UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

TRUY VẤN THÔNG TIN ĐA PHƯƠNG TIỆN
INFORMATION RETRIEVAL

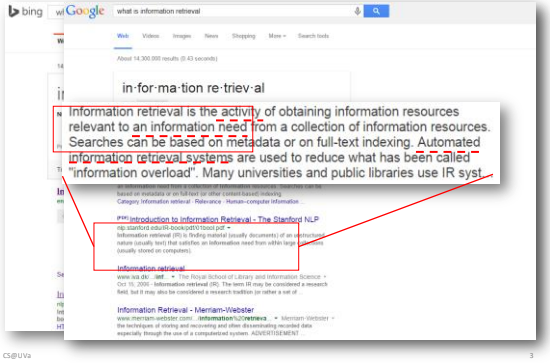
GIỚI THIỆU INFORMATION RETRIEVAL

1

Nội dung

1. IR là gì ?
2. Tại sao cần IR ?
3. Lịch sử IR
4. Bên trong một hệ thống tìm kiếm và truy vấn.
5. Một số lĩnh vực trong tìm kiếm và truy vấn.

Information Retrieval – IR là gì?



CS@UVa 3

Tại sao cần information retrieval ?

Bùng nổ dữ liệu

- *"It refers to the difficulty a person can have understanding an issue and making decisions that can be caused by the presence of too much information."* - wiki

CS@UVa 4

Tại sao cần Information Retrieval ?

- **Bùng nổ dữ liệu – Big Data**

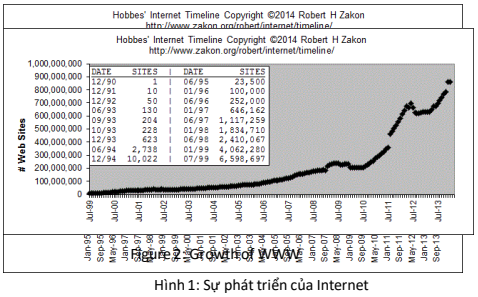
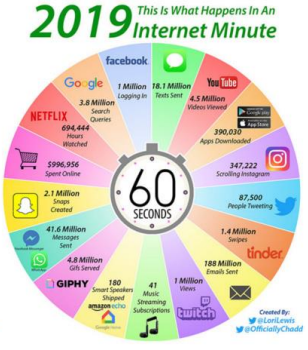


Figure 2 - Growth of WWW

Hình 1: Sự phát triển của Internet

5

6

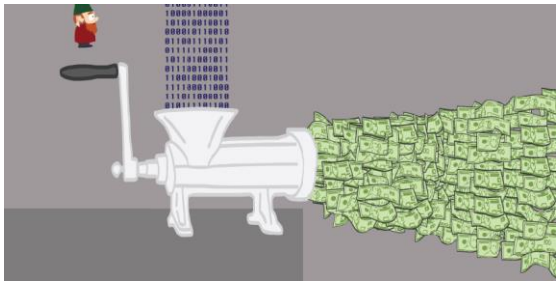


7

Trong **chương trình** cái gì quan **trọng nhất** ?

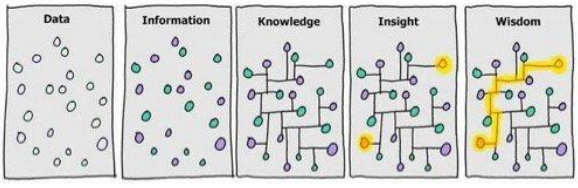
8

Data is money ?

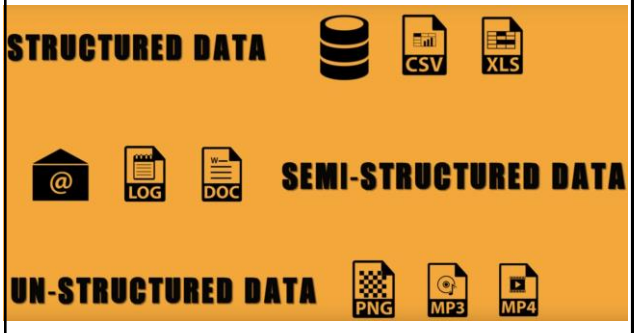


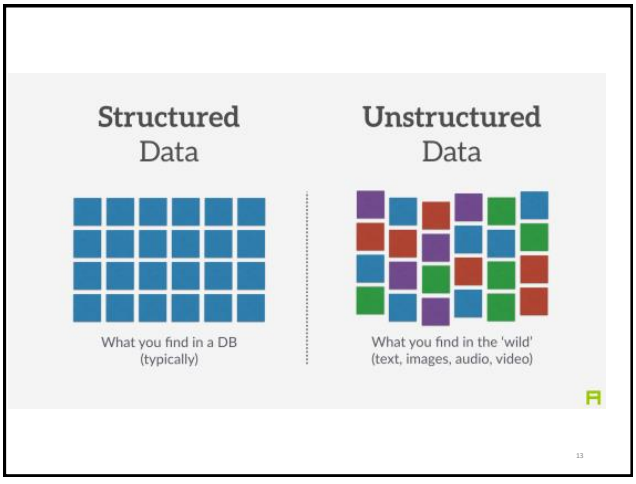
Data vs information ?

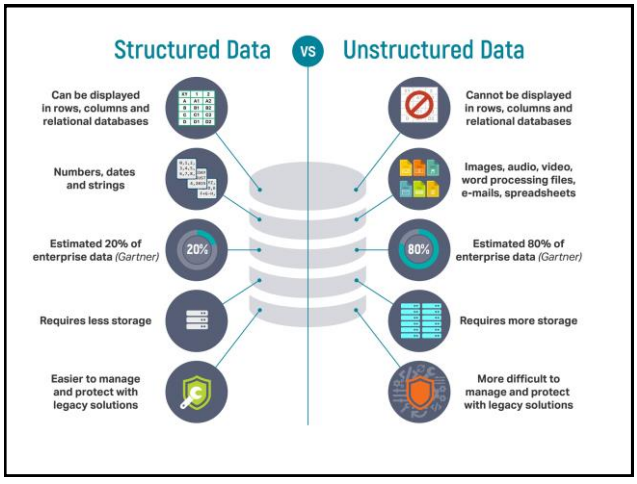
Data vs information ?



Có những loại data nào ?







Tại sao cần information retrieval

- Quản lý dữ liệu không cấu trúc unstructured data
  - Dữ liệu có ít
  - Unstructu
    - Text in
    - "85 per
    - Chỉ là

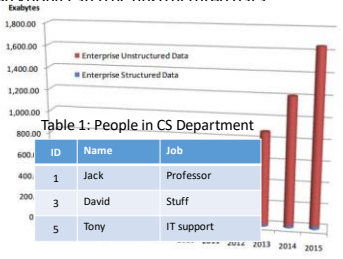


Table 1: People in CS Department

ID	Name	Job
1	Jack	Professor
3	David	Stuff
5	Tony	IT support

Total Enterprise Data Growth 2005-2015, IDC 2012

Tại sao cần information retrieval

An essential tool to deal with information overload

You are here!

Lịch sử information retrieval

- Idea popularized in the pioneer article "***As We May Think***" by Vannevar Bush, 1945
  - "*Wholly new forms of encyclopedias will appear, ready-made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified.*" **-> WWW**
  - "*A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility.*" **-> Search engine**

Lịch sử information retrieval

- Early days (late 1950s to 1960s): foundation of the field
  - Luhn's work on automatic indexing
  - Cleverdon's Cranfield evaluation methodology and index experiments
  - Salton's early work on SMART system and experiments
- 1970s-1980s: a large number of retrieval models
  - Vector space model
  - Probabilistic models
- 1990s: further development of retrieval models and new tasks
  - Language models
  - TREC evaluation
  - Web search
- 2000s-present: more applications, especially Web search and interactions with other fields
  - Learning to rank
  - Scalability (e.g., MapReduce)
  - Real-time search

Lịch sử information retrieval

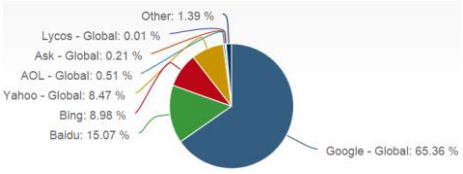
- Academia: Text Retrieval Conference (TREC) in 1992
  - "*Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.*"
  - "*… about one-third of the improvement in web search engines from 1999 to 2009 is attributable to TREC. Those enhancements likely saved up to 3 billion hours of time using web search engines.*"
  - Till today, it is still a major test-bed for academic research in IR

Lịch sử information retrieval

- Industry: web search engines
  - WWW unleashed explosion of published information and drove the innovation of IR techniques
  - First web search engine: "*Oscar Nierstrasz at the University of Geneva wrote a series of Perl scripts that periodically mirrored these pages and rewrote them into a standard format.*" Sept 2, 1993
  - Lycos (started at CMU) was launched and became a major commercial endeavor in 1994
  - Booming of search engine industry: *Magellan*, *Excite*, *Infoseek*, *Inktomi*, *Northern Light*, *AltaVista*, *Yahoo!*, *Google*, and *Bing*

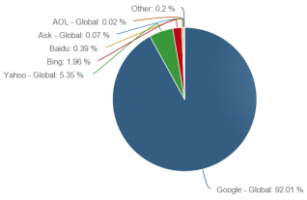
Thị phần trong lĩnh vực IR

- Global search engine market - desktop
  - By http://marketshare.hitslink.com/search-engine-market-share.aspx



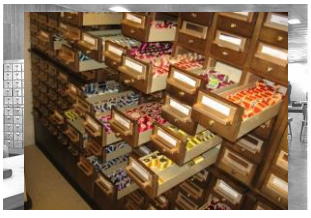
Thị phần trong lĩnh vực IR

- Global search engine market - mobile
  - By http://marketshare.hitslink.com/search-engine-market-share.aspx

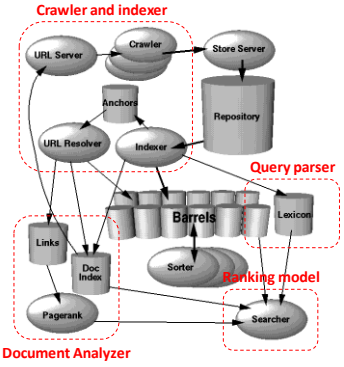


Kiến trúc hệ thống IR và Search

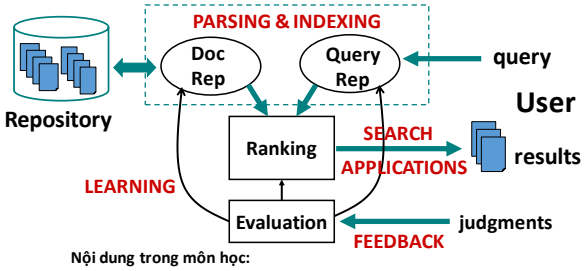
Thế giới thực khi chưa có IR & Search



Kiến trúc hệ thống IR và Search



Kiến trúc hệ thống IR và Search



Nội dung trong môn học:

1) Search engine architecture; 2)Retrieval models;
3) Retrieval evaluation; 4) Relevance feedback;
5) Link analysis; 6) Search applications.

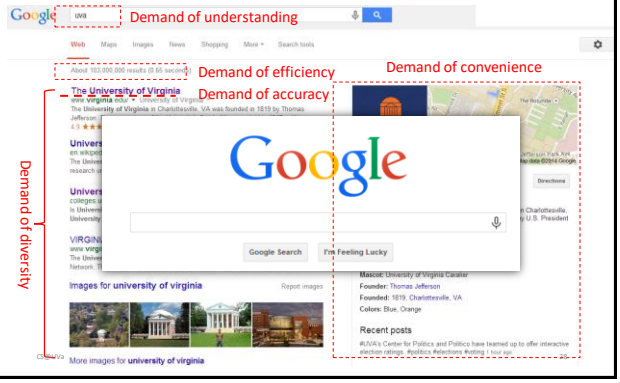
Một số điểm quan trọng trong IR

- **Biểu diễn câu truy vấn** -Query representation
  - Lexical gap: say v.s. said
  - Semantic gap: ranking model v.s. retrieval method
- **Biểu diễn dữ liệu** - Document representation
  - Specific data structure for efficient access
  - Lexical gap and semantic gap
- **Mô hình truy vấn** - Retrieval model
  - Algorithms that find the ***most relevant*** documents for the given information need

Một số search engine

*Yet Another **Hierarchical** Officious/Obstreperous/ Odiferous/Organized **Oracle***

Một số search engine

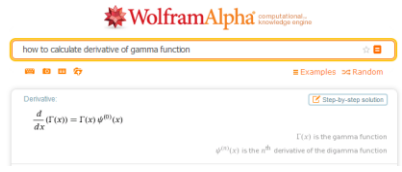


Lĩnh vực của IR

- Web search là một trong những mảng quan trong của information retrieval, tuy nhiên Information retrieval còn bao gồm:
  - Hệ thống khuyến nghị -Recommendation

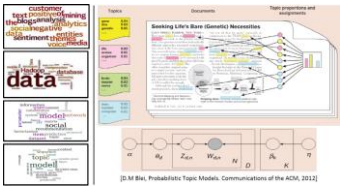
Lĩnh vực của IR

- Web search là một trong những mảng quan trong của information retrieval, tuy nhiên Information retrieval còn bao gồm:
- Hệ thống hỏi đáp - Question answering



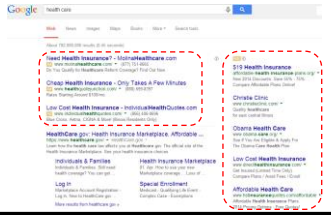
Lĩnh vực của IR

- Web search là một trong những mảng quan trong của information retrieval, tuy nhiên Information retrieval còn bao gồm:
- Khai mỏ dữ liệu - Text mining



Lĩnh vực của IR

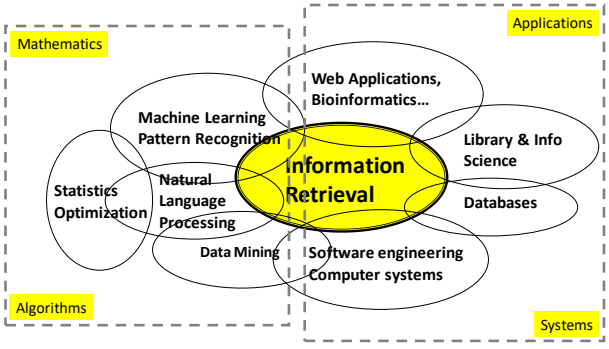
- Web search là một trong những mảng quan trong của information retrieval, tuy nhiên Information retrieval còn bao gồm:
- Quảng cáo trực tuyến- Online advertising



Lĩnh vực của IR

- Web search là một trong những mảng quan trong của information retrieval, tuy nhiên Information retrieval còn bao gồm:
- Hệ thống tìm kiếm - Enterprise search: web search + desktop search

Lĩnh vực IR



IR v.s. DBs

- Information Retrieval:
  - Unstructured data
  - Semantics of objects are subjective
  - Simple keyword queries
  - Relevance-drive retrieval
  - Effectiveness is primary issue, though efficiency is also important
- Database Systems:
  - Structured data
  - Semantics of each object are well defined
  - Structured query languages (e.g., SQL)
  - Exact retrieval
  - Emphasis on efficiency

IR and DBs are getting closer

- IR => DBs
  - Approximate search is available in DBs
  - Eg. in mySQL

```
mysql> SELECT * FROM articles
    -> WHERE MATCH (title,body)
       AGAINST ('database');
```

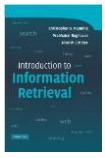
- DBs => IR
  - Use information extraction to convert unstructured data to structured data
  - Semi-structured representation: XML data; queries with structured information

IR v.s. NLP

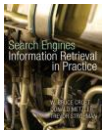
- Information retrieval
  - Computational approaches
  - Statistical (shallow) understanding of language
  - Handle large scale problems
- Natural language processing
  - Cognitive, symbolic and computational approaches
  - Semantic (deep) understanding of language
  - (often times) small scale problems

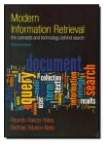
IR and NLP are getting closer

- IR => NLP
  - Larger data collections
  - Scalable/robust NLP techniques, e.g., translation models
- NLP => IR
  - Deep analysis of text documents and queries
  - Information extraction for structured IR tasks

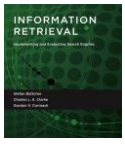


- ***Introduction to Information Retrieval***. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.
- ***Search Engines: Information Retrieval in Practice***. Bruce Croft, Donald Metzler, and Trevor Strohman, Pearson Education, 2009.

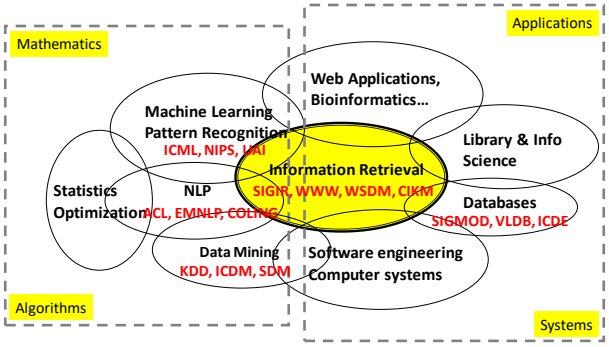




- ***Modern Information Retrieval.*** Ricardo Baeza-Yates and Berthier Ribeiro-Neto, Addison-Wesley, 2011.
- ***Information Retrieval:*** Implementing and Evaluating Search Engines. Stefan Buttcher, Charlie Clarke, Gordon Cormack, MIT Press, 2010.



What to read?



- Find more on course website for resource

IR in future

- Mobile search
  - Desktop search + location? Not exactly!!
- Interactive retrieval
  - Machine collaborates with human for information access
- Personal assistant
  - Proactive information retrieval
  - Knowledge navigator
- And many more
  - You name it!

Tài liệu tham khảo

Slide được tham khảo từ:

- http://www.cs.virginia.edu/~hw5x/Course/IR2015/_site/lectures/
- https://nlp.stanford.edu/IR-book/newslides.html
- https://course.ccs.neu.edu/cs6200s14/slides.html

Một số chủ đề Seminar

- Tìm hiểu **lucene** và minh họa.
- Tìm hiểu **Elasticsearch** và minh họa
- Tìm hiểu **Apache SOLR** và minh họa.
- Tìm hiểu **IRF framework** và minh họa.
- Tìm hiểu **Faiss** và minh họa.
- Tìm hiểu Apache Cassandravà minh họa.
- Tìm hiểu apche Hadoop và minh họa trong tìm kiếm, truy vấn.
- Tìm hiểu về **Scrapy Framework** và ví dụ minh họa cho một số loại dữ liệu khác nhau : Ảnh, Video, Text….
- Mô hình BOW và minh họa
- Mô hình BOF và minh họa

45