

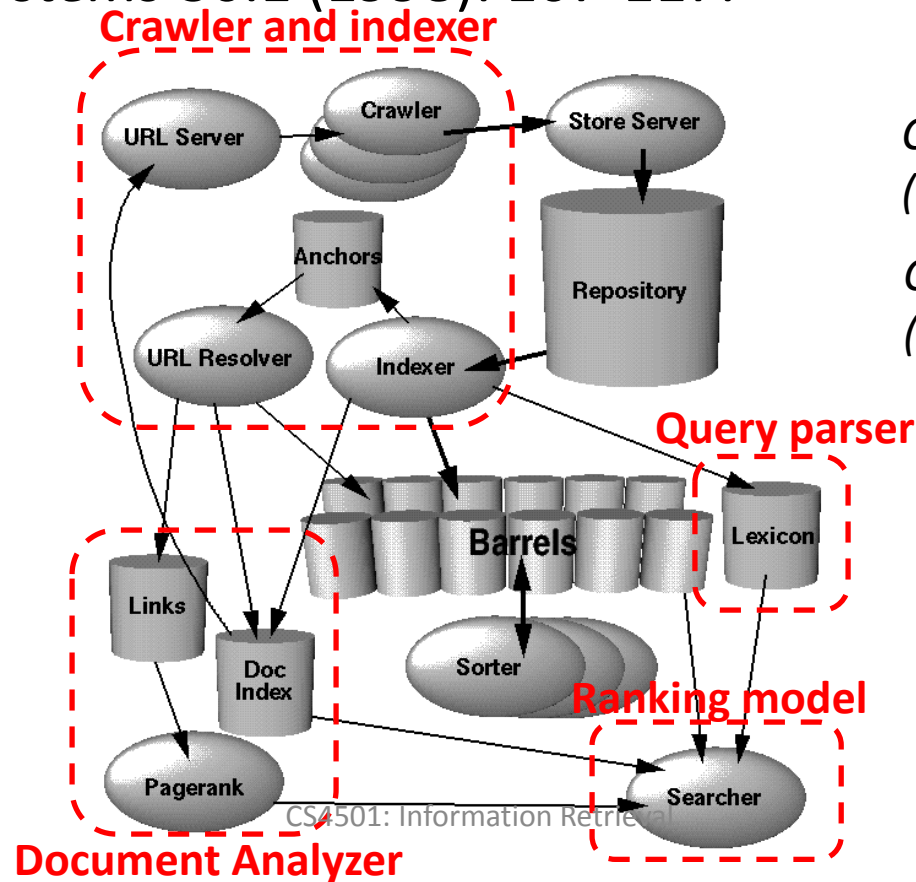
Search Engine Architecture

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

Result display

Query parser

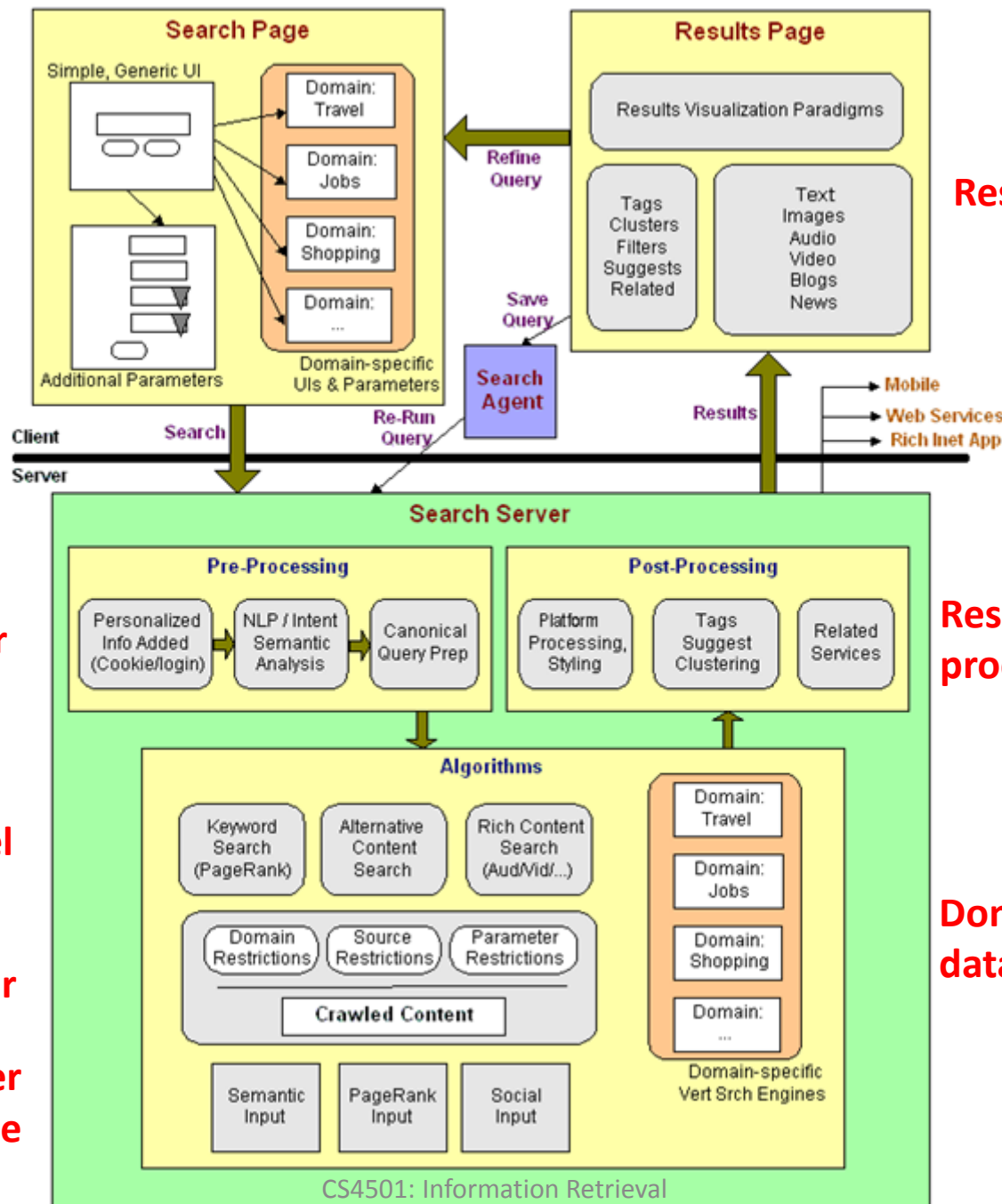
Ranking model

Crawler & Indexer

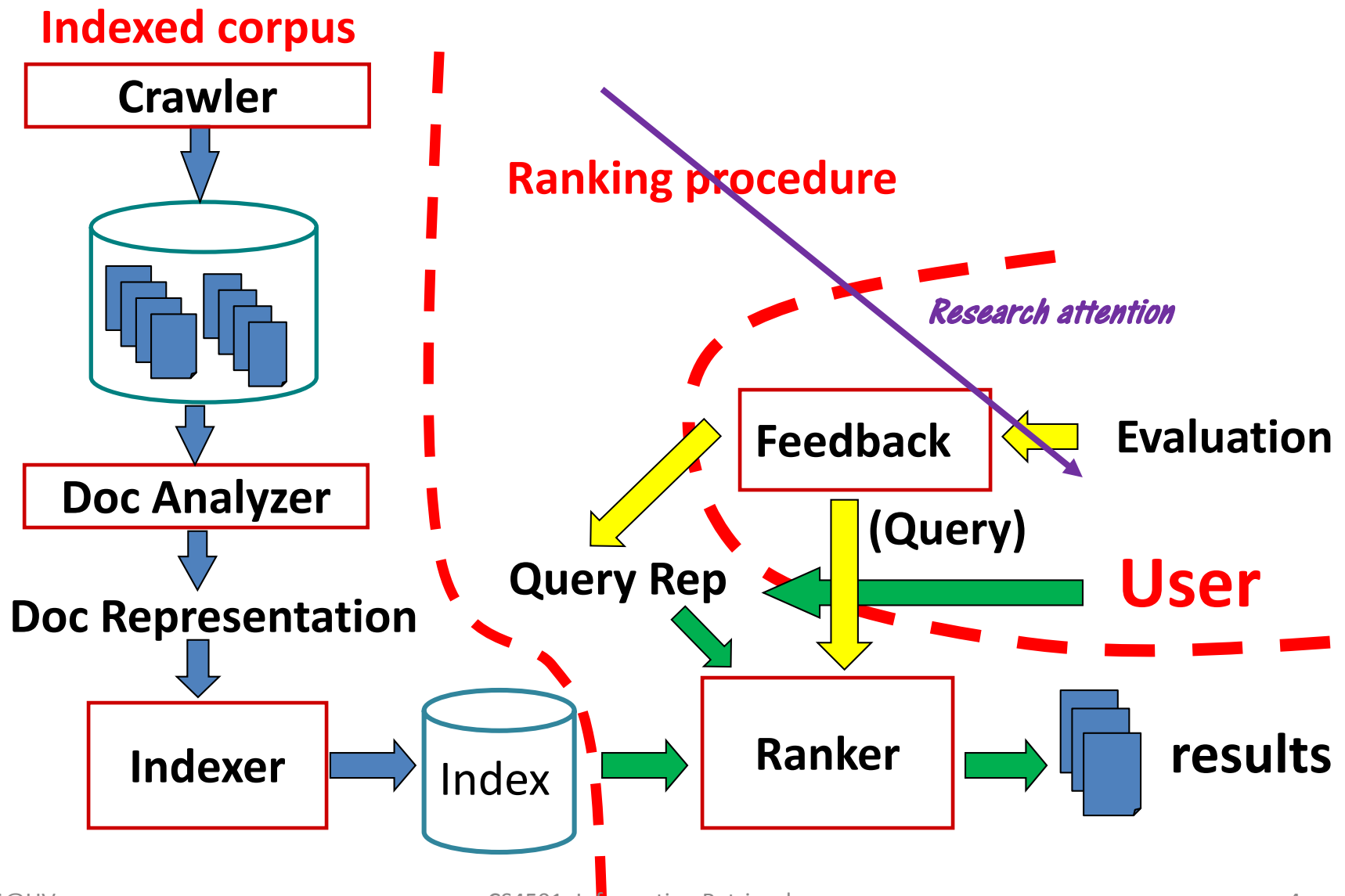
Document analyzer
& auxiliary database

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**
- Relevance ***documents by their relevance to***
 - Related ***the information need***
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 its 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 system informs the user about relevant documents and allows the user to follow the stream of documents
- Works well for browsing or does keyword searching (e.g., with a smartphone)

- Querying – what Google does

a (keyword) query. The system returns relevant documents when the user enters a query to the system. The user is passing her query to the system.



Pull vs. Push in Information Retrieval

- Pull mode – with query
- Push mode – without

The screenshot shows a Google search results page for the query "news about curfew in st louis". The search bar at the top displays the query and the Google logo. Below the search bar, there are tabs for "Web", "News", "Videos", "Images", "Shopping", and "More". The "Web" tab is selected. The search results show "About 6,140,000 results (0.34 seconds)". The first result is titled "News for news about curfew in st louis" and includes a thumbnail image of a person in a uniform. The second result is titled "1 shot, 7 arrested while police enforced Ferguson curfew" and includes a thumbnail image of a person in a uniform. The third result is titled "Police Begin to Impose Curfew in Ferguson" and includes a thumbnail image of a person in a uniform. The fourth result is titled "Gov. Nixon declares state of emergency, imposes curfew in ..." and includes a thumbnail image of a person in a uniform. The fifth result is titled "Police begin to impose curfew in St. Louis ..." and includes a thumbnail image of a person in a uniform. The sixth result is titled "Gov. Nixon taps National Guard to help bring calm in ..." and includes a thumbnail image of a person in a uniform. The search results are displayed in a list format with titles, snippets, and source information. On the right side of the page, there is a "Sign In" button and a "Mail" icon. Below these, there is a "Watch the show" button and a list of suggested videos. At the bottom right, there is a weather forecast for Saturday, showing a high of 89° and a low of 70°. The background of the slide shows a Yahoo! homepage with various links and a search bar.

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