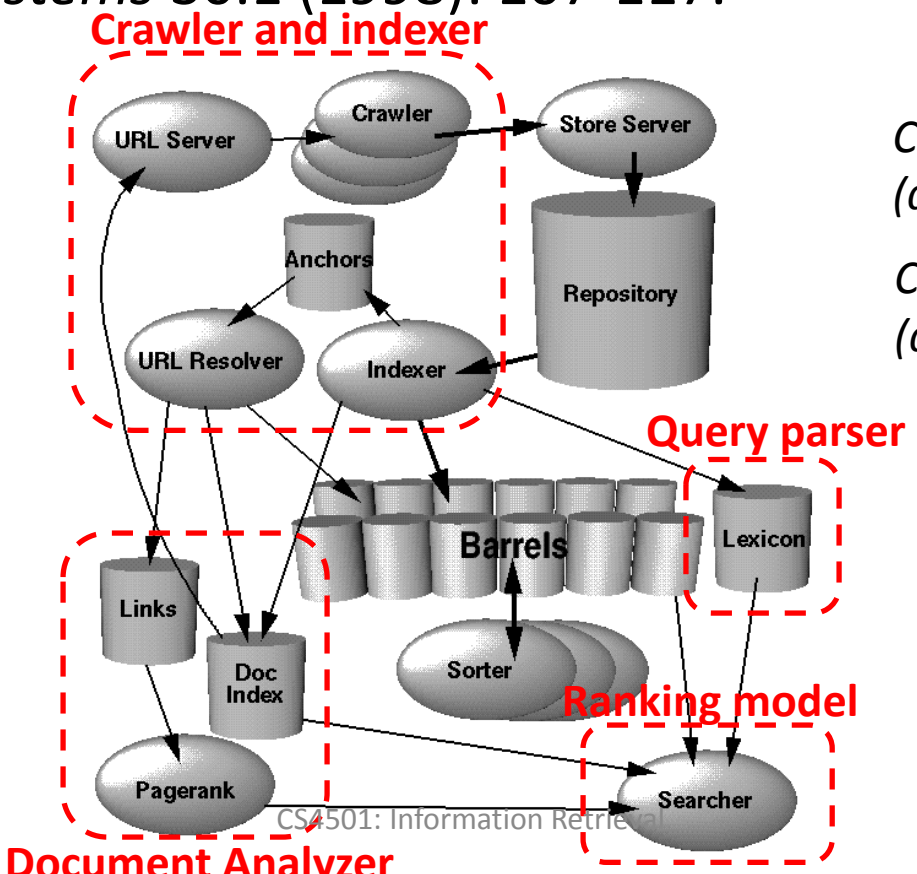
Search Engine Architecture

Hongning Wang
CS@UVa

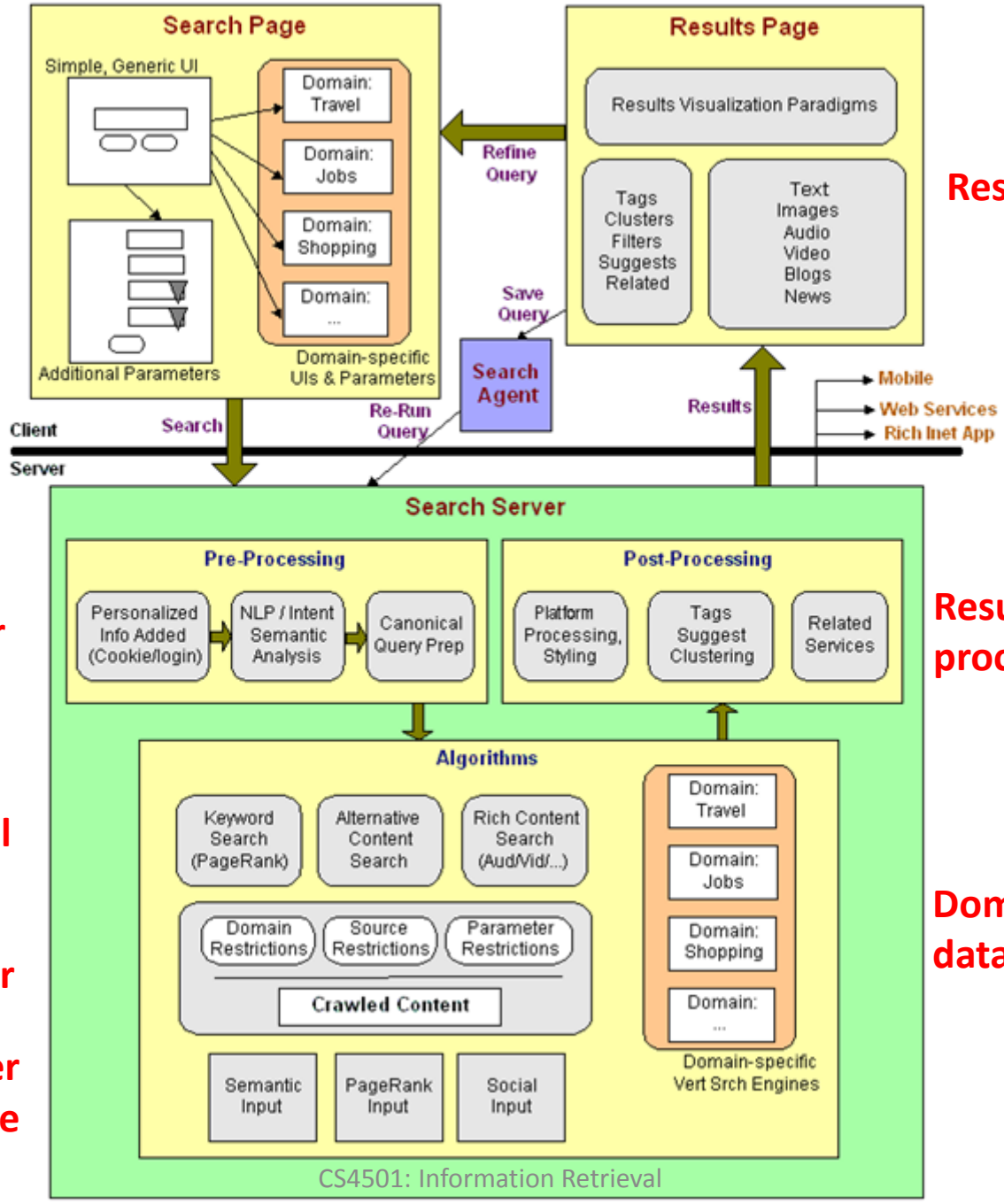
Classical search engine architecture

- ***"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: 12197 (as of Aug 27, 2014)

Citation count: 13727 (as of Aug 30, 2015)



User input

Query parser

Ranking model

Crawler & Indexer

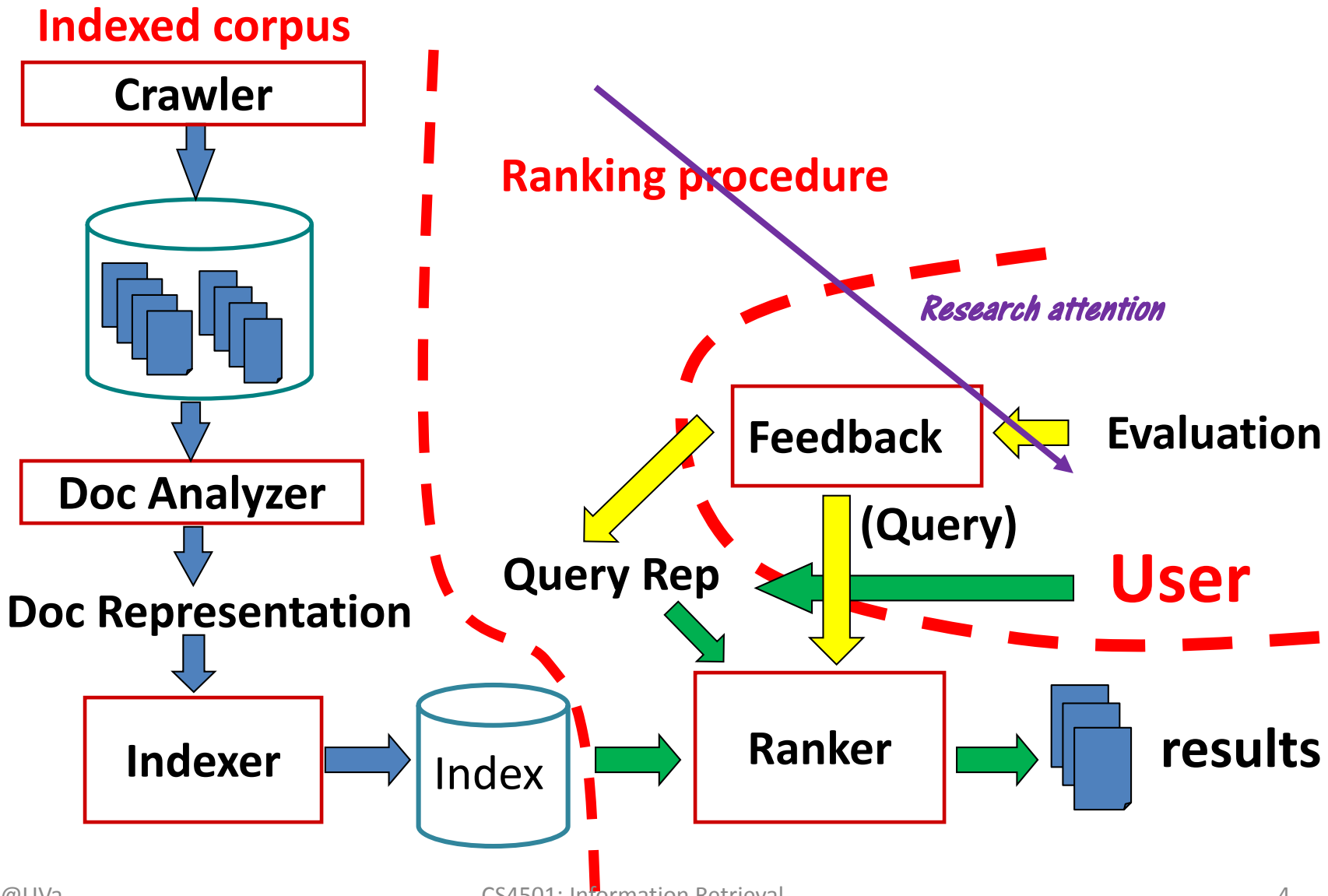
Document analyzer & auxiliary database

Result display

Result post-processing

Domain specific database

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 information need"***
- Releva
  - Relat information need
  - Multiple perspectives: topical, semantic, temporal, spatial, and etc.

Key components in a search engine

- Web crawler
  - A 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 return results
- Relevance feedback
  - Propagate the quality judgment back to the system for search result refinement

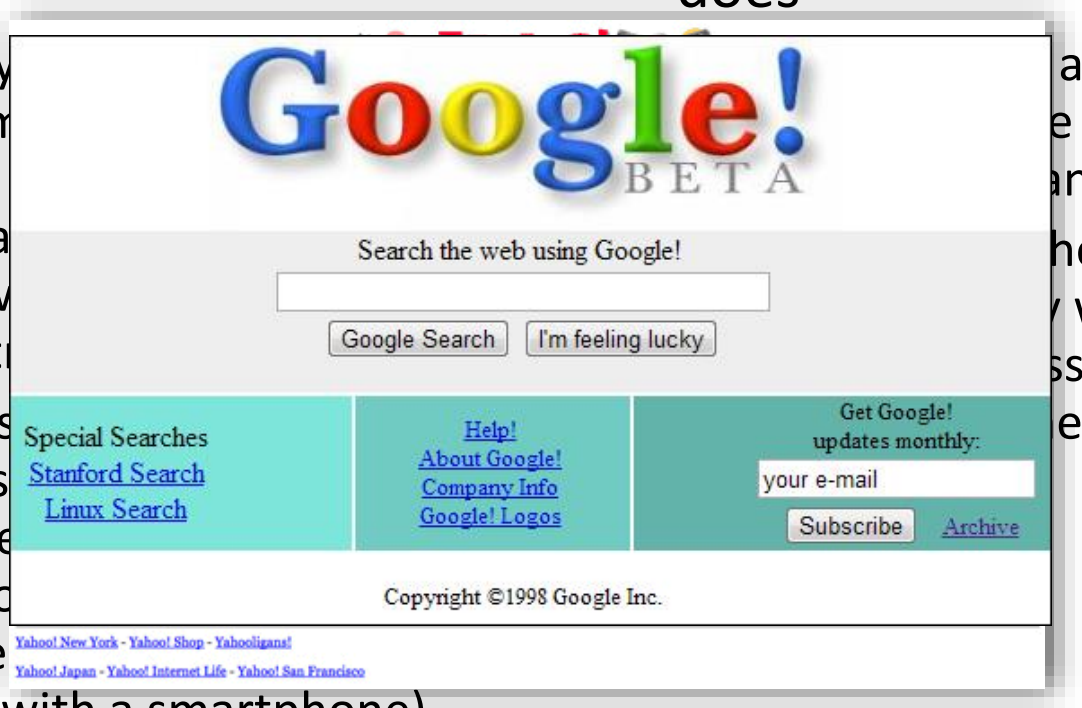
Key components in a search engine

- 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

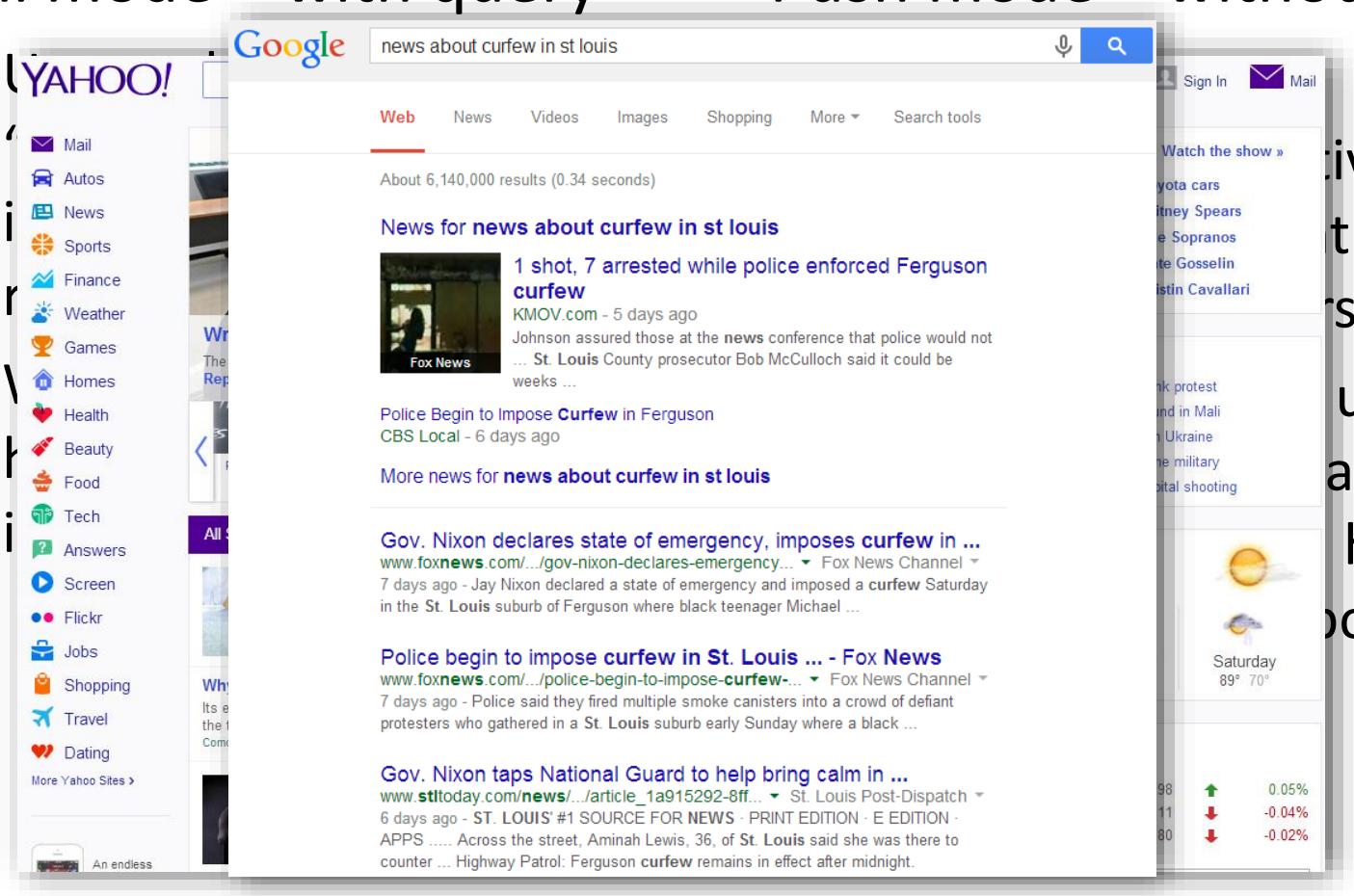
- Browsing – what Yahoo did before
  - The sy… a (keyword)
    inform… e system returns
    and a… ant documents
    releva… hen the user
    follow… what query to
    the st… ssing her
  - Works… eed
    wants…
    or doe…
    keywo…
    conve…
    (e.g., with a smartphone)

- Querying – what Google does



Pull vs. Push in Information Retrieval

- Pull mode – with query
  - l
  - \
- Push mode – without



What you should know

- Basic workflow and components in a 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