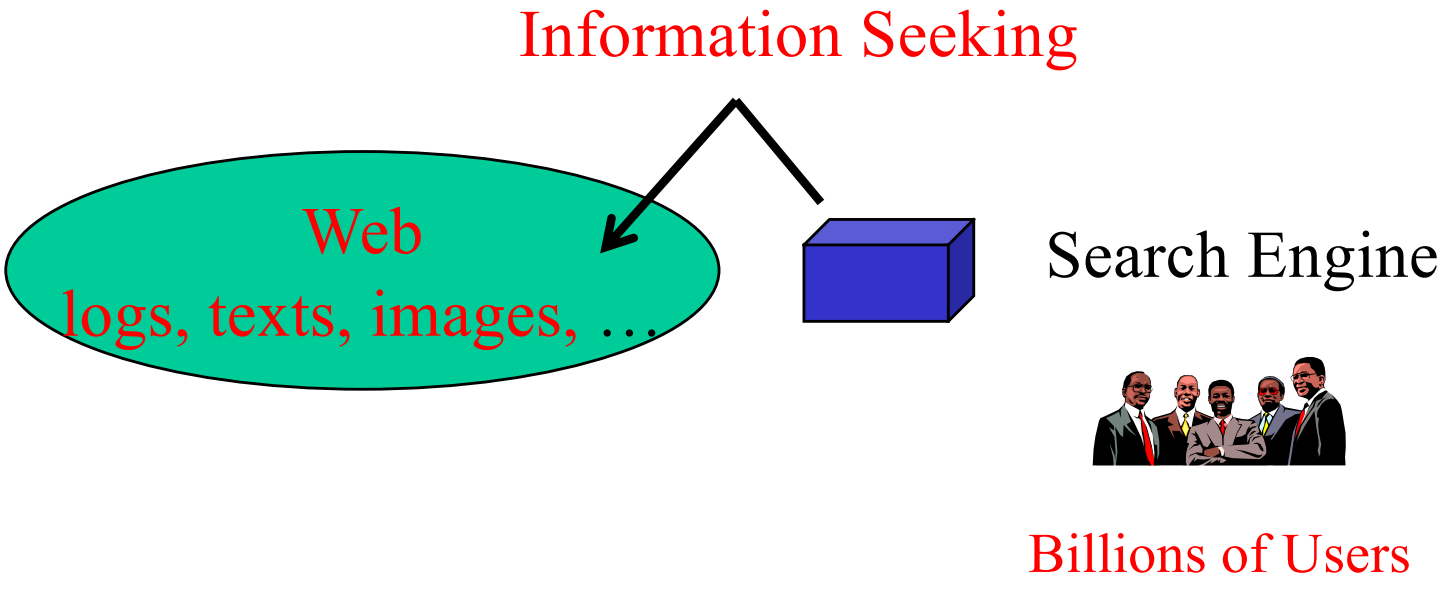
From World-Wide Web to Now (Cont.)

- 2001 Google (Getting noticed)
- 2004 Web 2.0 (Human-centered)
  Web as Platform (Services)
  Facebook
- 2008 Mobile Apps (Apple app store, Android market)
  Fintech (Bitcon)
- 2012 Deep Learning (Big Data+GPU)
- 2016 AlphaGo (Computer > Human: Go (10^{360}))
- 2017 Transformer (Self-Attention)
- 2020 GPT-3 (Text Generation)
- 2023 ChatGPT

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
- Web is both a technology artifact and a social environment
  - The Web has become the "new normal" in the way of life
  - Those who don't go online constitute an ever-shrinking minority
- 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

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

  to information items

  – **Information need**

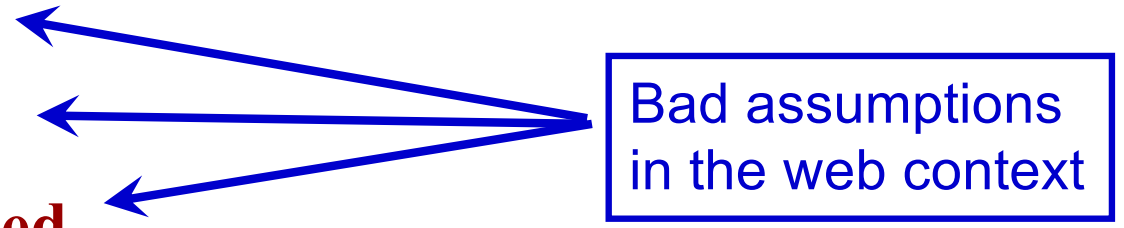
    full description, keyword-based query

  – **Information item**

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

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**



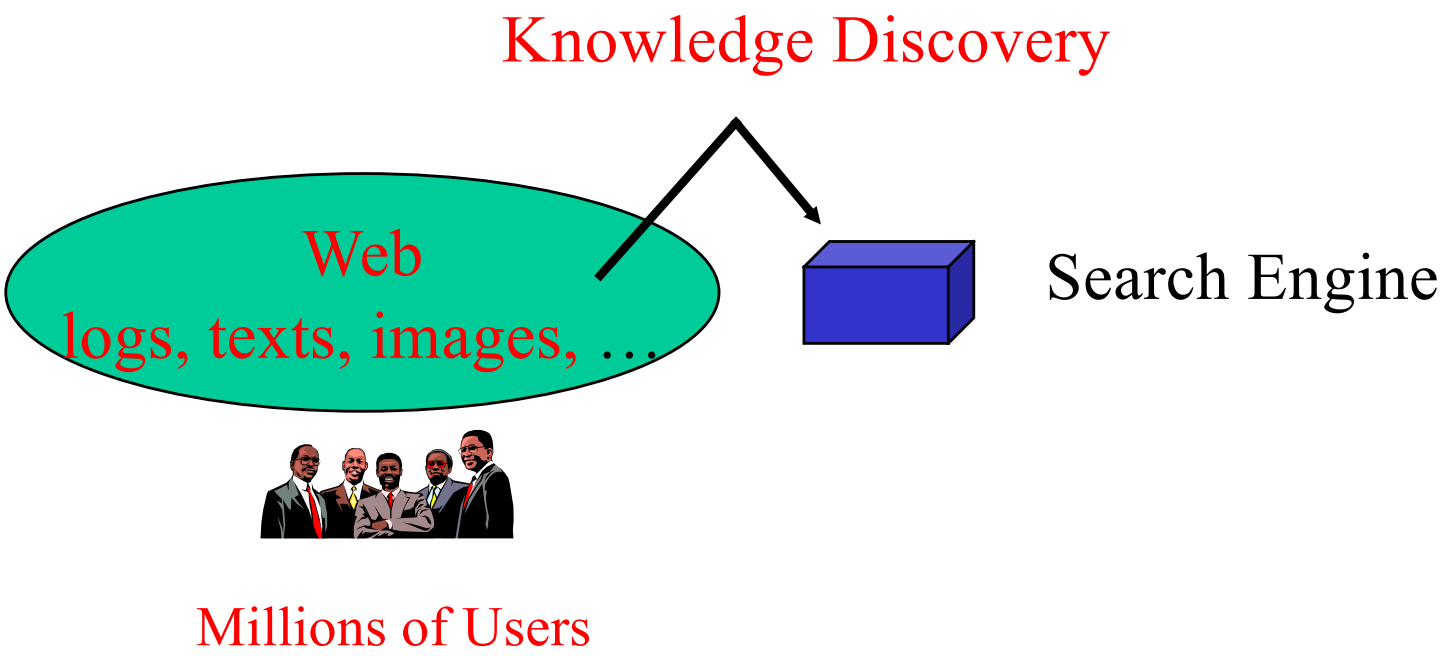
Web IR Goal

- **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, …**

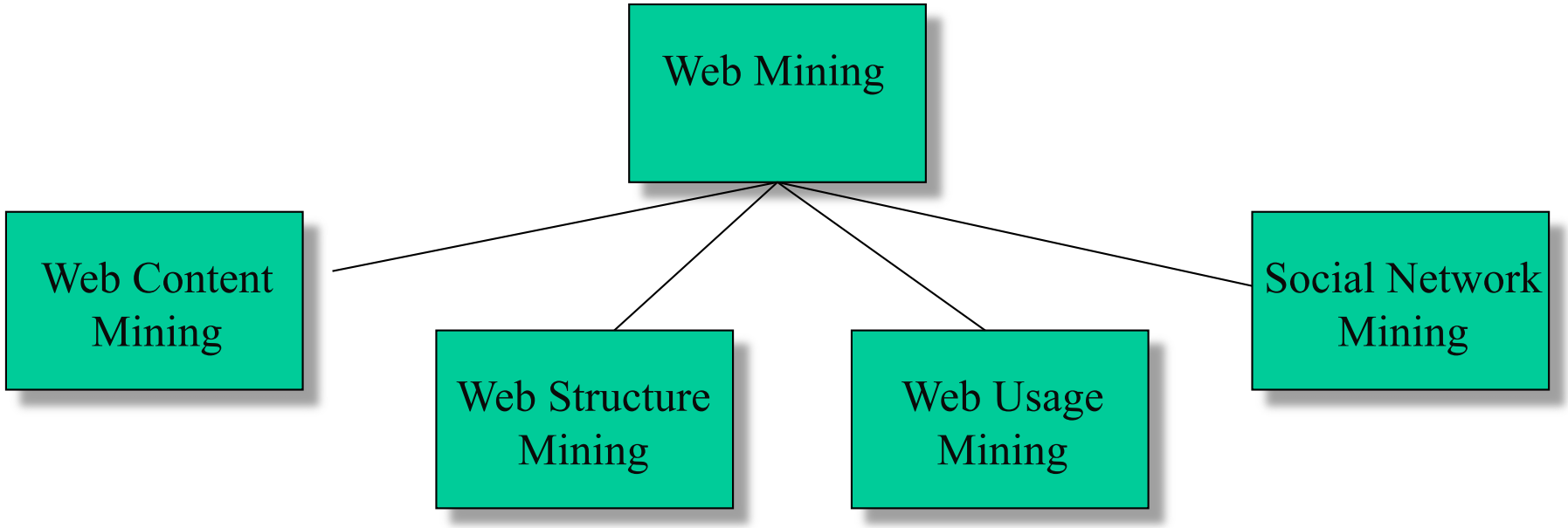
The Niche of

Information Overload

Web Mining


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

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		

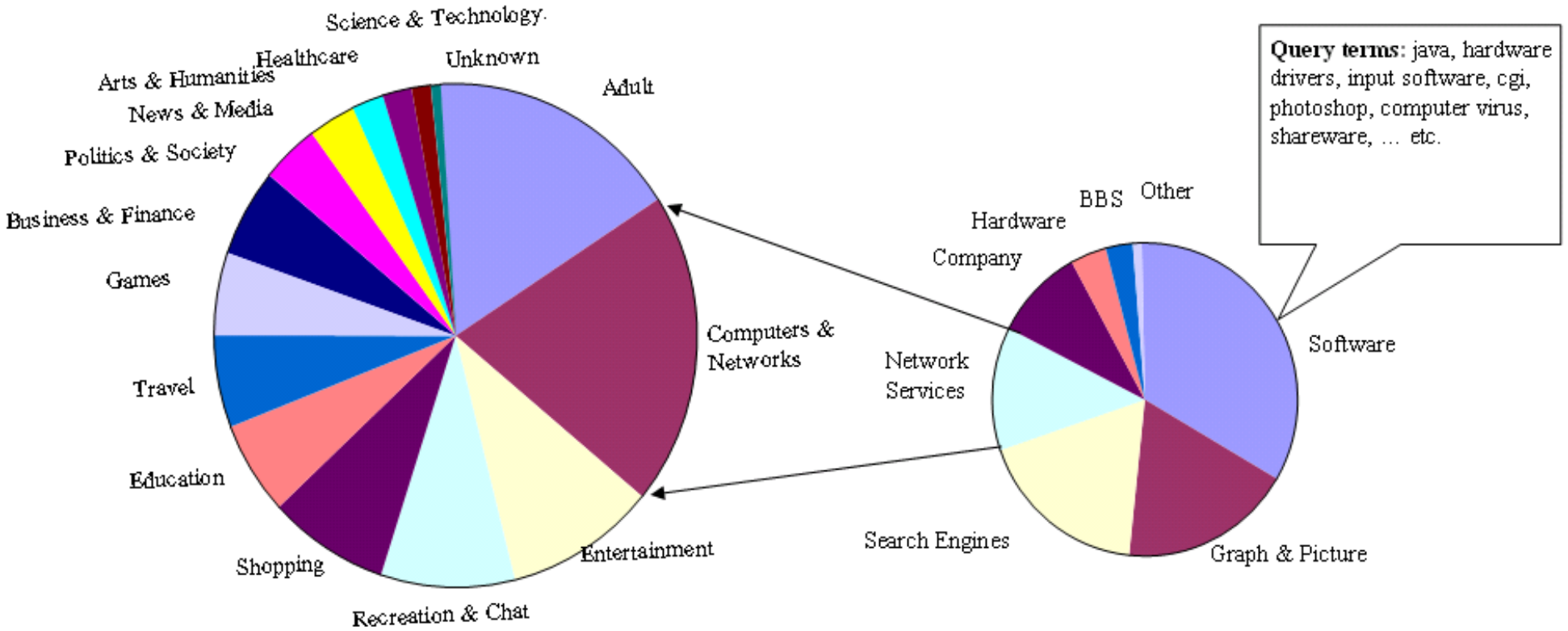
查詢語彙	查詢次數	比例
MP3	42561	1.95%
色情	24970	1.14%
情色	24363	1.12%
sex	20182	0.92%
模擬器	15071	0.69%
icq	13899	0.64%
同志	13622	0.62%
貼圖	12210	0.56%
桌面	12092	0.55%
桌面王	11680	0.53%
寫真集	11640	0.53%
蕃薯藤	10000	0.46%
情色文學	9817	0.45%
寫真	9530	0.44%
奇摩	9328	0.43%
bbs	8613	0.39%
kimo	8166	0.37%
104	7943	0.36%
小說	7456	0.34%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

Image Query Log

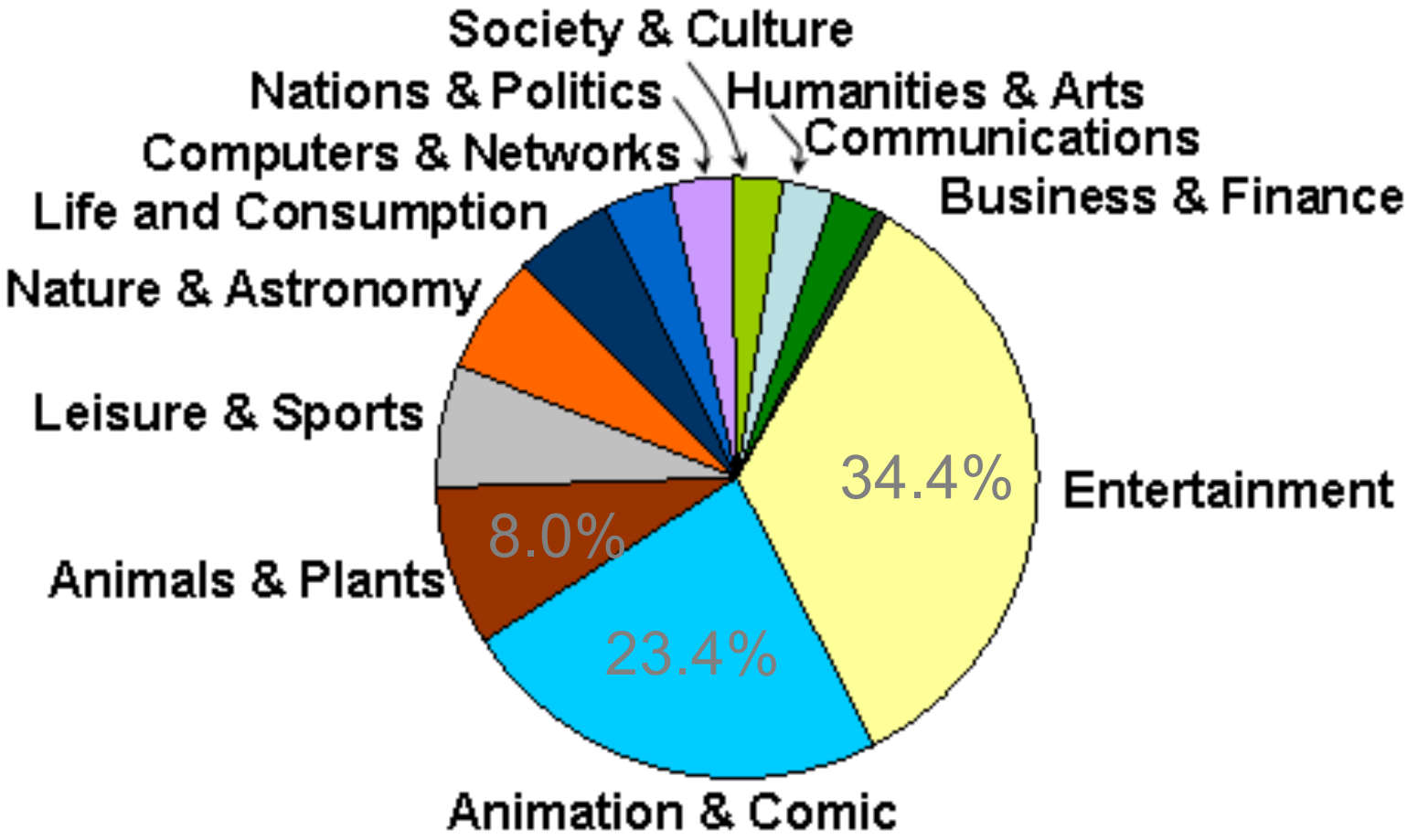
PCHome 2002/01~2002/03

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

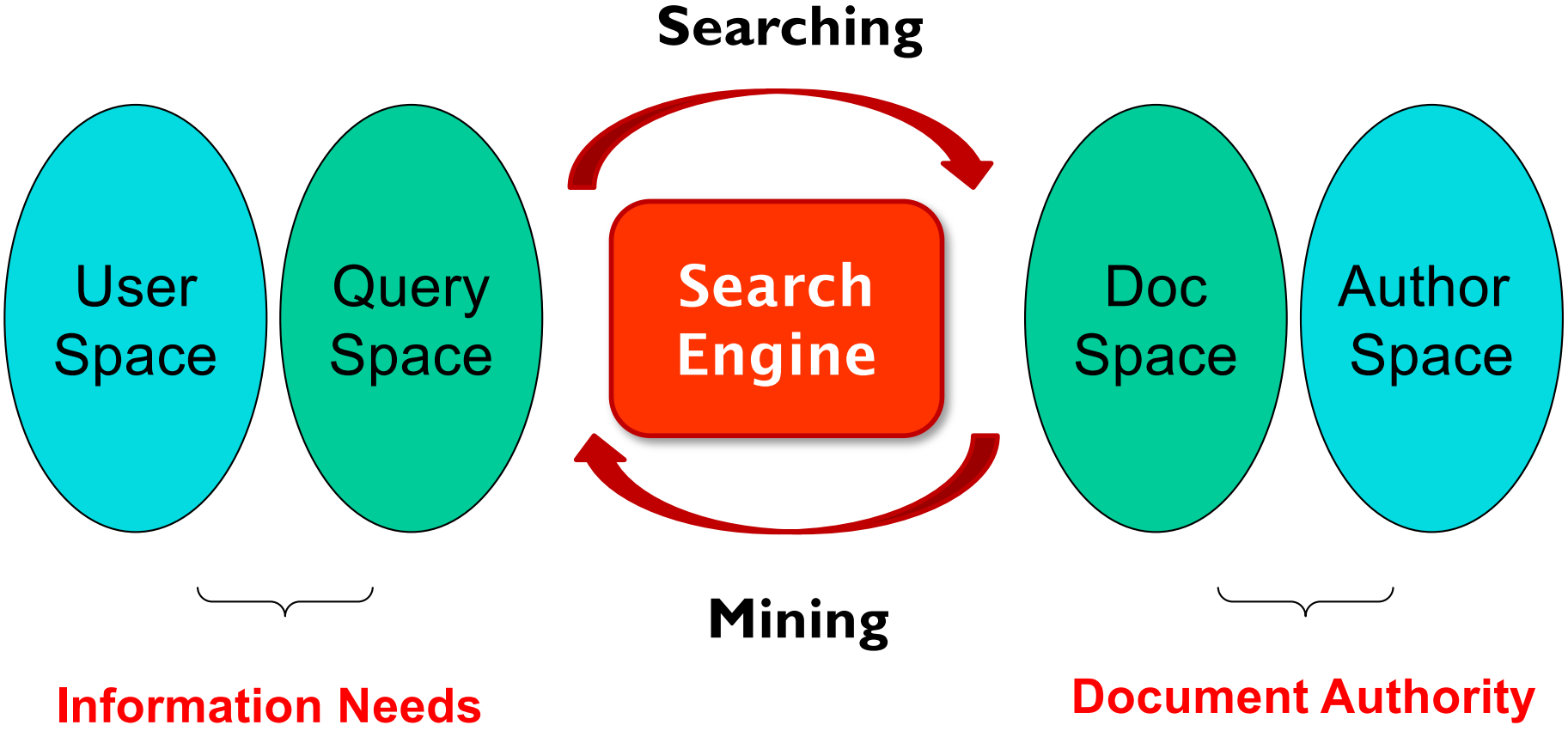


Common Interests in Web Images



Course Summary

Web Mining & Information Retrieval



Web Mining & Retrieval

Info. Retrieval

Information Access

- Generative IR
- Multi-modal IR
- RAG / QA
- Controllable NLG
- Recommendation

Multimedia Data

- Modality Alignment
- Info. Extraction
- Explanation
- Visualization
- Graph Learning

Web Mining

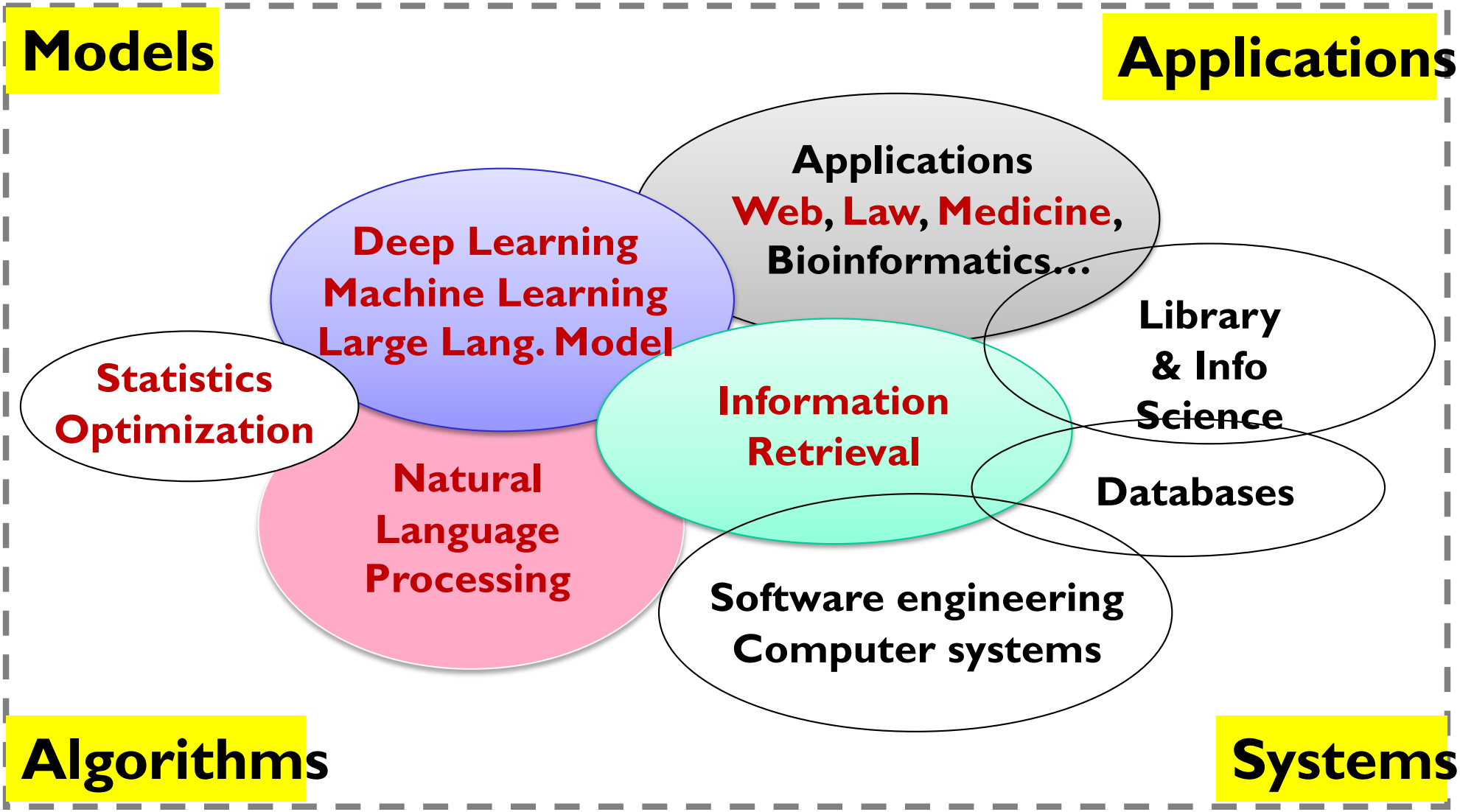
Knowledge Acquisition

Large Language Model

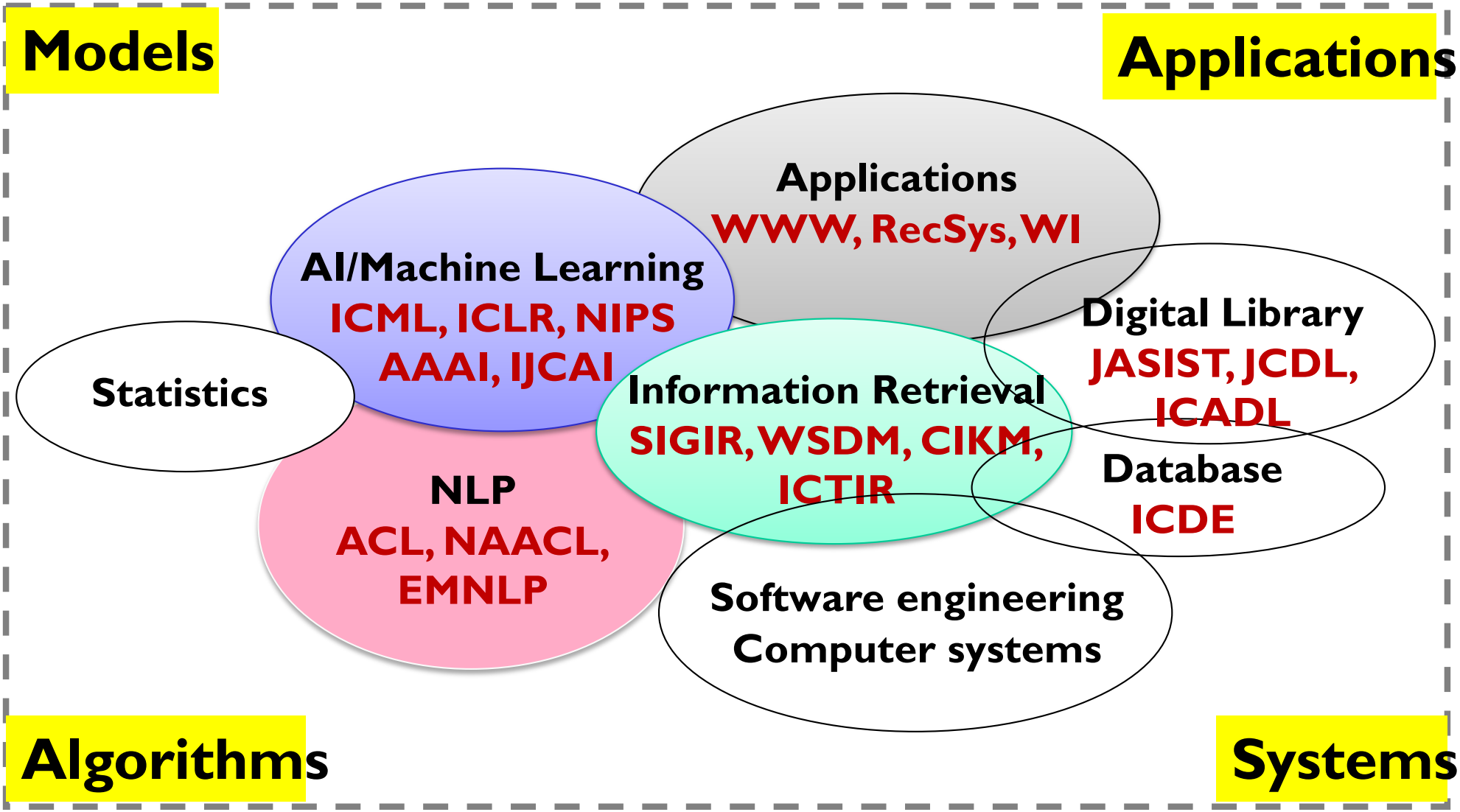
IR: Information Retrieval
NLG: Natural Language Generation

RAG: Retrieval-Augmented Generation
QA: Question Answering

Related Areas



Major Conferences



922 U3640

Web Retrieval and Mining

(Spring 2025)

Goal & Design

- Introduce "Web 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: Information Retrieval**
  - Indexing & Query Optimization
  - Retrieval Model & User Interaction
  - Evaluation
  - Link Analysis
  - Machine Learning for IR
  - Deep Learning for IR
- **Part II: SIG Study**
  - Recommendation
  - Opinion Mining, Sentiment Analysis (tentative)
  - Information Extraction & Filtering (tentative)

Some Relevant NTU CISE Courses

- **Information Retrieval**
- **Natural Language Processing**
- **Machine / Deep 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 (3~4 people, which depends on # of students)**
  - **Programming**
  - **Presentation & 4-pages Report**

    (including idea, literature review, method & experiment)

Grading

- **Assignments: 50% (hand-written, programming)**
- **Midterm Exam: 20%**
- **Term Project: 30%**

Readings

- ***Introduction to Information Retrieval*****,** *by Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze*, (Selected Chapters) **Available online!**
- ***Information retrieval : Implementing and Evaluating Search Engines***, by Stefan Büttcher, Charles L.A. Clarke, Gordon V. Cormack. (Selected Chapters)
- ***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?

Related Websites:

http://www.csie.ntu.edu.tw/~pjcheng/course/wm2025

NTU COOL (*https://cool.ntu.edu.tw/courses/44637*)

Office hours:

Tuesday 9:00am-11:00am, R220

Good Luck!