

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

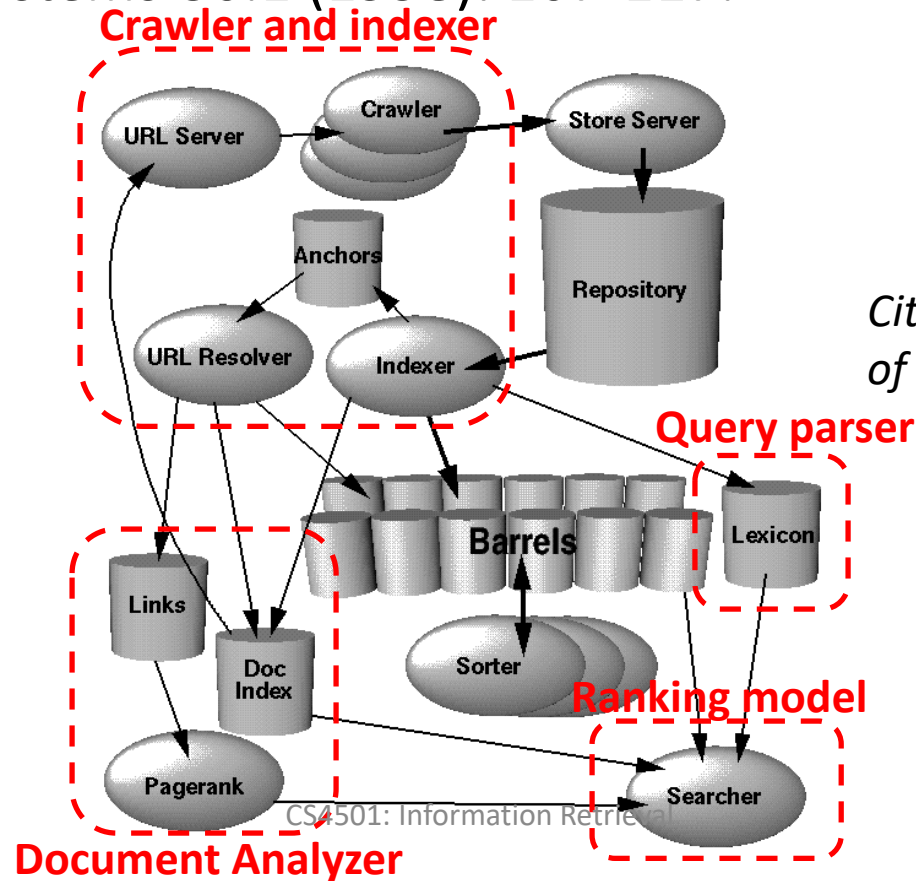


Recap: IR v.s. DBs

- 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

Result display

Query parser

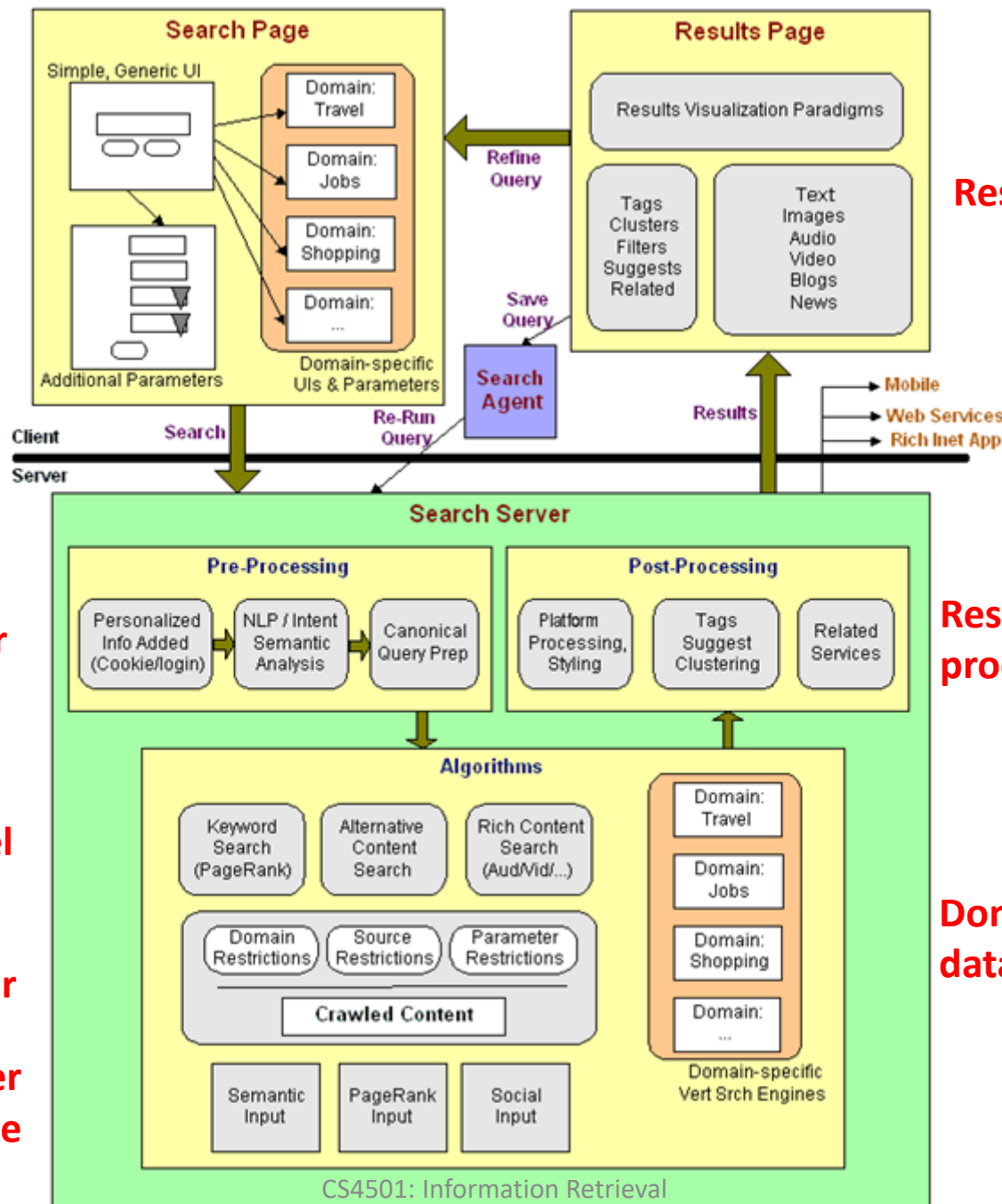
Ranking model

Crawler & Indexer

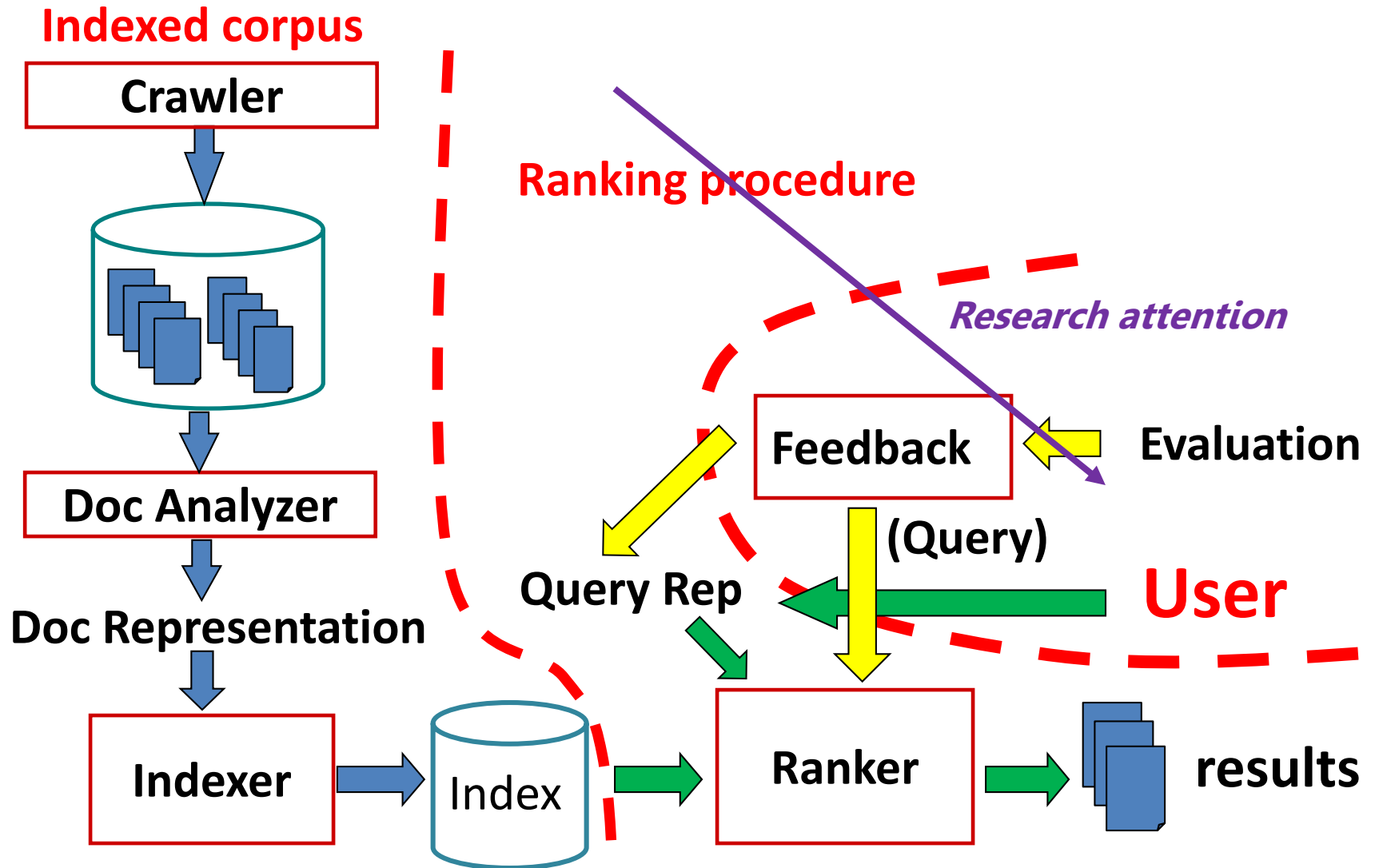
Document analyzer
& auxiliary database

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

- Querying – what Google does

a (keyword) query system returns relevant documents when the user enters a query to the system



Pull vs. Push in Information Retrieval

- Pull mode – with query
- Push mode – without

The screenshot shows a web browser window with a Google search bar at the top. The search query is "news about curfew in st louis". The search results are displayed in a pull mode, showing a list of news articles. The browser window also shows a Yahoo! sidebar on the left and a weather forecast on the right.

Search Results:

- News for news about curfew in st louis**
- 1 shot, 7 arrested while police enforced Ferguson curfew**
KMOV.com - 5 days ago
Johnson assured those at the news conference that police would not ... St. Louis County prosecutor Bob McCulloch said it could be weeks ...
- Police Begin to Impose Curfew in Ferguson**
CBS Local - 6 days ago
- More news for news about curfew in st louis**
- Gov. Nixon declares state of emergency, imposes curfew in ...**
www.foxnews.com/.../gov-nixon-declares-emergency...
7 days ago - Jay Nixon declared a state of emergency and imposed a curfew Saturday in the St. Louis suburb of Ferguson where black teenager Michael ...
- Police begin to impose curfew in St. Louis ... - Fox News**
www.foxnews.com/.../police-begin-to-impose-curfew-...
7 days ago - Police said they fired multiple smoke canisters into a crowd of defiant protesters who gathered in a St. Louis suburb early Sunday where a black ...
- Gov. Nixon taps National Guard to help bring calm in ...**
www.stltoday.com/news/.../article_1a915292-8ff...
6 days ago - ST. LOUIS' #1 SOURCE FOR NEWS · PRINT EDITION · E EDITION · APPS Across the street, Aminah Lewis, 36, of St. Louis said she was there to counter ... Highway Patrol: Ferguson curfew remains in effect after midnight.

Yahoo! Sidebar:

- Mail
- Autos
- News
- Sports
- Finance
- Weather
- Games
- Homes
- Health
- Beauty
- Food
- Tech
- Answers
- Screen
- Flickr
- Jobs
- Shopping
- Travel
- Dating
- More Yahoo! Sites >

Weather Forecast:

Saturday
89° 70°

98 ↑ 0.05%
11 ↓ -0.04%
80 ↓ -0.02%

Discussion: is Yelp a search engine?

The screenshot shows the Yelp homepage with a search bar containing 'Find pizza' and a location filter set to 'Near Charlottesville, VA, United States'. The navigation bar includes links for Restaurants, Nightlife, Home Services, Write a Review, Events, Talk, Sign Up, and Log In. The main heading is 'Best pizza in Charlottesville, VA' with a subtext 'Showing 1-10 of 96'. Below this are filter buttons for price (\$ to \$\$\$\$), 'Open Now', and 'All Filters'.

The search results list several pizza places:

- Ad Marco's Pizza**: 930 Olympia Dr, Charlottesville, VA 22911, (434) 465-6800. 3 reviews. Price: \$\$ - Sandwiches, Pizza, Chicken Wings. Description: 'We are officially open for business! Try our Ah!lthentic® Italian pizza made with fresh ingredients - dough made in store daily, a 3-cheese blend that is fresh, never frozen, and... [read more](#)'
- Ad The Carving Board Cafe**: 624 Albemarle Sq, Charlottesville, VA 22901, (434) 974-9004. 40 reviews. Price: \$ - Delis, Caterers, Sandwiches. Description: 'I just moved here and found this place on Yelp when I needed to satisfy a Reuben craving. Unfortunately, the corned beef had just come out of the oven wasn't totally ready, so I had... [read more](#)'
- 1. Lampo**: 205 Monticello Rd, Charlottesville, VA 22902, (434) 244-3226. 188 reviews. Price: \$\$ - Italian, Pizza.
- 2. Uncle Maddio's Pizza**: 3912 Lenox Ave, Charlottesville, VA 22901, (434) 234-3717. 11 reviews. Price: \$ - Pizza, Salad. Description: 'Hard to locate a decent gluten free crust...well Maddios NAILED the delivery!!! Perfect crunch with great sauce to boot!!! Will be returning! [read more](#)'

On the right side, there is a map showing the location of the search results in Charlottesville, VA. Below the map is an advertisement from the Rockefeller Foundation with the text: 'Do you agree? - Our core values—are threatened. See curated opportunities from the Rockefeller Foundation to make a difference.'

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