Search Engine Architecture

Hongning Wang
CS@UVa

Recap: why information retrieval

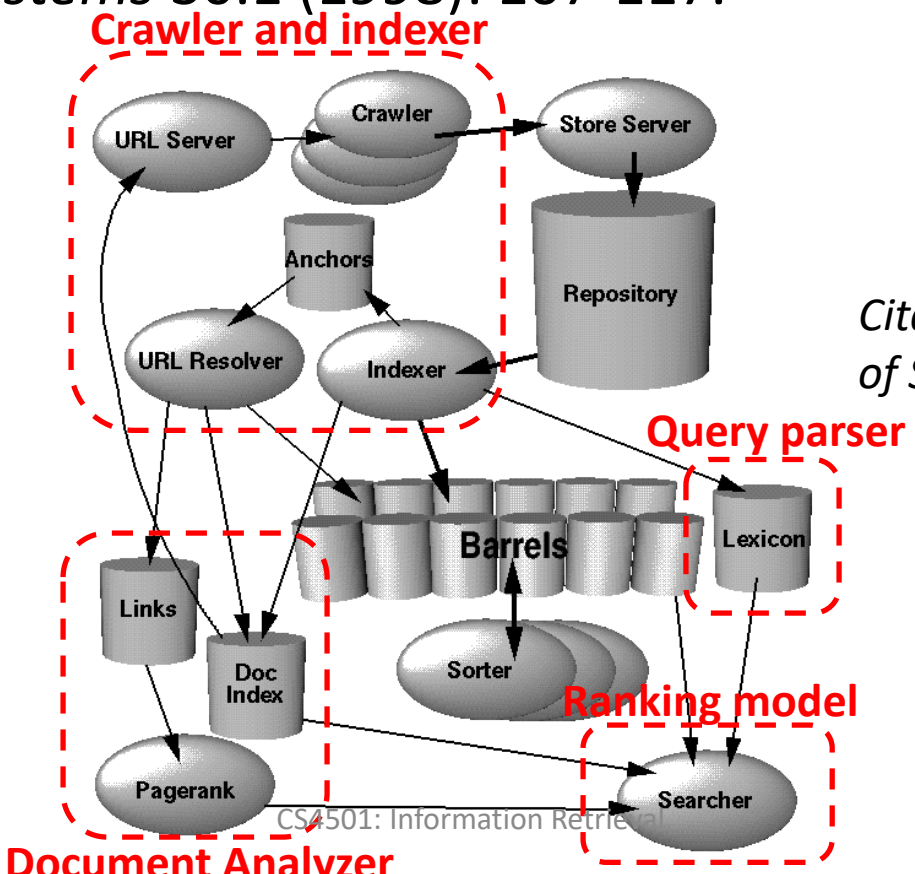
- Information overload
  - *"It refers to the difficulty a person can have understanding an issue and making decisions that can be caused by the presence of too much information."* - wiki

Recap: 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

Classical search engine architecture

- ***"The Anatomy of a Large-Scale Hypertextual Web Search Engine"*** - *Sergey Brin and* Lawrence Page, *Computer networks and ISDN systems* 30.1 (1998): 107-117.



Citation count: 17,463 (as of September 6, 2017)

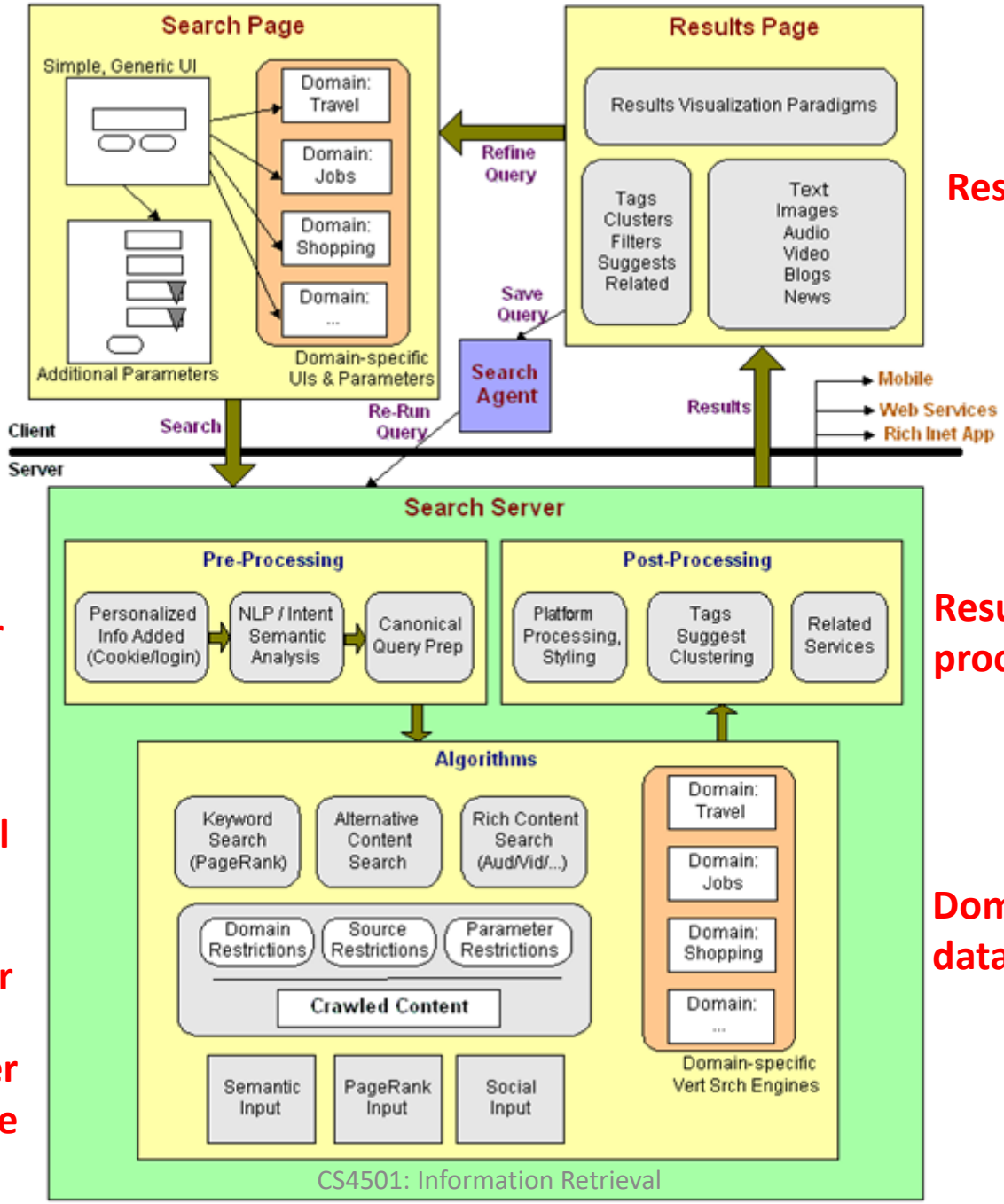
User input

Query parser

Ranking model

Crawler & Indexer

Document analyzer & auxiliary database

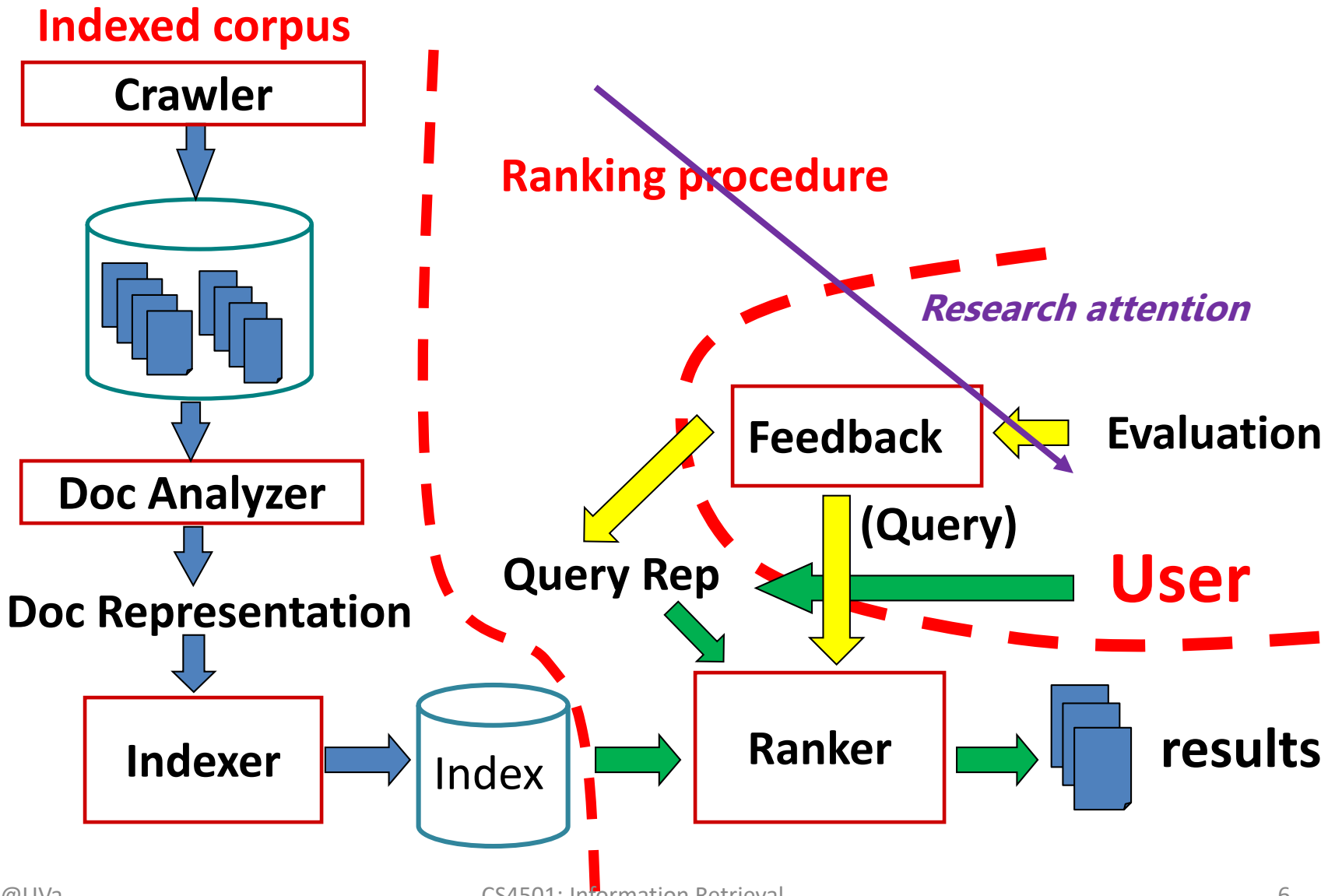


Result display

Result post-processing

Domain specific databases

Abstraction of search engine architecture



Core IR concepts

- Information need
  - "*an individual or group's desire to locate and obtain information to satisfy a conscious or unconscious need*" – wiki
  - An IR system is to satisfy users' information need
- Query
  - A designed representation of users' information need
  - In natural language, or some managed form

Core IR concepts

- Document
  - A representation of information that potentially satisfies users' information need
  - Text, **One sentence about IR - *"rank documents by their relevance to the user's information need"***
- Releva
  - Relat information need
  - Multiple perspectives: topical, semantic, temporal, spatial, and etc.

Key components in a search engine

- Web crawler
  - An automatic program that systematically browses the web for the purpose of Web content indexing and updating
- Document analyzer & indexer
  - Manage the crawled web content and provide efficient access of web documents

Key components in a search engine

- Query parser
  - Compile user-input keyword queries into managed system representation
- Ranking model
  - Sort candidate documents according to it relevance to the given query
- Result display
  - Present the retrieved results to users for satisfying their information need

Key components in a search engine

- Retrieval evaluation
  - Assess the quality of the returned results
- Relevance feedback
  - Propagate the quality judgment back to the system for search result refinement

Key components in a search engine

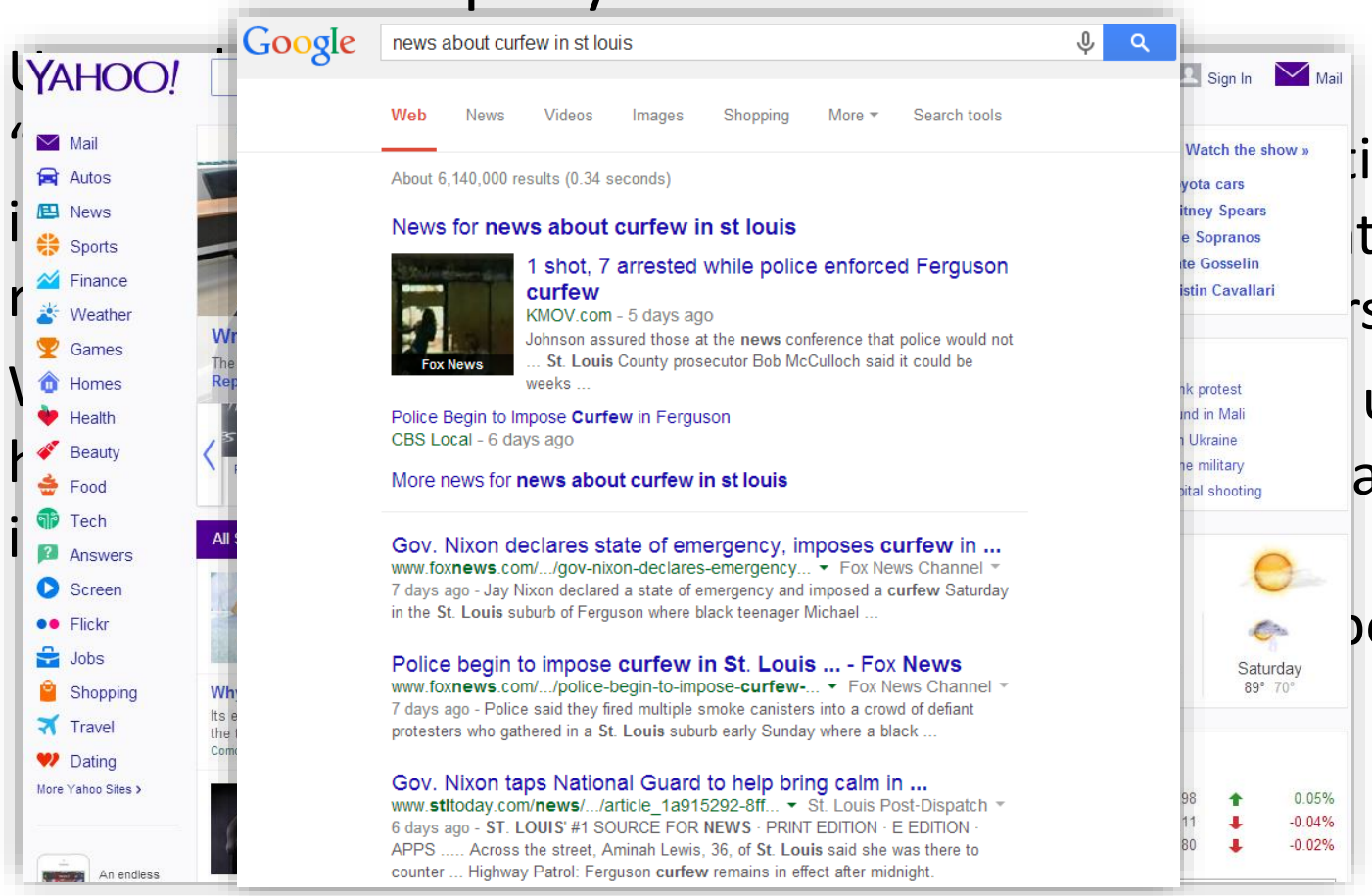
- Search query logs
  - Record users' interaction history with search engine
- User modeling
  - Understand users' longitudinal information need
  - Assess users' satisfaction towards search engine output

Discussion: Browsing v.s. Querying

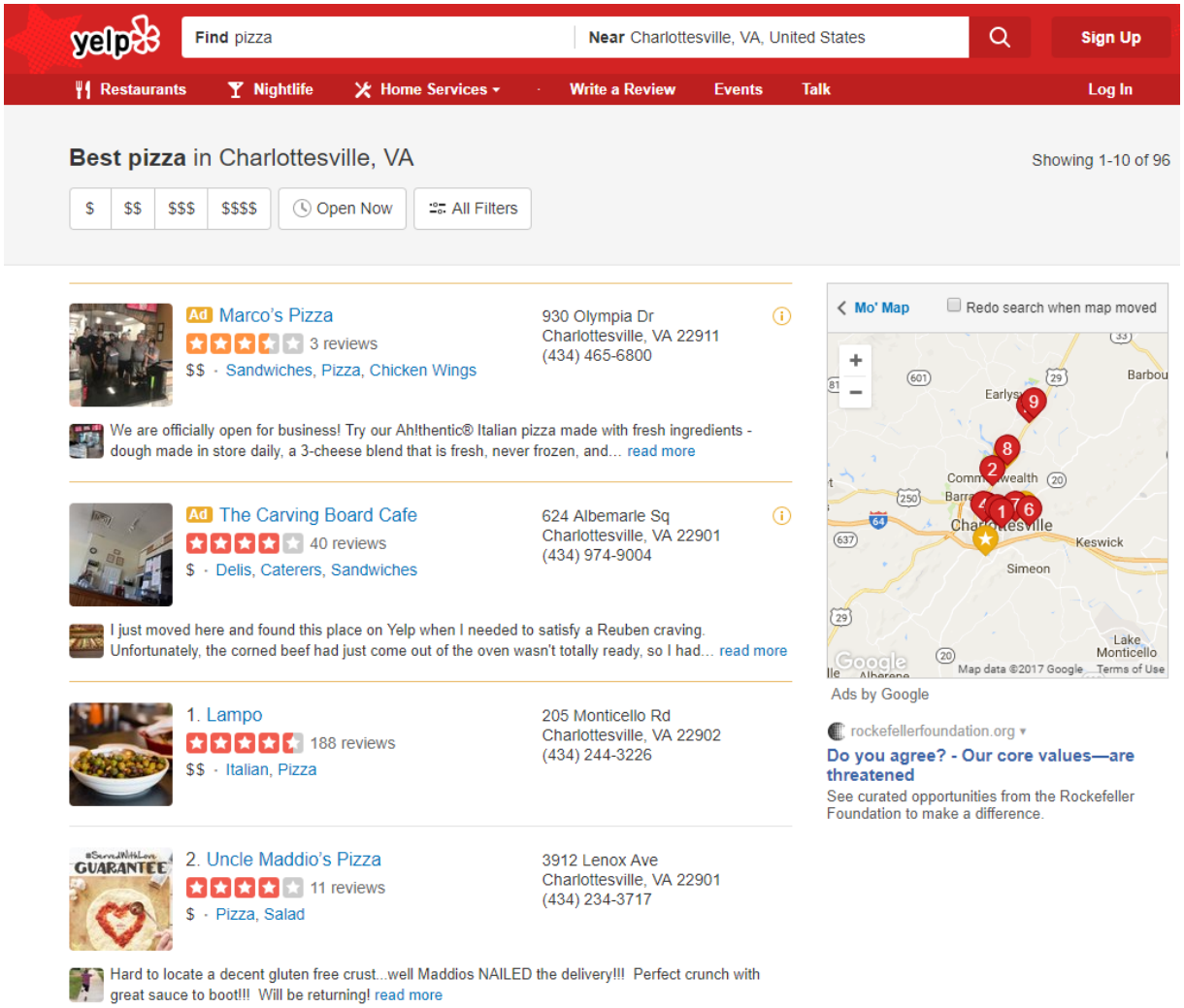
- Browsing – what Yahoo did before
  - The sy… a (keyword)
    inform… e system returns
    and a… ant documents
    releva… hen the user
    follow… what query to
    the st… ssing her
  - Works… eed
    wants…
    or doe…
    keywo…
    conve…
    (e.g., with a smartphone)
- Querying – what Google does

Pull vs. Push in Information Retrieval

- Pull mode – with query
  - U
  - i
  - r
  - V
  - h
  - i
- Push mode – without
  - tive
  - t
  - rs
  - user
  - ation
  - has
  - bout a



Discussion: is Yelp a search engine?



What you should know

- Basic workflow and components in an IR system
- Core concepts in IR
- Browsing v.s. querying
- Pull v.s. push of information

Today's reading

- Introduction to Information Retrieval
  - Chapter 19: Web search basics