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

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 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 organizes information 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 a set of relevant documents
  - Works well when the user knows exactly what query to express 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
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