Search Engine Architecture

Hongning Wang
CS@UVa
Classical search engine architecture


Citation count: 12197 (as of Aug 27, 2014)
Citation count: 13727 (as of Aug 30, 2015)
Citation count: 15627 (as of Jan 25, 2017)
Abstraction of search engine architecture

Indexed corpus

Crawler

Doc Analyzer

Doc Representation

Indexer

Index

Ranker

Query Rep

Feedback

(_Query)

User

Evaluation

Research attention

results

CS@UVa CS6501: Information Retrieval
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

• Releva documents by their relevance to the information need”
  – 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 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 organizes information with structures, and a user navigates into relevant information by following a path enabled by the structures.
  - Works well when the user wants to explore information or doesn't know what keywords to use, or can't conveniently enter a query (e.g., with a smartphone).

- **Querying** – what Google does
  - A user enters a (keyword) query, and the system returns relevant documents.
  - Works well when the user knows exactly what query to use for expressing her information need.
Pull vs. Push in Information Retrieval

• Pull mode – with query

  - Users take initiative and "pull" relevant information out from a retrieval system
  - Works well when a user has an ad hoc information need

• Push mode – without query

  - Systems take initiative and "push" relevant information to users
  - Works well when a user has a stable information need or the system has good knowledge about a user's need
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